Robust Meta-learning for Mixed Linear Regression with Small Batches

Weihao Kong[*] **Raghav Somani**[†] **Sham Kakade**[‡] **Sewoong Oh**[§]

Abstract

A common challenge faced in practical supervised learning, such as medical image processing and robotic interactions, is that there are plenty of tasks but each task cannot afford to collect enough labeled examples to be learned in isolation. However, by exploiting the similarities across those tasks, one can hope to overcome such data scarcity. Under a canonical scenario where each task is drawn from a mixture of k linear regressions, we study a fundamental question: can abundant small-data tasks compensate for the lack of big-data tasks? Existing second moment based approaches of [42] show that such a trade-off is efficiently achievable, with the help of medium-sized tasks with $\Omega(k^{1/2})$ examples each. However, this algorithm is brittle in two important scenarios. The predictions can be arbitrarily bad (i) even with only a few outliers in the dataset; or (ii) even if the medium-sized tasks are slightly smaller with $o(k^{1/2})$ examples each. We introduce a spectral approach that is simultaneously robust under both scenarios. To this end, we first design a novel outlier-robust principal component analysis algorithm that achieves an optimal accuracy. This is followed by a sum-of-squares algorithm to exploit the information from higher order moments. Together, this approach is robust against outliers and achieves a graceful statistical trade-off; the lack of $\Omega(k^{1/2})$-size tasks can be compensated for with smaller tasks, which can now be as small as $\mathcal{O}(\log k)$.

1 Introduction

Modern machine learning tasks and corresponding training datasets exhibit a long-tailed behavior [73], where a large number of tasks do not have enough training examples to be trained to the desired accuracy. Collecting high-quality labeled data can be time consuming or require expertise. Consequently, in domains such as annotating medical images or processing robotic interactions, there might be a large number of related but distinct tasks, yet each task is associated with only a small batch of training data. However, one can hope to meta-train across those tasks, exploiting their similarities, and collaboratively achieve accuracy far greater than what can be achieved for each task in isolation [29, 58, 41, 52, 68, 61]. This is the goal of meta-learning [62, 67].

Meta-learning is especially challenging under two practically important settings: (i) a few-shot learning scenario where each task is associated with an extremely small dataset; and (ii) an adversarial scenario where a fraction of those datasets are corrupted. We design a novel meta-learning approach that is robust to such data scarcity and adversarial corruption, under a canonical scenario where the tasks are linear regressions in d-dimensions and the model parameters are drawn from a discrete distribution of a support size k.

[*] `kweihao@gmail.com`. University of Washington

[†] `raghavs@cs.washington.edu`. University of Washington

[‡] `sham@cs.washington.edu`. University of Washington & Microsoft Research

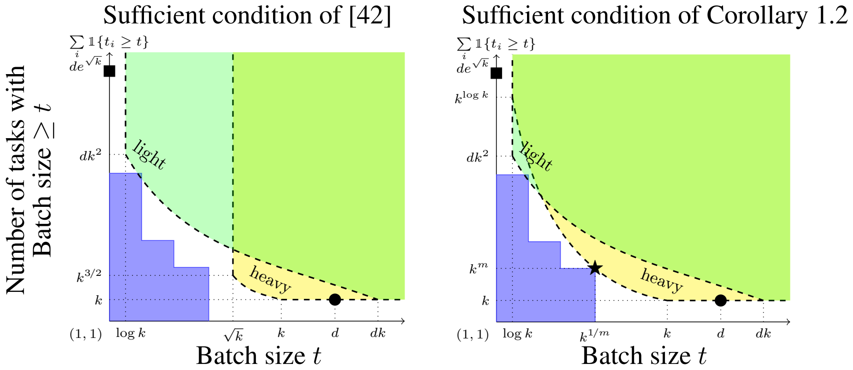
[§] `sewoong@cs.washington.edu`. University of Washington

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

First, consider a case where we have an *uncorrupted* dataset from a collection of n tasks, each with t training examples. Concretely, the i-th task for $i \in \{1, \dots, n\}$ is associated with a regression parameter $\beta_i \in \{w_1, \dots, w_k\}$ and a corresponding dataset $\{\mathbf{x}_{i,j} \in \mathbb{R}^d, y_{i,j} \in \mathbb{R}\}_{j=1}^t$ drawn from $y_{i,j} = \beta_i^\top \mathbf{x}_{i,j} + \epsilon_{i,j}$ for some noise $\epsilon_{i,j}$. A formal definition of the generative model is provided in § 1.1. If each task has a large enough training data with $t = \Omega(d)$ examples, it can be accurately learned in isolation. This is illustrated by solid circles in Fig. 1. On the opposite extreme, where each task has only a *single* example (i.e. $t = 1$), significant efforts have been made to make training statistically efficient [14, 77, 63, 78, 47, 16, 64]. However, even the best known result of [16] still requires exponentially many such tasks: $n = \Omega(de^{\sqrt{k}})$ (details in related work in §3). This is illustrated by solid squares in Fig. 1. Perhaps surprisingly, this can be significantly reduced to quasi-polynomial $n = \Omega(k^{\Theta(\log k)})$ sample complexity and quasi-polynomial run-time, with a slightly larger dataset that is only logarithmic in the problem parameters. This result is summarized in the following, with the algorithm and proof presented in §A of the supplementary material.

Corollary 1.1 (of our results with no corruption, informal)**.** *Given a collection of n tasks each associated with $t = \widetilde{\Omega}(1)$ labeled examples, if the effective sample size $nt = \widetilde{\Omega}(dk^2 + k^{\Theta(\log k)})$, then Algorithm 4 estimates the meta-parameters up to any desired accuracy of $\mathcal{O}(1)$ with high probability in time* $\text{poly}(d, k^{(\log k)^2})$*, under certain assumptions on the meta-parameters.*

This is a special case of a more general class of algorithms we design, tailored for the following practical scenario; the collection of tasks in hand are heterogeneous, each with varying sizes of datasets (illustrated by the blue bar graphs below in Fig. 1). Inspired by the seminal work of [71], we exploit such heterogeneity by separating the roles of *light* tasks that have smaller datasets and *heavy* tasks that have larger datasets. As we will show, the size of the heavy tasks determines the order of the higher order moments we can reliably exploit. Concretely, we first use the light tasks to *estimate the subspace* spanned by the regression parameters, and then *cluster* heavy tasks by projecting them on the estimated subspace. The first such attempt was taken in [42], where a linkage-based clustering was proposed. However, as this clustering method relies on the second moment statistics, it strictly requires heavy tasks with $\Omega(k^{1/2})$ examples (left panel in Fig. 1). In the absence of such heavy tasks, the abundant light tasks are wasted as no existing algorithm can harness their structural similarities. Such second moment barriers are common in even simpler problems, e.g. [43, 44].


Sufficient condition of [42]

Sufficient condition of Corollary 1.2

Figure 1: The blue bar graph summarizes the collection of tasks in hand, showing the cumulative count of tasks with more than t examples. Typically, this does not include extremely large data tasks (circle) and extremely large number of small data tasks (square), where classical approaches succeed. When any point in the light (green) region and any point in the heavy (yellow) region are both realized by the blue graph, the corresponding algorithm succeeds. On the left, the collection in blue cannot be learned by any existing methods including [42]. Our approach in Corollary 1.2 significantly extends the heavy region all the way down to $\log k$, leading to a successful meta-learning in this example.

We exploit higher order statistics to break this barrier, using computationally tractable tools from sum-of-squares methods [45]. This gives a class of algorithms parameterized by an integer m (for m-th order moment) to be chosen by the analyst tailored to the size of the heavy tasks in hand $t_H = \Omega(k^{1/m})$. This allows for a graceful trade-off between t_H and with the required number of heavy tasks n_H. We summarize the result below, with a proof in §A, and illustrate it in Fig. 1

(right). The choice of $m = \Theta(\log k)$ gives the minimum required batch size, as we highlighted in Corollary 1.1.

Corollary 1.2 (of our results with no corruption, informal)**.** *For any integer m, given two collections of tasks, first collection of light tasks with $t_L = \widetilde{\Omega}(1)$, $t_L n_L = \widetilde{\Omega}(dk^2)$, and the second collection of heavy tasks with $t_H = \widetilde{\Omega}(m\, k^{1/m})$, $t_H n_H = \widetilde{\Omega}(k^{\Theta(m)})$, the guarantees of Corollary 1.1 hold.*

Next, consider an adversarial scenario. Outliers are common in meta-learning as diverse sources contribute to the collection. Existing approaches are brittle to a few such outliers. [42] builds upon principal component analysis and linear regression, both of which are known to be brittle to outliers [40, 19]. For example, a single corrupted user can result in an arbitrarily bad subspace estimation in the first step of [42]. This causes the meta-learning algorithm to learn nothing about the true regression parameters, resulting in a completely random prediction in the subsequent step. A fundamental question of interest is, what can be meta-learned from past experience that is only partially trusted? Following robust learning literature [45, 25], we assume a general adversary who can adaptively corrupt any α fraction of the tasks, formally defined in Assumption 2. This parameter $\alpha \in [0, 1]$ captures how powerful an adversary is. Among all adversaries that can corrupt an α fraction of the dataset, we assume the strongest possible one that can adaptively select which samples to corrupt and replace them with arbitrary data points. We make both subspace estimation and clustering steps robust against adversarial corruption. The sum-of-squares approach is inherently robust, when used within an iterative clustering [45]. However, existing robust subspace estimation approaches are suboptimal, requiring $\widetilde{\mathcal{O}}(d^2)$ samples [24]. To this end, we introduce a novel algorithm, and prove its optimality in both accuracy and dependence in the dimension d. This resolves an open question posed in [64] on whether it is possible to robustly learn the subspace with $\widetilde{\mathcal{O}}(d)$ samples.

This achieves a similar sample complexity as the uncorrupted case in Corollary 1.2, while tolerating as much corruption as information theoretically possible: $\alpha = \widetilde{\mathcal{O}}(\epsilon/k)$ for an ϵ accuracy in parameter estimation. Such condition is necessary as otherwise the adversary can focus its attack on one of the mixtures, and incur $\Omega(\epsilon)$ error in estimating the parameter of that component.

Corollary 1.3 (of Theorem 1, informal)**.** *For any $\epsilon \in (0, 1/k^3)$ and $m \in \mathbb{N}$, given two collections of tasks, the first with $t_L = \widetilde{\Omega}(1), n_L t_L = \widetilde{\Omega}\big(dk\epsilon^{-2}\big)$, and the second with $t_H = \widetilde{\Omega}\big(mk^{1/m}\big), n_H t_H = \widetilde{\Omega}\big(k^{\mathcal{O}(m)}\big)$, if the fraction of corrupted tasks is $\alpha = \widetilde{\mathcal{O}}(\epsilon/k)$, Algorithm 1 achieves up to ϵ accuracy with high probability in time $\mathrm{poly}(d, k^{m^2}, \epsilon^{-1})$, under certain assumptions.*

We provide the algorithm (Algorithm 1) and the analysis (Theorem 1) under the adversarial scenario in the main text. When there is no corruption, the algorithm can be made statistically more efficient with tighter guarantees, which is provided in §A.

1.1 Problem formulation and notations

We present the probabilistic perspective on few-shot supervised learning following [31], but focusing on a simple yet canonical case where the tasks are linear regressions. A collection of n tasks are independently drawn according to some prior distribution. The i-th task is associated with a *model parameter* $\phi_i = (\beta_i \in \mathbb{R}^d, \sigma_i \in \mathbb{R}_+)$, and a *meta-train dataset* $\{(\mathbf{x}_{i,j}, y_{i,j}) \in \mathbb{R}^d \times \mathbb{R}\}_{j=1}^{t_i}$ of size t_i. Each example $(\mathbf{x}_{i,j}, y_{i,j}) \sim \mathbb{P}_{\phi_i}(y|\mathbf{x})\mathbb{P}(\mathbf{x})$ is independently drawn from a linear model, such that

$$y_{i,j} \;\; = \;\; \beta_i^\top \mathbf{x}_{i,j} + \epsilon_{i,j} \; , \tag{1}$$

where $\mathbf{x}_{i,j} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma_i^2)$. If $\mathbf{x}_{i,j}$ is from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, we assume to have enough $\mathbf{x}_{i,j}$'s (not necessarily labeled) for whitening, and $\mathbb{P}(\mathbf{x})$ can be made sufficiently close to isotropic.

The goal of meta-learning is to train a model for a new arriving task $\phi_{\text{new}} \sim \mathbb{P}_\theta(\phi)$ from a small size *training dataset* $\mathcal{D} = \{(\mathbf{x}_j^{\text{new}}, y_j^{\text{new}}) \sim \mathbb{P}_{\phi_{\text{new}}}(y|\mathbf{x})\mathbb{P}(\mathbf{x})\}_{j=1}^{\tau}$ of size τ. This is achieved by exploiting some structural similarities to the meta-train dataset, drawn from the same prior distribution $\mathbb{P}_\theta(\phi)$. To capture such structural similarities, we make a mild assumption that $\mathbb{P}_\theta(\phi)$ is a finite discrete distribution of a support size k. This is also known as mixture of linear experts [14]. Concretely, $\mathbb{P}_\theta(\phi)$ is fully defined by a *meta-parameter* $\theta = (\mathbf{W} \in \mathbb{R}^{d\times k}, \mathbf{s} \in \mathbb{R}_+^k, \mathbf{p} \in S^{k-1})$ with k candidate model parameters $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_k]$ and k candidate noise parameters $\mathbf{s} = [s_1, \ldots, s_k]$. The i-th task is drawn from $\phi_i \sim \mathbb{P}_\theta(\phi)$, where first a $z_i \sim \text{multinomial}(\mathbf{p})$ selects a component that the task belongs to, and training data is independently drawn from Eq. (1) with $\beta_i = \mathbf{w}_{z_i}$ and $\sigma_i = s_{z_i}$.

Following the definition of [31], the *meta-learning problem* refers to solving the following:

$$\theta^* \;\in\; \arg\max_{\theta} \;\log \mathbb{P}(\theta|\mathcal{D}_{\mathrm{meta-train}}) \;, \tag{2}$$

which estimates the most likely meta-parameter given *meta-training dataset* defined as $\mathcal{D}_{\mathrm{meta-train}} := \{\{(\mathbf{x}_{i,j}, y_{i,j}) \in \mathbb{R}^d \times \mathbb{R}\}_{j=1}^{t_i}\}_{i=1}^{n}$. This is a special case of empirical Bayes methods [13]. Our goal is to solve this meta-learning problem *robustly* against an adversarial corruption of $\mathcal{D}_{\mathrm{meta-train}}$ as formally defined in Assumption 2. Once meta-learning is solved, the model parameter of the newly arriving task can be estimated with a Maximum a Posteriori (MAP) or a Bayes optimal estimator:

$$\widehat{\phi}_{\mathrm{MAP}} \;\in\; \arg\max_{\phi} \log \mathbb{P}(\phi|\mathcal{D}, \theta^*) \;, \quad \text{and} \quad \widehat{\phi}_{\mathrm{Bayes}} \;\in\; \arg\min_{\phi} \mathbb{E}_{\phi' \sim \mathbb{P}(\phi|\mathcal{D},\theta^*)}[\ell(\phi, \phi')] \;, \tag{3}$$

for some choice of a loss $\ell(\cdot)$, which is straightforward. This is subsequently used to predict the label of a new data point $\mathbf{x}$ from ϕ_{new}. Concretely, $\widehat{y} \in \arg\max_y \mathbb{P}_{\widehat{\phi}_{\mathrm{MAP\,/\,Bayes}}}(y|\mathbf{x})$.

Notations. We define $[n] := \{1, \ldots, n\}$, $\forall\, n \in \mathbb{N}$; S^{k-1} as the standard k-dimensional probability simplex; $\|\mathbf{x}\|_p := (\sum_{i=1}^{d} |x_i|^p)^{1/p}$ as the standard vector ℓ_p-norm of a vector $\mathbf{x} \in \mathbb{R}^d \;\forall\; d \in \mathbb{N} \;\forall\, p \geq 1$; $\|\mathbf{A}\|_* := \sum_{i=1}^{\min\{n,m\}} \sigma_i(\mathbf{A})$, $\|\mathbf{A}\|_{\mathrm{F}} := (\sum_{i,j=1}^{m,n} A_{i,j}^2)^{1/2}$ as the standard nuclear norm and Frobenius norm of matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, where $\sigma_i(\mathbf{A})$ denotes the i-th singular value of $\mathbf{A}$ respectively; $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$; $\mathbb{1}\{\cdot\}$ is the indicator function. We define $\rho_i^2 := s_{z_i}^2 + \|\mathbf{w}_{z_i}\|_2^2$ as the variance of a label $y_{i,j}$ in the i-th task, and $\rho^2 := \max_i \rho_i^2$. We define $p_{\min} := \min_{j \in [k]} p_j$, and $\Delta := \min_{i,j \in [k], i \neq j} \|\mathbf{w}_i - \mathbf{w}_j\|_2$ and assume $p_{\min}, \Delta > 0$. We use $\widetilde{\mathcal{O}}$ and $\widetilde{\Omega}$ notations that are extensions of the standard $\mathcal{O}$ and Ω Bachmann–Landau notations to hide poly-logarithmic factors.

1.2 Algorithm and intuitions

Following the recipe of spectral algorithms for clustering [71] and few-shot learning [42], we propose the following approach consisting of three steps. Clustering step requires *heavy tasks*; each task has many labeled examples, but we need a smaller number of such tasks. Subspace estimation and classification steps require *light tasks*; each task has a few labeled examples, but we need a larger number of such tasks. Here, we provide the intuition behind each step and the corresponding requirements. The details are deferred to §B, where we emphasize robustness to corruption of the data. The estimated $\widehat{\theta} = \big(\widehat{\mathbf{W}}, \widehat{\mathbf{s}}, \widehat{\mathbf{p}}\big)$ is subsequently used in prediction, when a new task arrives.

Algorithm 1

Meta-learning

1. *Subspace estimation:* Compute subspace $\widehat{\mathbf{U}}$ which approximates $\mathrm{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$.
2. *Clustering:* Project the heavy tasks onto the subspace of $\widehat{\mathbf{U}}$, perform k clustering, and estimate $\widetilde{\mathbf{w}}_\ell$ for each cluster $\ell \in [k]$.
3. *Classification:* Perform likelihood-based classification of the light tasks using $\{\widetilde{\mathbf{w}}_\ell\}_{\ell=1}^{k}$ estimated from the *Clustering* step; compute refined estimates $\{\widehat{\mathbf{w}}_\ell, \widehat{s}_\ell, \widehat{p}_\ell\}_{\ell=1}^{k}$ of θ.

Prediction

4. *Prediction:* Perform MAP or Bayes optimal prediction using the estimated meta-parameter.

Subspace estimation. As $\boldsymbol{\Sigma} := \mathbb{E}[y_{i,j}^2 \mathbf{x}_{i,j} \mathbf{x}_{i,j}^\top] = c\mathbf{I}_d + 2\sum_{\ell=1}^{k} p_\ell \mathbf{w}_\ell \mathbf{w}_\ell^\top$ for some constant $c \geq 0$, the subspace spanned by the regression vectors, $\mathrm{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$, can be efficiently estimated by Principal Component Analysis (PCA), if we have uncorrupted data. This only requires $\widetilde{\Omega}(d)$ samples. With α-fraction of the tasks adversarially corrupted, existing approaches of outlier-robust PCA attempt to simultaneously estimate the principal subspace while *filtering out* the outliers [75]. This removes many uncorrupted data points, and hence can either only tolerate up to $\alpha = \mathcal{O}(1/k^8)$ fraction of corruption (assuming well-separated $\mathbf{w}_\ell$'s). We introduce a new approach in Algorithm 2 that uses a second filter to recover those erroneously removed data points. This improves the tolerance to $\alpha = \mathcal{O}(1/k^4)$ while requiring only $\widetilde{\Omega}(d)$ samples (see Remark B.2). We call this step *robust subspace estimation* (Algorithm 2 in §2.2).

Clustering. Once we have the subspace, we project the estimates of β_i's to the k-dimensional subspace and cluster those points to find the centers. As $k \ll d$ in typical settings, this significantly reduces the sample complexity from $\text{poly}(d)$ to $\text{poly}(k)$. Existing meta-learning algorithm of [42] proposed a linkage based clustering algorithm. This utilizes the bounded property of the second moment only. Hence, strictly requires *heavy tasks* with $t = \Omega(k^{1/2})$. We break this second moment barrier by exploiting the boundedness of higher order moments. The heavy tasks are now allowed to be much smaller, but at the cost of requiring a larger number of such tasks and additional computations.

One challenge is that the (empirical) higher order moments are tensors, and tensor norms are not efficiently computable. Hence boundedness alone does not give an efficient clustering algorithm. We need a stronger condition that the moments are Sum-of-Squares (SOS) bounded, i.e. there *exist* SOS proofs showing that the moments are bounded [45, 35]. This SOS boundedness is now tractable with a convex program, leading to a polynomial time algorithm that is also robust against outliers [45]. One caveat is that existing method in [45] requires data generated from a Poincaré distribution. As shown in Remark H.10, the distribution of our estimate $\widehat{\beta}_i = (1/t)\sum_{j=1}^{t} y_{i,j}\mathbf{x}_{i,j}$ is not Poincaré. Interestingly, as we prove in Lemma H.2, the higher order moments are still SOS bounded. This ensures that we can apply the robust clustering algorithm of [45]. We call this step *robust clustering* (Algorithm 7 in §H).

Classification and parameter estimation. Given rough estimates $\widetilde{\mathbf{w}}_\ell$'s as center of those clusters, we grow each cluster by classifying remaining light tasks. Classification only requires $t = \Omega(\log k)$. Once we have sufficiently grown each cluster, we can estimate the parameters to a desired level of accuracy. There are two reasons we need this refinement step. First, in the small corruption regime, where the fraction of corrupted tasks α is much smaller than the desired level of accuracy ϵ, this separation is significantly more sample efficient. The subspace estimation and clustering steps require only $\mathcal{O}(\Delta/\rho)$ accuracy, and the burden of matching the desired ϵ level of error is left to the final classification step, which is more sample efficient. Next, the classification step ensures an adaptive guarantee. As parameter estimation is done for each cluster separately, a cluster with small noise s_i can be more accurately estimated. This ensures a more accurate prediction for newly arriving tasks. We call this step *classification and robust parameter estimation* (Algorithm 9 in §I).

2 Main results

To give a more fine grained analysis, we assume there are two types of light tasks. In meta-learning, subspace estimation uses $\mathcal{D}_{L1}$, clustering uses $\mathcal{D}_H$, and classification uses $\mathcal{D}_{L2}$.

Assumption 1. *The heavy dataset $\mathcal{D}_H$ consists of n_H heavy tasks, each with at least t_H samples. The first light dataset $\mathcal{D}_{L1}$ consists of n_{L1} light tasks, each with at least t_{L1} samples. The second light dataset $\mathcal{D}_{L2}$ consists of n_{L2} tasks, each with at least t_{L2} samples. We assume $t_{L1}, t_{L2} < d$.*

The three batches of meta-train datasets are corrupted by an adversary.

Assumption 2. *From the datasets $\mathcal{D}_H$, $\mathcal{D}_{L1}$, and $\mathcal{D}_{L2}$, the adversary controls α_H, α_{L1}, and α_{L2} fractions of the tasks, respectively. The adversary is allowed to inspect all the examples, remove those examples associated with three subsets of tasks (of sizes at most $\alpha_H n_H$, $\alpha_{L1} n_{L1}$, and $\alpha_{L1} n_{L1}$ tasks from $\mathcal{D}_H$, $\mathcal{D}_{L1}$, and $\mathcal{D}_{L2}$), and replace the examples associated with those tasks with arbitrary points. The corrupted meta-train datasets are then presented to the algorithm.*

2.1 Meta-learning and prediction

We characterize the achievable accuracy in estimating the meta-parameters $\theta = (\mathbf{W}, \mathbf{s}, \mathbf{p})$.

Theorem 1. *For any $\delta \in (0, 1/2)$ and $\epsilon > 0$, given three batches of samples under Assumptions 1 and 2, the meta-learning step of Algorithm 1 achieves the following accuracy for all $i \in [k]$,*

$$\|\widehat{\mathbf{w}}_i - \mathbf{w}_i\|_2 \le \epsilon s_i \ , \quad \left|\widehat{s}_i^2 - s_i^2\right| \le \epsilon s_i^2/\sqrt{t_{L2}} \ , \quad \text{and} \quad |\widehat{p}_i - p_i| \le \epsilon\sqrt{t_{L2}/d}\, p_i \ + \ \alpha_{L2} \ ,$$

with probability $1-\delta$, if the numbers of tasks, samples in each task, and the corruption levels satisfy

$$\begin{aligned}
n_{L1} &= \widetilde{\Omega}\Big(\tfrac{dk^2}{\widetilde{\alpha} t_{L1}} + \tfrac{k}{\widetilde{\alpha}^2}\Big)\,, & t_{L1} &\geq 1\,, & \alpha_{L1} &= \mathcal{O}(\widetilde{\alpha})\,,\\
n_{H} &= \widetilde{\Omega}\Big(\tfrac{(km)^{\Theta(m)}}{p_{\min}} + \tfrac{\rho^4}{\Delta^4 p_{\min} t_H}\Big)\,, & t_H &= \Omega\bigg(\tfrac{\rho^2}{\Delta^2}\cdot\tfrac{m}{p_{\min}^{2/m}}\bigg)\,, & \alpha_H &= \widetilde{\mathcal{O}}\Big(p_{\min}\min\Big\{1, \sqrt{t_H}\cdot\tfrac{\Delta^2}{\rho^2}\Big\}\Big)\,,\\
n_{L2} &= \widetilde{\Omega}\Big(\tfrac{d}{t_{L2} p_{\min}\epsilon^2}\Big)\,, & t_{L2} &= \Omega\Big(\tfrac{\rho^4}{\Delta^4}\log\tfrac{kn_{L2}}{\delta}\Big)\,, & \alpha_{L2} &= \mathcal{O}(p_{\min}\epsilon/\log(1/\epsilon))\,,
\end{aligned}$$

where $\widetilde{\alpha} := \max\big\{\Delta^2\sigma_{\min}^2/(\rho^6 k^2), \Delta^6 p_{\min}^2/(k^2\rho^6)\big\}$, $\sigma_{\min}$ *is the smallest non-zero singular value of* $\sum_{j=1}^k p_j \mathbf{w}_j \mathbf{w}_j^\top$, *and* $m \in \mathbb{N}$ *is a parameter chosen by the analyst.*

We refer to §1.1 for the setup and notations, and provide key lemmas in §B and a complete proof in §C. We discuss each of the conditions in the following remarks assuming $\Delta = \Omega(\rho)$, for simplicity.

Remark 2.1 (Separating two types of light tasks)**.** *As* t_{L1} *can be as small as one, the conditions on* $\mathcal{D}_{L2}$ *does not cover the conditions for* $\mathcal{D}_{L1}$*. The conditions on* α_{L1} *and* n_{L1} *can be significantly more strict than what is required for* $\mathcal{D}_{L2}$*. Hence, we separate the analysis for* $\mathcal{D}_{L1}$ *and* $\mathcal{D}_{L2}$*.*

Remark 2.2 (Dependency in $\mathcal{D}_{L1}$)**.** *Since we are interested the large d small* t_{L1} *setting, the dominant term in* n_{L1} *is* $dk^2/\widetilde{\alpha} t_{L1}$*. The effective sample size* $n_{L1}t_{L1}$ *scaling as d is information theoretically necessary. The* $\min\{1/\sigma_{\min}^2, 1/p_{\min}^2\}$ *dependence of* $n_{L1}t_{L1}$ *allows sample efficiency even when* $\sigma_{\min}$ *is arbitrarily small, including zero. This is a significant improvement over the* $\mathrm{poly}(1/\sigma_k)$ *sample complexity of typical spectral methods, e.g. [14, 77], where* σ_k *is the* k*-th singular value of* $\sum_{\ell=1}^k p_\ell \mathbf{w}_\ell \mathbf{w}_\ell^\top$*. This critically relies on an extension of the gap-free spectral bound of [1, 47]. Our tolerance of* $\alpha_{L1} = \mathcal{O}(p_{\min}^2/k^2)$ *significantly improves upon the state-of-the-art guarantee of* $p_{\min}^4/k^4$ *as detailed in §2.2. Further, we show it is information theoretically optimal. This assumes only bounded fourth moment, which makes our analysis more generally applicable. However, this can be tightened under a stricter conditions of the distribution, as we discuss in §4.*

Remark 2.3 (Dependency in $\mathcal{D}_H$)**.** *Assuming* $p_{\min} = \Omega(1/k)$*, the dominant term in* n_H *is* $\widetilde{\Omega}((km)^{\Theta(m)}/p_{\min})$*, which is* $\widetilde{\Omega}(k^{\Theta(m)})$ *and the result is trivial when* $m \geq \log(k)$*. This implies a* $n_H = \widetilde{\Omega}(k^{\Theta(m)})$*,* $t_H = \Omega(m\cdot k^{2/m})$ *trade-off for any integer* m*, breaking the* $t_H = \Omega(k^{1/2})$ *barrier of [42]. In fact, for an optimal choice of* $m = \Theta(\log k)$ *to minimize the required examples, it can tolerate as small as* $t_H = \Omega(\log k)$ *examples, at the cost of requiring* $n_H = \widetilde{\Omega}(k^{\Theta(\log k)})$ *such heavy tasks. We conjecture* $t_H = \Omega(\log k)$ *is also necessary for any polynomial sample complexity. For the case of learning mixtures of isotropic Gaussians, [59] shows that super-polynomially many number of samples are information theoretically necessary when the centers are* $o(\sqrt{\log k})$ *apart. This translates to* $t = o(\log k)$ *in our setting. The requirement* $\alpha_H = \mathcal{O}(p_{\min})$ *is optimal. Otherwise, the adversary can remove an entire cluster.*

Remark 2.4 (Dependency in $\mathcal{D}_{L2}$)**.** *The requirement* $n_{L2}\cdot t_{L2} = \widetilde{\Omega}(d/p_{\min}\epsilon^2)$ *is optimal in* $d, p_{\min}$ *and* ϵ *due to the lower bound for linear regression. The requirement on* $\alpha_{L2} = \mathcal{O}(p_{\min}\epsilon/\log(1/\epsilon))$ *is also necessary upto a log factor, from lower bound on robust linear regression [26].*

At test time, we use the estimated $\widehat{\theta} = (\widehat{\mathbf{W}}, \widehat{\mathbf{s}}, \widehat{\mathbf{p}})$ to approximate the prior distribution on a new task. On a new arriving task with training data $\mathcal{D} = \{(\mathbf{x}_j^{\text{new}}, y_j^{\text{new}})\}_{j=1}^{\tau}$, we propose the standard MAP or Bayes optimal estimators to make predictions on this new task. The following guarantee is a corollary of Theorem 1 and [42, Theorem 2]. The term $\sum_{i\in[k]} p_i s_i^2$ is due to the noise in the test data $(\mathbf{x}, y)$ and cannot be avoided. We can get arbitrarily close to this fundamental limit with only $\tau = \Omega(\log k)$ samples. This is a minimax optimal sample complexity as shown in [42].

Corollary 2.5 (Prediction)**.** *Under the hypotheses of Theorem 1, the expected prediction errors of both the MAP and Bayes optimal estimators* $\widehat{\beta}(\mathcal{D})$ *defined in Eq.* (3) *are bound as* $\mathbb{E}[(\mathbf{x}^\top\widehat{\beta}(\mathcal{D}) - y)^2] \leq \delta + \big(1+\epsilon^2\big)\sum_{i=1}^k p_i s_i^2$*, if* $\tau = \Omega((\rho^4/\Delta^4)\log(k/\delta))$ *and* $\epsilon \leq \min\{\Delta/(10\rho), \Delta^2\sqrt{d}/(50\rho^2)\}$*, where the expectation is over the new task with model parameter* $\phi^{\text{new}} = (\beta^{\text{new}}, \sigma^{\text{new}}) \sim \mathbb{P}_\theta$*, training data* $(\mathbf{x}_j^{\text{new}}, y_j^{\text{new}}) \sim \mathbb{P}_{\phi^{\text{new}}}$*, and test data* $(\mathbf{x}, y) \sim \mathbb{P}_{\phi^{\text{new}}}$*.*

2.2 Novel robust subspace estimation

Our main result relies on making *each step* of Algorithm 1 robust, as detailed in §B. However, as our key innovation is a novel *robust subspace estimation* in the first step, we highlight it in this section.

We aim to estimate the subspace spanned by the true meta-parameters $\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$. As $\mathbf{\Sigma} := \mathbb{E}[\widehat{\beta}_{i,j}\widehat{\beta}_{i,j}^\top] = \{\sum_{\ell=1}^k p_\ell(s_\ell^2 + \|\mathbf{w}_\ell\|_2^2)\}\mathbf{I} + 2\sum_{\ell=1}^k p_\ell \mathbf{w}_\ell \mathbf{w}_\ell^\top$ for $\widehat{\beta}_{i,j}$ in Algorithm 2 line 2, we can

Algorithm 2 Robust subspace estimation

```
Input: Data $\mathcal{D}_{L1} = \{\{(\mathbf{x}_{i,j}, y_{i,j})\}_{j=1}^{t_{L1}}\}_{i=1}^{n_{L1}}$, $\alpha \in (0, 1/36]$, $\delta \in (0, 0.5)$, $k \in \mathbb{N}$, and $\nu \in \mathbb{R}_+$
$\widehat{\beta}_{i,j} \leftarrow y_{i,j}\mathbf{x}_{i,j}$ ,    for all $i \in [n_{L1}], j \in [t_{L1}]$
$S_0 \leftarrow \{\widehat{\beta}_{i,j}\widehat{\beta}_{i,j}^\top\}_{i\in[n_{L1}],j\in[t_{L1}]}$ , and $S_{\max} \leftarrow \emptyset$
for $\ell = 1, \ldots, \log_6(2/\delta)$ do
    $t \leftarrow 0$ and $S_{-1} \leftarrow \emptyset$
    while $t \le \lceil 9\alpha n \rceil$ and $S_t \neq S_{t-1}$ do
        $t \leftarrow t + 1$, and  $S_t \leftarrow$ Double-Filtering$(S_{t-1}, k, \alpha, \nu)$        [See Algorithm 3]
    if $|S_{\max}| < |S_t|$ then $S_{\max} \leftarrow S_t$
Output: $\widehat{\mathbf{U}} \leftarrow k$_SVD$\big( \sum_{\widehat{\beta}_{i,j}\in S_{\max}} \widehat{\beta}_{i,j}\widehat{\beta}_{i,j}^\top \big)$
```

use the k empirical principal components; this requires uncorrupted data. To remove the corrupted datapoints, we introduce *double filtering*. We repeat $\log_6(2/\delta)$ times for a high probability result.

Algorithm 3 Double-Filtering

```
Input: a set of PSD matrices $S = \{\mathbf{X}_i \in \mathbb{R}^{d\times d}\}_{i\in[n]}$, $k \in \mathbb{N}$, $\alpha \in (0, 1/36]$ and $\nu \in \mathbb{R}_+$
$\mathcal{S}_0 \leftarrow [n]$, $\mathbf{U}_0 \leftarrow k$_SVD$\big( \sum_{i\in\mathcal{S}_0} \mathbf{X}_i \big)$, and $z_i \leftarrow \mathrm{Tr}\big[\mathbf{U}_0^\top \mathbf{X}_i \mathbf{U}_0\big]$ for all $i \in \mathcal{S}_0$
$\mathcal{S}_G \leftarrow$ First-Filter$(\{z_i\}_{i\in\mathcal{S}_0}, \alpha)$        [Remove the upper and lower $2\alpha$ quantiles]
$\mu^{\mathcal{S}_0} \leftarrow (1/n)\sum_{i\in\mathcal{S}_0} z_i$  and  $\mu^{\mathcal{S}_G} \leftarrow (1/|\mathcal{S}_G|)\sum_{i\in\mathcal{S}_G} z_i$
if $\mu^{\mathcal{S}_0} - \mu^{\mathcal{S}_G} \le 48(\alpha\mu^{\mathcal{S}_G} + \nu\sqrt{k\alpha})$ then Output: $S$ [Sample mean not large, no need to filter.]
else        [Run a second filter if sample mean is corrupted]
    $Z \sim \mathcal{U}[0, 1]$,   $W \leftarrow Z \max\{z_i - \mu^{\mathcal{S}_G}\}_{i\in\mathcal{S}_0\setminus\mathcal{S}_G}$
    $\mathcal{S}_1 \leftarrow \mathcal{S}_G \cup \{i \in \mathcal{S}_0 \setminus \mathcal{S}_G \mid z_i - \mu^{\mathcal{S}_G} \le W\}$        [Add some removed points back.]
Output: $S' = \{\mathbf{X}_i\}_{i\in\mathcal{S}_1}$
```

If the adversarial examples have the outer product $\mathbf{X}_i = \widehat{\beta}_{i',j'}\widehat{\beta}_{i',j'}^\top$'s with small norms, then they are challenging to detect. However, such undetectable corruptions can only perturb the subspace by little. Hence, Algorithm 3 focuses on detecting large corruptions. Ideally, we want to find a subspace by

$$\widehat{\mathbf{U}} \;\leftarrow\; \operatorname*{arg\,max}_{\mathbf{U}\in\mathbb{R}^{d\times k}:\mathbf{U}^\top\mathbf{U}=\mathbf{I}_k} \;\operatorname*{minimize}_{\mathcal{S}'\subseteq[n]:|\mathcal{S}'|\ge(1-\alpha)n} \;\sum_{i\in\mathcal{S}'} \underbrace{\mathrm{Tr}[\mathbf{U}^\top\mathbf{X}_i\mathbf{U}]}_{:=z_i} ,$$

for $n = n_{L1}t_{L1}$, which is computationally intractable. This relies on the intuition that a good subspace preserves the second moment, even when large (potentially corrupted) points are removed.

We propose a *filtering* approach in Algorithm 3. At each iteration, we alternate between finding a candidate semi-orthogonal matrix $\mathbf{U}_0 \in \mathbb{R}^{d\times k}$ containing the top-k singular vectors using the k_SVD routine and then filtering out suspected corrupted data points, which have large trace norms in $\mathbf{U}_0$. Existing filtering approaches (e.g. [75]) use a *single filter* to remove examples with large trace norm (denoted by z_i in Algorithm 3). This suffers from removing too many *uncorrupted* examples. We give a precise comparison in Eq. (6). We instead use two filters to add back some of those mistakenly removed points. The First-Filter partitions the input set into a good set $\mathcal{S}_G$ and a bad set $\mathcal{S}_0 \setminus \mathcal{S}_G$. If the bad set contributed to a significant portion of the projected trace (this can be detected by the shift in the mean of the remaining points $\mu^{\mathcal{S}_G}$), a second filter is applied to the bad set, recovering some of the uncorrupted examples.

This algorithm and our analysis applies more generally to any random vector, and may be of independent interest in other applications requiring robust PCA. Under a mild assumption that $\mathbf{x}_i \sim \mathcal{P}$ has a bounded fourth-moment, we prove the following, with a proof in §D.1.

Proposition 2.6 (Robust PCA for general PSD matrices)**.** *Let* $S = \{\mathbf{x}_i \sim \mathcal{P}\}_{i=1}^n$ *where* $\mathbf{\Sigma} := \mathbb{E}_{\mathbf{x}\sim\mathcal{P}}\big[\mathbf{x}\mathbf{x}^\top\big]$ *is the second moment of* $\mathcal{P}$ *supported on* $\mathbb{R}^d$*. Given* $k \in \mathbb{N}$, $\delta \in (0, 0.5)$*, and a corrupted dataset* S' *with* $\alpha \in (0, 1/36]$ *fraction corrupted arbitrarily, if* $\mathcal{P}$ *has a bounded support such that* $\|\mathbf{x}\mathbf{x}^\top - \mathbf{\Sigma}\|_2 \le B$ *for* $\mathbf{x} \sim \mathcal{P}$ *with probability one, and a bounded* 4*-th moment such that* $\max_{\|\mathbf{A}\|_{\mathrm{F}}\le 1, \mathrm{rank}(\mathbf{A})\le k} \mathbb{E}_{\mathbf{x}\sim\mathcal{P}}\Big[\big(\langle \mathbf{A}, \mathbf{x}\mathbf{x}^\top - \mathbf{\Sigma}\rangle\big)^2\Big] \le \nu^2$*, and* $n = \Omega((dk^2 +$

$(B/\nu)\sqrt{k\alpha})\log(d/(\delta\alpha))/\alpha)$*, then with probability at least* $1-\delta$*,*

$$\mathrm{Tr}[\mathcal{P}_k(\mathbf{\Sigma})] - \mathrm{Tr}\Big[\widehat{\mathbf{U}}^\top \mathbf{\Sigma} \widehat{\mathbf{U}}\Big] = \mathcal{O}\Big(\alpha\, \mathrm{Tr}[\mathcal{P}_k(\mathbf{\Sigma})] + \nu\sqrt{k\alpha}\Big) \;, \tag{4}$$

$$\textit{and} \quad \Big\|\mathbf{\Sigma} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \mathbf{\Sigma} \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\Big\|_* \le \|\mathbf{\Sigma} - \mathcal{P}_k(\mathbf{\Sigma})\|_* + \mathcal{O}\Big(\alpha\|\mathcal{P}_k(\mathbf{\Sigma})\|_* + \nu\sqrt{k\alpha}\Big) \;. \tag{5}$$

where $\widehat{\mathbf{U}}$ *is the output of Algorithm 2, and* $\mathcal{P}_k(\cdot)$ *is the best rank-*k *approximation of a matrix in* ℓ_2*.*

The first term in the RHS of Eq. (5) is unavoidable, as we are outputting a rank-k subspace. In the setting of Theorem 1 in which we are interested in, the last term of $\nu\sqrt{k\alpha}$ in Equation (4) dominates the second term. We next show that this cannot be improved upon; no algorithm can learn the subspace with an additive error smaller than $\Omega(\nu\sqrt{k\alpha})$ under α fraction of corruption, even with infinite samples. In the following minimax lower bound, since the total variation distance $D_{\mathrm{TV}}(\mathcal{P},\mathcal{P}') \le \alpha$, the adversary can corrupted the datapoints from $\mathcal{P}'$ to match the distribution $\mathcal{P}$, by changing just the α fraction. It is impossible to tell if the corrupted samples came from $\mathcal{P}$ or $\mathcal{P}'$, resulting in an $\mathcal{O}(\nu\sqrt{k\alpha})$ error.

Proposition 2.7 (Information theoretic lower bound)**.** *Let* $\widehat{\mathbf{U}}(\{\mathbf{x}_i\}_{i=1}^n)$ *be any subspace estimator that takes* n *samples from distribution* $\mathcal{P}$ *as input, and estimates the* k *principal components of* $\mathbf{\Sigma} := \mathbb{E}_{\mathbf{x}\sim\mathcal{P}'}\big[\mathbf{x}\mathbf{x}^\top\big]$ *from another distribution* $\mathcal{P}'$ *that is* α*-close in total variation* D_{TV}*. Then,*

$$\inf_{\widehat{\mathbf{U}}} \max_{\mathcal{P}'\in\Theta_{\nu,B}} \max_{\mathcal{P}:D_{\mathrm{TV}}(\mathcal{P},\mathcal{P}')\le\alpha} \mathop{\mathbb{E}}_{\{\mathbf{x}_i\}_{i=1}^n\sim\mathcal{P}^n}\Big[\Big\|\mathbf{\Sigma} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{\Sigma}\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\Big\|_* - \|\mathbf{\Sigma} - \mathcal{P}_k(\mathbf{\Sigma})\|_*\Big] \;=\; \Omega(\nu\sqrt{k\alpha}),$$

for any $k \ge 16, d \ge k^2/\alpha$*, and* $B \ge 2d\nu$*, where* $\Theta_{\nu,B}$ *is a set of all distributions* $\mathcal{D}'$ *on* $\mathbb{R}^d$ *such that* $\max_{\|\mathbf{A}\|_{\mathrm{F}}\le 1}\mathbb{E}_{\mathbf{x}\sim\mathcal{D}'}[(\langle\mathbf{A},\mathbf{x}\mathbf{x}^\top - \mathbf{\Sigma}\rangle)^2] \le \nu^2$*, and* $\mathbb{P}_{\mathbf{x}\sim\mathcal{D}'}\big[\big\|\mathbf{x}\mathbf{x}^\top - \mathbb{E}[\mathbf{x}\mathbf{x}^\top]\big\|_2 \le B\big] = 1$.

Comparisons with [75]. Outlier-Robust Principal Component Analysis (ORPCA) [75, 28, 76] studies a similar problem under a Gaussian model. For comparison, we can modify the best known ORPCA estimator from [75] to our setting in Proposition 2.6, to get a semi-orthogonal $\widehat{\mathbf{U}}$ achieving

$$\Big\|\mathbf{\Sigma} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{\Sigma}\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\Big\|_* = \|\mathbf{\Sigma} - \mathcal{P}_k(\mathbf{\Sigma})\|_* + \mathcal{O}\big(\,\alpha^{1/2}\|\mathcal{P}_k(\mathbf{\Sigma})\|_* + \nu k\alpha^{1/4}\,\big) \;. \tag{6}$$

We significantly improve in the dominant third term (see Eq. (5)). Simulation results supporting our theoretical prediction are shown in Fig. 2. For the analysis and the experimental setup we refer to §K.

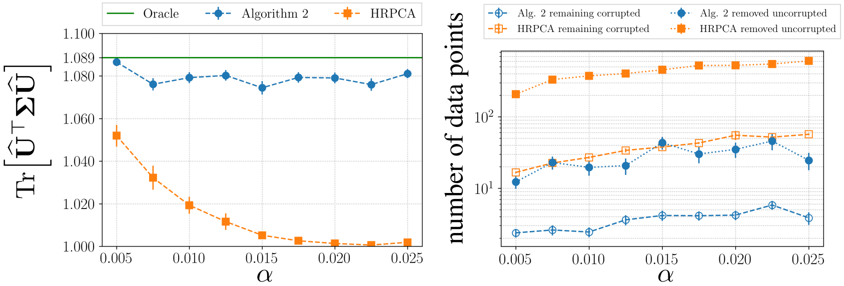


Figure 2: Algorithm 2 performs close to an oracle which knows the corrupted points, improving upon HRPCA of [75], by removing more corrupted points and less uncorrupted ones.

3 Related work

Mixed linear regression. Previous work on mixed linear regression focus on the setting where each task has only one sample, i.e. $t_i = 1$. As a consequence, all the previous work suffer from either the sample complexity or the running time that scale exponentially in k (specifically at least $\exp(\sqrt{k})$) [78, 47, 16, 64]. In other cases, such blow-up in complexity is hidden in the dependence of the inverse of k-th singular value of a moment matrix, which can be arbitrarily large [14, 77, 63].

Multi-task learning. [8, 3, 60, 53, 4, 32, 2, 54, 7, 11] address a similar problem as our setting, but focusing on finding a common low-dimensional linear representation, where all tasks can be accurately solved. Typically, the batch size is fixed and the performance is evaluated on the past tasks used in training. Close to ours are a few concurrent work [27, 69], but their focus is still on recovering the common subspace, and not the meta-parameters.

Robust regression. There are several work on robust linear regression and sparse regression problems, [10, 9, 6, 30, 55, 40, 23, 48, 39, 21, 51, 15, 57, 38]. The recent advances in the list-decodable setting [15, 57, 38] can potentially be applied to our mixture setting, but the sample complexity is exponentially large. Recently, [64] studies the robust mixed linear regression problem. In contrast to our setting which allows random noise on the label and adversarial corruption on both covariate $\mathbf{x}$ and label y, their setting assumes no noise on label y's, and the adversary is only allowed to corrupt α-fraction the label y's. Although their algorithm requires only $\widetilde{\mathcal{O}}(dk)$ samples, the running time is $\widetilde{\mathcal{O}}(k^k nd)$ and also requires a good estimate of the subspace spanned by $\{\mathbf{w}_\ell\}_{\ell=1}^k$.

Sum-of-squares algorithms. Our work is inspired by sum-of-squares algorithms that have recently been studied on many learning problems, including linear regression [38, 57], mixture models [35, 46, 37, 56, 22], mean estimation [34, 20], subspace estimation [5]. This provides the key building block of our approach, in breaking the second moment barrier of linkage-based clustering algorithms.

4 Conclusion

By exploiting similarities on a collection of related but different tasks, meta-learning predicts a newly arriving task with a far greater accuracy than what can be achieved in isolation. We ask two fundamental questions under a canonical model of k-mixed linear regression: (i) can we meta-learn from tasks with only a few training examples each?; and (ii) can we meta-learn from tasks when only part of the data can be trusted? We introduce a novel spectral approach that achieves both simultaneously, significantly improving the required batch size from $\Omega(k^{1/2})$ to $\Omega(\log k)$ while being robust to adversarial corruption. We use a sum-of-squares algorithm to exploit the higher order moments and design a novel robust subspace estimation algorithm that achieves optimal guarantees.

Closing the gap in robust subspace estimation. [75, 28, 76, 19] study robust PCA under the Gaussian assumption. For the reasons explained in §2.2, the rate is sub-optimal in α in comparisons to an information theoretic lower bound with a multiplicative factor of $(1-\Theta(\alpha))$. Applying the proposed Algorithm 2, it is possible to generalize Proposition 2.6 to this Gaussian setting and achieve an optimal upper bound. We leave this as a future research direction, and provide a sketch of how to adapt the proof of our algorithm to the exponential tail setting in §E.

Concretely, our analysis of Algorithm 2 assumes only a bounded 4-th moment of the input vector $\mathbf{z}_i$, of the form $\mathbb{P}\big[\big|(\mathbf{v}^\top \mathbf{z}_i)^2 - \mathbf{v}^\top \mathbf{\Sigma} \mathbf{v}\big| \geq t\big] \leq c\,t^{-2}$. Our current proof proceeds by focusing on that $1-\alpha$ probability mass, which falls in the interval $[-\sqrt{1/\alpha}, \sqrt{1/\alpha}]$. This is tight with only the second moment assumption. More generally, one can consider a family of distributions satisfying $\mathbb{P}\big[\big|(\mathbf{v}^\top \mathbf{z}_i)^2 - \mathbf{v}^\top \mathbf{\Sigma} \mathbf{v}\big| \geq \text{variance} \cdot t\big] \leq \exp(-t^\gamma)$. If we have such an exponential concentration, we can instead focus on the subset of examples with second moment $\big|(\mathbf{v}^\top \mathbf{z}_i)^2 - \mathbf{v}^\top \mathbf{\Sigma} \mathbf{v}\big|$ falling in the interval $[-\log^{1/\gamma}(1/\alpha), \log^{1/\gamma}(1/\alpha)]$. This bounded distribution has a sub-Gaussian norm $\sqrt{k}\log^{1/\gamma}(1/\alpha)$, and thus we can apply the sub-Gaussian filter (Proposition A.7 of [24]) to learn $\mathbb{E}\big[(\mathbf{v}^\top \mathbf{z}_i)^2\big]$ with error We can obtain an error of $\alpha\sqrt{k}\log^{1/\gamma}(1/\alpha)$. We provide a sketch of how to adapt the proof of our algorithm to the exponential tail setting in §E.

After submission, we became aware of an independent and concurrent result by Jambulapati et al. [36] which studies the robust PCA problem under the assumption that each datapoint $\mathbf{x}$ follows a sub-Gaussian distribution. Their algorithm is very similar to ours, except that it is only applied to estimating the top eigenvector of the covariance matrix, which corresponds to the $k=1$ special case in our setting. Their sample complexity and recovery guarantee are identical to ours in §E.

Removing the Gaussianity assumption. Our approach relies on the special structure of the 4-th moment of $\mathbf{x}_{i,j}$ and the SOS boundedness of higher order moments of $\mathbf{x}_{i,j}$. The approach in [42] is able to get around the 4-th moment requirement, and it is an interesting open problem to make the approach robust to outliers and still preserve the $\widetilde{\mathcal{O}}(d)$ sample complexity. while this class of SOS bounded distributions is fairly broad, as noted in [45], one could hope to establish sum-of-squares bounds for even broader families. For examples, it remains open that whether sum-of-squares certifies moment tensors for all sub-Gaussian distributions.

Broader Impact

One of the main contribution of this paper is to protect meta-learning approaches against data poisoning attacks. Such robustness encourages participation from data contributors, as they can collaborate without necessarily trusting the other data contributors. This facilitates participation of minor contributors who suffer from data scarcity. This fosters democratization of machine learning by allowing minor contributors to enjoy the benefit of big data through collaboration. Such ecosystem will also encourage data sharing, thus improving transparency.

The adaptive guarantee we provide in Theorem 1 is fair, in the sense that a group that provides low noise data will receive a model with better accuracy. However, one potential risk in fairness is that meta-learning might result in varying accuracy across the groups. This can be problematic as an under-represented group in training data could suffer from inaccurate prediction for that population. This is an active area of research in the fairness community, but there is no strong experimental evidence that this can be mitigated with algorithmic innovations that do not involve collecting more data from the under-represented population.

Another concern in meta-learning with data sharing is privacy. Without proper system to regulate the usage of shared data, sensitive information could be leaked or protected features could be inferred. One silver lining is that robust methods are naturally private, as the trained model is by definition not sensitive to any one particular data point. On the other hand, if the system relies on the participation of various individuals, then either a technological solution needs to be implemented with cryptographic or privacy preserving primitives, or a proper regulation must be enforced.

Acknowledgments and Disclosure of Funding

Sham Kakade acknowledges funding from the NSF Awards CCF-1637360, CCF-1703574, and CCF-1740551. Sewoong Oh acknowledges funding from Google Faculty Research Award, and NSF awards CCF-1927712, CCF-1705007, IIS-1929955, CNS-2002664, and 2019844.

References

[1] Zeyuan Allen-Zhu and Yuanzhi Li. Lazysvd: even faster svd decomposition yet without agonizing pain. In *NIPS*. arXiv:1607.03463, 2016.

[2] Yonatan Amit, Michael Fink, Nathan Srebro, and Shimon Ullman. Uncovering shared structures in multiclass classification. In *Proceedings of the 24th international conference on Machine learning*, pages 17–24, 2007.

[3] Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(Nov):1817–1853, 2005.

[4] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine learning*, 73(3):243–272, 2008.

[5] Ainesh Bakshi and Pravesh Kothari. List-decodable subspace recovery via sum-of-squares. *arXiv preprint arXiv:2002.05139*, 2020.

[6] S. Balakrishnan, S. S. Du, J. Li, and A. Singh. Computationally efficient robust sparse estimation in high dimensions. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017*, pages 169–212, 2017.

[7] Maria-Florina Balcan, Avrim Blum, and Santosh Vempala. Efficient representations for lifelong learning and autoencoding. In *Conference on Learning Theory*, pages 191–210, 2015.

[8] Jonathan Baxter. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

[9] K. Bhatia, P. Jain, P. Kamalaruban, and P. Kar. Consistent robust regression. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 2107–2116, 2017.

[10] Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pages 721–729, 2015.

[11] Brian Bullins, Elad Hazan, Adam Kalai, and Roi Livni. Generalize across tasks: Efficient algorithms for linear representation learning. In *Algorithmic Learning Theory*, pages 235–246, 2019.

[12] Emmanuel J Candes and Yaniv Plan. 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.

[13] Bradley P Carlin and Thomas A Louis. *Bayes and empirical Bayes methods for data analysis*. Chapman and Hall/CRC, 2010.

[14] Arun Tejasvi Chaganty and Percy Liang. Spectral experts for estimating mixtures of linear regressions. In *International Conference on Machine Learning (ICML)*, pages 1040–1048, 2013.

[15] Moses Charikar, Jacob Steinhardt, and Gregory Valiant. Learning from untrusted data. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 47–60, 2017.

[16] Sitan Chen, Jerry Li, and Zhao Song. Learning mixtures of linear regressions in subexponential time via Fourier moments. In *STOC*. `https://arxiv.org/pdf/1912.07629.pdf`, 2020.

[17] Yu Cheng, Ilias Diakonikolas, and Rong Ge. High-dimensional robust mean estimation in nearly-linear time. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2755–2771. SIAM, 2019.

[18] Yeshwanth Cherapanamjeri, Nicolas Flammarion, and Peter L Bartlett. Fast mean estimation with sub-gaussian rates. *arXiv preprint arXiv:1902.01998*, 2019.

[19] Yeshwanth Cherapanamjeri, Prateek Jain, and Praneeth Netrapalli. Thresholding based efficient outlier robust pca. *arXiv preprint arXiv:1702.05571*, 2017.

[20] Yeshwanth Cherapanamjeri, Sidhanth Mohanty, and Morris Yau. List decodable mean estimation in nearly linear time. *arXiv preprint arXiv:2005.09796*, 2020.

[21] Arnak Dalalyan and Philip Thompson. Outlier-robust estimation of a sparse linear model using ℓ_1-penalized huber's m-estimator. In *Advances in Neural Information Processing Systems*, pages 13188–13198, 2019.

[22] Ilias Diakonikolas, Samuel B Hopkins, Daniel Kane, and Sushrut Karmalkar. Robustly learning any clusterable mixture of gaussians. *arXiv preprint arXiv:2005.06417*, 2020.

[23] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, pages 1596–1606, 2019.

[24] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Ankur Moitra, and Alistair Stewart. Being Robust (in High Dimensions) Can Be Practical. *arXiv e-prints*, page arXiv:1703.00893, March 2017.

[25] Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. List-decodable robust mean estimation and learning mixtures of spherical gaussians. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1047–1060, 2018.

[26] Ilias Diakonikolas, Weihao Kong, and Alistair Stewart. Efficient algorithms and lower bounds for robust linear regression. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2745–2754. SIAM, 2019.

[27] Simon S Du, Wei Hu, Sham M Kakade, Jason D Lee, and Qi Lei. Few-shot learning via learning the representation, provably. *arXiv preprint arXiv:2002.09434*, 2020.

[28] Jiashi Feng, Huan Xu, and Shuicheng Yan. Robust pca in high-dimension: A deterministic approach. *arXiv preprint arXiv:1206.4628*, 2012.

[29] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1126–1135, 2017.

[30] Chao Gao et al. Robust regression via mutivariate regression depth. *Bernoulli*, 26(2):1139–1170, 2020.

[31] Erin Grant, Chelsea Finn, Sergey Levine, Trevor Darrell, and Thomas Griffiths. Recasting gradient-based meta-learning as hierarchical bayes. *arXiv preprint arXiv:1801.08930*, 2018.

[32] Zaid Harchaoui, Matthijs Douze, Mattis Paulin, Miroslav Dudik, and Jérôme Malick. Large-scale image classification with trace-norm regularization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3386–3393. IEEE, 2012.

[33] Samuel B Hopkins. Mean estimation with sub-gaussian rates in polynomial time. *arXiv preprint arXiv:1809.07425*, 2018.

[34] Samuel B Hopkins et al. Mean estimation with sub-gaussian rates in polynomial time. *Annals of Statistics*, 48(2):1193–1213, 2020.

[35] Samuel B Hopkins and Jerry Li. Mixture models, robustness, and sum of squares proofs. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1021–1034, 2018.

[36] Arun Jambulapati, Jerry Li, and Kevin Tian. Robust sub-gaussian principal component analysis and width-independent schatten packing. *arXiv preprint arXiv:2006.06980*, 2020.

[37] He Jia and Santosh Vempala. Robustly clustering a mixture of gaussians. *arXiv preprint arXiv:1911.11838*, 2019.

[38] Sushrut Karmalkar, Adam Klivans, and Pravesh Kothari. List-decodable linear regression. In *Advances in Neural Information Processing Systems*, pages 7423–7432, 2019.

[39] Sushrut Karmalkar and Eric Price. Compressed sensing with adversarial sparse noise via l1 regression. In *2nd Symposium on Simplicity in Algorithms*, 2019.

[40] Adam Klivans, Pravesh K Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, pages 1420–1430, 2018.

[41] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 2015.

[42] Weihao Kong, Raghav Somani, Zhao Song, Sham Kakade, and Sewoong Oh. Meta-learning for mixed linear regression. *arXiv e-prints*, page arXiv:2002.08936, February 2020.

[43] Weihao Kong and Gregory Valiant. Estimating learnability in the sublinear data regime. In *Advances in Neural Information Processing Systems*, pages 5455–5464, 2018.

[44] Weihao Kong, Gregory Valiant, and Emma Brunskill. Sublinear optimal policy value estimation in contextual bandits. *arXiv preprint arXiv:1912.06111*, 2019.

[45] Pravesh K. Kothari and Jacob Steinhardt. Better Agnostic Clustering Via Relaxed Tensor Norms. *arXiv e-prints*, page arXiv:1711.07465, November 2017.

[46] Pravesh K Kothari, Jacob Steinhardt, and David Steurer. Robust moment estimation and improved clustering via sum of squares. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1035–1046, 2018.

[47] Yuanzhi Li and Yingyu Liang. Learning mixtures of linear regressions with nearly optimal complexity. In *COLT*. arXiv preprint arXiv:1802.07895, 2018.

[48] Liu Liu, Yanyao Shen, Tianyang Li, and Constantine Caramanis. High dimensional robust sparse regression. *arXiv preprint arXiv:1805.11643*, 2018.

[49] Gábor Lugosi and Shahar Mendelson. Mean estimation and regression under heavy-tailed distributions: A survey. *Foundations of Computational Mathematics*, 19(5):1145–1190, 2019.

[50] Gabor Lugosi and Shahar Mendelson. Robust multivariate mean estimation: the optimality of trimmed mean. *arXiv preprint arXiv:1907.11391*, 2019.

[51] Bhaskar Mukhoty, Govind Gopakumar, Prateek Jain, and Purushottam Kar. Globally-convergent iteratively reweighted least squares for robust regression problems. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 313–322, 2019.

[52] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Advances in Neural Information Processing Systems*, pages 721–731, 2018.

[53] Alon Orlitsky. Supervised dimensionality reduction using mixture models. In *Proceedings of the 22nd international conference on Machine learning*, pages 768–775, 2005.

[54] Massimiliano Pontil and Andreas Maurer. Excess risk bounds for multitask learning with trace norm regularization. In *Conference on Learning Theory*, pages 55–76, 2013.

[55] A. Prasad, A. S. Suggala, S. Balakrishnan, and P. Ravikumar. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

[56] Prasad Raghavendra, Tselil Schramm, and David Steurer. High-dimensional estimation via sum-of-squares proofs. *arXiv preprint arXiv:1807.11419*, 6, 2018.

[57] Prasad Raghavendra and Morris Yau. List decodable learning via sum of squares. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 161–180. SIAM, 2020.

[58] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *International Conference on Representation Learning*, 2017.

[59] Oded Regev and Aravindan Vijayaraghavan. On learning mixtures of well-separated gaussians. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 85–96. IEEE, 2017.

[60] Irina Rish, Genady Grabarnik, Guillermo Cecchi, Francisco Pereira, and Geoffrey J Gordon. Closed-form supervised dimensionality reduction with generalized linear models. In *Proceedings of the 25th international conference on Machine learning*, pages 832–839, 2008.

[61] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018.

[62] Jürgen Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München, 1987.

[63] Hanie Sedghi, Majid Janzamin, and Anima Anandkumar. Provable tensor methods for learning mixtures of generalized linear models. In *Artificial Intelligence and Statistics (AISTATS)*, pages 1223–1231, 2016.

[64] Yanyao Shen and Sujay Sanghavi. Iterative least trimmed squares for mixed linear regression. In *Advances in Neural Information Processing Systems*, pages 6076–6086, 2019.

[65] Jacob Steinhardt. Lecture notes for stat260 (robust statistics).

[66] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[67] Sebastian Thrun and Lorien Pratt. *Learning to learn*. Springer Science & Business Media, 2012.

[68] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, and Hugo Larochelle. Meta-dataset: A dataset of datasets for learning to learn from few examples. *arXiv preprint arXiv:1903.03096*, 2019.

[69] Nilesh Tripuraneni, Chi Jin, and Michael I Jordan. Provable meta-learning of linear representations. *arXiv preprint arXiv:2002.11684*, 2020.

[70] Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

[71] Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

[72] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[73] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Advances in Neural Information Processing Systems*, pages 7029–7039, 2017.

[74] Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

[75] Huan Xu, Constantine Caramanis, and Shie Mannor. Outlier-robust pca: The high-dimensional case. *IEEE transactions on information theory*, 59(1):546–572, 2012.

[76] Wenzhuo Yang and Huan Xu. A unified framework for outlier-robust pca-like algorithms. In *ICML*, pages 484–493, 2015.

[77] Xinyang Yi, Constantine Caramanis, and Sujay Sanghavi. Solving a mixture of many random linear equations by tensor decomposition and alternating minimization. *arXiv preprint arXiv:1608.05749*, 2016.

[78] Kai Zhong, Prateek Jain, and Inderjit S Dhillon. Mixed linear regression with multiple components. In *Advances in neural information processing systems (NIPS)*, pages 2190–2198, 2016.

Appendix

A Proof of Corollary 1.2, Corollary 1.3, Corollary 1.1

Algorithm 4 Meta-learning without adversarial corruptions

```
Meta-learning
    1. Subspace estimation: Compute subspace $\widehat{\mathbf{U}}$ with [42, Algorithm 2] which approximates
       $\mathrm{span}\{\mathbf{w}_1, \ldots, \mathbf{w}_k\}$.
    2. Clustering: Project the heavy tasks onto the subspace of $\widehat{\mathbf{U}}$, perform $k$ clustering with
       Algorithm 7, and estimate $\widetilde{\mathbf{w}}_\ell$. Also estimate $\widetilde{r}_\ell^2$ using Algorithm 8 for each cluster $\ell \in [k]$.
    3. Classification: Perform likelihood-based classification of the light tasks using $\{\widetilde{\mathbf{w}}_\ell, \widetilde{r}_\ell^2\}_{\ell=1}^k$
       estimated from the Clustering step; compute refined estimates $\{\widehat{\mathbf{w}}_\ell, \widehat{s}_\ell, \widehat{p}_\ell\}_{\ell=1}^k$ of $\theta$ using [42,
       Algorithm 4].
Prediction
    4. Prediction: Perform MAP or Bayes optimal prediction using the estimated meta-parameter.
```

We assume that the meta-parameter satisfies $\Delta = \Theta(1)$, $\rho = \Theta(1)$, $p_{\min} = \Theta(1/k)$ for Corollary 1.2, Corollary 1.3, Corollary 1.1.

Proof of Corollary 1.1. Recall that for this corollary, we assume $\Delta = \Theta(1)$, $\rho = \Theta(1)$, $p_{\min} = \Theta(1/k)$ and there is no adversarial corruption. Thus we execute Algorithm 4 in this setting. We invoke [42, Lemma 5.1] and get that given $t = \Omega(1)$ and $tn = \widetilde{\Omega}(kn)$, the estimated subspace $\widehat{\mathbf{U}}$ satisfies that for all $i \in [k]$,

$$\left\|\left(\mathbf{I} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\right)\mathbf{w}_\ell\right\|_2 \;\le\; \Delta/(10) \quad \forall\, \ell \in [k] \;, \tag{7}$$

Then, we can invoke Lemma B.4 with $m = \Theta(\log k)$, $t = \Theta(\log k)$, $n = k^{\Theta(\log k)}$ and get that the estimated parameters $\widetilde{\mathbf{w}}_\ell$, $\widetilde{r}_\ell^2$ satisfy

$$\|\widetilde{\mathbf{w}}_\ell - \mathbf{w}_\ell\|_2 \le \frac{\Delta}{10} \;, \qquad \text{and} \qquad \left|\widetilde{r}_\ell^2 - r_\ell^2\right| \le r_\ell^2 \frac{\Delta^2}{50\rho^2} \;, \qquad \forall\, \ell \in [k] \;,$$

Finally, given $t = \widetilde{\Omega}(1)$ and $n = \Omega(dk)$, the output of the classification step satisfies

$$\|\widehat{\mathbf{w}}_\ell - \mathbf{w}_\ell\|_2 = \mathcal{O}(1) \;, \quad \left|\widehat{s}_\ell^2 - s_\ell^2\right| = \mathcal{O}\big(s_\ell^2\big) \;, \quad \text{and} \quad |\widehat{p}_\ell - p_\ell| = \mathcal{O}(p_\ell) \quad \forall\, \ell \in [k]. \tag{8}$$

To conclude, with $t = \widetilde{\Omega}(1), n = \widetilde{\Omega}(dk^2 + k^{\Theta(\log k)})$, Algorithm 4 can estimate model parameters θ with arbitrary small constant error. □

Proof of Corollary 1.2. The proof is the same as Corollary 1.2. □

Proof of Corollary 1.3. With the assumptions that $\Delta = \Omega(1)$, $\rho = \Omega(1)$, $p_{\min} = \Omega(1/k)$, $t_{L1} = t_{L2} = \widetilde{\Omega}(1)$, $t_H = \Omega(mk^{1/m})$ Theorem 1 can be simplified to that for all $i \in [k]$, with probability $1 - \delta$

$$\|\widehat{\mathbf{w}}_i - \mathbf{w}_i\|_2 \le \epsilon s_i \;, \quad \left|\widehat{s}_i^2 - s_i^2\right| \le \epsilon s_i^2 \;, \quad \text{and} \quad |\widehat{p}_i - p_i| \le \epsilon p_i \;+\; \alpha_{L2} \;,$$

as long as,

$$\begin{aligned} n_{L1} &= \widetilde{\Omega}\Big(\tfrac{dk^2}{\widetilde{\alpha}}\Big) \;, & \alpha_{L1} &= \mathcal{O}(\widetilde{\alpha}) \;, \\ n_H &= \widetilde{\Omega}\big((km)^{\Theta(m)}\big) \;, & \alpha_H &= \widetilde{\mathcal{O}}(1/k) \;, \\ n_{L2} &= \widetilde{\Omega}\big(\tfrac{dk}{\epsilon^2}\big) \;, & \alpha_{L2} &= \widetilde{\mathcal{O}}(\epsilon/k) \;, \end{aligned}$$

where $\widetilde{\alpha} := 1/k^4$. Using the assumption that $\epsilon \le 1/k^3$. this implies that as long as

$$\begin{aligned} n_{L1} &= \widetilde{\Omega}\big(\tfrac{dk}{\epsilon^2}\big) \;, & \alpha_{L1} &= \mathcal{O}(\epsilon/k) \;, \\ n_H &= \widetilde{\Omega}\big((km)^{\Theta(m)}\big) \;, & \alpha_H &= \widetilde{\mathcal{O}}(1/k) \;, \\ n_{L2} &= \widetilde{\Omega}\big(\tfrac{dk}{\epsilon^2}\big) \;, & \alpha_{L2} &= \widetilde{\mathcal{O}}(\epsilon/k) \;, \end{aligned}$$

Our algorithm can estimate the parameters up to error

$$\|\widehat{\mathbf{w}}_i - \mathbf{w}_i\|_2 \leq \epsilon s_i \ , \quad \left|\widehat{s}_i^2 - s_i^2\right| \leq \epsilon s_i^2 \ , \quad \text{and} \quad |\widehat{p}_i - p_i| \leq \epsilon / k \ ,$$

for all $i \in [k]$. □

B Details of the algorithm and the analyses

We explain and analyze each step in Algorithm 1, which imply our main result, as shown in §C.

B.1 Robust Subspace estimation

Building upon the tight guarantees of Proposition 2.6, Algorithm 2 achieves the following guarantee, first when $t_{L1} = 1$. The conditions depend on the ground truths meta-parameters. With $\widetilde{\mathcal{O}}(d)$ samples, we can tolerate up to $\mathcal{O}(\epsilon^2 p_{\min}^2/k^2)$ corruption, in an ideal case when $\mathbf{W} \in \mathbb{R}^{d\times k}$ is a semi-orthogonal matrix. For the worst case $\mathbf{W}$, it is $\mathcal{O}(\epsilon^6 p_{\min}^2/k^2)$.

Lemma B.1 (Learning the subspace)**.** *Under Assumptions 1 and 2, for any target probability* $\delta \in (0, 0.5)$*, and* $\epsilon > 0$*, if* $t_{L1} = 1$*,*

$$n_{L1} \ = \ \widetilde{\Omega}\left(dk^2 \min\left\{ \frac{\rho^4 k^2}{\epsilon^2 \sigma_{\min}^2}, \frac{k^2}{\epsilon^6 p_{\min}^2} \right\} \right) \ , \ \ \textit{and} \quad \alpha_{L1} \ = \ \mathcal{O}\left(\max\left\{ \frac{\epsilon^2 \sigma_{\min}^2}{\rho^4 k^2}, \frac{\epsilon^6 p_{\min}^2}{k^2} \right\} \right) \ , \tag{9}$$

then the semi-orthogonal matrix $\widehat{\mathbf{U}} \in \mathbb{R}^{d\times k}$ *from robust subspace estimation in Algorithm 2 achieves*

$$\left\| \left(\mathbf{I} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \right) \mathbf{w}_i \right\|_2 \ \leq \ \epsilon\rho \ , \tag{10}$$

with probability at least $1 - \delta$*, where* $\sigma_{\min}$ *is the minimum singular value of* $\sum_{j=1}^k p_j \mathbf{w}_j \mathbf{w}_j^\top$.

We provide a proof in § D. When $t_{L1} > 1$, we get the following sufficient condition. The dominant first term requires the effective sample size $n_{L1}\, t_{L1} = \widetilde{\Omega}(d\, \mathrm{poly}(k))$, which is linear in d.

Remark B.2 (Handling $t_{L1} > 1$)**.** *Lemma B.1 can be naturally generalized to the case where* $1 < t_{L1} < d$*, and the requirement on* n_{L1} *is*

$$n_{L1} \ = \ \widetilde{\Omega}\left(\frac{dk^2}{t_{L1}} \min\left\{ \frac{\rho^4 k^2}{\epsilon^2 \sigma_{\min}^2}, \frac{k^2}{\epsilon^6 p_{\min}^2} \right\} + k \min\left\{ \frac{\rho^8 k^4}{\epsilon^4 \sigma_{\min}^4}, \frac{k^4}{\epsilon^{12} p_{\min}^4} \right\} \right) \ , \tag{11}$$

Time complexity: $\mathcal{O}(n_{L1} t_{L1} d)$ for computing $\widehat{\beta}_{i,j}$'s and $\mathcal{O}(n_{L1}^2 t_{L1}^2 dk\alpha \log(1/\delta))$ for the filtering algorithm which uses k_SVD [1]. The running time is from the fact that Double-Filtering (Algorithm 3) executes at most $\mathcal{O}(n_{L1} t_{L1} \alpha)$ times, and the running time of each execution is dominated by k_SVD which takes $\mathcal{O}(n_{L1} t_{L1} dk)$ time.

Remark B.3 (Gaussianity assumptions)**.** *Although our robust PCA algorithm 2 only requires bounded fourth moment assumption, our robust subspace estimation succeeds with the fact that*

$$\mathbb{E}[y_{i,j}^2 \mathbf{x}_{i,j} \mathbf{x}_{i,j}^\top] = c\mathbf{I} + 2\sum_{\ell=1}^k p_\ell \mathbf{w}_\ell \mathbf{w}_\ell^\top$$

for some constant $c \geq 0$*. This depends on the fourth moment property of Gaussian (Fact J.7). [42] adopts a different approach by taking*

$$\mathbb{E}[y_{i,j} y_{i,j'} \mathbf{x}_{i,j} \mathbf{x}_{i,j'}^\top] = \sum_{\ell=1}^k p_\ell \mathbf{w}_\ell \mathbf{w}_\ell^\top$$

for $j \neq j'$*, which is able to handle general sub-Gaussian* $\mathbf{x}$*. While it is possible to make the approach in [42] robust to outliers with robust mean estimation techniques (Remark K.1), such approaches requires an sub-optimal* $\Omega(d^2)$ *samples complexity. How to make the approach in [42] robust with only linear dependency in* d *remains an open problem, and the key obstacle is that the random matrix* $y_{i,j} y_{i,j'} \mathbf{x}_{i,j} \mathbf{x}_{i,j'}^\top$ *is not PSD.*

B.2 Robust Clustering

Once we have the subspace, we use the Sum-of-Squares (SOS) algorithm of [45] to cluster the k-dimensional points $\widehat{\mathbf{U}}^\top \widehat{\beta}_i$'s where $\widehat{\beta}_i = (1/t_H) \sum_{j \in [t_H]} y_{i,j} \mathbf{x}_{i,j}$. For a value of m as discussed in Remark H.11, we can exploit the m-th order moment to filter out corrupted points and recover the clusters. This allows us to gracefully trade off t_H and n_H, breaking the barrier of $t = \Omega(k^{1/2})$ in [42]. Further, this approach is robust against adversarial corruption up to a $\mathcal{O}(p_{\min})$ fraction of the data. We explicitly write the algorithm in §G and provide a proof in §H.

Previous work [42] proposed a linkage based clustering algorithm that is able to correctly cluster $\widehat{\beta}_i$'s as long as $t_H = \Omega(\sqrt{k})$ and the second moment is bounded. However, the algorithm fails when $t_H = o(\sqrt{k})$, and it has been noticed in [42] that the failure of such kind of algorithms is inherent since it only relies some boundedness condition on the second moments. It is natural to ask whether it is possible to exploit the boundedness of higher order moments (larger than 2) of the distribution to obtain stronger clustering results. Assuming boundedness of higher order moments is not too restrictive, since typical distributional assumptions, e.g. sub-Gaussianity, often imply boundedness for arbitrarily high order moments.

It turns out in order to *efficiently* exploit the higher order moments assumptions for clustering, one need slight stronger condition than boundedness, that is the moments are sum-of-squares bounded, meaning there *exist* sum-of-squares proofs showing that the moments are bounded [45, 35]. It is also shown in [45] that a Poincaré distribution has sum-of-squares bounded moments, and thus their algorithm can be applied to clustering any Poincaré distributions.

It turns out that in our model, even assuming that $\mathbf{x}_{i,j}$ follows from isotropic Gaussian distribution, the distribution of $\widehat{\beta}_i$ is not Poincaré and thus preventing us from applying the result in [45] directly. Interestingly, as we showed in Lemma H.2, though $\widehat{\beta}_i$ is not Poincaré, the high order moments of $\widehat{\beta}_i$ is still sum-of-squares bounded under Gaussianity assumption of $\mathbf{x}_{i,j}$, and thus we can apply the result of [45] to efficiently cluster $\widehat{\beta}_i$'s, with guarantees formalized in Lemma B.4.

Lemma B.4 (Clustering and initial parameter estimation)**.** *Under Assumptions 1 and 2, if α_H fraction of the tasks are adversarially corrupted in $\mathcal{D}_H$, and given a semi-orthogonal matrix $\widehat{\mathbf{U}}$ satisfying Eq.* (10) *with $\epsilon = \mathcal{O}(\Delta/\rho)$, Algorithm 7 with a choice of $m \in \mathbb{N}$, and Algorithm 8 outputs $\left\{\widetilde{\mathbf{w}}_\ell \in \mathbb{R}^d, \widetilde{r}_\ell^2\right\}_{j\in[k]}$ satisfying*

$$\|\widetilde{\mathbf{w}}_\ell - \mathbf{w}_\ell\|_2 \le \frac{\Delta}{10}\ , \qquad \textit{and} \qquad \left|\widetilde{r}_\ell^2 - r_\ell^2\right| \le r_j^2 \frac{\Delta^2}{50\rho^2}\ , \qquad \forall\, \ell \in [k]\ , \tag{12}$$

with probability at least $1-\delta$, where $\widetilde{r}_\ell^2$ is the robust estimate of $r_\ell^2 := \|\widetilde{\mathbf{w}}_\ell - \mathbf{w}_\ell\|_2^2 + s_\ell^2$, if

$$\alpha_H < \min\left\{\frac{p_{\min}}{16}, \frac{C\Delta^2\sqrt{t_H}p_{\min}}{\rho^2\log\left(\rho^2/\Delta^2 t_H\right)}\right\}, \quad n_H = \widetilde{\Omega}\left(\frac{(km)^{\Theta(m)}}{p_{\min}} + \frac{\rho^4}{\Delta^4 t_H p_{\min}}\right)\ , \tag{13}$$

and $t_H = \Omega(m\rho^2/(p_{\min}^{2/m}\Delta^2))$ for some universal constant $C > 0$.

Assuming that $p_{\min} = 1/k$, we can therefore get the range of m, such that the condition on t_H holds. Which is when

$$\frac{2\log k}{-W_{-1}\left(-\frac{2c\rho^2\log k}{t_H\Delta^2}\right)} \le m \le \frac{2\log k}{-W_0\left(-\frac{2c\rho^2\log k}{t_H\Delta^2}\right)}$$

for some $c > 0$, where W_0 and W_{-1} are the Lambert W function, only if $t_H \ge 2ec\rho^2\Delta^{-2}\log k$.

Time complexity: The robust clustering algorithm runs in time $\mathcal{O}\big((n_H\, k)^{\mathcal{O}(m)}\log(1/\delta)\big)$.

Remark B.5 (Gaussianity assumption)**.** *The only distributional assumption required for the clustering algorithm is that $y_{i,j}\mathbf{x}_{i,j}$ is SOS bounded. It is noted in [46] that this is a general assumption, and the algorithm can be applied much more broadly than Gaussian.*

B.3 Classification and robust estimation

Once the $\mathbf{w}_\ell$'s are estimated from the robust clustering step, we can efficiently classify any light task with $t_{L2} = \Omega(\log(kn_{L2}))$. After classification, we use a robust linear regression method [26] on each group, which can tolerate up to $\mathcal{O}(p_{\min})$ fraction of corruption. It is critical that we separate the role of heavy and light tasks as initial estimation (robust clustering) and refinement (classification and robust estimation). This allows us to have abundant $n_{L2}p_{\min} = \widetilde{\Omega}(d/t_{L2})$ light tasks compensate for scarce $n_H p_{\min} = \Omega(\text{poly}(k))$ heavy tasks. We provide the algorithm and a proof in § I.

Lemma B.6 (Refined parameter estimation via classification)**.** *Under Assumptions 1 and 2 with α_{L2} fraction of the tasks are adversarially corrupted in $\mathcal{D}_{L2}$, and given estimated parameters $\widetilde{\mathbf{w}}_i$, $\widetilde{r}_i$ satisfying Eq.* (12)*, with probability $1-\delta$, for any accuracy $\epsilon > 0$, if $t_{L2} = \Omega\big(\rho^4 \log(kn_{L2}/\delta)/\Delta^4\big)$,*

$$\alpha_{L2} \;=\; \mathcal{O}\big(p_{\min}\epsilon \log^{-1}(1/\epsilon)\big) \;, \quad \textit{and} \quad n_{L2} \;=\; \widetilde{\Omega}\big(dt_{L2}^{-1}p_{\min}^{-1}\epsilon^{-2}\big) \;, \tag{14}$$

then Algorithm 9 outputs estimated parameters $\{\widehat{\mathbf{w}}_i\}_{i=1}^k, \{\widehat{s}_i\}_{i=1}^k, \{\widehat{p}_i\}_{i=1}^k$ such that for all $i \in [k]$,

$$\|\widehat{\mathbf{w}}_i - \mathbf{w}_i\|_2 \le \epsilon s_i \;, \quad \big|\widehat{s}_i^2 - s_i^2\big| \le \epsilon s_i^2/\sqrt{t_{L2}} \;, \quad \textit{and} \quad |\widehat{p}_i - p_i| \le \epsilon\, p_i \sqrt{t_{L2}/d} + \alpha_{L2} \;. \tag{15}$$

Time complexity: $\mathcal{O}\big(n_{L2}^2 t_{L2}^2 d\big)$. The running time of Algorithm 9 is dominated by the robust linear regression procedure ([26, Algorithm 2]), which takes at most $n_{L2}t_{L2}\alpha$ iterations and $\mathcal{O}(n_{L2}t_{L2}d)$ time for SVD per iteration.

C Proof of meta-learning in Theorem 1

Applying Lemma B.1 with $\epsilon = \Delta/(10\rho)$, we get a semi-orthogonal matrix $\widehat{\mathbf{U}}$ satisfying

$$\left\|\left(\mathbf{I} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\right)\mathbf{w}_i\right\|_2 \;\le\; \Delta/10 \;, \tag{16}$$

with $\widetilde{\alpha} \coloneqq \max\big\{\Delta^2\sigma_{\min}^2/(\rho^6k^2), \Delta^6p_{\min}^2/(k^2\rho^6)\big\}$ if

$$n_{L1} \;=\; \widetilde{\Omega}\left(\frac{dk^2}{\widetilde{\alpha}t_{L1}} + \frac{k}{\widetilde{\alpha}^2}\right) \;, \;\text{ and }\; \alpha_{L1} \;=\; \mathcal{O}(\widetilde{\alpha}) \;. \tag{17}$$

Since we have sufficiently accurate estimate of the k-dimensional subspace spanned by the columns of $\mathbf{W}$, we can cluster the tasks more efficiently in this lower dimensional space using robust a clustering algorithm. We use a SOS algorithm in Algorithm 7 with a choice of $m \in \mathbb{N}$ such that $t_H = \Omega\Big(m\rho^2/(p_{\min}^{2/m}\Delta^2)\Big)$, to get

$$\|\widetilde{\mathbf{w}}_j - \mathbf{w}_j\|_2 \;\le\; \Delta/10 \;, \text{ and } \big|\widetilde{r}_j^2 - r_j^2\big| \;\le r_j^2\Delta^2/(50\rho^2) \;, \tag{18}$$

for all $j \in [k]$, using Lemma B.4, which requires

$$n_H = \widetilde{\Omega}\left(\frac{(km)^{\Theta(m)}}{p_{\min}} + \frac{\rho^4}{\Delta^4 p_{\min} t_H}\right) \;, \;\text{ and } \alpha_H \;=\; \widetilde{\mathcal{O}}\left(p_{\min}\cdot\min\left\{1, \frac{\Delta^2\sqrt{t_H}}{\rho^2}\right\}\right) \;. \tag{19}$$

It follows from Lemma B.6 that for any desired accuracy $\epsilon \ge (\alpha_{L2}/p_{\min})\log(p_{\min}/\alpha_{L2})$ if

$$n_{L2} \;=\; \widetilde{\Omega}\left(\frac{d}{t_{L2}p_{\min}\epsilon^2}\right) \;, \qquad \text{and} \qquad t_{L2} = \Omega\big(\log(kn_{L2}/\delta)/\Delta^4\big) \tag{20}$$

the output of our algorithm achieves

$$\|\widehat{\mathbf{w}}_i - \mathbf{w}_i\|_2 \;\le\; \epsilon s_i \;, \tag{21}$$

$$\big|\widehat{s}_i^2 - s_i^2\big| \;\le\; \frac{\epsilon}{\sqrt{t_{L2}}}s_i^2 \;, \quad \text{and} \tag{22}$$

$$|\widehat{p}_i - p_i| \;\le\; \epsilon\sqrt{t_{L2}/d}\,p_i + \alpha_{L2}. \tag{23}$$

for all $i \in [k]$, as long as $\alpha_{L2} = \mathcal{O}\big(p_{\min}\epsilon/\log\frac{1}{\epsilon}\big)$.

D Proof of robust subspace estimation analysis in Lemma B.1

We first prove for the simple setting where $t_{L1} = 1$ and resolving a discrepancy in the independence when $t_{L1} > 1$ at the end of the section. Further, for notational convenience, we use i in place of (i, j), and n for $n_{L1} t_{L1}$.

First we compute the expectation of $\widehat{\beta}_i \widehat{\beta}_i^\top$. Using Lemma K.2, we have

$$\mathbf{M} := \mathbb{E}\Big[\widehat{\beta}_i \widehat{\beta}_i^\top\Big] = 2 \sum_{j=1}^{k} p_j \mathbf{w}_j \mathbf{w}_j^\top + \left(\sum_{j=1}^{k} p_j \left(\|\mathbf{w}_j\|^2 + s_j^2 \right) \right) \mathbf{I}_d ,$$

and we define $\bar{\rho}^2 = \left(\sum_{j=1}^{k} p_j (\|\mathbf{w}_j\|^2 + s_j^2) \right)$.

Since our goal is to recover the top k eigenspace of $\mathbf{M}$, we would like to apply Proposition 2.6 to $\mathbf{x}_i = \widehat{\beta}_i$, however $\widehat{\beta}_i$ does not satisfy the spectral norm bound requirement on $\left\|\mathbf{x}_i \mathbf{x}_i^\top - \mathbf{\Sigma}\right\|_2$. The following proposition shows that we can resolve this issue through conditioning on the event that $\widehat{\beta}_i$ is bounded.

Proposition D.1. *For any* $0 < \delta \le 1/2$, *define event*

$$\mathcal{E} := \left\{ \left\| \widehat{\beta}_i \right\|_2 \le \rho \sqrt{d} \log(nd/\delta) \ , \ \forall\, i \in [n] \right\} .$$

Then conditioned on event $\mathcal{E}$, *the distribution of* $\widehat{\beta}_i$ *satisfies the prerequisite of Proposition 2.6 with*

1. $\left\| \widehat{\beta}_i \widehat{\beta}_i^\top - \mathbb{E}\Big[\widehat{\beta}_i \widehat{\beta}_i^\top \Big| \mathcal{E}\Big] \right\|_2 \ \le \ \underbrace{5\rho^2 d \log^2(nd/\delta)}_{=B}$.

2. $\mathbb{E}\left[\mathrm{Tr}\Big[\mathbf{A} \Big(\widehat{\beta}_i \widehat{\beta}_i^\top - \mathbb{E}[\widehat{\beta}_i \widehat{\beta}_i^\top \,|\, \mathcal{E}] \Big) \Big]^2 \middle| \mathcal{E} \right] \le \underbrace{\mathcal{O}\big(k\rho^4\big)}_{=\nu^2(k)}$.

The mean shift is bounded under $\mathcal{E}$ *as:*

$$\left\| \mathbb{E}\Big[\widehat{\beta}_i \widehat{\beta}_i^\top\Big] - \mathbb{E}\Big[\widehat{\beta}_i \widehat{\beta}_i^\top \Big| \mathcal{E}\Big] \right\|_2 \ \le \ \mathcal{O}\Big(\rho^2 \sqrt{\delta}\Big) , \tag{24}$$

The proof is deferred to the end of the section.

Recall that $\mathbf{M} := \mathbb{E}\Big[\widehat{\beta}_i \widehat{\beta}_i^\top\Big]$, and let us define $\mathbf{M}' := \mathbb{E}\Big[\widehat{\beta}_i \widehat{\beta}_i^\top \Big| \mathcal{E}\Big]$ be the mean conditioned on $\mathcal{E}$. With Proposition D.1, we can apply Proposition 2.6 to obtain a nuclear norm guarantee on our our subspace estimation algorithm

Proposition D.2. *Given that*

$$n = \widetilde{\Omega}(dk^2/\alpha) ,$$

with probability $1 - 2\delta$, *then Algorithm 2 returns a rank-k semi-orthogonal matrix* $\widehat{\mathbf{U}} \in \mathbb{R}^{d \times k}$ *satisfying*

$$\left\| \widehat{\mathbf{U}} \widehat{\mathbf{U}}^\top \left(\sum_{j=1}^{k} p_j \mathbf{w}_j \mathbf{w}_j^\top \right) \widehat{\mathbf{U}} \widehat{\mathbf{U}}^\top - \sum_{j=1}^{k} p_j \mathbf{w}_j \mathbf{w}_j^\top \right\|_* = \mathcal{O}\big(\rho^2 k \sqrt{\alpha}\big)$$

for $\delta \in (0, 0.5)$.

Proof of Proposition D.2. We apply Proposition 2.6 to $\mathbf{x}_i = \widehat{\beta}_i$ conditioned on event $\mathcal{E}$ and get

$$\mathrm{Tr}\Big[\widehat{\mathbf{U}} \widehat{\mathbf{U}}^\top \mathbf{M}' \widehat{\mathbf{U}} \widehat{\mathbf{U}}^\top \Big] \ge (1 - \mathcal{O}(\alpha)) \, \mathrm{Tr}[\mathcal{P}_k(\mathbf{M}')] - \mathcal{O}\big(\rho^2 k \sqrt{\alpha}\big) \tag{25}$$

with probability $1 - 2\delta$, when

$$n = \widetilde{\Omega}\bigg(\Big(dk^2 + \frac{\rho^2 d}{\sqrt{k}\rho^2} \cdot \sqrt{k\alpha}\Big)/\alpha\bigg) \;= \widetilde{\Omega}(dk^2/\alpha).$$

WLOG, for the remaining analysis we will assume $\delta \leq {}^1\!/_{nd}$. The nuclear norm term in the proposition statement can be bounded as

$$\begin{aligned}
&\Big\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top(\mathbf{M} - \bar{\rho}^2\mathbf{I})\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - (\mathbf{M} - \bar{\rho}^2\mathbf{I})\Big\|_* \\
&= \operatorname{Tr}\big[\mathbf{M} - \bar{\rho}^2\mathbf{I}\big] - \operatorname{Tr}\Big[\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top(\mathbf{M} - \bar{\rho}^2\mathbf{I})\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\Big] \\
&= \operatorname{Tr}\big[\mathbf{M} - \bar{\rho}^2\mathbf{I}\big] - \operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top(\mathbf{M}' - \bar{\rho}^2\mathbf{I})\widehat{\mathbf{U}}\Big] - \operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top(\mathbf{M} - \mathbf{M}')\widehat{\mathbf{U}}\Big] \\
&\leq \operatorname{Tr}\big[\mathbf{M} - \bar{\rho}^2\mathbf{I}\big] - \operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top(\mathbf{M}' - \bar{\rho}^2\mathbf{I})\widehat{\mathbf{U}}\Big] + \rho^2 k\sqrt{\delta} \;\; \text{(Using Equation (24))} \\
&= \operatorname{Tr}\big[\mathbf{M} - \bar{\rho}^2\mathbf{I}\big] - \operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top\mathbf{M}'\widehat{\mathbf{U}}\Big] + k\bar{\rho}^2 + \rho^2 k\sqrt{\delta} \\
&\leq \operatorname{Tr}\big[\mathbf{M} - \bar{\rho}^2\mathbf{I}\big] - (1 - \mathcal{O}(\alpha))\operatorname{Tr}[\mathcal{P}_k(\mathbf{M}')] + \rho^2 k\sqrt{\alpha} + k\bar{\rho}^2 + \rho^2 k\sqrt{\delta} \qquad (26) \\
&\qquad\qquad \text{(Using Equation (25))}\,.
\end{aligned}$$

We need the following bound on $\operatorname{Tr}[\mathcal{P}_k(\mathbf{M}')]$ before proceeding:

$$\begin{aligned}
\operatorname{Tr}[\mathcal{P}_k(\mathbf{M}')] &= \operatorname{Tr}\big[\mathcal{P}_k(\mathbf{M}' - \bar{\rho}^2\mathbf{I})\big] + k\bar{\rho}^2 \\
&\geq \operatorname{Tr}\big[\mathcal{P}_k(\mathbf{M} - \bar{\rho}^2\mathbf{I})\big] - k\rho^2\sqrt{\delta} + k\bar{\rho}^2 \qquad (27) \\
&= \operatorname{Tr}\big[\mathbf{M} - \bar{\rho}^2\mathbf{I}\big] - k\rho^2\sqrt{\delta} + k\bar{\rho}^2,
\end{aligned}$$

where Equation (27) holds by the following matrix perturbation bound:

$$\begin{aligned}
\big|\operatorname{Tr}\big[\mathcal{P}_k(\mathbf{M} - \bar{\rho}^2\mathbf{I})\big] - \operatorname{Tr}\big[\mathcal{P}_k(\mathbf{M}' - \bar{\rho}^2\mathbf{I})\big]\big| &\leq \sum_{i=1}^k \big|\lambda_i(\mathbf{M}' - \bar{\rho}^2\mathbf{I}) - \lambda_i(\mathbf{M} - \bar{\rho}^2\mathbf{I})\big| \\
&\leq k\rho^2\sqrt{\delta} \qquad \text{(Using Equation (24))}.
\end{aligned}$$

Plugging Equation (27) back in Equation (26), we have Equation (26) bounded by

$$\begin{aligned}
&\leq \mathcal{O}\Big(\alpha\operatorname{Tr}\big[\mathbf{M} - \bar{\rho}^2\mathbf{I}\big] + \alpha k\bar{\rho}^2 + \rho^2 k\sqrt{\alpha} + \rho^2 k\sqrt{\delta}\Big) \\
&\leq \mathcal{O}\Big(\alpha\operatorname{Tr}\big[\mathbf{M} - \bar{\rho}^2\mathbf{I}\big] + \rho^2 k\sqrt{\alpha} + \rho^2 k\sqrt{\delta}\Big) \;\; \text{(Using } \delta \leq 1/nd \leq \alpha\text{)} \\
&\leq \mathcal{O}\big(\alpha\operatorname{Tr}\big[\mathbf{M} - \bar{\rho}^2\mathbf{I}\big] + \rho^2 k\sqrt{\alpha}\big)
\end{aligned}$$

Thus, we have obtained that

$$\begin{aligned}
&\left\|\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\left(\sum_{j=1}^k p_j\mathbf{w}_j\mathbf{w}_j^\top\right)\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top - \sum_{j=1}^k p_j\mathbf{w}_j\mathbf{w}_j^\top\right\|_* \\
=&\mathcal{O}\left(\alpha\operatorname{Tr}\left[\sum_{j=1}^k p_j\mathbf{w}_j\mathbf{w}_j^\top\right] + \rho^2 k\sqrt{\alpha}\right) \\
=&\mathcal{O}\big(\rho^2 k\sqrt{\alpha}\big). \qquad \square
\end{aligned}$$

The following lemma connects this nuclear norm bound to a subspace bound that we want.

Lemma D.3 (Gap-free spectral bound). *Given k vectors $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k \in \mathbb{R}^d$, we define $\mathbf{X}_i = \mathbf{x}_i\mathbf{x}_i^\top$ for each $i \in [k]$. For any $\gamma \geq 0$, $\sigma \in \mathbb{R}_+$, and any rank-k PSD matrix $\widehat{\mathbf{M}} \in \mathbb{R}^{d\times d}$ such that*

$$\left\|\widehat{\mathbf{M}} - \mathcal{P}_k\left(\sigma^2\mathbf{I} + \sum_{i=1}^k \mathbf{X}_i\right)\right\|_* \;\leq\; \gamma, \qquad (28)$$

where $\mathcal{P}_k(\cdot)$ is a best rank-k approximation of a matrix, we have

$$\sum_{i\in[k]} \left\| \mathbf{x}_i^\top \left(\mathbf{I} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\right) \right\|_2^2 \;\; \le \;\; \min\Big\{ \gamma^2 \sigma_{\max}/\sigma_{\min}^2 \,,\, 2\gamma^{2/3}\sigma_{\max}^{1/3} k^{2/3} \Big\}\,, \tag{29}$$

where $\sigma_{\min}$ is the smallest non-zero singular value of $\sum_{i\in[k]}\mathbf{X}_i$, and $\sigma_{\max} = \left\|\sum_{i\in[k]}\mathbf{X}_i\right\|_2$, and $\widehat{\mathbf{U}} \in \mathbb{R}^{d\times k}$ is the matrix consisting of the top-k singular vectors of $\widehat{\mathbf{M}}$. Further, for all $i \in [k]$, we have

$$\left\| \mathbf{x}_i^\top \left(\mathbf{I} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\right) \right\|_2^2 \;\; \le \;\; \min\Big\{ \gamma^2 \|\mathbf{x}_i\|_2^2/\sigma_{\min}^2 \,,\, 2\gamma^{2/3}\|\mathbf{x}_i\|_2^{2/3} \Big\}\,. \tag{30}$$

We provide a proof in Section D.6. Using this lemma with $\sigma = 0$,

$$\widehat{\mathbf{M}} = \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \left(\sum_{j=1}^{k} p_j \mathbf{w}_j \mathbf{w}_j^\top\right) \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top,$$

$\mathbf{X}_i = p_i \mathbf{w}_i \mathbf{w}_i^\top$ for all $i \in [k]$, and the nuclear norm bound in Proposition D.2, we get

$$\begin{aligned} p_i \left\| \mathbf{w}_i^\top \left(\mathbf{I} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\right) \right\|_2^2 &\le \min\left\{ \rho^2 \frac{\gamma^2}{\sigma_{\min}^2} p_i, 2\rho^{2/3}\gamma^{2/3} p_i^{1/3} \right\} \\ \implies \left\| \mathbf{w}_i^\top \left(\mathbf{I} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\right) \right\|_2^2 &\le \min\left\{ \rho^2 \frac{\gamma^2}{\sigma_{\min}^2}, 2\frac{\rho^{2/3}\gamma^{2/3}}{p_i^{2/3}} \right\} \\ &\lesssim \min\Big\{ \rho^6 k^2 \alpha/\sigma_{\min}^2, \rho^2 k^{2/3}\alpha^{1/3}/p_i^{2/3} \Big\}. \end{aligned} \tag{31}$$

Since we are aiming for $\left\| \mathbf{w}_i^\top \left(\mathbf{I} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\right) \right\|_2 = \epsilon\rho$ error, we need

$$\alpha = \mathcal{O}\left(\max\left\{ \frac{\epsilon^2\sigma_{\min}^2}{\rho^4 k^2}, \frac{\epsilon^6 p_{\min}^2}{k^2} \right\} \right)$$

and the sample complexity is

$$n \;=\; \widetilde{\Omega}\Big(dk^2 \min\left\{ \frac{\rho^4 k^2}{\epsilon^2\sigma_{\min}^2}, \frac{k^2}{\epsilon^6 p_{\min}^2} \right\} \Big).$$

In the analysis above, we assume that each example β_i is independently drawn. While this is true when $t_{L_1} = 1$, it is no longer the case when $t_{L_1} > 1$ where we have to break up the examples from each task into t_{L_1} different estimators. Recall that $\widehat{\mathbf{p}}$ is the vector of the empirical fractions of the examples that correspond to each linear regressor, and $\mathbf{p}$ is the population version of it. For given a pair of parameters n_{L_1}, t_{L_1} let us define

$$\widehat{\mathbf{p}} \sim \frac{1}{n_{L_1} t_{L_1}} \text{multinomial}(n_{L_1}, \mathbf{p}) \cdot t_{L_1},$$

and independently

$$\widehat{\mathbf{p}}^* \sim \frac{1}{n_{L_1} t_{L_1}} \text{multinomial}(n_{L_1} \cdot t_{L_1}, \mathbf{p}).$$

Notice that $\widehat{\mathbf{p}}$ corresponds to the setting where there are n_{L_1} tasks with each task having t_{L_1} examples, and $\widehat{\mathbf{p}}^*$ corresponds to the setting where there are $n_{L_1} \cdot t_{L_1}$ tasks, each with 1 example. By Proposition J.5, we know that when

$$n_{L_1} \ge \frac{2k\log(2/\delta)}{\alpha^2},$$

$\|\widehat{\mathbf{p}} - \widehat{\mathbf{p}}^*\|_1 \le \alpha$ with probability $1-\delta$. Hence if we denote the set $G = \{\beta_i\}_{i=1}^{n_{L_1}\cdot t_{L_1}}$ to be the set of the data coming from the model where each task has t_{L_1} examples. There exists a distribution of set L, E such that $G = (G' \setminus L') \cup E'$ with $|L| = |E|$, G' has data from $n_{L_1} \cdot t_{L_1}$ independent tasks with

1 example per task, and $|L| = |E| \leq \alpha$ with probability $1-\delta$. Thus we have obtain a reduction from t_{L_1} examples per task setting to the 1 example per task setting, in which case our algorithm receives a dataset with less than 2α fraction of corruption with probability δ. Since our previous proof applies to this setting, this concludes the proof of Lemma B.1 and the final sample complexity is

$$n_{L1} \;=\; \widetilde{\Omega}\Big(\frac{dk^2}{t_{L1}} \min\Big\{ \frac{\rho^4 k^2}{\epsilon^2 \sigma_{\min}^2}, \frac{k^2}{\epsilon^6 p_{\min}^2} \Big\} + k \min\Big\{ \frac{\rho^8 k^4}{\epsilon^4 \sigma_{\min}^4}, \frac{k^4}{\epsilon^{12} p_{\min}^4} \Big\} \Big) \;, \tag{32}$$

We are left to show $B = \mathcal{O}(\rho^2 d \log^2(nd/\delta))$, $\nu(k) = \mathcal{O}(\rho^2 k)$, and the mean shift bound in Eq. (24).

Proof of Proposition D.1. We first show $B = \mathcal{O}((\rho^2 d)\log^2(nd/\delta))$. From [42, Proposition A.1], we have $\|\widehat{\beta}_i\|_2^2 \leq (\rho^2 d)\log^2(nd/\delta)$ (i.e. event $\mathcal{E}$ happens) with probability at least $1-\delta$. Using this, we have

$$\begin{aligned} \left\| \widehat{\beta}_i \widehat{\beta}_i^\top - \mathbf{M} \right\|_2 &\leq \left\| \widehat{\beta}_i \right\|_2^2 + \|\mathbf{M}\|_2 \\ &\leq \rho^2 d \log^2(nd/\delta) + 3\rho^2 \\ &\leq 4\rho^2 d \log^2(nd/\delta) \;, \end{aligned} \tag{33}$$

for all $i \in [n]$ with probability at least $1-\delta$.

Second, we bound the mean shift conditioned on event $\mathcal{E}$

$$\left\| \mathbb{E}\left[\widehat{\beta}_i \widehat{\beta}_i \middle| \mathcal{E} \right] - \mathbf{M} \right\|_2 = \max_{\|\mathbf{v}\|_2 = 1} |\mathbb{E}[z_{\mathbf{v}} | \mathcal{E}]|, \tag{34}$$

where $z_{\mathbf{v}} := \left(\mathbf{v}^\top \widehat{\beta}_i\right)^2 - \mathbf{v}^\top \mathbf{M} \mathbf{v}$. The random variable $z_{\mathbf{v}}$ is centered with variance

$$\begin{aligned} \mathbb{E}\left[\left(\left(\mathbf{v}^\top \widehat{\beta}_i\right)^2 - \mathbf{v}^\top \mathbf{M} \mathbf{v} \right)^2 \right] &\leq \mathbb{E}\left[\left(\mathbf{v}^\top \widehat{\beta}_i\right)^4 \right] - \left(\mathbf{v}^\top \mathbf{M} \mathbf{v}\right)^2 \\ &= \mathcal{O}\big(\rho^4\big) \quad (\mathbf{v}^\top \widehat{\beta}_i \text{ is sub-exponential r.v.}) \;, \end{aligned} \tag{35}$$

Recall that $\mathbf{M}' := \mathbb{E}\left[\widehat{\beta}_i \widehat{\beta}_i \middle| \mathcal{E} \right]$, then using J.1, we have

$$\begin{aligned} \|\mathbf{M}' - \mathbf{M}\|_2 &= \max_{\|\mathbf{v}\|_2 = 1} |\mathbb{E}[z_{\mathbf{v}} | \mathcal{E}]| \\ &\leq \max_{\|\mathbf{v}\|_2 = 1} \frac{\mathbb{E}[z_{\mathbf{v}}] + \sqrt{(1 - \mathbb{P}(\mathcal{E})) \cdot \mathrm{Var}(z_{\mathbf{v}})}}{\mathbb{P}[\mathcal{E}]} \\ &\leq \mathcal{O}\left(\rho^2 \frac{\sqrt{\delta}}{\mathbb{P}[\mathcal{E}]} \right) \\ &\leq \mathcal{O}\Big(\rho^2 \sqrt{\delta} \Big). \end{aligned} \tag{36}$$

Finally, we show $\nu^2(k) = \mathcal{O}(k\rho^4)$. For any symmetric real matrix $\mathbf{A}$ with $\mathrm{rank}(\mathbf{A}) = k$, and $\|\mathbf{A}\|_{\mathrm{F}} \leq 1$,

$$\begin{aligned} \mathbb{E}\left[\mathrm{Tr}\left[\mathbf{A}\Big(\widehat{\beta}_i \widehat{\beta}_i^\top - \mathbf{M}' \Big) \right]^2 \middle| \mathcal{E} \right] &= \mathbb{E}\left[\mathrm{Tr}\left[\mathbf{A}\Big(\widehat{\beta}_i \widehat{\beta}_i^\top - \mathbf{M} + (\mathbf{M} - \mathbf{M}') \Big) \right]^2 \middle| \mathcal{E} \right] \\ &= \mathbb{E}\left[\mathrm{Tr}\left[\mathbf{A}\Big(\widehat{\beta}_i \widehat{\beta}_i^\top - \mathbf{M} \Big) \right]^2 \middle| \mathcal{E} \right] + \mathrm{Tr}[\mathbf{A}(\mathbf{M} - \mathbf{M}')]^2 \\ &\leq \mathbb{E}\left[\mathrm{Tr}\left[\mathbf{A}\Big(\widehat{\beta}_i \widehat{\beta}_i^\top - \mathbf{M} \Big) \right]^2 \middle| \mathcal{E} \right] + \mathcal{O}\big(\rho^4\big) \end{aligned} \tag{37}$$

where the last inequality is obtained using Equation (36). Considering the first term in Equation (37), we get

$$\begin{aligned}
&\mathbb{E}\left[\operatorname{Tr}\left[\mathbf{A}\left(\widehat{\beta}_i\widehat{\beta}_i^\top-\mathbf{M}\right)\right]^2\middle|\mathcal{E}\right]\\
=&\mathbb{E}\left[\left(\sum_{j=1}^k\lambda_j\mathbf{v}_j\mathbf{v}_j^\top\left(\widehat{\beta}_i\widehat{\beta}_i^\top-\mathbf{M}\right)\right)^2\middle|\mathcal{E}\right]\\
=&\mathbb{E}\left[\sum_{j,j'=1}^k\lambda_j\lambda_{j'}\left(\left(\widehat{\beta}_i^\top\mathbf{v}_j\right)^2\left(\widehat{\beta}_i^\top\mathbf{v}_{j'}\right)^2-\mathbf{v}_j^\top\mathbf{M}\mathbf{v}_j\mathbf{v}_{j'}^\top\mathbf{M}\mathbf{v}_{j'}\right)\middle|\mathcal{E}\right]\\
\leq&\sum_{j,j'=1}^k\lambda_j\lambda_{j'}\left(\sqrt{\mathbb{E}\left[\left(\widehat{\beta}_i^\top\mathbf{v}_j\right)^4\middle|\mathcal{E}\right]\mathbb{E}\left[\left(\widehat{\beta}_i^\top\mathbf{v}_{j'}\right)^4\middle|\mathcal{E}\right]}-\mathbf{v}_j^\top\mathbf{M}\mathbf{v}_j\mathbf{v}_{j'}^\top\mathbf{M}\mathbf{v}_{j'}\right) \quad \text{(Cauchy-Schwarz)}\\
\leq&\mathcal{O}\left(\left(\sum_{j=1}^k\lambda_j\right)^2\rho^4\right) \qquad \text{(4-th moment bound)}\\
\leq&\mathcal{O}\left(k\rho^4\right). \qquad (38)
\end{aligned}$$

Using Equation (38) in Equation (37), we get

$$\mathbb{E}\left[\operatorname{Tr}\left[\mathbf{A}\left(\widehat{\beta}_i\widehat{\beta}_i^\top-\mathbf{M}'\right)\right]^2\middle|\mathcal{E}\right]\leq\underbrace{\mathcal{O}\left(k\rho^4\right)}_{=\nu^2(k)}. \qquad \square$$

D.1 Proof of Proposition 2.6

The following main technical lemma guarantees that for any distribution $\mathbf{X}\sim\mathcal{P}$ with a bounded support and a bounded second moment, the filtering algorithm we introduce in Algorithm 2 robustly finds an accurate estimate of the principal subspace.

Lemma D.4 (Main Lemma for Algorithm 2)**.** *Let $\mathcal{P}$ be a distribution over $d\times d$ PSD matrices with the property that,*

$$\mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{P}}[\mathbf{X}]=\mathbf{M}\,,\quad \|\mathbf{X}-\mathbf{M}\|_2\leq B\,,\quad \textit{and}\quad \max_{\|\mathbf{A}\|_{\mathrm{F}}\leq1,\operatorname{rank}(\mathbf{A})\leq k}\mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{P}}\left[\operatorname{Tr}[\mathbf{A}(\mathbf{X}-\mathbf{M})]^2\right]\leq\nu(k)^2.$$

Let a set of n random matrices $G=\left\{\mathbf{X}_i\in\mathbb{R}^{d\times d}\right\}_{i\in[n]}$ where each $\mathbf{X}_i$ is independently drawn from $\mathcal{P}$, and the at most α fraction is corrupted by an adversary such that the input dataset $S=(G\backslash L)\cup E$ with $|E|=|L|\leq\alpha n$, $L\subset G$. There exists a numerical constant $c>0$ such that for any $0<\alpha<c$, if $n=\Omega((dk^2+(B/\nu)\sqrt{k\alpha})\log(d/(\delta\alpha))/\alpha)$, Algorithm 2 outputs a dataset $S'\subseteq S$ satisfying the following for $\widehat{\mathbf{M}}=\frac{1}{|S'|}\sum_{\mathbf{X}_i\in S'}\mathbf{X}_i$:

1. *for the top-k singular vectors $\widehat{\mathbf{U}}\in\mathbb{R}^{d\times k}$ of $\widehat{\mathbf{M}}$,*
$$\operatorname{Tr}\left[\widehat{\mathbf{U}}^\top\left(\widehat{\mathbf{M}}-\mathbf{M}\right)\widehat{\mathbf{U}}\right]\ \leq\ 48\alpha\operatorname{Tr}\left[\widehat{\mathbf{U}}^\top\mathbf{M}\widehat{\mathbf{U}}\right]+102\nu\sqrt{k\alpha}\,.$$
2. *for all rank-k semi-orthogonal matrices $\mathbf{V}\in\mathbb{R}^{d\times k}$, we have*
$$\operatorname{Tr}\left[\mathbf{V}^\top\left(\widehat{\mathbf{M}}-\mathbf{M}\right)\mathbf{V}\right]\ \geq\ -10\alpha\operatorname{Tr}\left[\mathbf{V}^\top\mathbf{M}\mathbf{V}\right]-8\nu\sqrt{k\alpha}\,.$$

We provide a proof in Section D.2.

The proof of Proposition 2.6 is straightforward given Lemma D.4. For the first claim, note that,

$$\begin{aligned}
\operatorname{Tr}\left[\widehat{\mathbf{U}}^\top\boldsymbol{\Sigma}\widehat{\mathbf{U}}\right]&\geq\operatorname{Tr}\left[\widehat{\mathbf{U}}^\top\widehat{\boldsymbol{\Sigma}}\widehat{\mathbf{U}}\right]-48\alpha\operatorname{Tr}[\mathcal{P}_k(\boldsymbol{\Sigma})]-102\nu\sqrt{k\alpha} &&\text{(Using Lemma D.4, part 1)}\\
&\geq\operatorname{Tr}\left[\mathbf{U}^\top\widehat{\boldsymbol{\Sigma}}\mathbf{U}\right]-48\alpha\operatorname{Tr}[\mathcal{P}_k(\boldsymbol{\Sigma})]-102\nu\sqrt{k\alpha} &&\text{(Property of SVD)}\\
&\geq\operatorname{Tr}[\mathcal{P}_k(\boldsymbol{\Sigma})]-58\alpha\operatorname{Tr}[\mathcal{P}_k(\boldsymbol{\Sigma})]-110\nu\sqrt{k\alpha} &&\text{(Using Lemma D.4, part 2).}
\end{aligned}$$

For the second claim, since $\mathbf{\Sigma} \succeq \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \mathbf{\Sigma} \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top$, we have

$$\begin{aligned}\left\|\mathbf{\Sigma} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \mathbf{\Sigma} \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\right\|_* &= \mathrm{Tr}\Big[\mathbf{\Sigma} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \mathbf{\Sigma} \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\Big] \\ &\leq \mathrm{Tr}[\mathbf{\Sigma}] - (1-58\alpha)\,\mathrm{Tr}[\mathcal{P}_k(\mathbf{\Sigma})] + 110\nu\sqrt{k\alpha} \qquad \text{(From the first claim).} \\ &= \mathrm{Tr}[\mathbf{\Sigma} - \mathcal{P}_k(\mathbf{\Sigma})] + 58\alpha\,\mathrm{Tr}[\mathcal{P}_k(\mathbf{\Sigma})] + 110\nu\sqrt{k\alpha} \\ &= \|\mathbf{\Sigma} - \mathcal{P}_k(\mathbf{\Sigma})\|_* + 58\alpha\|\mathcal{P}_k(\mathbf{\Sigma})\|_* + 110\nu\sqrt{k\alpha}.\end{aligned}$$

Similarly, we also get

$$\left\|\mathcal{P}_k(\mathbf{\Sigma}) - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \mathbf{\Sigma} \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\right\|_* \leq 58\alpha\|\mathcal{P}_k(\mathbf{\Sigma})\|_* + 110\nu\sqrt{k\alpha} \tag{39}$$

D.2 Proof of the main analysis of Algorithm 2 in Lemma D.4

The proof of Lemma D.4 is divided into two parts, for each statement of the lemma. Both parts are proven under the following good event. We provide a proof of this lemma in Section D.4.

Lemma D.5. *Under the hypotheses of Lemma D.4, when $n = \Omega((dk^2 + \frac{B}{\nu}\sqrt{k\epsilon})\log(d/(\epsilon\widetilde{\delta}))/\epsilon)$ with probability $1-\widetilde{\delta}$, the following events happen for all semi-orthogonal matrices $\mathbf{V} \in \mathbb{R}^{d\times k} s.t. \mathbf{V}^\top\mathbf{V} = \mathbf{I}_k$,*

1. *There exists $G_{\mathbf{V}} \subset G$ such that*
   (a) $|G_{\mathbf{V}}| \geq (1-\epsilon)n$,
   (b) $\left|\frac{1}{|G_{\mathbf{V}}|}\mathrm{Tr}\big[\sum_{\mathbf{X}_i \in G_{\mathbf{V}}}\big(\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big)\big]\right| \leq 1.01\nu\sqrt{k\epsilon}$, and
   (c) $\frac{1}{|G_{\mathbf{V}}|}\sum_{\mathbf{X}_i \in G_{\mathbf{V}}}\mathrm{Tr}\big[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big]^2 \leq 6.01k\nu^2$,
2. $\left|\dfrac{1}{n}\displaystyle\sum_{\mathbf{X}_i \in G}\mathrm{Tr}\big[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big]\right| \leq \nu\sqrt{k\epsilon}$,
3. *All subset $T \subset G$ such that $|T| \leq \epsilon n$ satisfies*
$$\sum_{\mathbf{X}_i \in T}\mathrm{Tr}\big[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big] \leq 7n\nu\sqrt{k\epsilon} + n\epsilon\,\mathrm{Tr}\big[\mathbf{V}^\top\mathbf{M}\mathbf{V}\big].$$

We provide the proof in Section D.4.

D.2.1 Part 1 of Lemma D.4

Proposition D.6 (Correctness of Algorithm 3)**.** *For a set G of n uncorrupted matrices defined as in the hypotheses of Lemma D.5 and for some $\epsilon \geq \alpha > 0$, suppose the that set S input to Algorithm 3 satisfies the following: $S = (G \setminus L) \cup E$ with $L \subset G$, $|E| \leq \alpha|G|$, $|L| \leq 9\alpha|G|$, $|S| \leq |G|$, and $E \cap G = \emptyset$. If the events in Lemma D.5 hold, then a single call of Algorithm 3 outputs a set $S' \subseteq S$ achieving one of the following two guarantees.*

1. *If Algorithm 3 returns $S' = S$ (unchanged), then*
$$\mathrm{Tr}\left[\widehat{\mathbf{U}}^\top\left(\frac{1}{|S'|}\sum_{\mathbf{X}_i \in S'}\mathbf{X}_i - \mathbf{M}\right)\widehat{\mathbf{U}}\right] \;\leq\; 48\alpha\,\mathrm{Tr}\Big[\widehat{\mathbf{U}}^\top\mathbf{M}\widehat{\mathbf{U}}\Big] + 102\nu\sqrt{k\alpha} \tag{40}$$
2. *If Algorithm 3 returns $S' \subset S$, then there exist two sets $L' \supseteq L$ and $E' \subseteq E$ such that $S' = (G \setminus L') \cup E'$ and $\mathbb{E}[2|L'| + |E'|] \leq 2|L| + |E|$.*

where $\widehat{\mathbf{U}}$ is the top-k singular matrix of $\frac{1}{|S'|}\sum_{\mathbf{X}_i \in S'}\mathbf{X}_i$.

We provide the proof in Section D.3. We are left to show that when n is sufficiently large, then Algorithm 2 terminates before removing too many points, with probability at least $1-\delta$. To get the logarithmic dependence on δ,we divide the meta-training dataset $\mathcal{D}_{L1}$ into $\log(1/\delta)$ partitions of an equal size, and apply the same routine of Algorithm 2.

We show that each run of Algorithm 2 succeeds with a strictly positive probability, and hence one of them is guaranteed to succeed with probability at least $1-\delta$. Further, we have a simple way of choosing the successful run; we select the one that outputs the largest set S'. Precisely, we call a routine successful if $|S'| \geq (1-8\alpha)|G|$.

First, we need to show that the conditions of Proposition D.6 hold, throughout the iterations. However, the condition that $|L| \leq 9\alpha|G|$ might be violated by chance, as we only have guarantee in expectation. We thus bound the probability that there exists a sub-routine of 1-D filtering that results in $|L'| \geq 8\alpha|G|$. The proof is similar to the one in [24]. Let L_i denote the removed set of uncorrupted points in $S_i = (G \backslash L_i) \cup E_i$. Notice that this event implies $|L_T| \geq 8\alpha|G|$ as L_i's are a monotonically increasing sequence of sets. From Markov's inequality, we have $\mathbb{P}(|L_T| \geq 8\alpha|G|) \leq \mathbb{E}[|L_T|]/(8\alpha|G|) \leq 1/6$, where in the last inequality we used the fact that $\mathbb{E}[|L_T|] \leq (1/2)\,\mathbb{E}[2|L_T| + |E_T|] \leq (1/2)(2|L_0| + |E_0|) \leq (3/2)\alpha|G|$. Hence, with probability at least $5/6$ one run succeeds (taking a union bound with Lemma D.5 for the good events with a choice of $\widetilde{\delta} = 1/6$). Out of $\log_6(2/\delta)$ runs, one succeeds with probability at least $1-\delta/2$.

D.2.2 Part 2 of Lemma D.4

There exists a set $T \subset G$ such that $S' \supseteq T$, and $|T| \geq (1-9\alpha)|G|$. Since S' contains a good fraction of points from G, then for any semi-orthogonal matrix $\mathbf{V} \in \mathbb{R}^{d \times k}$, we have

$$
\begin{aligned}
&\sum_{\mathbf{X}_i \in S'} \operatorname{Tr}\big[\mathbf{V}^\top \mathbf{X}_i \mathbf{V}\big] \\
&\geq \sum_{\mathbf{X}_i \in G} \operatorname{Tr}\big[\mathbf{V}^\top \mathbf{X}_i \mathbf{V}\big] - \sum_{\mathbf{X}_i \in G\backslash T} \operatorname{Tr}\big[\mathbf{V}^\top \mathbf{X}_i \mathbf{V}\big] \\
&\geq \sum_{\mathbf{X}_i \in G} \operatorname{Tr}\big[\mathbf{V}^\top \mathbf{X}_i \mathbf{V}\big] - \left(\sum_{\mathbf{X}_i \in G\backslash T} \operatorname{Tr}\big[\mathbf{V}^\top \mathbf{M} \mathbf{V}\big] + n\alpha \operatorname{Tr}\big[\mathbf{V}^\top \mathbf{M} \mathbf{V}\big] + 7n\nu\sqrt{k\alpha} \right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Using Lemma D.5, part 3)} \\
&\geq \sum_{\mathbf{X}_i \in G} \operatorname{Tr}\big[\mathbf{V}^\top \mathbf{M} \mathbf{V}\big] - n\nu\sqrt{k\alpha} - \left(\sum_{\mathbf{X}_i \in G\backslash T} \operatorname{Tr}\big[\mathbf{V}^\top \mathbf{M} \mathbf{V}\big] + n\alpha \operatorname{Tr}\big[\mathbf{V}^\top \mathbf{M} \mathbf{V}\big] + 7n\nu\sqrt{k\alpha} \right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Using Lemma D.5, part 2)} \\
&= \sum_{\mathbf{X}_i \in T} \operatorname{Tr}\big[\mathbf{V}^\top \mathbf{M} \mathbf{V}\big] - 8n\nu\sqrt{k\alpha} - n\alpha \operatorname{Tr}\big[\mathbf{V}^\top \mathbf{M} \mathbf{V}\big] \\
&\geq n(1-10\alpha) \operatorname{Tr}\big[\mathbf{V}^\top \mathbf{M} \mathbf{V}\big] - 8n\nu\sqrt{k\alpha} \qquad (\because |T| \geq (1-9\alpha)n)
\end{aligned}
$$

$$
\begin{aligned}
&\implies n \operatorname{Tr}\Big[\mathbf{V}^\top \widehat{\mathbf{M}} \mathbf{V}\Big] \geq |S'| \operatorname{Tr}\Big[\mathbf{V}^\top \widehat{\mathbf{M}} \mathbf{V}\Big] \geq n(1-10\alpha) \operatorname{Tr}\big[\mathbf{V}^\top \mathbf{M} \mathbf{V}\big] - 8n\nu\sqrt{k\alpha} \\
&\text{or, } \operatorname{Tr}\Big[\mathbf{V}^\top \Big(\widehat{\mathbf{M}} - \mathbf{M}\Big) \mathbf{V}\Big] \geq -10\alpha \operatorname{Tr}\big[\mathbf{V}^\top \mathbf{M} \mathbf{V}\big] - 8\nu\sqrt{k\alpha}
\end{aligned}
$$

when the good events of Lemma D.5 hold, which happens if $n = \Omega\Big(\frac{1}{\epsilon}\Big(k + \frac{B}{\nu}\sqrt{k\epsilon}\Big) \log\frac{d}{\delta}\Big)$ with probability at least $1-\delta$.

D.3 Proof of Proposition D.6

1. To show the first part of Proposition D.6, notice that by Lemma D.5, part 1, there exists a subset $G_{\widehat{\mathbf{U}}} \subset G$ such that

$$\begin{aligned} \left|G_{\widehat{\mathbf{U}}}\right| &\geq (1-\alpha)n \\ \left|\frac{1}{|G_{\widehat{\mathbf{U}}}|}\sum_{\mathbf{X}_i \in G_{\widehat{\mathbf{U}}}} \operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top(\mathbf{X}_i - \mathbf{M})\widehat{\mathbf{U}}\Big]\right| &\leq 1.01\nu\sqrt{k\alpha} \\ \frac{1}{|G_{\widehat{\mathbf{U}}}|}\sum_{\mathbf{X}_i \in G_{\widehat{\mathbf{U}}}} \operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top(\mathbf{X}_i - \mathbf{M})\widehat{\mathbf{U}}\Big]^2 &\leq 6.01k\nu^2. \end{aligned}$$

For the input to the First-Filter algorithm, $S_0 = (G \setminus L) \cup E$, we can reclassify the sample in $G \setminus G_{\widehat{\mathbf{U}}}$ to be the error and have $S_0 = (G_{\widehat{\mathbf{U}}} \setminus L) \cup E'$ where $|L| \leq 9\alpha n$ and $|E'| \leq |E| + \alpha n \leq 2\alpha n$. Hence the output of the First-Filter algorithm satisfies

$$\left|\frac{1}{|S_G|}\sum_{\mathbf{X}_i \in S_G} \operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top\mathbf{X}_i\widehat{\mathbf{U}}\Big] - \operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top\mathbf{M}\widehat{\mathbf{U}}\Big]\right| \leq 54\nu\sqrt{k\alpha}, \tag{41}$$

using Proposition J.4, and from **if** condition of Algorithm 3, we have

$$\frac{1}{n}\sum_{i=1}^{n} \operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top\mathbf{X}_i\widehat{\mathbf{U}}\Big] - \frac{1}{|S_G|}\sum_{\mathbf{X}_i \in S_G} \operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top\mathbf{X}_i\widehat{\mathbf{U}}\Big] \leq 48\Big(\nu\sqrt{k\alpha} + \alpha\operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top\mathbf{M}\widehat{\mathbf{U}}\Big]\Big). \tag{42}$$

Combining the above two inequalities, we get

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} \operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top\mathbf{X}_i\widehat{\mathbf{U}}\Big] - \operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top\mathbf{M}\widehat{\mathbf{U}}\Big] \\ =&\frac{1}{n}\sum_{i=1}^{n} \operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top\mathbf{X}_i\widehat{\mathbf{U}}\Big] - \frac{1}{|S_G|}\sum_{\mathbf{X}_i \in S_G} \operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top\mathbf{X}_i\widehat{\mathbf{U}}\Big] \\ &\qquad + \frac{1}{|S_G|}\sum_{\mathbf{X}_i \in S_G} \operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top\mathbf{X}_i\widehat{\mathbf{U}}\Big] - \operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top\mathbf{M}\widehat{\mathbf{U}}\Big] \\ \leq&\frac{1}{n}\sum_{i=1}^{n} \operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top\mathbf{X}_i\widehat{\mathbf{U}}\Big] - \frac{1}{|S_G|}\sum_{\mathbf{X}_i \in S_G} \operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top\mathbf{X}_i\widehat{\mathbf{U}}\Big] \\ &\qquad + \left|\frac{1}{|S_G|}\sum_{\mathbf{X}_i \in S_G} \operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top\mathbf{X}_i\widehat{\mathbf{U}}\Big] - \operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top\mathbf{M}\widehat{\mathbf{U}}\Big]\right| \\ \leq&102\nu\sqrt{k\alpha} + 48\alpha\operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top\mathbf{M}\widehat{\mathbf{U}}\Big]. \end{aligned} \tag{43}$$

2. We use x_i to denote $\mathrm{Tr}\big[\mathbf{U}^\top \mathbf{X}_i \mathbf{U}\big]$ and μ^P to denote $\mathrm{Tr}\big[\mathbf{U}^\top \mathbf{M} \mathbf{U}\big]$. Our goal is to show that our Algorithm removes much less good points, from the set $G \cap (S \setminus S_G)$. Notice that,

$$\begin{aligned}
\sum_{\substack{x_i \in S \setminus S_G, \\ x_i \geq \mu^{S_G}}} \big(x_i - \mu^{S_G}\big) &= \sum_{x_i \in S \setminus S_G} \big(x_i - \mu^{S_G}\big) - \sum_{\substack{x_i \in S \setminus S_G, \\ x_i < \mu^{S_G}}} \big(x_i - \mu^{S_G}\big) \\
&\geq \sum_{x_i \in S \setminus S_G} \big(x_i - \mu^{S_G}\big) \\
&= \sum_{x_i \in S} \big(x_i - \mu^{S_G}\big) - \sum_{x_i \in S_G} \big(x_i - \mu^{S_G}\big) \\
&= \sum_{x_i \in S} \big(x_i - \mu^{S_G}\big) \\
&\geq 48n\Big(\alpha\mu^{S_G} + \nu\sqrt{k\alpha}\Big) \qquad \text{(from the if clause in Algorithm 3)} \\
&\geq 48n\Big(\alpha\mu^{P} + (1 - 18\alpha)\nu\sqrt{k\alpha}\Big) \\
&\geq 48n\left(\alpha\mu^{P} + \frac{1}{2}\nu\sqrt{k\alpha}\right) \qquad (44)
\end{aligned}$$

where the last inequality is using $\alpha \leq 1/36$, and the second last from the guarantee from the First-Filter algorithm. (by the if clause in the algorithm).

On the other hand, note that,

$$\Big|\big\{x_i \in G \cap (S \setminus S_G) \,\big|\, x_i \geq \mu^{S_G}\big\}\Big| \leq |(S \setminus S_G)| \leq n\alpha. \qquad (45)$$

Therefore,

$$\begin{aligned}
\sum_{\substack{x_i \in G\cap(S \setminus S_G), \\ x_i \geq \mu^{S_G}}} \big(x_i - \mu^{S_G}\big) &\leq \sum_{\substack{x_i \in G\cap(S \setminus S_G), \\ x_i \geq \mu^{S_G}}} \big(x_i - \mu^{P}\big) + |S \setminus S_G|\big|\mu^P - \mu^{S_G}\big| \\
&\leq \sum_{\substack{x_i \in G\cap(S \setminus S_G), \\ x_i \geq \mu^{S_G}}} \big(x_i - \mu^{P}\big) + n\alpha\big|\mu^P - \mu^{S_G}\big| \\
&\leq n\Big(7\nu\sqrt{k\alpha} + \alpha\mu^P + 18\nu\alpha\sqrt{k\alpha}\Big) \\
&\qquad\qquad \text{(Lemma D.5, part 3 and Proposition J.4)} \qquad (46) \\
&\leq n\Big((7 + 18\alpha)\nu\sqrt{k\alpha} + \alpha\mu^P\Big) \\
&\leq n\Big(8\nu\sqrt{k\alpha} + \alpha\mu^P\Big) \qquad (\because \alpha \leq 1/36 \leq 1/18.) \\
&\leq n\Big(8\nu\sqrt{k\alpha} + 16\alpha\mu^P\Big) \\
&= 16n\left(\alpha\mu^P + \frac{1}{2}\nu\sqrt{k\alpha}\right) \qquad (47)
\end{aligned}$$

where the last inequality follows by using Lemma D.5, part 3, in conjunction with Equation (45), and using the guarantee provided in Proposition J.4. Hence we have shown that

$$\sum_{\substack{x_i \in S \setminus S_G, \\ x_i \geq \mu^{S_G}}} \big(x_i - \mu^{S_G}\big) \geq 3 \sum_{\substack{x_i \in G\cap(S \setminus S_G), \\ x_i \geq \mu^{S_G}}} \big(x_i - \mu^{S_G}\big). \qquad (48)$$

Applying Fact D.7, we get

$$\mathbb{E}[|S \setminus S'|] = \frac{\sum\limits_{\substack{i \in S\setminus S_G, \\ x_i \geq \mu^{S_G}}} \left(x_i - \mu^{S_G}\right)}{\max\{x_i - \mu^{S_G}\}_{i \in S \setminus S_G}} \tag{49}$$

$$\text{and, } \mathbb{E}[|G \cap (S \setminus S')|] = \frac{\sum\limits_{\substack{i \in G\cap(S\setminus S_G), \\ x_i \geq \mu^{S_G}}} \left(x_i - \mu^{S_G}\right)}{\max\{x_i - \mu^{S_G}\}_{i \in S \setminus S_G}}. \tag{50}$$

This combined with Equation (48) gives

$$\begin{aligned} 3\,\mathbb{E}[|G \cap (S \setminus S')|] &\leq \mathbb{E}[|S \setminus S'|] \\ &= \mathbb{E}[|(L' \setminus L) \cup (E \setminus E')|] \\ &\leq \mathbb{E}[|L' \setminus L|] + \mathbb{E}[|E \setminus E'|]. \end{aligned} \tag{51}$$

Since $|L' \setminus L| = |G \cap (S \setminus S')|$, we finally have

$$\begin{aligned} \mathbb{E}[|G \cap (S \setminus S')|] &\leq \frac{1}{2}\,\mathbb{E}[|E \setminus E'|], \\ \implies \mathbb{E}[2|L'| + |E'|] &\leq 2|L| + |E|. \end{aligned} \tag{52}$$

Hence, we can run the filter on $\{\mathbf{X}_i \in S \setminus S_G\}$, and guarantee that on every application of Algorithm 3, we remove at-least 2 times more corrupted points in expectation than good points.

This completes the proof.

Fact D.7 (Filter based on mean)**.** *Assuming that $a \leq x_1 \leq \ldots \leq x_n \leq b$, $t \sim \mathcal{U}[a, b]$, then*

$$\mathbb{E}\left[\sum_{i=1}^{n} \mathbb{1}\{x_i > t\}\right] = \sum_{i=1}^{n} (x_i - a)/(b - a)$$

D.4 Proof of Lemma D.5

D.4.1 Proof of Lemma D.5, part 1

The basic idea of the proof is the following. First we construct a ϵ-net on semi-orthogonal matrices $\mathbf{V} \in \mathbb{R}^{d\times k}$, and argue that for each semi-orthogonal matrices $\mathbf{V}$ on the net, the set $\left\{\mathbf{X}_i \in G \,\middle|\, \mathrm{Tr}\left[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\right] \leq \Theta(\nu\sqrt{k/\epsilon})\right\}$ satisfies the three conditions in the lemma. Then, for each matrix $\mathbf{V}'$ that is not on the set, we argue there exists a $G_{\mathbf{V}}$ with $\mathbf{V}$ on the ϵ-net which satisfies the three conditions under V' in the lemma. We show the three conditions for a matrix on the net as follows.

Lemma D.8. *Let $G = \{x_i \sim \mathcal{P}\}_{i=1}^{n}$ where μ^P is the mean, and σ_P^2 is the variance of a real distribution $\mathcal{P}$. For any real number $0 < \epsilon \leq 1/18$, define set $T := \left\{x_i \in G \,\middle|\, \left|x_i - \mu^P\right| \leq 3\sigma_P/\sqrt{\epsilon}\right\}$, and $\mu^T := \frac{1}{|T|}\sum_{x_i \in T} x_i$. Then with probability at least $1 - 3\exp(-\Theta(n\epsilon))$,*

1. *$|T| \geq (1 - \epsilon)n$,*
2. *$\left|\mu^T - \mu^P\right| \leq \sigma_P\sqrt{\epsilon}$, and*
3. *$\frac{1}{|T|}\sum_{x_i \in T}\left(x_i - \mu^P\right)^2 \leq 2\sigma_P^2$.*

We provide a proof in Section D.5.

Proposition D.9 (Covering Number for Low-Rank Matrices, Lemma 3.1 in [12])**.** *Let $S_r := \{\mathbf{X} \in \mathbb{R}^{n_1 \times n_2} : \mathrm{rank}(\mathbf{X}) \leq r, \|\mathbf{X}\|_{\mathrm{F}} = 1\}$. Then there exists an ϵ-net $\bar{S}_{r,\epsilon} \subset S_r$ with respect to Frobenius norm obeying*

$$\left|\bar{S}_{r,\epsilon}\right| \leq (9/\epsilon)^{(n_1+n_2+1)r}.$$

With a union bound over all the elements of the net, we show that for each matrix $\mathbf{V}$ in the net, set $\left\{\mathbf{X}_i \in G \,\middle|\, \operatorname{Tr}\left[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\right] \le \Theta(\nu\sqrt{k/\epsilon})\right\}$ satisfies the three conditions in the lemma. The following technical proposition will helps us connect an arbitrary semi-orthogonal matrix $\mathbf{V}'$ to the net.

Proposition D.10. *With probability* $1-\delta$*, all subset* $T \subset G$ *such that* $|T| \ge (1-\epsilon)n$ *satisfies* $\frac{1}{|T|}\sum_{\mathbf{X}_i \in T}\|\mathbf{X}_i - \mathbf{M}\|_\mathrm{F}^2 \le \nu^2 d^2/\delta(1-\epsilon)$

Proof. For each $i, j \in [d]$, define matrix $\mathbf{E}^{i,j} \in \mathbb{R}^{d\times d}$ such that

$$E^{i,j}_{i',j'} = \begin{cases} 1 & \text{if } i' = i \text{ and } j' = j, \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$\begin{aligned} \mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{P}}\left[\|\mathbf{X} - \mathbf{M}\|_\mathrm{F}^2\right] &= \sum_{i,j\in[d]} \mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{P}}\left[\operatorname{Tr}[\mathbf{E}_{i,j}(\mathbf{X}-\mathbf{M})]^2\right] \\ &\le \nu^2 d^2, \end{aligned}$$

where the last inequality follows from the assumption in Lemma D.5. By Markov's inequality,

$$\begin{aligned} \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \|\mathbf{X}_i - \mathbf{M}\|_\mathrm{F}^2 \ge \nu^2 d^2/\delta\right] &\le \frac{\delta}{\nu^2 d^2}\,\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \|\mathbf{X}_i - \mathbf{M}\|_\mathrm{F}^2\right] \\ &\le \delta. \end{aligned} \tag{53}$$

Then for any subset $T \subset G$, and $|T| \ge (1-\epsilon)n$,

$$\frac{1}{|T|}\sum_{\mathbf{X}_i\in T}\|\mathbf{X}_i - \mathbf{M}\|_\mathrm{F}^2 \le \frac{1}{|T|}\sum_{\mathbf{X}_i\in G}\|\mathbf{X}_i - \mathbf{M}\|_\mathrm{F}^2 \le \nu^2 d^2/\delta(1-\epsilon). \qquad \square$$

We show Lemma D.5, part 1 as follows. By Proposition D.9, there exists an δ'-net of size $(9\sqrt{k}/\delta')^{(k+d+1)k}$ over the set $S_k = \{\mathbf{V} \in \mathbb{R}^{d\times k} : \operatorname{rank}(\mathbf{V}) \le k, \|\mathbf{V}\|_\mathrm{F} = \sqrt{k}\}$. Since the set of rank k projection matrices $P_k := \{\mathbf{V} \in \mathbb{R}^{d\times k}, \mathbf{V}^\top\mathbf{V} = \mathbf{I}_k\} \subset S_k$, there exists a δ'-net $\bar{P}_k$ of size $\left(18\sqrt{k}/\delta'\right)^{(k+d+1)k} = \exp(\Theta(dk\log(k/\delta')))$ of the set P_k. Fixing a projection matrix $\mathbf{V}$, the distribution of $\mathbf{V}^\top\mathbf{X}_i\mathbf{V}$ satisfies $\mathbb{E}\left[\operatorname{Tr}\left[\mathbf{V}^\top\mathbf{X}_i\mathbf{V}\right]\right] = \operatorname{Tr}\left[\mathbf{V}^\top\mathbf{M}\mathbf{V}\right]$ and

$$\mathbb{E}\left[\operatorname{Tr}\left[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\right]^2\right] = k^2\,\mathbb{E}\left[\operatorname{Tr}\left[\frac{1}{k}\mathbf{V}\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\right]^2\right] \le k\nu^2.$$

Thus, we can apply Lemma D.8 with a union bound to show that, given that

$$n = \Omega(dk^2\log(k/\delta\delta')/\epsilon),$$

with probability $1-\delta$, for each projection matrix $\mathbf{V} \in \bar{P}_k$, there exist a subset of random matrices $G_\mathbf{V} \subset G$ such that $G_\mathbf{V} := \left\{X_i \in G \,\middle|\, \left|\operatorname{Tr}\left[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\right]\right| \le 3\nu\sqrt{k/\epsilon}\right\}$, and it satisfies

1. $|G_\mathbf{V}| \ge (1-\epsilon)n$,
2. $\left|\frac{1}{|G_\mathbf{V}|}\sum_{\mathbf{X}_i\in G_\mathbf{V}}\operatorname{Tr}\left[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\right]\right| \le \nu\sqrt{k\epsilon}$, and
3. $\frac{1}{|G_\mathbf{V}|}\sum_{\mathbf{X}_i\in G_\mathbf{V}}\operatorname{Tr}\left[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\right]^2 \le 2k\nu^2$.

From now on, let us set

$$\delta' = \sqrt{\epsilon\delta}/100d.$$

The remaining argument conditions on that the event in Proposition D.10, which happens with probability $1-\delta$, namely, for all $G \subset S$ such that $|G| \ge (1-\epsilon)n$,

$$\frac{1}{|G|}\sum_{\mathbf{X}_i\in G}\|\mathbf{X}_i - \mathbf{M}\|_\mathrm{F}^2 \le \nu^2 d^2/\delta(1-\epsilon). \tag{54}$$

The following proposition deal with the projections that are not in $\bar{P}_k$.

Proposition D.11. *For any arbitrary projection matrix $\mathbf{V}'$, which may not be an element of $\bar{P}_k$, define projection matrix $\mathbf{V} \in \bar{P}_k$ such that $\|\mathbf{V} - \mathbf{V}'\|_{\mathrm{F}} \le \delta$. Then $G_{\mathbf{V}}$ is still an ϵ-good set for projection $\mathbf{V}'$, namely*

1. $|G_{\mathbf{V}}| \ge (1-\epsilon)n$,
2. $\left|\frac{1}{|G_{\mathbf{V}}|}\sum_{\mathbf{X}_i \in G_{\mathbf{V}}} \mathrm{Tr}\Big[\mathbf{V}'^{\top}(\mathbf{X}_i - \mathbf{M})\mathbf{V}'\Big]\right| \le 1.01\nu\sqrt{k\epsilon}$*, and*
3. $\frac{1}{|G_{\mathbf{V}}|}\sum_{\mathbf{X}_i \in G_{\mathbf{V}}} \mathrm{Tr}\Big[\mathbf{V}'^{\top}(\mathbf{X}_i - \mathbf{M})\mathbf{V}'\Big]^2 \le 6.01k\nu^2$.

Proof.

1. $|G_{\mathbf{V}}| \ge (1-\epsilon)n$ holds trivially.
2. Observe that

$$\begin{aligned}
&\left|\frac{1}{|G_{\mathbf{V}}|}\sum_{\mathbf{X}_i \in G_{\mathbf{V}}} \mathrm{Tr}\Big[\mathbf{V}'^{\top}(\mathbf{X}_i - \mathbf{M})\mathbf{V}'\Big]\right| \\
&\le \frac{1}{|G_{\mathbf{V}}|}\left(\mathrm{Tr}\left[\sum_{\mathbf{X}_i \in G_{\mathbf{V}}}\Big(\mathbf{V}'^{\top}(\mathbf{X}_i - \mathbf{M})(\mathbf{V}' - \mathbf{V})\Big)\right]\right. \\
&\qquad \left. + \mathrm{Tr}\left[\sum_{\mathbf{X}_i \in G_{\mathbf{V}}}\Big((\mathbf{V}' - \mathbf{V})^{\top}(\mathbf{X}_i - \mathbf{M})\mathbf{V}\Big)\right] + \mathrm{Tr}\left[\sum_{\mathbf{X}_i \in G_{\mathbf{V}}}\Big(\mathbf{V}^{\top}(\mathbf{X}_i - \mathbf{M})\mathbf{V}\Big)\right]\right) \\
&\le \frac{2}{|G_{\mathbf{V}}|}\sum_{\mathbf{X}_i \in G_{\mathbf{V}}}\|\mathbf{X}_i - \mathbf{M}\|_{\mathrm{F}}\delta' + \nu\sqrt{k\epsilon} \\
&\le 2\delta'\sqrt{\frac{1}{|G_{\mathbf{V}}|}\sum_{\mathbf{X}_i \in G_{\mathbf{V}}}\|\mathbf{X}_i - \mathbf{M}\|_{\mathrm{F}}^2} + \nu\sqrt{k\epsilon} \\
&\le 1.01\nu\sqrt{k\epsilon}
\end{aligned}$$

3. Notice that

$$\begin{aligned}
&\frac{1}{|G_{\mathbf{V}}|}\sum_{\mathbf{X}_i \in G_{\mathbf{V}}} \mathrm{Tr}\Big[\Big(\mathbf{V}'^{\top}(\mathbf{X}_i - \mathbf{M})\mathbf{V}'\Big)\Big]^2 \\
&= \frac{1}{|G_{\mathbf{V}}|}\sum_{\mathbf{X}_i \in G_{\mathbf{V}}}\left(\mathrm{Tr}\big[(\mathbf{V}'^{\top}(\mathbf{X}_i - \mathbf{M})(\mathbf{V}' - \mathbf{V}))\big] + \mathrm{Tr}\Big[\Big((\mathbf{V}' - \mathbf{V})^{\top}(\mathbf{X}_i - \mathbf{M})\mathbf{V}\Big)\Big]\right. \\
&\qquad \left. + \mathrm{Tr}\big[(\mathbf{V}^{\top}(\mathbf{X}_i - \mathbf{M})\mathbf{V})\big]\right)^2 \\
&\le \frac{3}{|G_{\mathbf{V}}|}\sum_{\mathbf{X}_i \in G_{\mathbf{V}}}\left(\mathrm{Tr}\Big[\Big(\mathbf{V}'^{\top}(\mathbf{X}_i - \mathbf{M})(\mathbf{V}' - \mathbf{V})\Big)\Big]^2\right. \\
&\qquad \left. + \mathrm{Tr}\Big[\Big((\mathbf{V}' - \mathbf{V})^{\top}(\mathbf{X}_i - \mathbf{M})\mathbf{V}\Big)\Big]^2 + \mathrm{Tr}\big[(\mathbf{V}^{\top}(\mathbf{X}_i - \mathbf{M})\mathbf{V})\big]^2\right)
\end{aligned} \tag{55}$$

Notice that by Cauchy-Schwarz,

$$\mathrm{Tr}\Big[\Big(\mathbf{V}'^{\top}(\mathbf{X}_i - \mathbf{M})(\mathbf{V}' - \mathbf{V})\Big)\Big]^2 \le \Big\|(\mathbf{V}' - \mathbf{V})^{\top}\mathbf{V}'\Big\|_{\mathrm{F}}^2\|\mathbf{X}_i - \mathbf{M}\|_{\mathrm{F}}^2 \le \delta'^2\|\mathbf{X}_i - \mathbf{M}\|_{\mathrm{F}}^2,$$

and likewise

$$\operatorname{Tr}\Big[\Big((\mathbf{V}' - \mathbf{V})^\top (\mathbf{X}_i - \mathbf{M})\mathbf{V}\Big)\Big]^2 \le \delta'^2 \|\mathbf{X}_i - \mathbf{M}\|_{\mathrm{F}}^2.$$

By Equation (54), Equation (55) is bounded by

$$6\nu^2 k + \nu^2\epsilon/10000 \le 6.01k\nu^2. \qquad \square$$

This completes all the proofs.

D.4.2 Proof of Lemma D.5, part 2

Proof of Lemma D.5, part 2. Since $\|\mathbf{X} - \mathbf{M}\|_2 \le B$, and $\max_{\|\mathbf{A}\|_* \le 1} \mathbb{E}\Big[(\operatorname{Tr}[\mathbf{A}(\mathbf{X}_i - \mathbf{M})])^2\Big] \le \nu^2$ for all $i \in G$, therefore, from Bernstein inequality, we have

$$\mathbb{P}\left[\left\|\frac{1}{n}\sum_{i\in G}\mathbf{X}_i - \mathbf{M}\right\|_2 \ge \nu\sqrt{\epsilon/k}\right] \le 2d\exp\left[\frac{-n^2\nu^2\epsilon/(2k)}{nd\nu^2 + Bn\nu\sqrt{\epsilon/k}/3}\right] \le \delta \tag{56}$$

Define matrix $\mathbf{A}^{(1,\mathbf{v})}, \mathbf{A}^{(2,\mathbf{v})}, \ldots, \mathbf{A}^{(d,\mathbf{v})}$ such that $\mathbf{A}^{(i,\mathbf{v})}_{i,*} = \mathbf{v}$, and the other rows of $\mathbf{A}^{(i,\mathbf{v})}$ are 0.

$$\begin{aligned}
\left\|\mathbb{E}\left[\sum_{i=1}^n (\mathbf{X}_i - \mathbf{M})^2\right]\right\|_2 &= \max_{\|\mathbf{v}\|_2 = 1} \sum_{i=1}^n \mathbb{E}\big[\mathbf{v}^\top(\mathbf{X}_i - \mathbf{M})(\mathbf{X}_i - \mathbf{M})\mathbf{v}\big] \\
&= \max_{\|\mathbf{v}\|_2 = 1} \sum_{i=1}^n \sum_{j=1}^d \mathbb{E}\left[\operatorname{Tr}\Big[\mathbf{A}^{(j,\mathbf{v})}(\mathbf{X}_i - \mathbf{M})\Big]^2\right] \\
&\le nd\nu^2
\end{aligned}$$

if $n \ge \frac{2}{\epsilon}\Big(dk + \frac{B}{3\nu}\sqrt{\epsilon k}\Big)\log\frac{2d}{\delta}$. Therefore for any semi-orthogonal matrix $\mathbf{V} \in \mathbb{R}^{d\times k}$,

$$\begin{aligned}
\left|\frac{1}{n}\sum_{\mathbf{X}_i \in G} \operatorname{Tr}\big[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big]\right| &\le \left|\frac{1}{n}\sum_{\mathbf{X}_i \in G} \operatorname{Tr}[\mathcal{P}_k(\mathbf{X}_i - \mathbf{M})]\right| \\
&\le k\left\|\frac{1}{n}\sum_{i\in G}\mathbf{X}_i - \mathbf{M}\right\|_2 \\
&\le \nu\sqrt{k\epsilon}, \qquad (57)
\end{aligned}$$

if $n = \Omega\Big(\Big(\frac{1}{\epsilon}\Big(dk + \frac{B}{\nu}\sqrt{k\epsilon}\Big)\Big)\log\frac{d}{\delta}\Big)$. $\square$

D.4.3 Proof of Lemma D.5, part 3

Here we show that for any semi-orthogonal matrix $\mathbf{V} \in \mathbb{R}^{d\times k}$ on the net $\bar{P}_k$, and any subset T, $\sum_{\mathbf{X}_i \in T} \operatorname{Tr}\big[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big] \le n\nu\sqrt{k/\epsilon}$ condition on the event defined in Lemma D.5, part 1 happens.

Denote

$$\begin{aligned}
G_L &= \Big\{\mathbf{X}_i \in G \,\Big|\, \operatorname{Tr}\big[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big] < -3\nu\sqrt{k/\epsilon}\Big\}, \\
G_M &= \Big\{\mathbf{X}_i \in G \,\Big|\, -3\nu\sqrt{k/\epsilon} \le \operatorname{Tr}\big[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big] \le 3\nu\sqrt{k/\epsilon}\Big\}, \\
G_H &= \Big\{\mathbf{X}_i \in G \,\Big|\, \operatorname{Tr}\big[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big] > 3\nu\sqrt{k/\epsilon}\Big\}.
\end{aligned}$$

Given a subset T, with $|T| \le \epsilon n$ let $T_L = \Big\{\mathbf{X}_i \in T \,\Big|\, \operatorname{Tr}\big[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big] < 3\nu\sqrt{k/\epsilon}\Big\}$ and $T_H = T \setminus T_L$. By definition, $\sum_{\mathbf{X}_i \in T_L} \operatorname{Tr}\big[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big] \le 3\nu\sqrt{k/\epsilon}|T_L|$.

$$
\begin{aligned}
&\sum_{\mathbf{X}_i \in T_H} \operatorname{Tr}\big[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big] \\
&\leq \sum_{\mathbf{X}_i \in G_H} \operatorname{Tr}\big[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big] \qquad (\because T_H \subseteq G_H) \\
&\leq \sum_{\mathbf{X}_i \in G_H} \operatorname{Tr}\big[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big] + \sum_{\mathbf{X}_i \in G_L} \operatorname{Tr}\big[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big] + |G_L| \operatorname{Tr}\big[\mathbf{V}^\top\mathbf{M}\mathbf{V}\big] \\
&= \sum_{\mathbf{X}_i \in G} \operatorname{Tr}\big[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big] - \sum_{\mathbf{X}_i \in G_M} \operatorname{Tr}\big[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big] + |G_L| \operatorname{Tr}\big[\mathbf{V}^\top\mathbf{M}\mathbf{V}\big] \\
&\leq \sum_{\mathbf{X}_i \in G} \operatorname{Tr}\big[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big] - \sum_{\mathbf{X}_i \in G_M} \operatorname{Tr}\big[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big] + n\epsilon \operatorname{Tr}\big[\mathbf{V}^\top\mathbf{M}\mathbf{V}\big] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Using Lemma D.5, part 1(a))} \\
&\leq 3n\nu\sqrt{k\epsilon} + n\epsilon \operatorname{Tr}\big[\mathbf{V}^\top\mathbf{M}\mathbf{V}\big],
\end{aligned}
$$

where the last inequality holds since $\sum_{\mathbf{X}_i \in G} \operatorname{Tr}\big[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big] \leq n\nu\sqrt{k\epsilon}$ by Lemma D.5, part 2, and $\sum_{\mathbf{X}_i \in G_M} \operatorname{Tr}\big[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big] \leq 2n\nu\sqrt{k\epsilon}$ by Lemma D.5, part 1(b) with $G_M = G_{\mathbf{V}}$ when $\mathbf{V}$ is the net $\bar{P}_k$. Combining the bound on T_L and T_H yields that $\sum_{\mathbf{X}_i \in T} \operatorname{Tr}\big[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big] \leq 6n\nu\sqrt{k\epsilon} + n\epsilon \operatorname{Tr}\big[\mathbf{V}^\top\mathbf{M}\mathbf{V}\big]$.

For a projection matrix $\mathbf{V}'$ not on the δ'-net $\bar{P}_k$,

$$
\begin{aligned}
&\sum_{\mathbf{X}_i \in T} \operatorname{Tr}\Big[\mathbf{V}'^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}'\Big] \\
&= \left(\operatorname{Tr}\left[\sum_{\mathbf{X}_i \in T} \Big(\mathbf{V}'^\top(\mathbf{X}_i - \mathbf{M})(\mathbf{V}' - \mathbf{V})\Big)\right] + \operatorname{Tr}\left[\sum_{\mathbf{X}_i \in T} \Big((\mathbf{V}' - \mathbf{V})^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\Big)\right] \right. \\
&\qquad\qquad \left. + \operatorname{Tr}\left[\sum_{\mathbf{X}_i \in T} \big(\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big)\right] \right) \\
&\leq 2\sum_{\mathbf{X}_i \in T} \|\mathbf{X}_i - \mathbf{M}\|_F \delta' + 6n\sqrt{k\epsilon} + n\epsilon \operatorname{Tr}\big[\mathbf{V}^\top\mathbf{M}\mathbf{V}\big] \\
&\leq 2\delta'\sqrt{|T|}\sqrt{\sum_{\mathbf{X}_i \in T} \|\mathbf{X}_i - \mathbf{M}\|_F^2} + 6n\sqrt{k\epsilon} + n\epsilon \operatorname{Tr}\big[\mathbf{V}^\top\mathbf{M}\mathbf{V}\big] \\
&\leq \epsilon\nu n + 6\nu n\sqrt{k\epsilon} + n\epsilon \operatorname{Tr}\big[\mathbf{V}^\top\mathbf{M}\mathbf{V}\big] \\
&\leq 7\nu n\sqrt{k\epsilon} + n\epsilon \operatorname{Tr}\big[\mathbf{V}^\top\mathbf{M}\mathbf{V}\big] \qquad (58)
\end{aligned}
$$

D.5 Proof of Lemma D.8

1. Using Markov's inequality we get

$$
\begin{aligned}
\mathbb{P}_{x\sim\mathcal{P}}\big[\big|x - \mu^P\big| \geq z\big] &\leq \frac{\mathbb{E}_{x\sim\mathcal{P}}\big[(x - \mu^P)^2\big]}{z^2} \\
&\leq \frac{\sigma_P^2}{z^2} \\
\implies \mathbb{P}_{x\sim\mathcal{P}}\big[\big|x - \mu^P\big| \geq \sigma_P/\sqrt{\epsilon}\big] &\leq \epsilon.
\end{aligned}
$$

Let us define the indicator random variable $Z_i := \mathbb{1}\{|x_i - \mu^P| \geq \sigma_P/\sqrt{\epsilon}\}$, and let $p := \mathbb{E}[Z_i]$. Then from the Chernoff bound we get,

$$\mathbb{P}_{G\sim\mathcal{P}^n}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i \geq (1+z)p\right] \leq \exp\left[-\frac{z^2pn}{3}\right].$$

Set $(1+z)p = 2\epsilon$, we get

$$\begin{aligned}\mathbb{P}_{G\sim\mathcal{P}^n}\left[\frac{1}{n}\sum_{i=1}^{n} Z_i \geq 2\epsilon\right] &\leq \exp\left[-\frac{(2\epsilon/p-1)^2pn}{3}\right]\\ &\leq \exp\left[-\frac{\epsilon n}{3}\right],\end{aligned}$$

where the last inequality holds since $p \leq \epsilon$. This implies

$$\mathbb{P}[|T| \leq (1-2\epsilon)n] \leq \exp[-\epsilon n/3]$$

2. Define event $\mathcal{E} := \{x : |x - \mu^P| \leq \sigma_P/\sqrt{\epsilon}\}$. In order to show that $\mu^T = \frac{1}{|T|}\sum_{x_i\in T} x_i$ concentrates to μ^P, we will (1) apply Bernstein inequality to argue that μ^T concentrate around $\mu^{P'}$ which is the mean of $\mathcal{P}'$, the distribution of x conditioned on the event $\mathcal{E}$, and (2) argue that the mean $\mu^{P'}$ is close to μ^P, which, by triangle inequality, concludes the proof. First, we prove a bound on $\left|\mu^{P'} - \mu^P\right|$, thus finishing part (2) of the proof.

Proposition D.12.

$$\left|\mu^{P'} - \mu^P\right| \leq 2\sqrt{\epsilon}\sigma_P.$$

Proof. Notice that

$$\begin{aligned}\left|\mu^{P'} - \mu^P\right| &= \left|\mathop{\mathbb{E}}_{X\sim\mathcal{P}}[X \mid \mathcal{E}] - \mu^P\right|\\ &= \left|\mathop{\mathbb{E}}_{X\sim\mathcal{P}}[X - \mu^P \mid \mathcal{E}]\right|\\ &= \frac{1}{\mathbb{P}[\mathcal{E}]}\left|\mathop{\mathbb{E}}_{X\sim\mathcal{P}}\left[(X-\mu^P)\mathbb{1}\{|X-\mu^P| \leq \sigma_P/\sqrt{\epsilon}\}\right]\right|\\ &= \frac{1}{\mathbb{P}[\mathcal{E}]}\left|\mathop{\mathbb{E}}_{X\sim\mathcal{P}}\left[(X-\mu^P)\mathbb{1}\{|X-\mu^P| \geq \sigma_P/\sqrt{\epsilon}\}\right]\right|\\ &\leq \frac{1}{\mathbb{P}[\mathcal{E}]}\sqrt{\mathop{\mathbb{E}}_{X\sim\mathcal{P}}\left[(X-\mu^P)^2\right]\mathop{\mathbb{E}}_{X\sim\mathcal{P}}\left[\mathbb{1}\{|X-\mu^P| \geq \sigma_P/\sqrt{\epsilon}\}\right]} \qquad (59)\\ &\leq \frac{\sqrt{1-\mathbb{P}[\mathcal{E}]}}{\mathbb{P}[\mathcal{E}]}\sigma_P \qquad (60)\\ &\leq 2\sqrt{\epsilon}\sigma_P,\end{aligned}$$

where Equation (59) holds by Cauchy-Schwarz, and Equation (60) holds since

$$\mathop{\mathbb{E}}_{X\sim\mathcal{P}}\left[\left(X-\mu^P\right)^2\right] \leq \sigma_P^2.$$

□

To show part (1), let us first show the following simple fact due to Bernstein inequality

Proposition D.13. *Given n iid samples $X_1, \ldots, X_n$ from distribution $\mathcal{P}'$, it holds that*

$$\mathbb{P}_{X_i\sim\mathcal{P}'}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu^{P'}\right| \geq \sigma_P\sqrt{\epsilon}\right] \leq 2\exp(-n\epsilon/13)$$

Proof. First we bound the variance of $\mathcal{P}'$,

$$\begin{aligned}\mathop{\mathbb{E}}_{X\sim\mathcal{P}'}\left[\left(X-\mu^{P'}\right)^2\right] &= \frac{1}{\mathbb{P}[\mathcal{E}]}\mathop{\mathbb{E}}_{X\sim\mathcal{P}}\left[\left(X-\mu^{P'}\right)^2\mathbb{1}\left\{\left|X-\mu^P\right|\le\sigma_P/\sqrt{\epsilon}\right\}\right]\\ &\le \frac{1}{\mathbb{P}[\mathcal{E}]}\mathop{\mathbb{E}}_{X\sim\mathcal{P}}\left[\left(X-\mu^{P'}\right)^2\right]\\ &= \frac{1}{\mathbb{P}[\mathcal{E}]}\mathop{\mathbb{E}}_{X\sim\mathcal{P}}\left[\left(X-\mu^{P}\right)^2+\left(\mu^P-\mu^{P'}\right)^2\right]\\ &\le \frac{1}{\mathbb{P}[\mathcal{E}]}\left(\sigma_P^2+4\epsilon\sigma_P^2\right)\\ &\le 6\sigma_P^2.\end{aligned}$$

Hence, we can apply Bernstein inequality (Proposition J.2) and get

$$\begin{aligned}\mathbb{P}_{X_i\sim\mathcal{P}'}\left[\left|\frac{1}{n}\sum_{i=1}^n X_i-\mu^{P'}\right|\ge\sigma_P\sqrt{\epsilon}\right] &\le 2\exp\left[\frac{-n^2\sigma_P^2\epsilon/2}{6n\sigma_P^2+n\sigma_P^2/3}\right]\\ &\le 2\exp(-n\epsilon/13).\end{aligned}$$ □

Notice that condition on the size of set T, each $X_i\in T$ follows from distribution $\mathcal{P}'$ independently. Hence, we have that condition on the size of T, with probability at least $1-2\exp(-|T|\epsilon/13)$, it holds that $\left|\frac{1}{|T|}\sum_{X_i\in T}X_i-\mu^{P'}\right|\le\sigma_P\sqrt{\epsilon}$. Since we have shown that $|T|\ge n(1-2\epsilon)$ with probability $1-\exp(-n\epsilon/3)$, by a union bound, we conclude that with probability at-least $1-3\exp(-n\epsilon/13)$,

$$\left|\frac{1}{|T|}\sum_{X_i\in T}X_i-\mu^{P'}\right|\le\sigma_P\sqrt{\epsilon}.$$

Combining Proposition D.12 with triangle inequality yields that with probability $1-3\exp(-n\epsilon/13)$,

$$\left|\mu^G-\mu^P\right|=\left|\frac{1}{|T|}\sum_{X_i\in T}X_i-\mu^P\right|\le 3\sigma_P\sqrt{\epsilon}.$$

3. Let us define the function $y(x):=\left(x-\mu^P\right)\cdot\mathbb{1}\left\{\left|x-\mu^P\right|\le\sigma_P/\sqrt{\epsilon}\right\}$. This implies $|y(X)|\le\sigma_P/\sqrt{\epsilon}$. Then,

$$\mathop{\mathbb{E}}_{X\sim\mathcal{U}(G)}\left[y(X)^2\right]\le\left|\mathop{\mathbb{E}}_{X\sim\mathcal{U}(G)}\left[y(X)^2\right]-\mathop{\mathbb{E}}_{X\sim\mathcal{P}}\left[y(X)^2\right]\right|+\mathop{\mathbb{E}}_{X\sim\mathcal{P}}\left[y(X)^2\right] \tag{61}$$

Looking at the above two terms individually, with the second term

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim\mathcal{P}}\left[y(x)^2\right] &= \mathop{\mathbb{E}}_{x\sim\mathcal{P}}\left[\left(x-\mu^P\right)^2\cdot\mathbb{1}\left\{\left|x-\mu^P\right|\le\sigma_P/\sqrt{\epsilon}\right\}\right]\\ &\le \mathop{\mathbb{E}}_{X\sim\mathcal{P}}\left[\left(x-\mu^P\right)^2\right]\\ &\le \sigma_P^2. \qquad (62)\end{aligned}$$

For the first term we apply [72, Lemma 5.44.] on the random variable $y(x)$ for $x\sim\mathcal{P}$, and get that with probability $1-\exp(-C\epsilon n)$ for a constant $C>0$,

$$\left|\mathop{\mathbb{E}}_{X\sim\mathcal{U}(G)}\left[y(x)^2\right]-\mathop{\mathbb{E}}_{x\sim\mathcal{P}}\left[y(x)^2\right]\right|\le\frac{\sigma_P^2}{2}. \tag{63}$$

Applying the fact that $|T|\ge(1-2\epsilon)n$ holds with probability $1-\exp(-\epsilon n/3)$, and plugging Equation (62) and (63) in Equation (61), we get that with probability $1-\exp(-\Theta(\epsilon n))$

$$\frac{1}{|T|}\sum_{X_i\in T}\left(X_i-\mu^P\right)^2=\frac{n}{|T|}\mathop{\mathbb{E}}_{X\sim\mathcal{U}(G)}\left[y(X)^2\right]\le 2\sigma_P^2. \tag{64}$$

Taking a union bound of the probability of the three conditions and replacing ϵ by $\epsilon/9$ yield the statement of the lemma.

D.6 Proof of Lemma D.3

We claim that

$$\left\|\left(\mathbf{I}-\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\top}\right)\mathbf{V}_j\right\|_* \;\leq\; \gamma/\sigma_j \;, \tag{65}$$

where $\mathbf{V}_j = [\mathbf{v}_1 \,\ldots\, \mathbf{v}_j]$ is the matrix consisting of the j singular vectors of $\mathbf{M} = \sigma^2\mathbf{I} + \sum_{i'\in[k]} \mathbf{X}_{i'}$ corresponding to the top j singular values, and σ_j is the j-th singular value. This follows from the proof of the gap-free Wedin's theorem in [1, Lemma B.3], which proves a similar bound on the spectral norm. Concretely, let $\widehat{\mathbf{U}}_\perp \in \mathbb{R}^{d\times(d-k)}$ denote an orthogonal matrix spanning the null space of $\widehat{\mathbf{U}}^\top$. We can write the singular value decomposition as

$$\widehat{\mathbf{M}} \;=\; \widehat{\mathbf{U}}\widehat{\mathbf{D}}\widehat{\mathbf{U}}^\top \;, \quad \mathbf{B} = \mathbf{V}_j\mathbf{D}\mathbf{V}_j^\top + \mathbf{V}'_j\mathbf{D}'\mathbf{V}'^{\top}_j \;, \tag{66}$$

where $\mathbf{V}'_j$ spans the subspace orthogonal to $\mathbf{V}_j$, and $\mathbf{B} = \sum_{i=1}^k \mathbf{X}_i + \sigma^2\mathbf{U}\mathbf{U}^\top$. Let $\mathbf{R} = \widehat{\mathbf{M}} - \mathbf{B}$, and we get

$$\begin{aligned}\widehat{\mathbf{U}}_\perp^\top\widehat{\mathbf{M}}\mathbf{V}_j &= \widehat{\mathbf{U}}_\perp^\top\mathbf{R}\mathbf{V}_j + \widehat{\mathbf{U}}_\perp^\top\mathbf{B}\mathbf{V}_j \\ &= \widehat{\mathbf{U}}_\perp^\top\mathbf{R}\mathbf{V}_j + \widehat{\mathbf{U}}_\perp^\top\mathbf{V}_j\mathbf{D} \;.\end{aligned}$$

Since $\widehat{\mathbf{U}}_\perp^\top\widehat{\mathbf{M}} = \mathbf{0}$, taking nuclear norm and applying the triangular inequality,

$$\begin{aligned}\left\|\widehat{\mathbf{U}}_\perp^\top\mathbf{V}_j\right\|_* &= \left\|\widehat{\mathbf{U}}_\perp^\top\mathbf{R}\mathbf{V}_j\mathbf{D}^{-1}\right\|_* \\ &\leq \gamma/\sigma_j \;.\end{aligned}$$

To get the first term on the upper bound (29), we follow the analysis of [47, Lemma 5]. Notice that $\mathbf{x}_i$ lie on the subspace spanned by $\mathbf{V}_j$ where j is the rank of $\sum_{i'\in[k]}\mathbf{X}_{i'}$. It follows from $\left\|\left(\mathbf{I}-\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\top}\right)\mathbf{V}_j\right\|_{\rm F} \leq \left\|\left(\mathbf{I}-\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\top}\right)\mathbf{V}_j\right\|_* \leq \gamma/\sigma_j$ with a choice of $j = k$ that

$$\begin{aligned}\sum_{i\in[k]}\left\|\left(\mathbf{I}-\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\top}\right)\mathbf{V}_j\mathbf{V}_j^\top\mathbf{x}_i\right\|_2^2 &= \operatorname{Tr}\left[\left(\mathbf{I}-\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\top}\right)\mathbf{V}_j\mathbf{V}_j^\top\left(\sum_{i\in[k]}\mathbf{x}_i\mathbf{x}_i^\top\right)\mathbf{V}_j\mathbf{V}_j^\top\left(\mathbf{I}-\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\top}\right)\right] \\ &\leq \left\|\sum_{i\in[k]}\mathbf{x}_i\mathbf{x}_i^\top\right\|_2 \left\|\mathbf{V}_j\mathbf{V}_j^\top\left(\mathbf{I}-\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\top}\right)\left(\mathbf{I}-\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\top}\right)\mathbf{V}_j\mathbf{V}_j^\top\right\|_* \\ &\leq \sigma_{\max}\gamma^2/\sigma_{\min}^2 \;.\end{aligned}$$

Next, we optimize over this choice of j to get the tightest bound that does not depend on the singular values. Applying a similar bound as the above series of inequalities, we get

$$\begin{aligned}\sum_{i\in[k]}\left\|\left(\mathbf{I}-\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\top}\right)\mathbf{x}_i\right\|_2^2 &= \sum_{i\in[k]}\left\|\left(\mathbf{I}-\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\top}\right)\mathbf{V}_j\mathbf{V}_j^\top\mathbf{x}_i\right\|_2^2 + \sum_{i\in[k]}\left\|\left(\mathbf{I}-\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\top}\right)(\mathbf{I}-\mathbf{V}_j\mathbf{V}_j^\top)\mathbf{x}_i\right\|_2^2 \\ &\leq \left(\sigma_{\max}\gamma^2/\sigma_j^2\right) + (k-j)\sigma_{j+1} \;,\end{aligned}$$

where we used the fact that

$$\begin{aligned}\sum_{i\in[k]}\left\|\left(\mathbf{I}-\widehat{\mathbf{U}}\widehat{\mathbf{U}}^{\top}\right)(\mathbf{I}-\mathbf{V}_j\mathbf{V}_j^\top)\mathbf{x}_i\right\|_2^2 &\leq \operatorname{Tr}\left[(\mathbf{I}-\mathbf{V}_j\mathbf{V}_j^\top)\left(\sum_{i\in[k]}\mathbf{x}_i\mathbf{x}_i^\top\right)(\mathbf{I}-\mathbf{V}_j\mathbf{V}_j^\top)\right] \\ &\leq (k-j)\sigma_{j+1} \;.\end{aligned}$$

A good choice of j to approximately minimizes the upper bound is for the two terms to be of similar orders. Precisely, we choose j to be the largest index such that $\sigma_j \geq \gamma^{2/3}\sigma_{\max}^{1/3}(k-j)^{-1/3}$ (we take $j=0$ if $\sigma_1 \leq \gamma^{2/3}\sigma_{\max}^{1/3}k^{-1/3}$). This gives an upper bound of $2\gamma^{2/3}\sigma_{\max}^{1/3}k^{2/3}$.

The second upper bound in (29) follows from a similar argument, and is a direct corollary of [42, Lemma A.11].

E Proof sketch of the adaption to exponential tail setting

In this setting, we give a sketch of how to adapt the proof of Algorithm 2 to the setting where the distribution has exponential like tail.

Closing the gap in Outlier-Robust PCA (ORPCA). [75, 28, 76, 19] study robust PCA under the assumption that each sample $\mathbf{z}_i = \mathbf{A}\mathbf{x}_i + \mathbf{v}_i$ where $\mathbf{x}_i, \mathbf{v}_i$ are drawn from isotropic Gaussian distribution, and the goal is to learn the top-k eigenspace of $\mathbf{A}\mathbf{A}^\top$. When n samples are observed, α fraction of which are corrupted by an adversary, [75] introduces a filtering algorithm to find a subspace $\widehat{\mathbf{U}}$ achieving:

$$\left\|\widehat{\mathbf{U}}^\top \mathbf{A}\mathbf{A}^\top \widehat{\mathbf{U}}\right\|_* \geq (1 - c\sqrt{\alpha}\log(1/\alpha)) \left\|\mathbf{A}\mathbf{A}^\top\right\|_* ,$$

for some $c > 0$ when $n/k(\log k)^5 \to \infty$. For the reasons explained in §2.2, this is sub-optimal in α in comparisons to an information theoretic lower bound with a multiplicative factor of $(1 - c\,\alpha)$. Applying the proposed Algorithm 2, it is possible to generalize Proposition 2.6 to this Gaussian setting and achieve an optimal upper bound. To get some intuition, notice that with our assumption on the second moment of the projected variance $\mathrm{Tr}\left[\mathbf{U}^\top \mathbf{X}_i \mathbf{U}\right]$, our current proof proceeds by focusing on that $1 - \alpha$ probability mass, which falls in the interval $[-\Theta(\sqrt{1/\alpha}), \Theta(\sqrt{1/\alpha})]$. This is tight with only the second moment assumption. However, if we assume exponential tail on $\mathrm{Tr}\left[\mathbf{U}^\top \mathbf{X}_i \mathbf{U}\right]$, namely $\mathbb{P}\left[\mathrm{Tr}\left[\mathbf{U}^\top \mathbf{X}_i \mathbf{U}\right] > t\right] \leq \exp(-t^p)$, for some $p > 0$, we can instead focus on the probability mass in $[-\Theta(\log^{1/p}(1/\alpha)), \Theta(\log^{1/p}(1/\alpha))]$ which is also at least $1 - \alpha$. The would give an error of $\alpha \log^{1/p}(1/\alpha)$. We provide a sketch of how to adapt the proof of our algorithm to the exponential tail setting as follows.

Lemma E.1 (Main Lemma for Algorithm 2, adaption from Lemma D.4)**.** *Let $\mathcal{P}$ be a distribution over $d \times d$ PSD matrices with the property that,*

$$\mathop{\mathbb{E}}_{\mathbf{X}\sim\mathcal{P}}[\mathbf{X}] = \mathbf{M} , \quad \|\mathbf{X} - \mathbf{M}\|_2 \leq B ,$$

$$\textit{and} \qquad \max_{\|\mathbf{A}\|_F \leq 1, \mathrm{rank}(\mathbf{A}) \leq k} \mathbb{P}_{\mathbf{X}\sim\mathcal{P}}[|\mathrm{Tr}[\mathbf{A}(\mathbf{X} - \mathbf{M})]| \geq \nu(k)t] \leq \exp(-t^p) ,$$

for some $p > 0$. Let a set of n random matrices $G = \left\{\mathbf{X}_i \in \mathbb{R}^{d\times d}\right\}_{i\in[n]}$ where each $\mathbf{X}_i$ is independently drawn from $\mathcal{P}$, and the at most α fraction is corrupted by an adversary such that the input dataset $S = (G \setminus L) \cup E$ with $|E| = |L| \leq \alpha n$, $L \subset G$. There exists a numerical constant $c > 0$ such that for any $0 < \alpha < c$, if $n = \widetilde{\Omega}((dk^2 + (B/\nu)\sqrt{k}\alpha)/\alpha^2)$, Algorithm 2 outputs a dataset $S' \subseteq S$ satisfying the following for $\widehat{\mathbf{M}} = \frac{1}{|S'|}\sum_{\mathbf{X}_i \in S'} \mathbf{X}_i$:

1. *for the top-k singular vectors $\widehat{\mathbf{U}} \in \mathbb{R}^{d\times k}$ of $\widehat{\mathbf{M}}$,*

$$\mathrm{Tr}\left[\widehat{\mathbf{U}}^\top \left(\widehat{\mathbf{M}} - \mathbf{M}\right)\widehat{\mathbf{U}}\right] \quad \leq \quad \mathcal{O}\left(\alpha\,\mathrm{Tr}\left[\widehat{\mathbf{U}}^\top \mathbf{M} \widehat{\mathbf{U}}\right] + \nu\sqrt{k}\alpha \log(1/\alpha)^{1/p}\right) .$$

2. *for all rank-k semi-orthogonal matrices $\mathbf{V} \in \mathbb{R}^{d\times k}$, we have*

$$\mathrm{Tr}\left[\mathbf{V}^\top \left(\widehat{\mathbf{M}} - \mathbf{M}\right)\mathbf{V}\right] \quad \geq \quad -\mathcal{O}\left(\alpha\,\mathrm{Tr}\left[\mathbf{V}^\top \mathbf{M}\mathbf{V}\right] + \nu\sqrt{k}\alpha \log(1/\alpha)^{1/p}\right) .$$

Notice that the probability mass beyond $\left|\mathrm{Tr}\left[\sum_{\mathbf{X}_i \in G_\mathbf{V}} \left(\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\right)\right]\right| \geq \nu\sqrt{k}\log(1/\epsilon)^{1/p}$ is less than ϵ by the exponential tail bound. Hence similar to Lemma D.5, by letting $G_\mathbf{V} \subset G$ to contain all the points in G such that $\left|\mathrm{Tr}\left[\sum_{\mathbf{X}_i \in G_\mathbf{V}} \left(\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\right)\right]\right| \leq \nu\sqrt{k}\log(1/\epsilon)^{1/p}$, we have part 1 of the following lemma. Part 2 of the following lemma follows from matrix Bernstein inequality, and part 3 can be shown with the same argument of Lemma D.5.

Lemma E.2 (Adaption from Lemma D.5)**.** *Under the hypotheses of Lemma E.1, when $n = \widetilde{\Omega}((dk^2 + \frac{B}{\nu}\sqrt{k}\epsilon)/\epsilon^2)$ with probability $1 - \widetilde{\delta}$, the following events happen for all semi-orthogonal matrices $\mathbf{V} \in \mathbb{R}^{d\times k} s.t. \mathbf{V}^\top\mathbf{V} = \mathbf{I}_k$,*

1. *There exists $G_\mathbf{V} \subset G$ such that*

   - *(a)* $|G_{\mathbf{V}}| \geq (1-\epsilon)n$,
   - *(b)* $\left|\frac{1}{|G_{\mathbf{V}}|}\operatorname{Tr}\big[\sum_{\mathbf{X}_i \in G_{\mathbf{V}}}\big(\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big)\big]\right| \leq \nu\sqrt{k}\epsilon\log(1/\epsilon)^{1/p}$, *and*
   - *(c)* $\left|\operatorname{Tr}\big[\sum_{\mathbf{X}_i \in G_{\mathbf{V}}}\big(\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big)\big]\right| \leq \nu\sqrt{k}\log(1/\epsilon)^{1/p}$,

2. $\left|\frac{1}{n}\sum_{\mathbf{X}_i \in G}\operatorname{Tr}\big[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big]\right| \leq \nu\sqrt{k}\epsilon\log(1/\epsilon)^{1/p}$,

3. *All subset $T \subset G$ such that $|T| \leq \epsilon n$ satisfies*
$$\sum_{\mathbf{X}_i \in T}\operatorname{Tr}\big[\mathbf{V}^\top(\mathbf{X}_i - \mathbf{M})\mathbf{V}\big] \leq 7n\nu\sqrt{k}\epsilon\log(1/\epsilon)^{1/p} + n\epsilon\operatorname{Tr}\big[\mathbf{V}^\top\mathbf{M}\mathbf{V}\big].$$

Thus by changing line 5 of Algorithm 3 to $48(\alpha\mu^{S_G} + \nu\sqrt{k}\alpha\log(1/\alpha)^{1/p})$, the statement in Proposition D.6 can be changed to that either the output of Algorithm 3 satisfies

$$\operatorname{Tr}\left[\widehat{\mathbf{U}}^\top\left(\frac{1}{|S'|}\sum_{\mathbf{X}_i \in S'}\mathbf{X}_i - \mathbf{M}\right)\widehat{\mathbf{U}}\right] \;\leq\; \mathcal{O}\Big(\alpha\operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top\mathbf{M}\widehat{\mathbf{U}}\Big] + \nu\sqrt{k}\alpha\log(1/\alpha)^{1/p}\Big) \tag{67}$$

or the algorithm removes more corrupted data points than uncorrupted data points in expectation. Finally, similar to Proposition 2.6, we get

$$\begin{aligned}
\operatorname{Tr}[\mathcal{P}_k(\mathbf{\Sigma})] - \operatorname{Tr}\Big[\widehat{\mathbf{U}}^\top\mathbf{\Sigma}\widehat{\mathbf{U}}\Big] &= \mathcal{O}\Big(\alpha\operatorname{Tr}[\mathcal{P}_k(\mathbf{\Sigma})] + \nu\sqrt{k}\alpha\log(1/\alpha)^{1/p}\Big)\,, \\
\text{and}\quad \Big\|\mathbf{\Sigma} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{\Sigma}\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\Big\|_* &\leq \|\mathbf{\Sigma} - \mathcal{P}_k(\mathbf{\Sigma})\|_* + \mathcal{O}\Big(\alpha\|\mathcal{P}_k(\mathbf{\Sigma})\|_* + \nu\sqrt{k}\alpha\log(1/\alpha)^{1/p}\Big)\,.
\end{aligned}$$

F Lower bound for robust PCA, proof of Proposition 2.7

In this section we show that under the setting of Proposition 2.6, it is information theoretically impossible to learn subspace $\widehat{\mathbf{U}}$ such that $\|\mathbf{\Sigma} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\mathbf{\Sigma}\widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\| = o\Big(\nu(k)\sqrt{k\alpha}\Big)$.

Definition F.1. *Given a subset $I \subset [d], |I| = k$, define distribution $\mathcal{P}_I$ as follows: Suppose random variable $\mathbf{x} \sim \mathcal{P}_I$, and each coordinate $x_i, i \in I$ is sampled independently such that*

$$x_i = \begin{cases} \sqrt{\nu(k)} & \textit{with probability } (1-\alpha/k)/2 \\ -\sqrt{\nu(k)} & \textit{with probability } (1-\alpha/k)/2 \\ (\nu(k)^2k/\alpha)^{1/4} & \textit{with probability } \alpha/2k \\ -(\nu(k)^2k/\alpha)^{1/4} & \textit{with probability } \alpha/2k \end{cases}.$$

The other coordinates $x_i, i \notin I$ is sampled independently such that

$$x_i = \begin{cases} \sqrt{\nu(k)} & \textit{with probability } 1/2 \\ -\sqrt{\nu(k)} & \textit{with probability } 1/2 \end{cases}.$$

The second moment matrix $\mathbf{\Sigma}^I := \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_I}\big[\mathbf{x}\mathbf{x}^\top\big]$ of $\mathcal{P}_I$ satisfies

$$\Sigma^I_{i,j} = \begin{cases} \nu(k)(1-\alpha/k+\sqrt{\alpha/k}) & i=j, i \in I \\ \nu(k) & i=j, i \notin I \\ 0 & i \neq j \end{cases}.$$

It is clear that with probability 1, $\big\|\mathbf{x}\mathbf{x}^\top - \mathbf{\Sigma}^I\big\|_2 \leq \|\mathbf{x}\|_2^2 + \big\|\mathbf{\Sigma}^I\big\|_2 \leq \nu(k)\big(d + 2k^{3/2}/\sqrt{\alpha}\big) \leq 2\nu(k)d = B$. Next we verify the fourth moment condition in Proposition 2.6. WLOG, let us assume $A_{i,j} = 0$ for any $j < i$, hence $\mathbf{A}$ is a upper triangular.

$$\begin{aligned}
\mathbb{E}\Big[\operatorname{Tr}\big[\mathbf{A}\big(\mathbf{x}\mathbf{x}^\top - \mathbf{\Sigma}^I\big)\big]^2\Big] &= \mathbb{E}\left[\left(\sum_{i,j}A_{i,j}\big(x_ix_j - \Sigma^I_{i,j}\big)\right)^2\right] \\
&= \mathbb{E}\left[\sum_{i,j}\sum_{i',j'}A_{i,j}\big(x_ix_j - \Sigma^I_{i,j}\big)A_{i',j'}\big(x_{i'}x_{j'} - \Sigma^I_{i',j'}\big)\right]
\end{aligned}$$

Based on the number of distinct values i, j, i', j' take, the terms inside the summation can be classified into 4 difference cases

1. i, j, i', j' takes 4 difference values. In this case,
$$\mathbb{E}\big[A_{i,j}\big(x_i x_j - \Sigma^I_{i,j}\big)A_{i',j'}\big(x_{i'}x_{j'} - \Sigma^I_{i',j'}\big)\big] = 0$$
2. i, j, i', j' takes 3 difference values. In this case,
$$\mathbb{E}\big[A_{i,j}\big(x_i x_j - \Sigma^I_{i,j}\big)A_{i',j'}\big(x_{i'}x_{j'} - \Sigma^I_{i',j'}\big)\big] = 0$$
3. i, j, i', j' takes 2 difference values. In this case, if $i = j$, $i' = j'$ and $i \neq i'$,
$$\mathbb{E}\big[A_{i,i}\big(x_i^2 - \Sigma^I_{i,i}\big)A_{i',i'}\big(x_{i'}^2 - \Sigma^I_{i',i'}\big)\big] = 0.$$
If $i = i'$, $j = j'$ and $i \neq j$,
$$\mathbb{E}\Big[A_{i,j}^2\big(x_i x_j - \Sigma^I_{i,j}\big)^2\Big] = \mathbb{E}\big[A_{i,j}^2(x_i x_j)^2\big] = \mathbb{E}\big[A_{i,j}^2\Sigma^I_{i,i}\Sigma^I_{j,j}\big].$$
If $i = j'$, $j = i'$ and $i \neq j$, $A_{i,j}$ or $A_{j,i}$ must be 0, hence the expectation is 0.
4. i, j, i', j' takes 1 value. In this case
$$\mathbb{E}\Big[A_{i,i}^2\big(x_i^2 - \Sigma^I_{i,i}\big)^2\Big] = \begin{cases} 0 & i \notin I \\ A_{i,i}^2\nu(k)^2(2-\alpha/k) & i \in I \end{cases}$$

Taking summation over the above cases yields the following bound
$$\sum_{i<j} A_{i,j}^2\Sigma^I_{i,i}\Sigma^I_{j,j} + \sum_{i\in I} A_{i,i}^2(2-\alpha/k) \leq 2\|\mathbf{A}\|_{\mathrm{F}}^2\nu(k)^2 \leq 2\nu(k)^2.$$

Here we have shown that each distribution $\mathcal{P}_I$ satisfies
$$\max_{\|\mathbf{A}\|_{\mathrm{F}}\leq 1} \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{P}_I}\Big[\big(\big\langle \mathbf{A}, \mathbf{x}\mathbf{x}^\top - \mathbb{E}\big[\mathbf{x}\mathbf{x}^\top\big]\big\rangle\big)^2\Big] \leq 2\nu(k)^2.$$

Then we define the base case distribution $\mathcal{P}_\emptyset$ as follows:

Definition F.2. *Suppose random variable $\mathbf{x} \sim \mathcal{P}_\emptyset$, and each coordinate $x_i, i \in I$ is sampled independently*
$$x_i = \begin{cases} \sqrt{\nu(k)} & \textit{with probability } 1/2 \\ -\sqrt{\nu(k)} & \textit{with probability } 1/2 \end{cases}.$$

It is clear that $D_{\mathrm{TV}}(\mathcal{P}_I, \mathcal{P}_\emptyset) = \alpha$. Thus we have shown that each pair $(\mathcal{P}_\emptyset, \mathcal{P}_I) \in \Theta_{\nu(k),\alpha}$. Now let us fix an estimator $\widehat{\mathbf{U}}$, and let $\widehat{\mathbf{U}}_\emptyset = \widehat{\mathbf{U}}(\{\mathbf{x}_i\}_{i=1}^n)$, $\{\mathbf{x}_i\}_{i=1}^n \sim \mathcal{P}_\emptyset^n$ denote the random subspace when the datapoints are drawn from $\mathcal{P}_\emptyset$. WLOG, let us assume d is a multiple of k, and let $I_1, \ldots, I_{d/k}$ be a partition of $[n]$. Notice that
$$\mathbb{E}\left[\sum_{i=1}^{d/k}\mathrm{Tr}\Big[\widehat{\mathbf{U}}_\emptyset^\top\Sigma_{I_i}\widehat{\mathbf{U}}_\emptyset\Big]\right] = \nu(k)\cdot k\cdot\Big(d/k + 1 - \alpha/k + \sqrt{\alpha/k}\Big),$$
and hence there exists a i^* such that
$$\begin{aligned}\mathbb{E}\Big[\mathrm{Tr}\Big[\widehat{\mathbf{U}}_\emptyset^\top\Sigma_{I_{i^*}}\widehat{\mathbf{U}}_\emptyset\Big]\Big] &\leq \nu(k)\cdot k\cdot\Big(d/k + 1 - \alpha/k + \sqrt{\alpha/k}\Big)\cdot(d/k)^{-1} \\ &\leq \nu(k)\cdot k\cdot\left(1 + \frac{1+\sqrt{\alpha/k}}{d/k}\right).\end{aligned}$$

The sub-optimality can be expressed as
$$\begin{aligned}&\mathbb{E}\Big[\Big\|\mathbf{\Sigma}_{I_{i^*}} - \widehat{\mathbf{U}}_\emptyset\widehat{\mathbf{U}}_\emptyset^\top\mathbf{\Sigma}_{I_{i^*}}\widehat{\mathbf{U}}_\emptyset\widehat{\mathbf{U}}_\emptyset^\top\Big\|_* - \|\mathbf{\Sigma}_{I_{i^*}} - \mathcal{P}_k(\mathbf{\Sigma}_{I_{i^*}})\|_*\Big] \\ =&\nu(k)k\Big(1 - \alpha/k + \sqrt{\alpha/k}\Big) - \mathbb{E}\Big[\mathrm{Tr}\Big[\widehat{\mathbf{U}}_\emptyset^\top\Sigma_{I_{i^*}}\widehat{\mathbf{U}}_\emptyset\Big]\Big] \\ \geq&\nu(k)\left(\sqrt{\alpha k} - \alpha - \frac{k^2(1+\sqrt{\alpha/k})}{d}\right) \text{ (Using } k \geq 16,\, d \geq k^2/\alpha) \\ \geq&\frac{1}{16}\nu(k)\sqrt{\alpha k}\end{aligned}$$

This implies that for any subspace estimator $\widehat{\mathbf{U}}$, we can find distribution $\mathcal{D} = \mathcal{P}_{\emptyset}, \mathcal{D}' = \mathcal{P}_{I_{i^*}}$ such that

1. $\underset{\{\mathbf{x}_i\}_{i=1}^n \sim \mathcal{D}^n}{\mathbb{E}} \left[\left\| \mathbf{\Sigma}_{I_{i^*}} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \mathbf{\Sigma}_{I_{i^*}} \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \right\|_* - \left\| \mathbf{\Sigma}_{I_{i^*}} - \mathcal{P}_k(\mathbf{\Sigma}_{I_{i^*}}) \right\|_* \right] \geq \frac{1}{16}\nu(k)\sqrt{k\alpha}$,
2. $D_{\mathrm{TV}}(\mathcal{D}, \mathcal{D}') \leq \alpha$,
3. $\max_{\|\mathbf{A}\|_{\mathrm{F}} \leq 1} \underset{\mathbf{x} \sim \mathcal{D}'}{\mathbb{E}} \left[\left(\langle \mathbf{A}, \mathbf{x}\mathbf{x}^\top - \mathbb{E}[\mathbf{x}\mathbf{x}^\top] \rangle \right)^2 \right] \leq 2\nu(k)^2$.

The proof is complete.

G Robust clustering algorithm

Definition G.1. *Pseudo-distributions are generalizations of probability distributions except for the fact that they need not be non-negative. A level-$2m$ pseudo-distribution ξ, for $m \in \mathbb{N} \cup \{\infty\}$, is a measurable function that must satisfy*

$$\int_{\mathbb{R}^d} q(\mathbf{x})^2 d\xi(\mathbf{x}) \geq 0 \qquad \text{for all polynomials } q \text{ of degree at-most } m\text{, and} \tag{68}$$

$$\int_{\mathbb{R}^d} d\xi(\mathbf{x}) = 1. \tag{69}$$

A straightforward polynomial interpolation argument shows that every level-∞ pseudo-distribution ξ is non-negative, and thus are actual probability distributions.

Definition G.2. *A pseudo-expectation $\widetilde{\mathbb{E}}_\xi[f(\mathbf{x})]$ of a function f on $\mathbb{R}^d$ with respect to a pseudo-expectation ξ, just like the usual expectation, is denoted as*

$$\widetilde{\mathbb{E}}_\xi[f(\mathbf{x})] = \int_{\mathbb{R}^d} f(\mathbf{x}) d\xi(\mathbf{x}).$$

Definition G.3. *The SOS ordering $\preceq_{SOS}$ between two finite dimensional tensors $\mathcal{T}_1$, and $\mathcal{T}_2$, i.e., $\mathcal{T}_1 \preceq_{SOS} \mathcal{T}_2$, means $\langle \mathcal{T}_1, \mathbf{v}^{\otimes 2m} \rangle \preceq_{2m} \langle \mathcal{T}_2, \mathbf{v}^{\otimes 2m} \rangle$ as polynomial in $\mathbf{v}$.*

Given $\left\{\widehat{\beta}_i\right\}_{i=1}^n$, we want to find $\{\gamma_i\}_{i=1}^n$ such that $\frac{1}{n}\sum_{i=1}^n \widetilde{\mathbb{E}}_\xi\left[\left\langle \widehat{\beta}_i - \gamma_i, \mathbf{v}\right\rangle^{2m}\right]$ is small for all pseudo-distributions ξ of $\mathbf{v}$ over the sphere. Since $\gamma_i = \widehat{\beta}_i \ \forall\ i \in [n]$ would be an over-fit, therefore to avoid it, it turns out that the natural way is to introduce the term $\sum_{i=1}^n \langle \gamma_i - \beta_i, \gamma_i \rangle^{2m}$ which must be small at the same time. On the other hand, we know that if $\sum_{i \in \mathcal{G}} \widetilde{\mathbb{E}}_\xi\left[\left\langle \widehat{\beta}_i - \gamma_i, \mathbf{v}\right\rangle^{2m}\right]$, and $\sum_{i \in \mathcal{G}} \widetilde{\mathbb{E}}_\xi\left[\left\langle \widehat{\beta}_i - \beta_i, \mathbf{v}\right\rangle^{2m}\right]$ (from the SOS proof) are small, then from the Minkowski's inequality, $\sum_{i \in \mathcal{G}} \widetilde{\mathbb{E}}_\xi\left[\langle \gamma_i - \beta_i, \mathbf{v}\rangle^{2m}\right]$ will also be small. To make this hold, it is sufficient to impose that whenever $\{\mathbf{z}_i\}_{i=1}^n$ are such that $\sum_{i=1}^n \widetilde{\mathbb{E}}_\xi\left[\langle \mathbf{z}_i, \mathbf{v}\rangle^{2m}\right] \leq 1$ for all pseudo-distributions ξ over the unit sphere, then $\sum_{i=1}^n \langle \mathbf{z}_i, \gamma_i \rangle^{2m}$ is also small. This, however, is not efficiently imposable, but there is a standard SOS way of relaxing this, which is to require $\sum_{i=1}^n \widetilde{\mathbb{E}}_{\zeta_i}\left[\langle \mathbf{z}_i, \gamma_i\rangle^{2m}\right]$ to be small whenever $\sum_{i=1}^n \widetilde{\mathbb{E}}_{\zeta_i}\left[\mathbf{z}_i^{\otimes 2m}\right] \preceq_{\mathrm{SOS}} \mathcal{I}$ for all pseudo-distributions $\{\zeta_i(\mathbf{z}_i)\}_{i=1}^n$, where $\mathcal{I}$ is the identity tensor of the appropriate dimension.

Therefore, we need to find $\{\gamma_i\}_{i=1}^n$ such that $\sum_{i=1}^n \widetilde{\mathbb{E}}_\xi\left[\left\langle \widehat{\beta}_i - \gamma_i, \mathbf{v}\right\rangle^{2m}\right]$, and $\sum_{i=1}^n \widetilde{\mathbb{E}}_{\zeta_i}\left[\langle \mathbf{z}_i, \gamma_i\rangle^{2m}\right]$ are small whenever $\sum_{i=1}^m \widetilde{\mathbb{E}}_{\zeta_i}\left[\mathbf{z}_i^{\otimes 2m}\right] \preceq_{\mathrm{SOS}} \mathcal{I}$ for all pseudo-distributions ξ over the unit sphere, and $\{\zeta_i\}_{i=1}^n$.

Algorithm 5 Basic clustering relaxation [45, Adaptation of Algorithm 1]

```
Input: $\{\widehat{\beta}_i\}_{i\in[n]}$, $\{c_i \in [0,1]\}_{i=1}^n$, $m \in \mathbb{N}$, multiplier $\lambda \geq 0$, threshold $\Gamma \geq 0$.
Define $\tau_i(\gamma_i, \xi, \zeta_i) := \widetilde{\mathbb{E}}_{\mathbf{v}\sim\xi}\left[\left\langle \widehat{\beta}_i - \gamma_i, \mathbf{v}\right\rangle^{2m}\right] + \lambda\widetilde{\mathbb{E}}_{\mathbf{z}_i\sim\zeta_i}\left[\langle \gamma_i, \mathbf{z}_i\rangle^{2m}\right]$.
Find $\{\gamma_i^*\}_{i=1}^n$ such that $\sum_{i=1}^n c_i\tau_i(\gamma_i^*, \xi, \zeta_i) \leq 2\Gamma$, for all $\xi$ over unit sphere, and
        for all $\{\zeta_i\}_{i=1}^n$ satisfying $\sum_{i=1}^n \widetilde{\mathbb{E}}_{\zeta_i}\left[\mathbf{z}_i^{\otimes 2m}\right] \preceq_{\text{sos}} \mathcal{I}$.
Or else, find $\xi^*$, and $\{\zeta_i^*\}_{i=1}^n$ such that $\forall\ \{\gamma_i\}_{i=1}^n$, $\sum_{i=1}^n c_i\tau_i(\gamma_i, \xi^*, \zeta_i^*) \geq \Gamma$ .
Output: $\{\gamma_i^*\}_{i=1}^n$, or $\left\{\xi^*, \{\zeta_i^*\}_{i=1}^n\right\}$.
```

If there is no solution that makes the desired quantities small, then from duality there must exist pseudo-distributions ξ^*, and $\{\zeta_i^*\}_{i=1}^n$ such that the objective cannot be small for any choice of $\{\gamma_i\}_{i=1}^n$. Since elements in $\{\gamma_i\}_{i=1}^n$ are independent of each other in the objective for a fixed ξ and $\{\zeta_i\}_{i=1}^n$, and the objective can made small on the good set $\mathcal{G}$, therefore we can look at $\min_\gamma \widetilde{\mathbb{E}}_{\xi^*}\left[\left\langle \widehat{\beta}_i - \gamma, \mathbf{v}\right\rangle^{2m}\right]$ or $\min_\gamma \widetilde{\mathbb{E}}_{\zeta_i^*}\left[\langle \mathbf{z}_i, \gamma\rangle^{2m}\right]$ if they are large for any $i \in [n]$. Such tasks can be removed and the process can be repeated. It is shown in [45] that the procedure after a finite number of iterations can remove all the outliers and eventually the sum of the desired quantities can be made small.

Algorithm 6 Outlier Removal Algorithm [45, Adaptation of Algorithm 2]

```
Input: $\left\{\widehat{\beta}_i\right\}_{i\in[n]}$, $B \geq 0$, $m \in \mathbb{N}$, $p_{\min}$, and $\rho \geq 0$.
Initialize $\mathbf{c} = \mathbf{1}_n$, and set $\lambda = p_{\min} n (B/\rho)^{2m}$.
while true do
    Run Algorithm 5 with $\{\widehat{\beta}_i\}_{i\in[n]}$, $\mathbf{c}$, $\lambda$, and threshold $\Gamma = 4\left(nB^{2m} + \lambda\rho^{2m}/p_{\min}\right)$ to obtain
    $\{\gamma_i^*\}_{i=1}^n$, or $\left\{\xi^*, \{\zeta_i^*\}_{i=1}^n\right\}$.
    if $\{\gamma_i^*\}_{i=1}^n$ are obtained then
        Output: $\{\gamma_i^*\}_{i=1}^n$, $\mathbf{c}$
    else if $\left\{\xi^*, \{\zeta_i^*\}_{i=1}^n\right\}$ is obtained then
        $\tau_i^* \leftarrow \min_\gamma \tau_i(\gamma, \xi^*, \zeta_i^*)$   $\forall\ i \in [n]$, as defined in Algorithm 5
        $c_i \leftarrow c_i\left(1 - \tau_i^*/\max_{j\in[n]} \tau_j^*\right)$   $\forall\ i \in [n]$
    end if
end while
```

Algorithm 6 uses a down-weighting way of reducing the weight of possible outlier tasks, and [45] show that the re-weighting step down-weights the outlier tasks more than the tasks in the ℓ-th set. They also show that the returned $\{\gamma_i^*\}_{i=1}^n$ constitute a good clustering such that one of the clusters is centered close to some true mean $\mathbf{w}_j$ for some $j \in [k]$.

Algorithm 7 repeatedly uses Algorithm 6 on re-centered data to find the individual clusters. The set S_j obtained in Algorithm 7 is almost entirely a subset of one of the true good sets. After obtaining M centers from Algorithm 7 we can re-consolidate the sets into k new sets $\{\mathcal{C}_\ell\}_{\ell=1}^k$ (by merging together all S_j whose means are within distance $B\widetilde{p}_{\min}^{-1/m}/4$.), and can shown to obey the desired guarantee using [45, Theorem 5.4.].

H Proof of Lemma B.4 for robust clustering

We use [45, Theorem 1.2], to analyze Algorithm 7 with input $\left\{\widehat{\mathbf{U}}^\top\widehat{\beta}_i = \frac{1}{t}\sum_{j=1}^t y_{i,j}\widehat{\mathbf{U}}^\top\mathbf{x}_{i,j}\right\}_{i\in[n]}$ for $t = t_H$ and $n = n_H$.

Algorithm 7 Algorithm for re-clustering $\{\gamma_i\}_{i=1}^n$ [45, Adaptation of Algorithm 3]

```
**Input:** $\mathcal{D}_H = \{\{(\mathbf{x}_{i,j}, y_{i,j})\}_{j\in[t_H]}\}_{i\in[n_H]}$, $\widehat{\mathbf{U}}$, $B \geq 0$, $m, k \in \mathbb{N}$, $p_{\min}$, and $\rho \geq 0$.
Initialize $R = \rho$, set $W = \{\mathbf{0}\}$, and $\widehat{\beta}_i \leftarrow (1/t_H)\sum_{\ell=1}^{t_H} \widehat{\mathbf{U}}^\top \mathbf{x}_{i,j} y_{i,j}$ for all $i \in [n]$
$\rho_{\text{final}} \leftarrow \Theta\Big(Bp_{\min}^{-1/m}\Big)$, $M = \lceil 4/p_{\min} \rceil$
**while** $R \geq \rho_{\text{final}}$ **do**
    $b \leftarrow 1$
    **for** $\mathbf{w}' \in W$ **do**
        Let $\{\gamma_i^{(b)}\}_{i=1}^n, \mathbf{c}^{(b)}$ be the output of Algorithm 6 with $\{\widehat{\beta}_i - \mathbf{w}'\}_{i=1}^n$, $B$, $p_{\min}$, $R$ as input.
        $b \leftarrow b + 1$
    **end for**
    Let $\{\mathcal{C}_j\}_{j=1}^M$ be the maximal covering derived from $\{\gamma_i^{(j)}\}_{i=1,j=1}^{n,M}$, $\big\{\mathbf{c}^{(j)}\big\}_{j=1}^M$
    $W \leftarrow \big\{\mathbf{w}'_j\big\}_{j=1}^M$ where $\mathbf{w}'_j$ is the mean of points in $\mathcal{C}_j$ $\forall\, j \in [M]$
    $R \leftarrow C'\left(\sqrt{RBp_{\min}^{-1/m}} + Bp_{\min}^{-1/m}\right)$
**end while**
**Output:** $W, \{\mathcal{C}_\ell\}_{\ell=1}^M$
```

Algorithm 8 Estimating $\big\{r_\ell^2\big\}_{\ell=1}^k$

```
**Input:** $\mathcal{D}_H = \{\{(\mathbf{x}_{i,j}, y_{i,j})\}_{j\in[t_H]}\}_{i\in[n_H]}$, $\{\widetilde{\mathbf{w}}_\ell\}_{\ell=1}^k$, $\alpha \geq 0$, $\delta \in (0,1)$
**for** $\ell \in [k]$ **do**
    $r_{\ell,i}^2 \leftarrow t_H^{-1}\sum_{j\in[t_H]}\big(y_{i,j} - \mathbf{x}_{i,j}^\top \widetilde{\mathbf{w}}_\ell\big)^2$ for all $i \in \mathcal{C}_\ell$
    **if** $\alpha > 0$ **then**
        $\widetilde{r}_\ell^2 \leftarrow$Univariate_Mean_Estimator$\left(\left\{r_{\ell,i}^2\right\}_{i\in\mathcal{C}_\ell}, \alpha, \delta\right)$                [ [50]]
    **else**
        $\widetilde{r}_\ell^2 \leftarrow \frac{1}{|\mathcal{C}_\ell|}\sum_{i\in\mathcal{C}_\ell} r_{\ell,i}^2$
    **end if**
**end for**
**Output:** $\big\{\widetilde{r}_\ell^2\big\}_{\ell=1}^k$
```

Theorem H.1 ([45, Theorem 1.2])**.** *Suppose $\Big\{\widehat{\mathbf{U}}^\top \widehat{\beta}_i \in \mathbb{R}^k\Big\}_{i\in[n]}$ can be partitioned into sets $\{\mathcal{G}_\ell\}_{\ell=1}^k \cup \mathcal{H}$, where $\mathcal{H}$ is the set of outliers, of size αn. Suppose $|\mathcal{G}_\ell| = n\widetilde{p}_\ell$, and has mean $\mathbf{w}_\ell$, that its $2m$-th moment $M_{2m}(\mathcal{G}_\ell) \preceq_{2m} B$. Also suppose that $\alpha \leq \widetilde{p}_{\min}/8$. Finally suppose the separation $\Delta \geq C_{\text{sep}}B/\widetilde{p}_{\min}^{1/m}$ with $C_{\text{sep}} \geq C_0$ for a universal constant C_0. Then Algorithm 7 runs in time $\mathcal{O}((nk)^{\mathcal{O}(m)})$ and outputs estimates $\{\widetilde{\mathbf{w}}'_\ell\}_{\ell=1}^k$ such that $\Big\|\widetilde{\mathbf{w}}'_\ell - \widehat{\mathbf{U}}^\top \mathbf{w}_\ell\Big\|_2 \leq \mathcal{O}\Big(B\cdot\big(\alpha/\widetilde{p}_{\min} + C_{\text{sep}}^{-2m}\big)^{1-1/2m}\Big)$.*

It shows how the accuracy depends on the choice of m and the SOS proof of an upper bound B. We will show that if $B = \rho\sqrt{2mC/t}$ then the SOS proof holds, in which case the condition $\Delta \geq C_{\text{sep}}B/\widetilde{p}_{\min}^{1/m}$ in Theorem H.1 translates into

$$t \geq \frac{2mC\rho^2}{\Delta^2}\cdot\frac{C_{\text{sep}}^2}{\widetilde{p}_{\min}^{2/m}}. \tag{70}$$

Further, to get $\Big\|\widetilde{\mathbf{w}}_j - \widehat{\mathbf{U}}^\top \mathbf{w}_j\Big\|_2 = \mathcal{O}(\Delta)$ we need

$$t \gtrsim \frac{2mC\rho^2}{\Delta^2}\cdot\max\left\{\frac{1}{C_{\text{sep}}^{4m-2}}, \left(\frac{\alpha}{\widetilde{p}_{\min}}\right)^{2-\frac{1}{m}}\right\}. \tag{71}$$

Combining the two conditions, we finally get

$$t \gtrsim \frac{2mC\rho^2}{\Delta^2} \cdot \max\left\{ \frac{C_{\text{sep}}^2}{\widetilde{p}_{\min}^{2/m}}, \frac{1}{C_{\text{sep}}^{4m-2}}, \left(\frac{\alpha}{\widetilde{p}_{\min}}\right)^{2-\frac{1}{m}} \right\}. \tag{72}$$

We are left to show that SOS proof exists for the choice of $B = \rho\sqrt{2mC/t}$. The following lemma, whose proof is in §H.1, gives

$$\frac{1}{|\mathcal{G}_\ell|} \sum_{i \in \mathcal{G}_\ell} \left\langle \widehat{\mathbf{U}}^\top \left(\widehat{\beta}_i - \beta_i\right), \mathbf{v} \right\rangle^{2m} \preceq_{2m} \rho^{2m} \|\mathbf{v}\|_2^{2m} (2m)^m \frac{C^m}{t^m} \leq (2m)^m \frac{C^m}{t^m} \tag{73}$$

for all $\|\mathbf{v}\|_2 \leq 1$ with probability at least $7/8$ for all $\ell \in [k]$.

Lemma H.2 (SOS proof exists with high probability)**.** *Given $t \geq 2m$, for $m, t \in \mathbb{N}$, there exists a constant $C > 0$ such that for any $\ell \in [k]$, and $np_\ell \geq (km)^{\Theta(m)}/\delta$, with probability at least $1 - \delta$, it holds that*

$$\frac{1}{n\widetilde{p}_\ell} \sum_{i \in \mathcal{G}_\ell} \left\langle \widehat{\mathbf{U}}^\top (\widehat{\beta}_i - \beta_i), \mathbf{v} \right\rangle^{2m} \preceq_{2m} \rho^{2m} \|\mathbf{v}\|_2^{2m} (2m)^m \frac{C^m}{t^m},$$

for all $\mathbf{v} \in \mathbb{R}^k$.

From [42, Proposition D.7.] we have that if $n = \Omega\left(\frac{\log k}{p_{\min}}\right)$, then $\widetilde{p}_{\min} \geq p_{\min}/2$ with probability at least $7/8$. To further simplify the conditions, note that $\alpha < \widetilde{p}_{\min}/8$, and fix $C_{\text{sep}} = \Theta(1)$, then we simply need

$$t \gtrsim \frac{m\rho^2}{p_{\min}^{2/m} \Delta^2}. \tag{74}$$

Using a median of means algorithm from [49, Proposition 1] and [33, 18], by repeatedly and independently estimating $M = \Omega\left(\log \frac{1}{\delta}\right)$ number of estimates, $\left\{ \left\{ \widetilde{\mathbf{w}}_j^{\prime(\ell)} \right\}_{j=1}^k \right\}_{\ell=1}^M$ we can compute the improved estimates $\left\{ \widetilde{\mathbf{w}}_j \in \mathbb{R}^d \right\}_{j=1}^k$ by applying back $\widehat{\mathbf{U}}$ that satisfy

$$\left\| \widetilde{\mathbf{w}}_j - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \mathbf{w}_j \right\|_2 \lesssim \Delta \quad \forall\, j \in [k] \tag{75}$$

With the assumption that

$$\left\| \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \mathbf{w}_j - \mathbf{w}_j \right\|_2 \lesssim \Delta$$

and Equation (75), we have

$$\|\widetilde{\mathbf{w}}_j - \mathbf{w}_j\|_2 \leq \left\| \widetilde{\mathbf{w}}_j - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \mathbf{w}_j \right\|_2 + \left\| \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \mathbf{w}_j - \mathbf{w}_j \right\|_2 \lesssim \Delta. \tag{76}$$

We next show a bound on the error in estimating r_ℓ.

Proposition H.3 (Estimating r_ℓ)**.** *If $n_H = \widetilde{\Omega}\left(\frac{\rho^4}{\Delta^4 t_H p_{\min}}\right)$, we can estimate r_ℓ^2 as $\widetilde{r}_\ell^2$ satisfying*

$$\left|\widetilde{r}_\ell^2 - r_\ell^2\right| \leq r_\ell^2 \frac{\Delta^2}{50\rho^2} \tag{77}$$

with probability at least $1 - \delta$, for all $\ell \in [k]$ using Algorithm 8, where $r_\ell^2 := \|\widetilde{\mathbf{w}}_\ell - \mathbf{w}_\ell\|_2^2 + s_\ell^2$, if $\alpha_H = \mathcal{O}\left(\frac{\Delta^2 \sqrt{t_H} p_{\min}}{\rho^2 \log\left(\frac{\rho^2}{\Delta^2 t_H}\right)}\right)$.

Proof. Define $r_{\ell,i}^2 = t_H^{-1} \sum_{j=1}^{t_H} \left(y_{i,j} - \widetilde{\mathbf{w}}_\ell^\top \mathbf{x}_{i,j}\right)^2 \sim \frac{r_\ell^2}{t_H} \chi^2(t_H)$ for all $i \in \mathcal{C}_\ell, \ell \in [k]$. Since we compute r_ℓ^2 for each cluster $\mathcal{C}_\ell$ independently, the maximum corruption in the ℓ-th cluster is bounded by α_H/p_ℓ. Using Corollary J.3, we can compute an estimator using Algorithm 8, that given $\widetilde{\alpha}_\ell = \alpha_H/p_\ell$ corrupted samples, returns $\widetilde{r}_\ell^2$ for all $\ell \in [k]$ satisfying

$$\begin{aligned}
\left|r_\ell^2 - \widetilde{r}_\ell^2\right| &= \mathcal{O}\left(r_\ell^2 \cdot \widetilde{\alpha}_\ell \cdot \max\left\{\frac{\log(1/\widetilde{\alpha}_\ell)}{t_H}, \sqrt{\frac{\log(1/\widetilde{\alpha}_\ell)}{t_H}}\right\}\right) \\
&= \mathcal{O}\left(r_\ell^2 \cdot \frac{\alpha_H}{p_\ell} \cdot \max\left\{\frac{\log(p_\ell/\alpha_H)}{t_H}, \sqrt{\frac{\log(p_\ell/\alpha_H)}{t_H}}\right\}\right) \\
&= \mathcal{O}\left(r_\ell^2 \cdot \frac{\alpha_H}{p_\ell} \cdot \frac{\log(p_\ell/\alpha_H)}{\sqrt{t_H}}\right)
\end{aligned}$$

when $n_H p_\ell = \widetilde{\Omega}\Big(\frac{1}{\widetilde{\alpha}_\ell^2}\Big) = \widetilde{\Omega}\Big(\frac{p_\ell^2}{\alpha_H^2}\Big)$ for all $\ell \in [k]$. Using the fact that

$$\frac{\beta}{\log\frac{1}{\beta}} \geq \frac{e}{e-1}\alpha \implies \alpha \log\frac{1}{\alpha} \leq \beta$$

for $\alpha, \beta \in (0,1)$, we have that for $\alpha_H \leq C\frac{\Delta^2\sqrt{t_H}p_{\min}}{\rho^2\log\left(\frac{\rho^2}{\Delta^2 t_H}\right)}$ for some $C > 0$,

$$\left|r_\ell^2 - \widetilde{r}_\ell^2\right| = r_\ell^2 \cdot \frac{\Delta^2}{50\rho^2} \tag{78}$$

with probability at-least $1-\delta$, when $n_H = \widetilde{\Omega}\Big(\frac{\rho^4}{\Delta^4 t_H p_{\min}}\Big)$. □

H.1 Proof of Lemma H.2 for Sum-of-Squares proof $y\mathbf{x}$

We combine the following Proposition H.4 and Lemma H.8 to yield the desired SOS proof.

Proposition H.4 (see the proof of Lemma 4.1 in [35])**.** *Let* $Z_i = \frac{1}{t}\sum_{j=1}^t y_{i,j}\mathbf{x}_{i,j} - \beta_i$ *for* $i \in \mathcal{G}_\ell$*. If* $np_\ell \geq (km)^{\Theta(m)}/\delta \;\forall\; \ell \in [k]$*, then with probability at least* $1-\delta$*,*

$$\sum_{i=1}^n \langle Z_i, \mathbf{v}\rangle^{2m} - \mathbb{E}\Big[\langle Z_i, \mathbf{v}\rangle^{2m}\Big] \preceq_{2m} \frac{1}{4}\|\mathbf{v}\|_2^{2m}\frac{C^m}{t^m}$$

Proof. We show in Lemma H.8 that the distribution of $\frac{\sqrt{t}}{\sqrt{C}}Z_i$ is $2m$-explicitly bounded with variance proxy 1. In the proof of Lemma 4.1 in [35], it is shown that for a $2m$-explicitly bounded distribution, given $n \geq (km)^{\Theta(m)}/\delta$ samples, with probability at least $1-\delta$ (Fact 7.6 in [35]),

$$\frac{t^m}{C^m}\left(\sum_{i=1}^n \langle Z_i, \mathbf{v}\rangle^{2m} - \mathbb{E}\Big[\langle Z_i, \mathbf{v}\rangle^{2m}\Big]\right) \preceq_{2m} \frac{1}{4}\|\mathbf{v}\|_2^{2m},$$

which implies the propositions. □

Fact H.5 (Claim A.9. in [26])**.** *Let* $\sigma_y^2 = \|\beta\|_2^2 + \sigma^2$*.For any* $v \in \mathbb{R}^d$*, we have that*

$$(\mathbf{v}^\top\mathbf{x})y = \left(\frac{\mathbf{v}^\top\beta + \|\mathbf{v}\|_2\sigma_y}{2}\right)Z_1^2 + \left(\frac{\mathbf{v}^\top\beta - \|\mathbf{v}\|_2\sigma_y}{2}\right)Z_2^2\,,$$

where $Z_1, Z_2 \sim \mathcal{N}(0,1)$ *and* Z_1, Z_2 *are independent.*

We say that polynomial $p(v) \succeq q(v)$ if $p(x) - q(x)$ can be written as a sum of squares of polynomials. We write this as $p(v) \succeq_{2m} q(v)$ if we want to emphasize that the proof only involves polynomials of degree at most $2m$.

Fact H.6 (Basic facts about SOS proofs)**.**

- $\left(\mathbf{v}^\top\beta\right)^2 \preceq_2 \|\mathbf{v}\|_2^2\|\beta\|_2^2$ *(Cauchy-Schwarz),*

- $p_1 \succeq_{m_1} p_2 \succeq_{m_1} 0$ *, and* $q_1 \succeq_{m_2} q_2 \succeq_{m2} 0 \implies p_1 p_2 \succeq_{m_1+m_2} q_1 q_2$.

Definition H.7. *A distribution $\mathcal{D}$ over $\mathbb{R}^d$ with mean $\mu \in \mathbb{R}^d$ is $2m$-explicitly bounded for $m \in \mathbb{N}$, if $\forall\, i \leq 2m$, we have*

$$\mathop{\mathbb{E}}_{X \sim \mathcal{D}}\Big[\langle X - \mu, \mathbf{v}\rangle^i\Big] \preceq_i (\sigma i)^{i/2} \|\mathbf{v}\|_2^i$$

for variance proxy $\sigma^2 \in \mathbb{R}_+$.

Lemma H.8. *Given that $t \geq 2m$, for $m, t \in \mathbb{N}$, there exists a constant $C > 0$ such that*

$$\mathbb{E}\left[\left\langle \frac{1}{t}\sum_{i=1}^{t} y_i \mathbf{x}_i - \beta, \mathbf{v}\right\rangle^{2m}\right] \preceq_{2m} \rho^{2m}\|\mathbf{v}\|_2^{2m}(2m)^m \frac{C^m}{t^m}.$$

Proof.

$$\begin{aligned}
&\mathbb{E}\Big[\Big\langle \tfrac{1}{t}\textstyle\sum_{i=1}^{t} y_i \mathbf{x}_i - \beta, \mathbf{v}\Big\rangle^{2m}\Big] \\
&= \mathbb{E}\Big[\Big(\Big(\tfrac{\mathbf{v}^\top\beta + \|\mathbf{v}\|_2\sigma_y}{2}\Big)(S_1 - 1) + \Big(\tfrac{\mathbf{v}^\top\beta - \|\mathbf{v}\|_2\sigma_y}{2}\Big)(S_2 - 1)\Big)^{2m}\Big] \\
&= \textstyle\sum_{i=0}^{2m}\binom{2m}{i}\Big(\tfrac{\mathbf{v}^\top\beta + \|\mathbf{v}\|_2\sigma_y}{2}\Big)^i\Big(\tfrac{\mathbf{v}^\top\beta - \|\mathbf{v}\|_2\sigma_y}{2}\Big)^{2m-i} M_i M_{2m-i} \\
&= \textstyle\sum_{i=0}^{m}\binom{2m}{i}\Big(\Big(\tfrac{\mathbf{v}^\top\beta + \|\mathbf{v}\|_2\sigma_y}{2}\Big)^{2m-2i} + \Big(\tfrac{\mathbf{v}^\top\beta - \|\mathbf{v}\|_2\sigma_y}{2}\Big)^{2m-2i}\Big)\Big(\tfrac{(\mathbf{v}^\top\beta)^2 - \|\mathbf{v}\|_2^2\sigma_y^2}{4}\Big)^i M_i M_{2m-i}
\end{aligned}$$

where $S_1, S_2 \sim \frac{\chi^2(t)}{t}$, and $M_i = \mathbb{E}_{Z_j \sim \mathcal{N}(0,1)}\Big[\Big(\frac{1}{t}\sum_{j=1}^{t}(Z_j^2 - 1)\Big)^i\Big]$.

First, notice that

$$\begin{aligned}
&\left(\frac{\mathbf{v}^\top\beta + \|\mathbf{v}\|_2\sigma_y}{2}\right)^{2m-2i} + \left(\frac{\mathbf{v}^\top\beta - \|\mathbf{v}\|_2\sigma_y}{2}\right)^{2m-2i} \\
&= 2^{2i-2m} \cdot 2 \cdot \sum_{j=0}^{m-i}\binom{2m-2i}{2j}\left(\mathbf{v}^\top\beta\right)^{2j}(\|\mathbf{v}\|_2\sigma_y)^{2m-2i-2j} \\
&\preceq_{2m-2i} 2^{2i-2m}2^{2m-2i}\|\mathbf{v}\|_2^{2m-2i}\rho^{2m-2i} \\
&= \|\mathbf{v}\|_2^{2m-2i}\rho^{2m-2i}.
\end{aligned} \tag{79}$$

The SOS order hold by the fact that $\left(\mathbf{v}^\top\beta\right)^2 \preceq_2 \|\beta\|_2^2\|\mathbf{v}\|_2^2 \preceq_2 \sigma_y^2\|\mathbf{v}\|_2^2$ and Fact H.6. Secondly, notice that

$$\begin{aligned}
\left(\frac{\left(\mathbf{v}^\top\beta\right)^2 - \|\mathbf{v}\|_2^2\sigma_y^2}{4}\right)^i &= 2^{-2i}\sum_{j=0}^{i}\binom{i}{j}(-1)^{i-j}(\mathbf{v}^\top\beta)^{2j}(\|\mathbf{v}\|_2\sigma_y)^{2i-2j} \\
&\preceq_{2i} 2^{-i}\|\mathbf{v}\|_2^{2i}\rho^{2i} \\
&\preceq_{2i} \|\mathbf{v}\|_2^{2i}\rho^{2i}
\end{aligned} \tag{80}$$

Combining Equation (79) and Equation (80) yields

$$\mathbb{E}\left[\left\langle \frac{1}{t}\sum_{i=1}^{t} y_i \mathbf{x}_i - \beta, \mathbf{v}\right\rangle^{2m}\right] \preceq_{2m} \rho^{2m}\|\mathbf{v}\|_2^{2m}\sum_{i=0}^{m}\binom{2m}{i}M_i M_{2m-i}$$

Fact H.9. *Given that $t \geq 2m$, it holds that*

$$M_i \leq 2e^{2i} i^{i/2}/t^{i/2}\ ,$$

Proof. By Bernstein inequality, $\sum_{j=1}^{t}(Z_j^2 - 1)$ is a combination of sub-Gaussian with norm $\Theta\big(\sqrt{t}\big)$ and sub-exponential with norm $\Theta(1)$ with disjoint supports. Hence

$$\begin{aligned}\frac{1}{t^i}\,\mathbb{E}\left[\left(\sum_{j=1}^{t}(Z_j^2-1)\right)^{i}\right] &\lesssim \frac{e^i(ti)^{i/2}+e^{2i}i^i}{t^i}\\ &= e^i\frac{i^{i/2}}{t^i}\left(t^{i/2}+e^i i^{i/2}\right)\,,\\ &= e^i\frac{i^{i/2}}{t^{i/2}}\left(1+e^i\frac{i^{i/2}}{t^{i/2}}\right)\\ &\le 2e^{2i}\frac{i^{i/2}}{t^{i/2}}\,.\end{aligned}$$
□

Applying Fact H.9 gives

$$\mathbb{E}\left[\left\langle \frac{1}{t}\sum_{i=1}^{t} y_i\mathbf{x}_i - \beta, \mathbf{v}\right\rangle^{2m}\right] \preceq_{2m} \rho^{2m}\|\mathbf{v}\|_2^{2m}(2m)^m\frac{C^m}{t^m}$$

for some $0 < C < e^6$. This concludes the proof. □

H.2 The distribution of $y\mathbf{x}$ is $\Omega(d)$-Poincaré

We show that even for the simplest parameter setting where the regression vector $\beta = 0$, and the noise has variance 1, the distribution of $y\mathbf{x}$ is Poincaré with parameter $\Omega(d)$, and thus applying the result on Poincaré distribution from [45] naively would only yield trivial guarantee.

Remark H.10. *Suppose $\mathbf{x} \sim \mathcal{N}(0, \mathbf{I})$, $y \sim \mathcal{N}(0, 1)$, and function $f(\mathbf{z}) = \|\mathbf{z}\|_2^2$. Then $\mathrm{Var}[f(\mathbf{z})] = (2d + 6)\,\mathbb{E}\big[\|\nabla f(\mathbf{z})\|_2^2\big]$.*

Proof.

$$\begin{aligned}\mathrm{Var}[f(\mathbf{z})] &= \mathbb{E}\big[y^4\|\mathbf{x}\|_2^4\big] - \mathbb{E}\big[y^2\|\mathbf{x}\|_2^2\big]^2\\ &= 3d(d+2) - d^2\\ &= 2d^2 + 6d.\end{aligned}$$

Since $\nabla f(\mathbf{z}) = 2\mathbf{z}$, we have

$$\mathbb{E}\big[\|\nabla f(\mathbf{z})\|_2^2\big] = \mathbb{E}\big[4y^2\|\mathbf{x}\|_2^2\big] = 4d.$$

Hence

$$\mathrm{Var}[f(\mathbf{z})] = \frac{1}{2}(d+3)\,\mathbb{E}\big[\|\nabla f(\mathbf{z})\|_2^2\big].$$
□

Remark H.11. *The choice of m can be made from t appropriately by considering the analytical inverse map. For any $c > 0$, if $y = \frac{t\Delta^2}{2c\rho^2\log(1/p_{\min})}$, and $x = \frac{2\log(1/p_{\min})}{m}$, then it is clear that the condition $t \ge c\cdot\frac{m}{p_{\min}^{2/m}}\cdot\frac{\rho^2}{\Delta^2}$, can be written as $y \ge e^x/x$, or $-1/y \ge (-x)e^{-x}$. Therefore, this condition is satisfied when*

$$\frac{2\log(1/p_{\min})}{-W_{-1}\Big(-\frac{2c\rho^2\log(1/p_{\min})}{t\Delta^2}\Big)} \le m \le \frac{2\log(1/p_{\min})}{-W_0\Big(-\frac{2c\rho^2\log(1/p_{\min})}{t\Delta^2}\Big)} \tag{81}$$

if $t \ge 2ec\cdot\frac{\rho^2}{\Delta^2}\log\frac{1}{p_{\min}}$, where W_0 and W_{-1} are the Lambert W functions.

Algorithm 9 Classification and robust estimation

```
Input: data $\mathcal{D}_{L2} = \{(\mathbf{x}_{i,j}, y_{i,j})\}_{i\in[n_{L2}], j\in[t_{L2}]}$, $\{\mathcal{C}_\ell,\ \widetilde{\mathbf{w}}_\ell,\ \widetilde{r}_\ell^2\}_{\ell\in[k]}$, $\alpha > 0$, $\delta \in (0,1)$.
compute for all $i \in [n_{L2}]$
    $$h_i \leftarrow \arg\min_{\ell\in[k]} \frac{1}{2\widetilde{r}_\ell^2} \sum_{j\in[t_{L2}]} \big(y_{i,j} - \mathbf{x}_{i,j}^\top \widetilde{\mathbf{w}}_\ell\big)^2 + t_{L2}\log \widetilde{r}_\ell$$
    $\mathcal{C}_{h_i} \leftarrow \mathcal{C}_{h_i} \cup \{(\mathbf{x}_{i,j}, y_{i,j})\}_{j=1}^{t_{L2}}$
compute for all $\ell \in [k]$,
    $\widehat{\mathbf{w}}_\ell \leftarrow$ Robust_Least_Squares($\mathcal{C}_\ell$)                    [ [26, Algorithm 2]]
    $r_{\ell,i}^2 \leftarrow t_{L2}^{-1} \sum_{j\in[t_{L2}]} \big(y_{i,j} - \mathbf{x}_{i,j}^\top \widehat{\mathbf{w}}_\ell\big)^2$ for all $i \in \mathcal{C}_\ell$
    $\widehat{s}_\ell^2 \leftarrow$Univariate_Mean_Estimator$\Big(\Big\{r_{\ell,i}^2\Big\}_{i\in\mathcal{C}_\ell}, \alpha, \delta\Big)$                    [ [50]]
    $\widehat{p}_\ell \leftarrow |\mathcal{C}_\ell|/n_{L2}$
Output: $\big\{\mathcal{C}_\ell,\ \widehat{\mathbf{w}}_\ell,\ \widehat{s}_\ell^2,\ \widehat{p}_\ell\big\}_{\ell=1}^k$
```

I Proof of Lemma B.6, Classification and robust estimation

Lemma I.1 (Lemma A.15 in [42]). *Given estimated parameters satisfying $\|\widetilde{\mathbf{w}}_i - \mathbf{w}_i\|_2 \le \Delta/10$, $(1-\Delta^2/50)\widetilde{r}_i^2 \le s_i^2 + \|\widetilde{\mathbf{w}}_i - \mathbf{w}_i\|_2^2 \le (1+\Delta^2/50)\widetilde{r}_i^2$ for all $i \in [k]$, and a new task with $t_{\text{out}} \ge \Theta\big(\log(k/\delta)/\Delta^4\big)$ samples whose true regression vector is $\beta = \mathbf{w}_h$, Algorithm 9 predicts h correctly with probability $1-\delta$.*

Since the set G contains n_{L2} i.i.d. samples from our data generation model, by the assumption that $n_{L2} = \widetilde{\Omega}\Big(\frac{d}{p_{\min}\epsilon^2 t_{L2}}\Big) = \Omega\Big(\frac{\log(k/\delta)}{p_{\min}}\Big)$ and from Proposition J.6, it holds that the number of tasks such that $\beta = \mathbf{w}_i$ is $n_{L2}\widehat{p}_i \ge \frac{1}{2}n_{L2}p_i$ with probability at least $1-\delta$. Hence, with this probability, there exists at least $n_{L2}p_i/2$ i.i.d. examples in G for estimating $\mathbf{w}_i$ and s_i^2. Lemma B.6 guarantee that our algorithm correctly classified all the tasks in G, which implies that there are at least $(p_i/2 - \alpha_{L2})n_{L2}$ uncorrupted tasks, and at most $\alpha_{L2}n_{L2}$ corrupted tasks, and hence the corruption level is at most $\frac{\alpha_{L2}n}{(p_i/2-\alpha_{L2})n} \le \frac{4\alpha_{L2}}{p_i}$ since $\alpha_{L2} \le p_{\min}/4$. We can apply robust linear regression algorithm to each cluster separately, and the error of the algorithm is bounded by the following lemma.

Lemma I.2 (Robust Linear Regression, Lemma 1.3 in [26]). *Let S' be an α-corrupted set of labeled samples of size $\Omega((d/\alpha^2)\operatorname{poly}\log(d/(\alpha\tau)))$. There exists an efficient algorithm that on input S' and $\alpha > 0$, returns a candidate vector $\widehat{\beta}$ such that with probability at least $1-\tau$ it holds $\Big\|\widehat{\beta} - \beta\Big\|_2 = \mathcal{O}(\sigma\alpha\log(1/\alpha))$.*

For a single regressor $i \in [k]$, Lemma I.2 implies that for any $\epsilon \ge \Omega\Big(\frac{\alpha_{L2}}{p_i}\log(p_i/\alpha_{L2})\Big)$, given that $n_{L2}t_{L2}p_i \ge \widetilde{\Omega}(d/\epsilon^2)$, it holds that with probability $1-\delta$, our estimation satisfies

$$\|\widehat{\mathbf{w}}_i - \mathbf{w}_i\|_2 \le \epsilon s_i.$$

Using Corollary J.3, we have that our robust variance estimator guarantees that

$$\begin{aligned}\Big|\widehat{s}_i^2 - \Big(s_i^2 + \|\widehat{\mathbf{w}}_i - \mathbf{w}_i\|_2^2\Big)\Big| &\le \frac{\epsilon}{\sqrt{t_{L2}}}\Big(s_i^2 + \|\widehat{\mathbf{w}}_i - \mathbf{w}_i\|_2^2\Big)\\ &\le \frac{\epsilon}{\sqrt{t_{L2}}}s_i^2,\end{aligned}$$

with probability $1-\delta$. Taking a union bound over k regressors, we have that for any $\epsilon \ge \Omega\Big(\frac{\alpha}{p_{\min}}\log(p_{\min}/\alpha)\Big)$, given that

$$n_{L2} \ge \widetilde{\Omega}\bigg(\frac{d}{t_{L2}p_{\min}\epsilon^2}\bigg),$$

for all $i \in [k]$, our estimators satisfy

$$\begin{aligned} \|\widehat{\mathbf{w}}_i - \mathbf{w}_i\|_2^2 &\le \epsilon s_i, \quad \text{and} \\ \left|\widehat{s}_i^2 - s_i^2\right| &\le \frac{\epsilon}{\sqrt{t_{L2}}} s_i^2. \text{ (Applying Corollary J.3)} \end{aligned}$$

By Proposition J.6, it holds that

$$\begin{aligned} |\widehat{p}_i - p_i| &\le \sqrt{\frac{3\log(k/\delta)}{n_{L2}} p_i} + \alpha_{L2} \\ &\le \min\left\{p_{\min}/10, \epsilon p_i \sqrt{t_{L2}/d}\right\} + \alpha_{L2}. \end{aligned}$$

The condition on ϵ can be converted in to a condition on α_H by the fact that $\frac{\beta}{\log \frac{1}{\beta}} \ge \frac{e}{e-1}\alpha \implies \alpha \log \frac{1}{\alpha} \le \beta$ for $\alpha, \beta \in (0, 1)$. This completes the proof.

J Proofs of technical lemmas and remarks

J.1 Auxiliary Lemmas

Fact J.1 (ϵ-tail bound for distributions with bounded second moment)**.** *Suppose random variable z with probability density function $p(\cdot)$, satisfies $\mathbb{E}\left[z^2\right] \le \sigma^2$, then for any event $\mathcal{E}$ with $\mathbb{P}[\mathcal{E}] \ge 1 - \epsilon$, it holds that*

$$|\mathbb{E}[z] - \mathbb{P}[\mathcal{E}]\, \mathbb{E}[z|\mathcal{E}]| \le \sqrt{\epsilon}\sigma.$$

Proof. Notice that

$$\begin{aligned} & |\mathbb{E}[z] - \mathbb{P}[\mathcal{E}]\, \mathbb{E}[z|\mathcal{E}]| \\ =& \left|\mathbb{P}\left[\bar{\mathcal{E}}\right] \mathbb{E}\left[z|\bar{\mathcal{E}}\right]\right| \\ =& \left|\int_{-\infty}^{\infty} \mathbb{1}\left\{z \in \bar{\mathcal{E}}\right\} z p(z) dz\right| \\ \le& \sqrt{\int_{-\infty}^{\infty} \mathbb{1}\left\{z \in \bar{\mathcal{E}}\right\} p(z) dz \cdot \int_{-\infty}^{\infty} z^2 p(z) dz} \quad \text{(Using Cauchy–Schwarz)} \\ \le& \sqrt{\epsilon}\sigma. \end{aligned}$$

□

Proposition J.2 (Matrix Bernstein inequality, Theorem 1.6.2 in [70])**.** *Let $\mathbf{S}_1, \ldots, \mathbf{S}_n$ be independent, centered random matrices with common dimension $d_1 \times d_2$, and assume that each one is uniformly bounded $\mathbb{E}[\mathbf{S}_k] = \mathbf{0}$ and $\|\mathbf{S}_k\|_2 \le L \;\forall\; k = 1, \ldots, n$.*

Introduce the sum

$$\mathbf{Z} \coloneqq \sum_{k=1}^{n} \mathbf{S}_k$$

and let $v(\mathbf{Z})$ denote the matrix variance statistic of the sum:

$$v(\mathbf{Z}) \coloneqq \max\left\{\left\|\mathbb{E}\left[\mathbf{Z}\mathbf{Z}^\top\right]\right\|_2, \left\|\mathbb{E}\left[\mathbf{Z}^\top\mathbf{Z}\right]\right\|_2\right\}$$

Then

$$\mathbb{P}[\|\mathbf{Z}\|_2 \ge t] \le (d_1 + d_2) \exp\left\{\frac{-t^2/2}{v(\mathbf{Z}) + Lt/3}\right\}$$

for all $t \ge 0$.

Corollary J.3 (Robust mean estimation of chi-square distribution)**.** *Let $G = \{x_i\}_{i=1}^n$ where each x_i is drawn independently from a scaled chi-square distribution $\frac{\sigma^2}{t}\chi^2(t)$ for $t \in \mathbb{N}$. Suppose $S = (G \setminus L) \cup E$ where $|L| \le \epsilon n$ and $|E| \le \epsilon n$. Assuming that $n = \widetilde{\Omega}(\frac{1}{\epsilon^2})$, the trimmed mean estimator define in [50] takes S as input and output an estimate $\widehat{\sigma}^2$, with probability $1 - \delta$, that satisfies*

$$\left|\widehat{\sigma}^2 - \sigma^2\right| = \mathcal{O}\left(\sigma^2 \epsilon \max\left\{\frac{\log(1/\epsilon)}{t}, \sqrt{\frac{\log(1/\epsilon)}{t}}\right\}\right).$$

Proof. We show the corollary by applying Theorem 1 in [50] to the chi-square setting. First we bound the $\mathcal{E}(4\epsilon, X)$ term in [50, Theorem 1]. For a zero mean random variable X, define quantile

$$Q_p(X) = \sup\{M \in \mathbb{R} : \mathbb{P}[X \geq M] \geq 1 - p\},$$

and $\mathcal{E}(\epsilon, X)$ is defined as

$$\mathcal{E}(\epsilon, X) \coloneqq \max\big\{\mathbb{E}\big[|X|\mathbb{1}\big\{X \leq Q_{\epsilon/2}\big\}\big], \mathbb{E}\big[|X|\mathbb{1}\big\{X \geq Q_{1-\epsilon/2}\big\}\big]\big\}.$$

where we denote $Q_p(X)$ by Q_p simply. Let $X = \frac{x_i}{\sigma^2} - 1$. Note that under chi-square distribution for x_i, for small ϵ, we can assume $Q_{\epsilon/2} \leq 0$ and $Q_{1-\epsilon/2} \geq 0$. By Bernstein's inequality, we have that for any $u \geq 0$,

$$\mathbb{P}\Big[\Big|\frac{x_i}{\sigma^2} - 1\Big| \geq u\Big] \leq 2\exp\big(-ct\min\big\{u^2, u\big\}\big).$$

If $\log(2/\epsilon) \leq ct$, let $u_\epsilon = \sqrt{\log(2/\epsilon)/ct}$ such that $2\exp\big(-ct\min\big\{u_\epsilon^2, u_\epsilon\big\}\big) = \epsilon$. We have

$$\begin{aligned}
\mathbb{E}\big[|X|\mathbb{1}\big\{X \geq Q_{1-\epsilon/2}\big\}\big] &= \int_{Q_{1-\epsilon}}^{\infty} \mathbb{P}[X \geq u]du \\
&= \int_{Q_{1-\epsilon}}^{u_\epsilon} \mathbb{P}[X \geq u]du + \int_{u_\epsilon}^{\infty} \mathbb{P}[X \geq z]dz \\
&\leq \epsilon \cdot u_\epsilon + \frac{\epsilon}{ct}\sqrt{\log\frac{2}{\epsilon}} \\
&= \mathcal{O}\left(\frac{\epsilon\sqrt{\log(1/\epsilon)}}{\sqrt{t}}\right).
\end{aligned}$$

Otherwise, let $u_\epsilon = \log(2/\epsilon)/ct$ such that $2\exp(-ct\min(u_\epsilon^2, u_\epsilon)) = \epsilon$. Then,

$$\begin{aligned}
\mathbb{E}[|X|\mathbb{1}\{X \geq Q_{1-\epsilon}\}] &= \int_{Q_{1-\epsilon}}^{\infty} \mathbb{P}[X \geq u]du \\
&= \int_{Q_{1-\epsilon}}^{u_\epsilon} \mathbb{P}[X \geq u]du + \int_{u_\epsilon}^{\infty} \mathbb{P}[X \geq u]du \\
&\leq \epsilon \cdot u_\epsilon + \frac{2}{ct}\exp(-ctu_\epsilon) \\
&= \mathcal{O}\Big(\frac{\epsilon\log(1/\epsilon)}{t}\Big).
\end{aligned}$$

Combing the two term we get

$$Q_{1-\epsilon/2}(X) = \mathcal{O}\left(\epsilon\max\left\{\frac{\log(1/\epsilon)}{t}, \sqrt{\frac{\log(1/\epsilon)}{t}}\right\}\right),$$

which implies

$$\mathcal{E}(4\epsilon, X) = Q_{1-\epsilon/2}(X) = \mathcal{O}\left(\epsilon\max\left\{\frac{\log(1/\epsilon)}{t}, \sqrt{\frac{\log(1/\epsilon)}{t}}\right\}\right).$$

The variance of X is bounded as

$$\sigma_X^2 = \mathrm{Var}\Big[\frac{x_i}{\sigma^2} - 1\Big] = \mathcal{O}\Big(\frac{1}{t}\Big).$$

Hence, with the assumption that $n = \widetilde{\Omega}(1/\epsilon^2)$, Theorem 1 in [50] guarantee to estimate σ^2 with error

$$\frac{\big|\widehat{\sigma}^2 - \sigma^2\big|}{\sigma^2} = \mathcal{O}\left(\epsilon\max\left\{\frac{\log(1/\epsilon)}{t}, \sqrt{\frac{\log(1/\epsilon)}{t}}\right\}\right). \qquad \square$$

Proposition J.4 (Trimmed mean estimator for distributions with bounded variances (see, e.g. Proposition 2.2 in [65]))**.** *Suppose a multi-set $S = \{x_i\}_{i=1}^n$, $0 < \epsilon \leq 1/8$, satisfies $S = (G \setminus L) \cup E$, $|E| \leq \epsilon|G|$, $|L| \leq \epsilon|G|$, and set G satisfies*

$$\left|\frac{1}{|G|}\sum_{x_i \in G} x_i - \mu\right| \leq \sqrt{\epsilon}$$
$$\frac{1}{|G|}\sum_{x_i \in G} (x_i - \mu)^2 \leq 1$$

Let R be the set containing the lower and upper 2ϵ quantiles from S, then set $S' = S \setminus R = (G \setminus L') \cup E'$ satisfies

$$\left|\frac{1}{|S'|}\sum_{x_i \in S'} x_i - \mu\right| \leq 18\sqrt{\epsilon},$$

Proof. The result is well-known. We provide a proof here for completeness. First note that all the datapoints in E that exceed the ϵ-quantile of G must lie in R. By Chebyshev's inequality, $\frac{1}{|G|}\sum_{x_i \in G} \mathbb{1}\left\{|x_i - \mu| \geq \sqrt{\frac{1}{\epsilon}} + \sqrt{\epsilon}\right\} \leq \epsilon$. Therefore $|x_i - \mu| \leq \sqrt{\frac{1}{\epsilon}} + \sqrt{\epsilon}$ for all $x_i \in E'$. Second, since $|L'| \leq 5\epsilon$, by Fact J.1, the mean of $G \setminus L'$ lies within $\sqrt{\frac{10\epsilon}{1-5\epsilon}}$ of μ. Finally, the difference between μ and the mean of $(G \setminus L') \cup E'$ is bounded by

$$\begin{aligned}\left|\frac{1}{|S'|}\sum_{x_i \in S'} x_i - \mu\right| &\leq \frac{1}{|S'|}\left(|E'|(\sqrt{\frac{1}{\epsilon}} + \sqrt{\epsilon}) + |G \setminus L'|\sqrt{\frac{10\epsilon}{1-5\epsilon}}\right)\\ &\leq 18\sqrt{\epsilon} \text{ (Assuming that } \epsilon \leq 1/8) .\end{aligned}$$

□

Proposition J.5 (ℓ_1 deviation bound of multinomial distributions [74])**.** *Let $\mathbf{p} = \{p_1, \ldots, p_k\}$ be a vector of probabilities (i.e. $p_i \geq 0$ for all $i \in [k]$ and $\sum_{i=1}^k p_i = 1$). Let $\mathbf{x} \sim \mathrm{multinomial}(n, \mathbf{p})$ follow a multinomial distribution with n trials and probability $\mathbf{p}$. Then given $n \geq 2k\log(2/\delta)/\alpha^2$ with probability $1 - \delta$,*

$$\left\|\frac{1}{n}\mathbf{x} - \mathbf{p}\right\|_1 \leq \alpha,$$

Proposition J.6 (ℓ_∞ deviation bound of multinomial distributions, Proposition D.7 in [42])**.** *Let $\mathbf{p} = \{p_1, \ldots, p_k\}$ be a vector of probabilities (i.e. $p_i \geq 0$ for all $i \in [k]$ and $\sum_{i=1}^k p_i = 1$). Let $\mathbf{x} \sim \mathrm{multinomial}(n, \mathbf{p})$ follow a multinomial distribution with n trials and probability $\mathbf{p}$. Then with probability $1 - \delta$, for all $i \in [k]$,*

$$\left|\frac{1}{n}x_i - p_i\right| \leq \sqrt{\frac{3\log(k/\delta)}{n}p_i},$$

which implies

$$\left\|\frac{1}{n}\mathbf{x} - \mathbf{p}\right\|_\infty \leq \sqrt{\frac{3\log(k/\delta)}{n}}.$$

for all $i \in [k]$.

Fact J.7 (Gaussian 4-th moment conditions)**.** *Let $\mathbf{v}$, $\mathbf{u}$, and $\mathbf{w}$ denote three fixed vectors, we have*

1. $\underset{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}{\mathbb{E}}\left[\left(\mathbf{v}^\top\mathbf{x}\right)^2\left(\mathbf{u}^\top\mathbf{x}\right)^2\right] = \|\mathbf{u}\|_2^2 \cdot \|\mathbf{v}\|_2^2 + 2\langle\mathbf{u}, \mathbf{v}\rangle^2$

2. $\underset{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}{\mathbb{E}}\left[\left(\mathbf{v}^\top\mathbf{x}\right)^3\left(\mathbf{u}^\top\mathbf{x}\right)\right] = 3\|\mathbf{v}\|_2^2 \cdot \langle\mathbf{v}, \mathbf{u}\rangle$

3. $\underset{\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}{\mathbb{E}}\left[\left(\mathbf{u}^\top\mathbf{x}\right)\left(\mathbf{v}^\top\mathbf{x}\right)\left(\mathbf{w}^\top\mathbf{x}\right)^2\right] = \|\mathbf{w}\|_2^2\langle\mathbf{u}, \mathbf{v}\rangle + 2\langle\mathbf{u}, \mathbf{w}\rangle\langle\mathbf{v}, \mathbf{w}\rangle$

K Outlier robust principal component analysis

We provide comparisons of Algorithm 2 to state-of-the-art baselines, in both theory and numerical experiments.

K.1 Theoretical comparisons

Comparisons with [75]. Outlier-Robust Principal Component Analysis (ORPCA) [75, 28, 76] studies a similar problem under a Gaussian model. For comparison, we can modify the best known ORPCA estimator from [75] to our setting in Proposition 2.6, to get a semi-orthogonal $\widehat{\mathbf{U}}$ achieving

$$\left\|\mathbf{\Sigma} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \mathbf{\Sigma} \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\right\|_* = \|\mathbf{\Sigma} - \mathcal{P}_k(\mathbf{\Sigma})\|_* + \mathcal{O}\big(\, \alpha^{1/2}\|\mathcal{P}_k(\mathbf{\Sigma})\|_* + \nu k \alpha^{1/4}\,\big) \ . \tag{82}$$

We significantly improve in the dominant third term (see Eq. (5)).

This is due to our double filtering in Algorithm 3, which guarantees that we remove more corrupted examples than uncorrupted examples. On the other hand, ORPCA estimator in [75] uses a single filter that removes a single example per iteration. The removed example is a corrupted one with probability at least $\gamma \in (0, 1)$. This filter runs for $\Theta(n\alpha/\gamma)$ iterations to ensure that the corrupted examples are sufficiently removed (remaining corrupted examples only contribute to γ-fraction of the second moment). However, this filter also remove roughly a $\Theta(\alpha/\gamma)$-fraction of good examples. Under our setting, this causes a $\Theta(\alpha/\gamma)$ multiplicative error and $\Theta(k\sqrt{\alpha/\gamma})$ additive deviation by Lemma D.5, part 3. This achieves

$$\mathrm{Tr}\big[\mathbf{U}^\top \mathbf{\Sigma} \mathbf{U}\big] \;\; \geq \;\; (1 - \Theta(\gamma + \alpha/\gamma))\, \mathrm{Tr}[\mathcal{P}_k(\mathbf{\Sigma})] - \mathcal{O}\Big(\nu k \sqrt{\alpha/\gamma}\Big) \ ,$$

for any $\gamma > 0$. Setting $\gamma = \sqrt{\alpha}$ gives Eq. (82), following a similar line of analysis as in §D.1.

Comparisons with filters based on the second moment of z_i's. Popular recent results on robust estimation are based on filters that rely on the second moment [24, 66]. One might wonder if it is possible to apply these filters to z_i's to remove the corrupted samples. Such approaches fail when $n = \mathcal{O}(d)$, even when $\mathbf{x}_i$ are standard Gaussian; this is immediate from the fact that the empirical second moment of z_i's does not concentrate until we have $n = \mathcal{O}(d^2)$ uncorrupted samples.

Comparisons with robust mean estimation [17]. Another approach is to use the existing off-the-shelf robust mean estimators, such as [17], directly to estimate the second moment matrix $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$, as remarked in [64]. However, the application of [17, Theorem 1.3] does not take advantage of the spiked low-rank structure of $\mathbf{\Sigma}$. This results in a larger sample complexity scaling as $n = \widetilde{\Omega}(d^2/\alpha)$ to achieve the following bound similar to Eq. (5).

$$\left\|\mathbf{\Sigma} - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top \mathbf{\Sigma} \widehat{\mathbf{U}}\widehat{\mathbf{U}}^\top\right\|_* = \|\mathbf{\Sigma} - \mathcal{P}_k(\mathbf{\Sigma})\|_* + \mathcal{O}(\nu\, k\sqrt{\alpha}\, \|\mathbf{\Sigma}\|_2) \tag{83}$$

Remark K.1. *Given $\alpha \in (0, 1/3)$ fraction corrupted tasks, there exists an algorithm [17] that can robustly estimate the matrix $\mathbf{M} = \sum_{j=1}^{k} p_j \mathbf{w}_j \mathbf{w}_j^\top$ with $n = \Omega\Big(\frac{d^2}{\alpha}\log\frac{d}{\delta}\Big)$, and time $\widetilde{\mathcal{O}}\big(nd^2/\mathrm{poly}(\alpha)\big)$. The algorithm returns $\bar{\mathbf{M}}$ that satisfies*

$$\big\|\bar{\mathbf{M}} - \mathbf{M}\big\|_2 \lesssim \rho^2\sqrt{\alpha}$$

with probability at least $1 - \delta$ for $\delta \in (0, 1)$.

K.1.1 Proof of robust mean estimation for the second moment in Remark K.1

From the definition of $\widehat{\mathbf{M}}$,

$$\widehat{\mathbf{M}} = (2n)^{-1}\sum_{i=1}^{n}\Big(\widehat{\beta}_i^{(1)}\widehat{\beta}_i^{(2)\top} + \widehat{\beta}_i^{(2)}\widehat{\beta}_i^{(1)\top}\Big), \tag{84}$$

which is the empirical mean of the matrices $\widehat{\beta}_i^{(1)}\widehat{\beta}_i^{(2)\top} + \widehat{\beta}_i^{(2)}\widehat{\beta}_i^{(1)\top}$. Let us consider the d^2-length vectors $\widehat{\mathbf{m}}_i$, and $\mathbf{m}_i$ constructed by unrolling the matrix $\widehat{\beta}_i^{(1)}\widehat{\beta}_i^{(2)\top}$, and $\beta_i\beta_i^\top$ respectively. Then

$(2n)^{-1}\mathbf{B}\sum_{i=1}^{n}\widehat{\mathbf{m}}_i$ will be the unrolled vector corresponding to the matrix $\widehat{\mathbf{M}}$ where $\mathbf{B} \coloneqq \mathbf{I}_d + \mathbf{P}_d$, and $\mathbf{P}_d \in \mathbb{R}^{d^2 \times d^2}$ is the permutation matrix corresponding to the transposition of $d \times d$ matrices. The covariance matrix of the samples $\mathbf{B}\widehat{\mathbf{m}}$ is therefore

$$\mathbf{B}\,\mathbb{E}\big[\widehat{\mathbf{m}}\widehat{\mathbf{m}}^\top\big]\mathbf{B} - \mathbf{B}\mathbf{m}\mathbf{m}^\top\mathbf{B}, \tag{85}$$

and its norm is bounded by

$$\big\|\mathbb{E}\big[\widehat{\mathbf{m}}\widehat{\mathbf{m}}^\top\big] - \mathbf{m}\mathbf{m}^\top\big\|_2. \tag{86}$$

Returning back to the matrix notation, we therefore essentially need to bound the operator norm of the covariance tensor of the samples $\widehat{\beta}_i^{(1)}\widehat{\beta}_i^{(2)\top}$, where $\beta_i = \mathbf{w}_j$ with probability p_j. This can be bounded as

$$\begin{aligned}
&\sup_{\|\mathbf{A}\|_{\mathrm{F}}=1}\ \mathop{\mathbb{E}}_{i\sim\mathbf{p},\mathbf{x},y}\left[\mathrm{Tr}\Big[\mathbf{A}\Big(\widehat{\beta}_i^{(1)}\widehat{\beta}_i^{(2)\top} - \mathbf{M}\Big)\Big]^2\right] \\
&= \sup_{\|\mathbf{A}\|_{\mathrm{F}}=1}\ \mathop{\mathbb{E}}_{i\sim\mathbf{p},\mathbf{x},y}\left[\Big(\widehat{\beta}_i^{(2)\top}\mathbf{A}\widehat{\beta}_i^{(1)} - \mathrm{Tr}[\mathbf{A}\mathbf{M}]\Big)^2\right] \\
&= \sup_{\|\mathbf{A}\|_{\mathrm{F}}=1}\ \mathop{\mathbb{E}}_{i\sim\mathbf{p},\mathbf{x},y}\left[\Big(\widehat{\beta}_i^{(2)\top}\mathbf{A}\widehat{\beta}_i^{(1)}\Big)^2 - (\mathrm{Tr}[\mathbf{A}\mathbf{M}])^2\right] \\
&= \sup_{\|\mathbf{A}\|_{\mathrm{F}}=1}\ \mathop{\mathbb{E}}_{i\sim\mathbf{p},\mathbf{x},y}\left[\Big(\widehat{\beta}_i^{(2)\top}\mathbf{A}\widehat{\beta}_i^{(1)}\Big)^2 - \big(\mathrm{Tr}\big[\mathbf{A}\beta_i\beta_i^\top\big]\big)^2 + \big(\mathrm{Tr}\big[\mathbf{A}\beta_i\beta_i^\top\big]\big)^2 - (\mathrm{Tr}[\mathbf{A}\mathbf{M}])^2\right] \\
&\le \sup_{\|\mathbf{A}\|_{\mathrm{F}}=1}\ \mathop{\mathbb{E}}_{i\sim\mathbf{p},\mathbf{x},y}\left[\Big(\widehat{\beta}_i^{(2)\top}\mathbf{A}\widehat{\beta}_i^{(1)}\Big)^2 - \big(\beta_i^\top\mathbf{A}\beta_i\big)^2\right] + \sup_{\|\mathbf{A}\|_{\mathrm{F}}=1}\ \mathop{\mathbb{E}}_{i\sim\mathbf{p}}\left[\big(\mathrm{Tr}\big[\mathbf{A}\beta_i\beta_i^\top\big]\big)^2 - (\mathrm{Tr}[\mathbf{A}\mathbf{M}])^2\right] \\
&\le \mathop{\mathbb{E}}_{i\sim\mathbf{p}}\left[\sup_{\|\mathbf{A}\|_{\mathrm{F}}=1}\ \mathop{\mathbb{E}}_{\mathbf{x},y}\left[\Big(\widehat{\beta}_i^{(2)\top}\mathbf{A}\widehat{\beta}_i^{(1)}\Big)^2 - \big(\beta_i^\top\mathbf{A}\beta_i\big)^2\Big|\, i\right]\right] \\
&\qquad\qquad + \sup_{\|\mathbf{A}\|_{\mathrm{F}}=1}\ \mathop{\mathbb{E}}_{i\sim\mathbf{p}}\left[\big(\mathrm{Tr}\big[\mathbf{A}\beta_i\beta_i^\top\big]\big)^2 - (\mathrm{Tr}[\mathbf{A}\mathbf{M}])^2\right]
\end{aligned} \tag{87}$$

Considering the first term of Equation (87), where for a fixed i, we compute the inner expectation

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{x},y}\left[\Big(\widehat{\beta}_i^{(2)\top}\mathbf{A}\widehat{\beta}_i^{(1)}\Big)^2 - \big(\beta_i^\top\mathbf{A}\beta_i\big)^2\right] \\
&= \mathop{\mathbb{E}}_{\mathbf{x},y}\left[\widehat{\beta}_i^{(2)\top}\mathbf{A}\widehat{\beta}_i^{(1)}\widehat{\beta}_i^{(1)\top}\mathbf{A}^\top\widehat{\beta}_i^{(2)} - \beta_i^\top\mathbf{A}\beta_i\beta_i^\top\mathbf{A}^\top\beta_i\right] \\
&= \mathop{\mathbb{E}}_{\widehat{\beta}^{(1)}}\left[\mathop{\mathbb{E}}_{\mathbf{x},y}\left[\widehat{\beta}_i^{(2)\top}\mathbf{A}\widehat{\beta}_i^{(1)}\widehat{\beta}_i^{(1)\top}\mathbf{A}^\top\widehat{\beta}_i^{(2)} - \beta_i^\top\mathbf{A}\beta_i\beta_i^\top\mathbf{A}^\top\beta_i\ \Big|\ \widehat{\beta}_i^{(1)}\right]\right]
\end{aligned} \tag{88}$$

Define the PSD matrix $\mathbf{B} \coloneqq \mathbf{A}\widehat{\beta}_i^{(1)}\widehat{\beta}_i^{(1)\top}\mathbf{A}^\top$, and note that its expectation is

$$\begin{aligned}
\mathbb{E}\Big[\mathbf{A}\widehat{\beta}_i^{(1)}\widehat{\beta}_i^{(1)\top}\mathbf{A}^\top\Big] &\le \mathbf{A}\left(\left(1 + \frac{1}{t}\right)\beta_i\beta_i^\top + \frac{\rho^2}{t}\mathbf{I}\right)\mathbf{A}^\top \\
&= \left(1 + \frac{1}{t}\right)\mathbf{A}\beta_i\beta_i^\top\mathbf{A}^\top + \frac{\rho^2}{t}\mathbf{A}\mathbf{A}^\top
\end{aligned} \tag{89}$$

from which we get

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{x},y}\left[\left(\widehat{\beta}_i^{(2)\top}\mathbf{A}\widehat{\beta}_i^{(1)}\right)^2-\left(\beta_i^\top\mathbf{A}\beta_i\right)^2\right]\\
&=\mathop{\mathbb{E}}_{\widehat{\beta}_i^{(1)}}\left[\mathop{\mathbb{E}}_{\mathbf{x},y}\left[\widehat{\beta}_i^{(2)\top}\mathbf{B}\widehat{\beta}_i^{(2)}-\beta_i^\top\mathbf{A}\beta_i\beta_i^\top\mathbf{A}^\top\beta_i\;\middle|\;\widehat{\beta}_i^{(1)}\;\right]\right]\\
&=\mathop{\mathbb{E}}_{\widehat{\beta}_i^{(1)}}\left[\mathrm{Tr}\left[\mathbf{B}\mathop{\mathbb{E}}_{\mathbf{x},y}\left[\widehat{\beta}_i^{(2)}\widehat{\beta}_i^{(2)\top}\right]\right]\right]-\beta_i^\top\mathbf{A}\beta_i\beta_i^\top\mathbf{A}^\top\beta_i\\
&\le\mathop{\mathbb{E}}_{\widehat{\beta}_i^{(1)}}\left[\mathrm{Tr}\left[\mathbf{B}\left(1+\frac{1}{t}\right)\beta_i\beta_i^\top+\frac{\rho^2}{t}\mathbf{B}\right]\right]-\beta_i^\top\mathbf{A}\beta_i\beta_i^\top\mathbf{A}^\top\beta_i\\
&=\mathop{\mathbb{E}}_{\widehat{\beta}_i^{(1)}}\left[\left(1+\frac{1}{t}\right)\beta_i^\top\mathbf{B}\beta_i+\frac{\rho^2}{t}\,\mathrm{Tr}[\mathbf{B}]\right]-\beta_i^\top\mathbf{A}\beta_i\beta_i^\top\mathbf{A}^\top\beta_i\\
&=\left(1+\frac{1}{t}\right)\beta_i^\top\left(\left(1+\frac{1}{t}\right)\mathbf{A}\beta\beta^\top\mathbf{A}^\top+\frac{\rho^2}{t}\mathbf{A}\mathbf{A}^\top\right)\beta_i\\
&\qquad+\frac{\rho^2}{t}\,\mathrm{Tr}\left[\left(1+\frac{1}{t}\right)\mathbf{A}\beta_i\beta_i^\top\mathbf{A}^\top+\frac{\rho^2}{t}\mathbf{A}\mathbf{A}^\top\right]-\beta_i^\top\mathbf{A}\beta_i\beta_i^\top\mathbf{A}^\top\beta_i\\
&=\left[\left(1+\frac{1}{t}\right)^2-1\right]\left(\beta_i^\top\mathbf{A}\beta_i\right)^2+\frac{\rho^2}{t}\left(1+\frac{1}{t}\right)\beta_i^\top\left(\mathbf{A}\mathbf{A}^\top+\mathbf{A}^\top\mathbf{A}\right)\beta_i+\frac{\rho^4}{t^2}\|\mathbf{A}\|_{\mathrm{F}}^2\\
&=\left(\frac{2}{t}+\frac{1}{t^2}\right)\left(\mathrm{Tr}\left[\mathbf{A}\beta_i\beta_i^\top\right]\right)^2+\frac{\rho^2}{t}\left(1+\frac{1}{t}\right)\left(\mathrm{Tr}\left[\beta_i^\top\mathbf{A}\mathbf{A}^\top\beta_i\right]+\mathrm{Tr}\left[\beta_i^\top\mathbf{A}^\top\mathbf{A}\beta_i\right]\right)+\frac{\rho^4}{t^2}\|\mathbf{A}\|_{\mathrm{F}}^2\\
&\le\left(\frac{2}{t}+\frac{1}{t^2}\right)d\|\mathbf{A}\|_{\mathrm{F}}^2\|\beta_i\|_2^4+\frac{2\rho^2}{t}\left(1+\frac{1}{t}\right)\|\mathbf{A}\|_{\mathrm{F}}^2\|\beta_i\|_2^2+\frac{\rho^4}{t^2}\|\mathbf{A}\|_{\mathrm{F}}^2\\
&\le\frac{3\rho^4}{t}+\frac{4\rho^4}{t}+\frac{\rho^4}{t^2}\\
&=\frac{8\rho^4}{t}.
\end{aligned}
\tag{90}
$$

Plugging back Equation (90) in Equation (87) we get

$$
\begin{aligned}
\sup_{\|\mathbf{A}\|_{\mathrm{F}}=1}\mathop{\mathbb{E}}_{i\sim\mathbf{p},\mathbf{x},y}\left[\mathrm{Tr}\left[\mathbf{A}\left(\widehat{\beta}_i^{(1)}\widehat{\beta}_i^{(2)\top}-\mathbf{M}\right)\right]^2\right]&\lesssim\frac{\rho^4}{t}+\sup_{\|\mathbf{A}\|_{\mathrm{F}}=1}\mathop{\mathbb{E}}_{i\sim\mathbf{p}}\left[\left(\mathrm{Tr}\left[\mathbf{A}\beta_i\beta_i^\top\right]\right)^2-\left(\mathrm{Tr}[\mathbf{A}\mathbf{M}]\right)^2\right]\\
&\lesssim\frac{\rho^4}{t}+\rho^4\\
&\lesssim\rho^4\\
\implies\left\|\mathbf{B}\,\mathbb{E}\left[\widehat{\mathbf{m}}\widehat{\mathbf{m}}^\top\right]\mathbf{B}-\mathbf{B}\mathbf{m}\mathbf{m}^\top\mathbf{B}\right\|_2&\lesssim\rho^4.
\end{aligned}
\tag{91}
$$

Using [17, Theorem 1.3.], we finally get that the robust mean estimate $\bar{\mathbf{M}}$ of $\mathbf{M}$ can be computed using $n=\Omega\left(\frac{d^2}{\alpha}\log\frac{d}{\delta}\right)$ independent tasks in time $\mathcal{O}\left(nd^2/\mathrm{poly}(\alpha)\right)$, and satisfies

$$
\left\|\bar{\mathbf{M}}-\mathbf{M}\right\|_2\le\left\|\bar{\mathbf{M}}-\mathbf{M}\right\|_{\mathrm{F}}\le\mathcal{O}\left(\rho^2\sqrt{\alpha}\right)
\tag{92}
$$

with probability at-least $1-\delta$ for $\delta\in(0,1)$ using [17, Algorithm 2].

Lemma K.2. *Let $\beta\in\mathbb{R}^d$ be the true model for a task which gets to observe t samples $\mathbf{X}\in\mathbb{R}^{t\times d}$, and provides labels $\mathbf{y}\sim\mathcal{N}\left(\mathbf{X}\beta,\sigma^2\mathbf{I}_t\right)$ where $y_j\in\mathbb{R}$ is the label for $\mathbf{x}_j\sim\mathcal{N}(\mathbf{0},\mathbf{I}_d)$. The estimator for β would be $\widehat{\beta}:=\frac{1}{t}\mathbf{X}^\top\mathbf{y}$, and will satisfy*

$$
\mathbb{E}\left[\widehat{\beta}\widehat{\beta}^\top\right]-\beta\beta^\top=\frac{1}{t}\beta\beta^\top+\frac{\|\beta\|_2^2+\sigma^2}{t}\mathbf{I}\,.
\tag{93}
$$

Proof.

$$\begin{aligned}
\widehat{\beta} &= \frac{1}{t}\mathbf{X}^\top\mathbf{y} \\
&= \frac{1}{t}\mathbf{X}^\top\mathbf{X}\beta + \frac{1}{t}\mathbf{X}^\top\epsilon \qquad (\text{where } \epsilon \sim \mathcal{N}\big(\mathbf{0}, \sigma^2\mathbf{I}_t\big)) \\
\implies \widehat{\beta} - \beta &= \left(\frac{1}{t}\mathbf{X}^\top\mathbf{X} - \mathbf{I}_d\right)\beta + \frac{1}{t}\mathbf{X}^\top\epsilon
\end{aligned}$$

Let $\mathbf{z} := \widehat{\beta} - \beta$, then

$$\begin{aligned}
\mathbb{E}[\mathbf{z}] &= \mathbf{0}, \quad \text{and} \\
\mathbb{E}\big[\mathbf{z}\mathbf{z}^\top\big] &= \mathbb{E}\left[\left(\left(\frac{1}{t}\mathbf{X}^\top\mathbf{X} - \mathbf{I}_d\right)\beta + \frac{1}{t}\mathbf{X}^\top\epsilon\right)\left(\left(\frac{1}{t}\mathbf{X}^\top\mathbf{X} - \mathbf{I}_d\right)\beta + \frac{1}{t}\mathbf{X}^\top\epsilon\right)^\top\right] \\
&= \frac{\sigma^2}{t}\mathbf{I}_d + \frac{1}{t^2}\,\mathbb{E}\big[\mathbf{X}^\top\mathbf{X}\beta\beta^\top\mathbf{X}^\top\mathbf{X}\big] - \beta\beta^\top \\
&= \frac{1}{t}\Big(\|\beta\|_2^2 + \sigma^2\Big)\mathbf{I}_d + \frac{1}{t}\beta\beta^\top \qquad (\text{Using Fact J.7}) \qquad (94)
\end{aligned}$$

completing the proof. □

K.2 Experimental comparisons

We demonstrate the comparison between Algorithm 2 and HRPCA by considering a distribution class. Consider the distribution of the uncorrupted samples $\mathbf{x} \in \mathbb{R}^d$ to be as follows:

- $x_1 \sim \mathcal{N}(0, 1.1)$,
- $x_2 = z \cdot x_1/\sqrt{1.1}$, where z is an independent Rademacher random variable,
- $\mathbf{x}_{3:d} \sim \mathcal{N}(\mathbf{0}_{d-2}, \mathbf{I}_{d-2})$,

and α is the corruption level. We sample $n \geq 5d/\alpha$ points from this distribution. Note that $\mathbf{\Sigma} := \mathbb{E}\big[\mathbf{x}\mathbf{x}^\top\big] = \mathbf{I}_d + 0.1\mathbf{e}_1\mathbf{e}_2^\top$. The adversary then corrupts a point $\mathbf{x}$ with probability α as

$$\mathbf{x}' \leftarrow \begin{bmatrix} 0 & z' \cdot 2\alpha^{1/4} & \mathbf{x}_{3:d} \end{bmatrix},$$

where z' is an independent Rademacher random variable.

We run Algorithm 2 and HRPCA, on the above described setup by choosing $d = 10$, $\alpha \in \{0.005, 0.01, \ldots, 0.025\}$, $k = 1$, and $n = 10^4$. To evaluate the performance of both the algorithms, we compute the variance captured: $\mathrm{Tr}\Big[\widehat{\mathbf{U}}^\top\mathbf{\Sigma}\widehat{\mathbf{U}}\Big]$, where $\widehat{\mathbf{U}} \in \mathbb{R}^{d\times k}$ is the output of either algorithms. We also compute the best oracle solution which is $\mathrm{Tr}\big[\mathbf{U}_n^\top\mathbf{\Sigma}\mathbf{U}_n\big]$ where $\mathbf{U}_n \in \mathbb{R}^{d\times k}$ is the top k singular vector matrix of $\widehat{\mathbf{\Sigma}}_n := \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$ where $\{\mathbf{x}_i\}_{i=1}^n$ is the original uncorrupted sample set. Notice that this estimator is the optimal subspace estimator in the absence of extra structural assumptions about the subspaces.

We demonstrate the variance captured, the number of corrupted points left, and the number of uncorrupted points removed by the algorithms in Figure 2. A random guess of the subspace will capture roughly variance 1. The oracle estimator has variance ≈ 1.0886. The best rank-1 subspace is spanned by $\mathbf{e}_1$ whose captured variance is 1.1. We show the average performance of Algorithm 2 over 100 independent trials. The HRPCA algorithm is very slow since one need to pick the best subspace from the $n/2$ iterations while each iteration requires eigen-decomposition once. Thus we only take the average over 10 trials, which is enough to see the trend of its performance.