Statistically and Computationally Efficient Linear Meta-representation Learning

Kiran Koshy Thekumparampil[†], **Prateek Jain**[‡], **Praneeth Netrapalli**[‡], **Sewoong Oh**[±] [*]
[†]University of Illinois at Urbana-Champaign, [‡]Google Research, India,
[±]University of Washington, Seattle

Abstract

In typical few-shot learning, each task is not equipped with enough data to be learned in isolation. To cope with such data scarcity, meta-representation learning methods train across many related tasks to find a shared (lower-dimensional) representation of the data where all tasks can be solved accurately. It is hypothesized that any new arriving tasks can be rapidly trained on this low-dimensional representation using only a few samples. Despite the practical successes of this approach, its statistical and computational properties are less understood. Recent theoretical studies either provide a highly suboptimal statistical error, or require many samples for every task, which is infeasible in the few-shot learning setting. Moreover, the prescribed algorithms in these studies have little resemblance to those used in practice or they are computationally intractable. To understand and explain the success of popular meta-representation learning approaches such as ANIL [34], MetaOptNet [27], R2D2 [7], and OML [25], we study a alternating gradient-descent minimization (AltMinGD) method (and its variant alternating minimization (AltMin)) which underlies the aforementioned methods. For a simple but canonical setting of shared linear representations, we show that AltMinGD achieves nearly-optimal estimation error, requiring only $\Omega(\mathrm{polylog}\, d)$ samples per task. This agrees with the observed efficacy of this algorithm in the practical few-shot learning scenarios.

1 Introduction

Common real world tasks follow a long tailed distribution where most of the tasks only have a small number of labeled examples [42]. Collecting more clean labels is often costly (e.g., medical imaging). As each task does not have enough examples to be learned in isolation under this few-shot learning scenario, meta-learning attempts to jointly learn across a large number of tasks to exploit some structural similarities among those tasks.

One popular approach is to learn a *shared representation*, where new arriving tasks can be solved accurately [36]. The premise is that (i) there is a shared low-dimensional representation $f_U(x) \in \mathbb{R}^r$ represented by a task-independent meta-parameter U and (ii) a simple linear model of $\langle v_i, f_U(x)\rangle$ can make accurate prediction on the i-th task with a task-specific parameter v_i. Once the representation f_U has been learnt, we can rapidly adapt to new arriving tasks as the representation dimension r is much smaller than the dimension d of the input data. This approach is becoming increasingly popular with a growing list of recent applications [25, 27, 7, 33, 34, 19, 38, 10] and has been empirically shown to achieve the state-of-the-art performances on benchmark few-shot learning datasets [38, 10, 34].

These successes rely on a simple but effective training algorithm which alternately updates U and $\{v_i\}$ which we call AltMinGD (Alternating Minimization and Gradient Descent). Suppose we are

[*][†]thekump2@illinois.edu, [†]{prajain,pnetrapalli}@google.com, and [±]sewoong@cs.washington.edu.

Preprint. Under review.

given t tasks, and the i-th task is associated with a dataset $\{(x_j^{(i)} \in \mathbb{R}^d, y_j^{(i)})\}_{j=1}^m$ of size m. In this paper, we closely follow the formulation of [38], which solves for a function $f_U : \mathbb{R}^d \to \mathbb{R}^r$ (typically a deep neural network) and a task-specific linear model $v_i \in \mathbb{R}^r$ on a choice of a loss $\ell(\cdot, \cdot)$:

$$\min_U \left\{ \sum_{i\in[t]} \min_{v_i \in \mathbb{R}^r} \sum_{j\in[m]} \ell(\langle v_i, f_U(x_j^{(i)})\rangle, y_j^{(i)}) \right\}, \tag{1}$$

by alternately applying a (stochastic) gradient descent step of U in the outer loop (for given v_i's) and numerically finding the optimal solution v_i in the inner loop (for a given U). Several closely related algorithms have been proposed, including separating training-set used for the inner loop and the validation-set used for the outer-loop [34, 27, 7], early stopping the inner-loop [25], applying to datasets with imbalanced data sizes [33, 10], and proposing new architectures and regularizers [19]. There is an increasing list [38, 10, 34] of numerical evidences showing that these meta representation learning improves upon competing approaches including MAML [14] and its variants [15, 24, 30]. Further, [34] provides experimental evidences that shared representation is the dominant component in the efficacy of MAML [14], even though MAML does not explicitly seek a shared representation.

In this paper, we analyze the computational and statistical properties of AltMinGD and its variant AltMin under the simple but canonical setting of learning a shared linear representation for linear regression tasks [39]. The fundamental question of interest is: as the number of tasks grow, does AltMinGD learn the underlying r-dimensional shared representation (subspace) more accurately, and consequently make more accurate predictions on new tasks? This question is critical in explaining the empirical success in few-shot learning where the number of tasks in the training set is large while each of those tasks is data starved. Further, in settings like crowdsourcing or bioinformatics, collecting more data on new tasks is easier than collecting more data on existing tasks.

Contributions. We analyze the widely adopted AltMinGD and prove a nearly optimal error rate. We show that AltMinGD requires only $m = \Omega(\log t + \log\log(1/\varepsilon))$ samples per task to achieve an error of ε in estimating the representation U when we have a large enough number t of tasks in the training data and assuming a constant dimensionality $r = O(1)$ of the representation. Under this condition, AltMinGD achieves an error decaying as $\widetilde{O}(\sigma\sqrt{d/mt})$, which nearly matches the fundamental lower bound. Together, these analyses imply that AltMinGD is able to compensate for having only a few samples per task (small m) by having many few-shot tasks (large t), significantly improving the state-of-the-art (see Table 1). Note that the $\log\log(1/\varepsilon)$ dependence of m is hidden in the $\widetilde{\Omega}$ notation and is not explicitly visible from our main theorems or the table. A fine grained analysis showing this dependence is provided in Theorem 9 in the appendix.

We follow the proof strategy of alternating minimization algorithms for matrix sensing [22, 29], but there are important differences making the analysis challenging. First, the meta-learning dataset does not satisfy Restricted Isometry Property (RIP) central in the existing matrix sensing analysis, and hence none of the technical lemmas can be directly applied. We leverage on the *task diversity property* in Assumption 2, to prove all necessary concentration bounds. Next, there is an inherent asymmetry in the problem; we require accurate estimation of U for generalization to new arriving tasks (which is the primary goal of meta-learning), but we do not necessarily require accurate estimation of v_i's. We exploit this to ensure accurate estimation of U with a small m.

Our analysis of AltMinGD leads to a fundamental theoretical question: is the condition $m = \Omega(\log t)$ necessary? We introduce a variation AltMinGD-S, which at each iteration selects a subset of tasks that are well-behaved (covering the r-dimensional subspace of current estimated U) and uses (the empirical risk of) only those tasks in the update. While $\log t$ dependence is unavoidable if we require all t tasks to be well-behaved, ensuring a large fraction to be well-behaved requires smaller m. When the noise is sufficiently small with variance $O(1/\log t)$, we show that AltMinGD-S requires only $m = \Omega(\log\log(1/\varepsilon))$ (with no dependence on t) to estimate the shared representation accurately.

Inspired by a long line of successes in matrix completion and matrix sensing [22], we also analyze a variation of AltMinGD that alternately applies minimization for U and $\{v_i\}$ updates, which we call AltMin, and prove a slightly improved guarantees at an extra computational cost of a factor of dr^2.

Notations: $[n] = \{1, 2, \ldots, n\}$. $\|A\|$ and $\|A\|_F$ denote the spectral and Frobenius norms of a matrix A. $\langle A, B\rangle$ denotes the inner product. $A^\dagger$ is the Moore-Penrose pseudoinverse. $x \sim \mathcal{N}(0, \mathbf{I}_{d\times d})$ means that x is a d dimensional standard isotropic Gaussian random vector. $\widetilde{O}$, $\widetilde{\Omega}$ and $\widetilde{\Theta}$ hide logarithmic terms in dimension d, rank r, tolerance ε and other problems parameters.

1.1 Related work

There is a large body of work in meta-learning since the seminal work in learning to learn [37], inductive bias learning [6], and multitask learning [9]. One popular approach starting from [20, 5] is to learn a shared low-dimensional representation for a set of related tasks. This is becoming increasingly popular with empirical successes in the few-shot learning scenarios [25, 27, 7, 33, 34, 19, 38, 10].

Linear representation learning. In this paper, we show that the popular AltMinGD algorithm for solving meta representation learning, achieves near-optimal error rate and sample complexity when applied to recovering linear representations, i.e. $f_U(x) = U^T x$. This problem has been studied in [3, 35, 31] and Nuclear-norm minimization approaches are proposed in [4, 16, 2, 32] but they do not provide subspace/generalization error guarantees and suffer from large training time. Closest to our work are [39, 26, 13] which propose new algorithms with statistical guarantees. We also point out a concurrent and independent work [11], which also analyzes AltMinGD but for a special case of the noiseless setting. Authors empirically showed that AltMinGD performs better than other baseline federated learning algorithms for neural meta-representation learning on some datasets. We compare these results with our guarantees in Section 4.1.

Matrix sensing. Starting from matrix sensing and completion problems [8, 28, 22], recovering a low-rank matrix from linear measurements has been a popular topic of research. Linear meta-learning is a special case of matrix sensing, but with a non-standard sensing operator of the form $\mathcal{A}(UV^T) = [A_1(UV^T), \ldots, A_{mt}(UV^T)]$ where $A_{ij}(UV^T) = \langle x_{ij} e_i^\top, UV^\top \rangle$. This operator does not satisfy restricted isometry property in general, and existing matrix sensing results do not apply. Similar sensing operators have been studied in [21, 43] which gives $m = \Omega(d)$. We provide a significantly tightened analysis to require only $m = \Omega(\log t + \log\log(1/\varepsilon))$.

2 Problem Formulation: Meta-learning of Shared Representation

We focus on the meta-learning problem with a shared linear representation for linear regression tasks. Let t denote the number of tasks. The i-th task is associated with m samples $\{(x_j^{(i)} \in \mathbb{R}^d, y_j^{(i)} \in \mathbb{R})\}_{j=1}^m$. We assume there is a common *low-dimensional representation* $(U^*)^T x$ of each data point x, parameterized by $U^* \in \mathbb{R}^{d \times r}$ where $r \ll d$. The corresponding observation y is sampled by regressing over the low-dimensional representation $(U^*)^T x$. Now, in general, learning U^* is NP-hard [17]. Instead, similar to [39], we study the problem in the following tractable random design setting.

Assumptions 1. *Let $U^* \in \mathbb{R}^{d\times r}$ be an orthonormal matrix. For a task $i \in [t]$, with task specific parameter vector $v^{*(i)} \in \mathbb{R}^r$ and j-th example $x_j^{(i)} \sim \mathcal{N}(0, \mathbf{I}_{d\times d})$, its observation is:*

$$y_j^{(i)} \;=\; \langle x_j^{(i)}, U^* v^{*(i)} \rangle + \varepsilon_j^{(i)} \;, \tag{2}$$

where $\varepsilon_j^{(i)} \sim \mathcal{N}(0, \sigma^2)$ is the measurement noise which is independent of $x_j^{(i)}$. We denote by $\widetilde{v}^{*(i)} = U^* v^{*(i)}$ the model parameter vector for each regression task in d-dimensions. We denote the matrix of these parameters as: $\widetilde{V}^* = U^*(V^*)^T$ where $(V^*)^T = [v^{*(1)}, \ldots, v^{*(t)}]$.

The difficulty of estimating U^* still depends on the *diversity* or incoherence of the tasks.

Assumptions 2. *Let λ_1^* and λ_r^* denote the largest and smallest eigenvalues of the task diversity matrix $(r/t)(V^*)^T V^* \in \mathbb{R}^{r\times r}$ respectively. Let $\kappa = \lambda_1^*/\lambda_r^*$. We say that V^* is μ-incoherent if*

$$\max_{i\in[t]} \|v^{*(i)}\|^2 \leq \; \mu\, \lambda_r^* \;. \tag{3}$$

To estimate the subspace U, we minimize the empirical risk of the t tasks in the training data, over the meta-parameter $U \in \mathbb{R}^{d\times r}$ and the task-specific model parameters $V = [v^{(1)}, \ldots, v^{(t)}]^T \in \mathbb{R}^{t\times r}$:

$$\mathcal{L}(U, V) \;=\; \sum_{i=1}^{t} \sum_{j=1}^{m} \frac{1}{2} \left(y_j^{(i)} - \langle U v^{(i)}, x_j^{(i)} \rangle \right)^2 \;. \tag{4}$$

The problem is non-convex due to the bi-linearity of $U$and V. We are interested in the few-shot learning setting where the goal is to learn the representation accurately despite a small number of samples per task in the training data. Now, even if the representation U^* is known a priori, we would require $O(r)$ samples per task to learn the parameter v. Furthermore, information theoretically the total number of samples $m \cdot t$ should scale at most linearly with the data dimension d.

3 Alternating minimization

We focus on AltMinGD from [38], which learns a shared parameterized representation $f_U(\cdot)$ as in (1). Several variations of this algorithm are widely used, for example [34, 27, 7]. AltMinGD alternately updates the matrix of regression parameters V using exact minimization with fixed U, and updates the representation parameter U using the standard gradient descent step. Concretely,

$$v^{(i)} \in \arg\min_{v\in\mathbb{R}^r} \sum_{j\in[m]} \ell(\langle v, f_U(x_j^{(i)})\rangle, y_j^{(i)}) \quad ,\forall i \in [t]\,,$$
$$U \leftarrow U - \eta \sum_{i\in[t]}\sum_{j\in[m]} \nabla_U \ell(\langle v^{(i)}, f_U(x_j^{(i)})\rangle, y_j^{(i)})\,.$$

As $\ell(\cdot,\cdot)$ is typically a convex function, we can estimate $v^{(i)}$ efficiently for a fixed U. For the linear representation learning problem specified in Section 2, the above updates reduce to the following:

$$v^{(i)} \in \arg\min_v \sum_j (y_j^{(i)} - \langle x_j^{(i)}, Uv\rangle)^2\,,\ \text{for all } i\in[t]\,,$$
$$U \leftarrow U - \eta\nabla_U \mathcal{L}(U,V) \;=\; U + \eta\sum_{i,j}(y_j^{(i)} - \langle x_j^{(i)}, Uv^{(i)}\rangle)\, x_j^{(i)}(v^{(i)})^\top\,.$$

Given U, we can efficiently estimate each of the low r-dimensional regression parameters $v^{(i)}$'s *separately and in parallel* using standard least squares regression. Our analysis requires that when we update V for current U, U should be independent from the training points. Similarly, during the update for U, V should be independent of the data points. We ensure the independence using two strategies: (a) similar to standard online meta-learning settings [14], we randomly select (previously unseen) tasks to update U and V, (b) within each task, we divide the datapoints into two sets to update V and U separately. But in our experiments, we re-used all the samples at each iteration. Algorithm 1 presents a pseudo-code of AltMinGD applied to Problem (4). Note that in Algorithm 1, we apply QR-decomposition on U after every U update to ensure that magnitude of U and V does not stray far away from that of true U^* and V^*, respectively. Otherwise, the sample complexity requirements of the algorithm increase in the condition number factors.

Run-time and memory usage: Exact update for $v^{(i)}$ has a time complexity of $O(mr^2 + r^3)$, which can be brought down to $O(m\cdot r)$ by using gradient descent for solving the least squares. Our analysis shows that under the sample complexity assumptions of Theorem 1, each of the least squares problem has a constant condition number. So, the total number of iterations for this update scale as $\log\frac{1}{\epsilon}$ to achieve ϵ error. If we set $\epsilon = 1/poly(t,\sigma)$, then using standard error analysis, we should be able to obtain the optimal error rate in Theorem 9. The gradient descent update for U requires $O(mt\cdot dr)$ time assuming large enough mt. Furthermore space complexity of AltMinGD is $O(dr + t\cdot r^2)$.

We provide an estimate of the statistical efficiency of the AltMinGD in Theorem 1. We also provide an analysis of the traditional Alternating Minimization algorithm (AltMin) which uses exact minimization for updating U in the Appendix A. We obtain a slightly improved statistical guarantee for AltMin in terms of the condition number but its run-time is slower than that of AltMinGD.

3.1 Subset Selection

Algorithm 1 operates over all the tasks in a batch, each of which are generated using a random process. Now, if the number of tasks t is large, then there is a non-trivial probability that some of the tasks are *outliers*, i.e., they have a large amount of error. This might lead to an arbitrary poor solution due to the outlier tasks. This is reflected in our analysis of AltMinGD (see Theorem 1), where the number of samples per task grows logarithmically with t which is non-intuitive as typically larger number of tasks should not hurt the sample complexity.

In more general representation learning problems, when the number of tasks t is large, there is more chance that some of them are outlier tasks. Ideally we want to design an estimator for shared representation U that is robust to a few outlier tasks. For the linear representation learning problem, we observe that to ensure small error for a task, we require the Hessian to be well-conditioned. So, we compute the Hessian for each task with the current U, and select only tasks where the condition number of the Hessian is at most 4. Algorithm 2 applies this criteria to select tasks in each iteration,

Algorithm 1 AltMinGD : Meta-learning linear regression parameters via alternating minimization gradient descent

Required: Data: $\{(x_j^{(i)} \in \mathbb{R}^d, y_j^{(i)} \in \mathbb{R})\}_{j=1}^m$ for all $1 \le i \le t$, K: number of steps, η: stepsize.

1 Initialize $U \leftarrow U_{\text{init}}$
2 Randomly shuffle the tasks {1,...,t}
for $1 \le k \le K$ **do**
3 $\quad \mathcal{T}_k \leftarrow [1 + \frac{t(k-1)}{K}, \frac{tk}{K}]$
$\quad$ **for** $i \in \mathcal{T}_k$ **do**
4 $\quad\quad v^{(i)} \leftarrow \arg\min_{\widehat{v} \in \mathbb{R}^r} \sum_{j \in [m/2]} \left(y_j^{(i)} - \langle U\widehat{v}, x_j^{(i)} \rangle\right)^2$
$\quad$ **end**
5 $\quad U \leftarrow U + \eta \sum_{i=\mathcal{T}_k} \sum_{j=1+\frac{m}{2}}^{m} \left(y_j^{(i)} - \langle Uv^{(i)}, x_j^{(i)} \rangle\right) x_j^{(i)} (v^{(i)})^\top$
6 $\quad U \leftarrow \text{QR}(U)$
end
7 **return** U

Algorithm 2 AltMinGD-S : Meta-Learning regression parameters via AltMinGD over task subsets

Required: Data: $\{(x_j^{(i)} \in \mathbb{R}^d, y_j^{(i)} \in \mathbb{R})\}_{j=1}^m$ for all $1 \le i \le t$, K: number of steps, η: stepsize.
Use the same steps as AltMinGD (Algorithm 1), but replace Line 3 with:

3 $\mathcal{T}_k \leftarrow \Big\{ i \in [1 + \frac{t(k-1)}{K}, \frac{tk}{K}] \;\Big|\; \sigma_{\max}(\widetilde{U}^\top S^{(i)} \widetilde{U}) \le 2;\ \sigma_{\min}(\widetilde{U}^\top S^{(i)} \widetilde{U}) \ge \frac{1}{2};$
$\quad\quad$ where $\;S^{(i)} = \frac{2}{m} \sum_{j \in [m/2]} x_j^{(i)} (x_j^{(i)})^\top \;$ and $\; \widetilde{U} = \text{QR}(U) \Big\}$

and then use the standard AltMinGD updates on those selected tasks. This leads to an improved dependence on t as we show in Theorem 3.

Run-time: On top of the run-time complexity of AltMinGD, the subset selection scheme adds an additive $O(mt \cdot dr + t \cdot r^3)$ term. This arises due to $O(r^3)$ eigen-decompositions of $\widetilde{U}^\top S^{(i)} \widetilde{U} \in \mathbb{R}^{r \times r}$ for each task.

4 Statistical guarantees for Alternating Minimization algorithms

We reiterate that $\widetilde{O}$ and $\widetilde{\Omega}$ hide logarithmic terms in d and r and other problem parameters. We analyze our algorithms using the rescaled Frobenius norm error: $\|(\mathbf{I} - U^*(U^*)^\top)U\|_F/\sqrt{r} \in [0, 1]$ between the rank-r subspaces corresponding to the true U^* and the output of the algorithm U. We first provide our main results analyzing AltMinGD and AltMinGD-S, and present detailed comparisons to previous results in Section 4.1

AltMinGD: We first present our main result for the AltMinGD method (Algorithm 1), applied to the linear representation learning problem described in Section 2.

Theorem 1. *Let there be t linear regression tasks, each with m samples satisfying Assumptions 1, 2. Let $\kappa := \lambda_1^*/\lambda_r^*$ and let,*

$$m \ge \widetilde{\Omega}(r^2 + r\log t + \kappa \cdot (\sigma/\sqrt{\lambda_r^*})^2 r^2 \log t),\ t \ge \widetilde{\Omega}(\kappa \cdot \mu^2 r^3),\ \textit{ and}$$
$$mt \ge \widetilde{\Omega}(\kappa \cdot \mu d r^2 + \kappa^3 \cdot \mu d r^2 (\sigma/\sqrt{\lambda_r^*})^2).$$

Then AltMinGD (Algorithm 1), initialized at U_{init} s.t. $\|(\mathbf{I} - U^*(U^*)^\top)U_{\text{init}}\|_F \le \min(21/121, \widetilde{O}(1/\kappa))$ and run for $K = \Omega(\lceil \kappa \log(mt/(\kappa \cdot \mu dr \cdot (\sigma/\sqrt{\lambda_r^*}))\rceil)$ iterations with the stepsize $\eta = \frac{r/t}{2\lambda_1^*}$, outputs U so that the following holds (w.p. $\ge 1 - K/(dr)^{10}$):

$$\frac{\|(\mathbf{I} - U^*(U^*)^\top)U\|_F}{\sqrt{r}} \;\le\; \widetilde{O}\left(\sqrt{\kappa}\left(\frac{\sigma}{\sqrt{\lambda_r^*}}\right)\sqrt{\frac{\mu d\, r}{m\, t}}\right). \tag{5}$$

Remark 1 (Initialization): Our result holds if the initial point $U_{\rm init}$ is reasonably accurate. One choice of initialization is to use the Method-of-Moments (MoM) [39]. Due to sub-optimality of MoM approach ([39, Theorem 3], also provided in Theorem 12 in Appendix), we get an additional sample complexity requirement of $mt \geq \widetilde{\Omega}(\kappa^2 dr^2 \, (\mu\kappa + r(\sigma/\sqrt{\lambda_r^*})^4)$. Note that this does not degrade the asymptotic error rate, $\widetilde{O}(\sqrt{dr/mt})$ when $\varepsilon = \widetilde{O}(\sqrt{dr/mt}) \to 0$. In our experiments, we observed that random initialization works just as well. Such a requirement of a good initialization is common in theoretical analyses of alternating update methods [22, 29], where it has been widely observed that random initialization works well in practice.

Remark 2 (Generalization in few-shot learning): Learning a shared representation helps in generalizing to new arriving tasks in few-shot learning. Suppose we run Algorithm 1, under the conditions of Theorem 1 to get an estimated subspace U. Let a new task, whose task specific regression parameter v^{*+} lie in U^*, be introduced with m^+ samples. Now, we can apply the step 4 of Algorithm 1, with U and the new samples, to meta-learn an estimate v^+ of v^{*+}. Then the mean-squared-error (MSE) of the estimated parameter is $\widetilde{O}((\sigma/\sqrt{\lambda_r^*})(\mu dr^2/mt + r/m^+))$. Therefore, as long as mt is large enough, we only need $m^+ = \Omega(r)$ additional samples to get an arbitrarily small MSE, as opposed to $m^+ = \Omega(d)$ of the trivial baseline of solving the new task by itself. We also improve upon other baselines from [39] in terms of dependence on σ and t; see Section 4.1 and Table 1 for more details.

Remark 3 (Near-optimality of the error rate): We note that our error rate matches – up to $poly(\kappa, \mu)$ factors – the information theoretic lower bound given in Corollary 2.

Corollary 2. *[39, Theorem 5] Let $r \leq d/2$ and $mt \geq r(d-r)$, then for all V^*, w.p. $\geq 1/2$*

$$\inf_{\widehat{U}} \sup_{U \in \mathrm{Gr}_{r,d}} \frac{\|(\mathbf{I} - U^*(U^*)^\top)\widehat{U}\|_F}{\sqrt{r}} \;\geq\; \Omega\Big(\frac{1}{\kappa}\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{d\,r}{m\,t}}\Big)\,, \tag{6}$$

where $G_{r,d}$ is the Grassmannian manifold of r-dimensional subspaces in $\mathbb{R}^d$, the infimum for $\widehat{U}$ is taken over the set of all measurable functions that takes mt samples in total from the model in Section 2 satisfying Assumption 1 and 2.

However, the sufficient conditions on mt in Theorem 1 has a factor r gap from the necessary condition above, which we discuss with a concrete example in the next remark.

Remark 4 (Gaussian example): Let us interpret our result using a concrete example. Consider independent Gaussian parameters $v^{*(i)} \sim \mathcal{N}(0, (1/r)\mathbf{I}_{r\times r})$ such that the signal-to-noise ratio (i.e., $x^T U^* v^{*(i)}/\sigma^2$) is independent of r. Then with high probability $\|v^{*(i)}\| = \widetilde{\Theta}(1)$ and $\lambda_1^* = \lambda_r^* = \widetilde{\Theta}(1)$. It follows that as per Assumption 2 the condition number $\kappa = \widetilde{\Theta}(1)$ and $\mu = \widetilde{\Theta}(1)$. To estimate U^* up to an ε error, AltMinGD needs a total of $mt = \widetilde{O}(dr^2 + \sigma^2 dr/\varepsilon^2)$ samples. The second term is dominant for small ε and is optimal, which follows from the near-optimality in Remark 2. However, it is an open question if the first term is necessary, as the best known lower bound in the noiseless case will require $mt = \Omega(dr)$. In this well-behaved Gaussian case, AltMinGD requires $m \geq \widetilde{\Omega}(r^2 + (1+\sigma^2) r \log t)$ per task samples.

Remark 5 (Dependence on the minimum eigenvalue): Notice that in the limit of $\lambda_r^* \to 0$, V^* is rank deficient, thus making it impossible to recover the entire subspace of U^*. This is reflected in our Theorem 3 where the error-rate approaches the maximum possible value of one as λ_r^* approaches zero (the LHS of Eq. (5) is at most one). However, for prediction error, smaller rank of V^* implies smaller dimensional representation to be learned, thus the *prediction error* bound should *improve* with lower λ_r^* (and also smaller rank of V^*). Proving a tight guarantee in the prediction error is more challenging and most of the existing results in matrix sensing literature [21] also provide guarantees in parameter estimation error.

On the contrary, the lower-bound in (6) becomes zero as λ_r^* decreases, implying that the lower-bound is significantly weaker in λ_r^*. This is expected since the lower-bound is derived through a lower-bound for the corresponding subspace regression loss. Intuitively when $\lambda_r^* = 0$ the tasks become less diverse (more homogeneous), and therefore the regression becomes easier. We also note that such condition number mismatch in upper and lower-bounds are common in low-rank literature [22].

***Task subset selection (AltMinGD-S)*:** One downside of Algorithm 1 is that m needs to increase with t (i.e., $m = \Omega(\log t)$). We introduce AltMinGD-S in Algorithm 2 to study a fundamental question of whether this $\log t$ dependence is necessary. We show that when the noise is sufficiently small, AltMinGD-S achieves a per task sample complexity that does not increase with t.

Theorem 3. *Consider the setting of Theorem 1. Let $\kappa := \lambda_1^*/\lambda_r^*$.*

$$m \geq \widetilde{\Omega}(r^2 + \kappa \cdot (\sigma/\sqrt{\lambda_r^*})^2 r^2 \log t), \quad t \geq \widetilde{\Omega}(\kappa \cdot \mu^2 r^3), \text{ and}$$
$$mt \geq \widetilde{\Omega}(\kappa \cdot \mu d r^2 + \kappa^3 \cdot \mu d r^2 (\sigma/\sqrt{\lambda_r^*})^2).$$

Then AltMinGD (Algorithm 2), initialized at U_{init} s.t. $\|(\mathbf{I} - U^*(U^*)^\top)U_{\text{init}}\|_F \leq \min(21/121, \widetilde{O}(\frac{1}{\kappa}))$ and run for $K = \Omega(\lceil \kappa \log(mt/(\kappa \cdot \mu d r \cdot (\sigma/\sqrt{\lambda_r^*}))\rceil)$ iterations using the stepsize $\eta = \frac{r/t}{2\lambda_1^*}$, outputs U so that the following holds (w.p. $\geq 1 - K/(dr)^{10}$):

$$\frac{\|(\mathbf{I} - U^*(U^*)^\top)U\|_F}{\sqrt{r}} \;\leq\; \widetilde{O}\left(\sqrt{\kappa}\left(\frac{\sigma}{\sqrt{\lambda_r^*}}\right)\sqrt{\frac{\mu d\, r}{m\, t}}\right). \tag{7}$$

Remark 6 (When noise is small enough): Note that when the noise variance σ is small enough, i.e. $\sigma^2 \ll O(1/\log t)$, AltMinGD-S only needs $m \geq \widetilde{\Omega}(r^2)$ samples per task, assuming suitable initialization (see Remark 1). Furthermore, since AltMin-S selects a fraction of tasks to perform updates and the selection process requires only $O(mt \cdot dr + t \cdot r^3)$, the time-complexity of the method remains same as that of AltMinGD, up to constant factors. Note that in the noiseless setting, AltMinGD-S removes the dependence of m on t completely, as shown below.

Corollary 4. *Let there be t linear regression tasks, each with m samples satisfying Assumptions 1, 2,*

$$m \geq \widetilde{\Omega}(r^2), \; t \geq \widetilde{\Omega}(\mu^2 r^3 K), \text{ and } mt \geq \widetilde{\Omega}(\mu d r^2 K).$$

Additionally assume that the observations are noiseless, i.e. $\sigma = 0$. Then AltMinGD-S (Algorithm 2), initialized at U_{init} s.t. $\|(\mathbf{I} - U^*(U^*)^\top)U_{\text{init}}\|_F \leq \min(21/121, \widetilde{O}(1/\kappa))$ and run for K iterations using the stepsize $\eta = \frac{1}{2\lambda_1^*}$, outputs U so that the following holds (w.p. $\geq 1 - K/(dr)^{10}$):

$$\frac{\|(\mathbf{I} - U^*(U^*)^\top)U\|_F}{\sqrt{r}} \;\leq\; (1 - \frac{1}{6\kappa})^K \widetilde{O}\,(\kappa)\;. \tag{8}$$

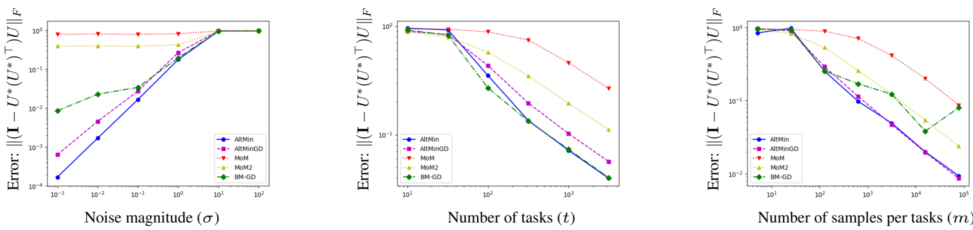
The above corollary shows that in the noiseless setting, the per-task sample complexity for AltMinGD-S does not grow with t, and is nearly optimal. Also note that that even for noiseless setting, techniques like Method-of-Moments (MoM) still incur error of $\sqrt{dr/mt}$, ignoring κ terms. In contrast, AltMinGD-S when initialized using MoM method (see Remark 1)), incurs just $\widetilde{O}(\exp(-t/\kappa))$ error.

Proofs of Theorems 1 & 3 are in Appendix C.1 & Appendix D.1.

4.1 Sample complexity comparison

To the best of our knowledge, Theorems 1 and 3 presents the first analysis of an efficient method for achieving optimal error rate in σ, d and r. [39] is most relevant that analyzes the landscape of the Empirical Risk Minimization (ERM) with Burer-Monteiro factorization. It shows that ERM can achieve a rescaled Frobenius norm error of ε with t tasks (assuming $t \geq d$), when $m \geq \widetilde{\Omega}(r^4 \log(t) + r^2 \log(t)\sigma^2/\varepsilon^2)$. We stress that this is highly sub-optimal as for small estimation ε, more tasks do not help improve the per-task sample complexity. This also does not reconcile with practice where more tasks tend to help accuracy and helps overcome small number of samples per-task. In contrast, AltMinGD requires $m \geq \widetilde{\Omega}(r^2(1+\sigma^2)\log(t) + (r^2\sigma^2/\varepsilon^2)(d/t))$ where small error ε can be achieved by collecting more tasks and increasing t. [13] studies the global minimizer of the non-convex ERM optimization in Eq. (4), without providing an efficient algorithm to solve it. The authors show that non-convex ERM achieves a small generalization error if $m = \widetilde{\Omega}(d)$, which is impractical in the few-shot learning setting.

Another approach is Method-of-Moments (MoM), which estimates U by finding the principal directions of a particular 4th moment of the data [39, 26]. MoM can indeed trade-off smaller error ε by increasing the number of tasks t. But the algorithm is inexact, i.e., even for $\sigma = 0$, we need $m \to \infty$ to achieve exact recovery of U^*. This is in a stark contrast with our approach where for



(a) (b) (c)

Figure 1: (a): AltMin and AltMinGD achieves vanishing error as noise σ decreases, whereas the error of the two Method-of-Moments (MoM, MoM2) stay bounded away from zero. BM-GD, which seems unstable and hard to tune, achieves an intermediate level of error. (b), (c): AltMin, AltMinGD and BM-GD incurs significantly smaller error in estimation of true subspace U^* than MoM and MoM2, both for growing number of tasks (t) and for growing number of samples per task (m).

noiseless case, we can find U^* *exactly*, as long as $m = O(r \log t + r^2)$ and $t = O(dr)$; see Figure 1a for an illustration. We consolidate these comparisons in Table 1.

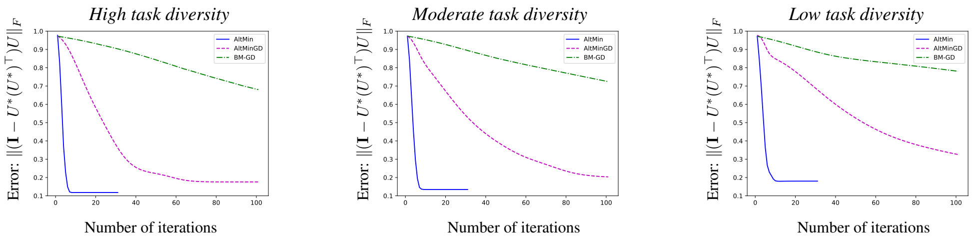
Finally, a *concurrent and independent* work by [11] also analyzes AltMinGD but only for the special case when there is no noise, i.e., $\sigma = 0$. We show tighter results that are more generally applicable: (i) our analysis applies to general noise σ that is not necessarily zero, (ii) even in the noiseless case, our analysis of AltMinGD is tighter and shows a smaller sample complexity, and (iii) we present novel AltMinGD-S that further improves the sample complexity. Precisely, in the noiseless case, [11] proves that $m = \widetilde{\Omega}(\kappa^2 \cdot r^3 \, \log t)$ is sufficient for finding U^* with a large enough t. Our tighter analysis shows that $m = \widetilde{\Omega}(r \log t + r^2)$ (Theorem 1) is sufficient with no dependence in κ. Note that the condition number $\kappa > 1$ and can be arbitrarily large depending on the problem instance. Further, we present a novel algorithm, AltMinGD-S , that only requires $m \geq \widetilde{\Omega}(r^2)$ with no dependence in κ or t (Corollary 4).

Table 1: Comparison of per-task sample complexity results $m(t, \varepsilon)$ to reach ε error when solving linear meta-representation learning with t tasks, d dimensions, subspace rank $r = O(1)$ and noise variance σ^2 (Sections 4, 2); let $t > d$. We also report if the prescribed algorithm is computationally tractable and extendable to practical neural-net setting. AltMinGD-S relies on the eigen values of the data when projected onto U and cannot be directly applied to neural networks.

Analysis	**Per-task sample complexity** $m(t, \varepsilon)$	**Tractable?**	**Practical?**
Non-convex ERM [13]	$\widetilde{\Omega}(d + \log(t) + \frac{\sigma^2}{\varepsilon^2})$	No	–
Burer-Monteiro ERM [39]	$\widetilde{\Omega}(\log(t) + \frac{\sigma^2}{\varepsilon^2})$	Yes	Yes
Method-of-Moments [39, 26]	$\Omega(1 + \frac{d}{t\varepsilon^2} + \frac{\sigma^2 d}{t\varepsilon^2})$	Yes	No
AltMinGD (Theorem 1)	$\widetilde{\Omega}((1 + \sigma^2) \log t + \frac{\sigma^2 d}{t\varepsilon^2})$	Yes	Yes
AltMinGD-S (Theorem 3)	$\widetilde{\Omega}(1 + \sigma^2 \log(t) + \frac{\sigma^2 d}{t\varepsilon^2})$	Yes	No
Lower-bound [39]	$\widetilde{\Omega}(1 + \frac{\sigma^2 d}{t\varepsilon^2})$	–	–

5 Experimental results

In this section we empirically compare the performance of our methods AltMinGD (Algorithm 1) and its exact minimization variant AltMin (Algorithm 3 in Appendix), two different versions of Method-of-Moments (MoM [39], MoM2 [26]), and simultaneous gradient descent on (U, V) using the Burer-Monteiro factorized loss (4) (BM-GD [39]). In all the figures, the magenta dashed line with square marker represents AltMinGD, the blue straight line with circular marker denotes the AltMin , the red dotted line with downwards pointing triangular marker denotes the MoM, the yellow dotted line with upwards pointing triangular marker represents the MoM2, and the green dashed and dotted line with diamond marker represents the BM-GD. In all the figures we plot the subspace estimation error of the output U of the algorithms. The error is calculated using the rescaled Frobenius norm $\|(\mathbf{I} - U^*(U^*)^\top)U\|_F/\sqrt{r}$, which takes a value in the interval $[0, 1]$. All results are averaged over multiple runs, and details of the experiments and extra plots are provided in Appendix G.



(a) (b) (c)

Figure 2: AltMin converges using fewer iterations than AltMinGD and BM-GD, but AltMinGD can be faster in practice due to its computationally cheaper iteration. While the performance of all the methods degrade as task diversity decreases, AltMin appears to be most robust to changes in diversity.

Figure 1a plots subspace distance against the standard deviation σ of the regression noise, $\varepsilon_j^{(i)} \sim \mathcal{N}(0, \sigma^2)$; see (2). Clearly, as predicted by Theorems 1 and 5 (in Appendix), our methods AltMinGD and AltMin achieve smaller error than MoM methods for small noise regime. Here the error of AltMinGD and AltMin are linearly proportional to σ. However as predicted the error of MoM and MoM is a constant multiple of $\sqrt{dr^3/mt} = \sqrt{r}$ for all values of σ, and it does not improve when σ decreases (see Table 1). While BM-GD does not have any known algorithmic guarantees, it still performs better than MoM methods. However, BM-GM becomes unstable and challenging to tune at low noise regime. Figure 1b plots the subspace error against the number of tasks t. In Figure 1c, we plot the the error against the number samples per tasks m. In both of these figures, we observe that, AltMinGD, AltMin and BM-GD achieve much smaller subspace error than the MoM and MoM2. Furthermore, as predicted, the squared error of all these methods decrease linearly in m and t. We again note that BM-GD is unstable and hard to tune, especially for large t.

In Figure 2, we plot the subspace estimation error against the number of iterations of AltMinGD, AltMin, and BM-GD for varying levels of task diversity/incoherence (Assumption 2). We observe that AltMin takes significantly fewer iterations to converge than AltMinGD and BM-GD, and AltMinGD converges earlier than BM-GD. However, each iteration of AltMin is very slow as it needs $O(d^3)$ operations, where as AltMinGD and BM-GD need only $O(d)$ operations per iteration. Therefore, AltMinGD could be the fastest in practical high-dimensional setting. BM-GD seems to be slower than AltMinGD because BM-GD seems to need a smaller stepsize than AltMinGD to stabilize its convergence. While all the methods perform worse when the task diversity decreases ($(a) \to (b) \to (c)$), we see that AltMin is more robust than others. This may be attributable to AltMin's tighter dependence on the condition number of the problem (Theorem 5, in Appendix) when compared to AltMinGD (Theorem 1). We give more details of these experiments in Appendix G.

6 Conclusion

When learning a shared representation for multiple tasks, a common approach is to alternate between finding the best linear model for each task on the current representation, and taking one gradient descent step to update the shared representation. This algorithm, AltMinGD, has been widely used in meta-representation learning with little theoretical understanding. We provide insights into the empirical success of AltMinGD by studying it in the simple but canonical problem of linear meta-learning. We showed that, AltMinGD can provide a nearly optimal error rate, along with nearly optimal per-task and overall sample complexities in their dependence in the dimensionality d of the data. To the best of our knowledge, this is the first such optimal error rate that scales appropriately with the noise in observations, while still ensuring per-task sample complexity to be nearly independent of the dimensionality d. Latter is a key requirement in meta-learning as individual tasks are data-starved. The limitations of our results are: (i) the analysis does not extend to non-linear representations, (ii) the dependence on the rank r of the shared subspace, the incoherence μ, and the condition number κ may not be tight; and (iii) our analysis is "local" and requires a good initialization.

We also proposed and analyzed a task subset selection-based method (AltMinGD-S) that further improves the per-task sample complexity and ensures that it is *independent* of the number of tasks in small noise regime. However, the subset selection scheme heavily relies on the linearity of the shared representation. Therefore, this scheme cannot be directly applied to more practical neural network training. It also remains an open question if it is possible to achieve a per-task sample complexity that does not depend on the number of tasks t, even in the large noise setting.

Our work leads to several interesting future directions and questions. For the non-linear version of the problem, ensuring optimal error rate with optimal per-task sample complexity is an interesting open question. Finally, analyzing alternating minimization methods with stochastic gradients and streaming tasks is another promising direction. Our proof techniques could be combined with that of recent results in efficient one-pass SGD [23] to design a nearly optimal stochastic algorithm.

References

[1] Ali R Amir-Moéz et al. Extreme properties of eigenvalues of a hermitian transformation and singular values of the sum and product of linear transformations. *Duke Mathematical Journal*, 23(3):463–476, 1956.

[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] Jonathan Baxter. Learning internal representations. In *Proceedings of the eighth annual conference on Computational learning theory*, pages 311–320, 1995.

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

[7] Luca Bertinetto, Joao F Henriques, Philip Torr, and Andrea Vedaldi. Meta-learning with differentiable closed-form solvers. In *International Conference on Learning Representations*, 2018.

[8] Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

[9] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[10] Yinbo Chen, Xiaolong Wang, Zhuang Liu, Huijuan Xu, and Trevor Darrell. A new meta-baseline for few-shot learning. *arXiv preprint arXiv:2003.04390*, 2020.

[11] Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. *arXiv preprint arXiv:2102.07078*, 2021.

[12] I. Csiszár and G. Tusnady. Information geometry and alternating minimization procedure. *Statistics and Decision*, 1984.

[13] 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.

[14] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400*, 2017.

[15] Fusheng Hao, Fengxiang He, Jun Cheng, Lei Wang, Jianzhong Cao, and Dacheng Tao. Collect and select: Semantic alignment metric learning for few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8460–8469, 2019.

[16] 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.

[17] Moritz Hardt, Raghu Meka, Prasad Raghavendra, and Benjamin Weitz. Computational limits for matrix completion. In *Conference on Learning Theory*, pages 703–725. PMLR, 2014.

[18] Harold V Henderson and Shayle R Searle. On deriving the inverse of a sum of matrices. *Siam Review*, 23(1):53–60, 1981.

[19] Shaoli Huang and Dacheng Tao. All you need is a good representation: A multi-level and classifier-centric representation for few-shot learning. *arXiv preprint arXiv:1911.12476*, 2019.

[20] Nathan Intrator and Shimon Edelman. Making a low-dimensional representation suitable for diverse tasks. In *Learning to learn*, pages 135–157. Springer, 1996.

[21] Prateek Jain and Inderjit S Dhillon. Provable inductive matrix completion. *arXiv preprint arXiv:1306.0626*, 2013.

[22] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674, 2013.

[23] Prateek Jain, Sham Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18, 2018.

[24] Muhammad Abdullah Jamal and Guo-Jun Qi. Task agnostic meta-learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11719–11727, 2019.

[25] Khurram Javed and Martha White. Meta-learning representations for continual learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[26] Weihao Kong, Raghav Somani, Zhao Song, Sham Kakade, and Sewoong Oh. Meta-learning for mixed linear regression. *arXiv preprint arXiv:2002.08936*, 2020.

[27] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10649–10657. IEEE Computer Society, 2019.

[28] Raghu Meka, Prateek Jain, and Inderjit S Dhillon. Guaranteed rank minimization via singular value projection. *arXiv preprint arXiv:0909.5457*, 2009.

[29] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. *IEEE Transactions on Signal Processing*, 63(18):4814–4826, 2015.

[30] Boris N Oreshkin, Pau Rodriguez, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. *arXiv preprint arXiv:1805.10123*, 2018.

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

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

[33] Siyuan Qiao, Chenxi Liu, Wei Shen, and Alan L. Yuille. Few-shot image recognition by predicting parameters from activations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[34] Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of MAML. *arXiv preprint arXiv:1909.09157*, 2019.

[35] 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.

[36] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[37] Sebastian Thrun and Lorien Pratt. Learning to learn: Introduction and overview. In *Learning to learn*, pages 3–17. Springer, 1998.

[38] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? *arXiv preprint arXiv:2003.11539*, 2020.

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

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

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

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

[43] Kai Zhong, Prateek Jain, and Inderjit S Dhillon. Efficient matrix sensing using rank-1 gaussian measurements. In *International conference on algorithmic learning theory*, pages 3–18. Springer, 2015.

Appendix

This appendix contains additional results, proofs for all the claims, and details of the experiments. Section A provides the alternating (exact) minimization algorithm (AltMin) and its task subset selection-based variant, and their statistical and computational guarantees. Sections C and D contain the analyses of Algorithm 1 and 2, respectively. Sections B and D contain the analyses of Algorithm 3 and 4, respectively. Section E contains corollaries of some known results. Section F contains some general technical lemmas used in this paper. Section G provides the details of the experiments.

A Alternating Exact Minimization algorithms

A.1 Alternating minimization (AltMin) algorithm

In this section we discuss the alternating (exact) minimization algorithm (Algorithm 3), which has been widely studied in different but related problems including matrix completion, tensor completion, phase retrieval, and matrix sensing. The algorithm follows the standard alternating minimization procedure [22, 12] where we update the representation matrix U and regression parameters V alternately. Note that, given U, we can estimate each of the parameters vector $v^{(i)}$ *separately* using standard least squares regression, i.e.,

$$v^{(i)} = \arg\min_{v} \sum_{j} (y_j^{(i)} - \langle x_j^{(i)}, Uv \rangle)^2 \, .$$

Similarly, given the updated regression parameter vectors $v^{(i)}$'s, we can now update U as:

$$U = \arg\min_{U} \sum_{i,j} (y_j^{(i)} - \langle x_j^{(i)}, Uv^{(i)} \rangle)^2 \, .$$

To ensure certain normalization, we analyze a modification of the algorithm where the next iterate for U is the orthonormal subspace containing $\widehat{U}$, which we can obtain using the QR-decomposition of $\widehat{U}$.

Similar to AltMinGD, our analysis requires that when we update V using current U, we require U to be independent from the training data points. Similarly, during the update for U, we require V to be independent of the data points. Again this is ensured using the same two strategies: a) similar to standard online meta-learning settings [14], we select random (previously unseen) tasks and update U and V, b) within each task, we divide the data points into two sets to update V and U separately.

Run-time and memory usage: Although AltMin is a conceptually simpler algorithm than AltMinGD (Algorithm 1), per iteration cost of AltMin is larger than AltMinGD due to the exact minimization on U. Our update for $v^{(i)}$ require $O(mr^2 + r^3)$ time complexity, which can be brought down to $O(m \cdot r)$ by using gradient descent for solving the least squares. Our analysis shows that under the sample complexity assumptions of Theorem 5, each of the least squares problem has a constant condition number. So, the total number of iterations scale as $\log(1/\epsilon)$ to achieve ϵ error. If we set $\epsilon = 1/poly(t, \sigma)$, then using standard error analysis, we should be able to obtain the optimal error rate in Theorem 8. Similarly, *exact* update for U requires $O((dr)^3 + mt \cdot (dr)^2)$ time, that decreases to $O(mt \cdot d \cdot r)$ when using gradient descent updates.

A.1.1 Subset Selection (AltMin-S)

Similar to AltMinGD-S (Algorithm 2), to reduce the per-task sample complexity, we also provide an algorithm AltMin-S (Algorithm 4) based on selecting a subset of tasks. This uses the same subset selection scheme as AltMinGD-S. Since, AltMin-S selects a fraction of tasks to perform updates, it has the same run-time and memory complexities as AltMin.

A.2 Statistical guarantees for Alternating Exact Minimization algorithms

***Alternating minimization (AltMin)*:** We first present our main result for a standard alternating minimization method (Algorithm 3) when applied to the meta-learning linear regression problem in the problem setting described in Section 2.

Algorithm 3 AltMin : Meta-Learning linear regression parameters via Alternating Minimization

Required: Data: $\{(x_j^{(i)} \in \mathbb{R}^d, y_j^{(i)} \in \mathbb{R})\}_{j=1}^m$ for all $1 \le i \le t$, K: number of steps.

1 Initialize $U \leftarrow U_{\text{init}}$
2 Randomly shuffle the tasks {1,...,t}
for $1 \le k \le K$ **do**
3 $\quad \mathcal{T}_k \leftarrow [1 + \frac{t(k-1)}{K}, \frac{tk}{K}]$
$\quad$ **for** $i \in \mathcal{T}_k$ **do**
4 $$v^{(i)} \leftarrow \arg\min_{\widehat{v} \in \mathbb{R}^r} \sum_{j \in [m/2]} \left(y_j^{(i)} - \left\langle U\widehat{v}, x_j^{(i)} \right\rangle \right)^2$$
$\quad$ **end**
5 $$\widehat{U} \leftarrow \arg\min_{\widehat{U} \in \mathbb{R}^{d \times r}} \sum_{i = \mathcal{T}_k} \sum_{j=1+\frac{m}{2}}^{m} \left(y_j^{(i)} - \left\langle \widehat{U} v^{(i)}, x_j^{(i)} \right\rangle \right)^2$$
6 $\quad U \leftarrow \mathrm{QR}(\widehat{U})$
end
return U

Algorithm 4 AltMin-S : Meta-Learning regression parameters via AltMin over task subsets

Required: Data: $\{(x_j^{(i)} \in \mathbb{R}^d, y_j^{(i)} \in \mathbb{R})\}_{j=1}^m$ for all $1 \le i \le t$, K: number of steps, η: stepsize.
Use the same steps as AltMinGD (Algorithm 3), but replace Line 3 with:

3 $$\mathcal{T}_k \leftarrow \Big\{ i \in [1 + \tfrac{t(k-1)}{K}, \tfrac{tk}{K}] \;\Big|\; \sigma_{\max}(\widetilde{U}^\top S^{(i)} \widetilde{U}) \le 2;\ \sigma_{\min}(\widetilde{U}^\top S^{(i)} \widetilde{U}) \ge \tfrac{1}{2}; \text{ where } S^{(i)} = \tfrac{2}{m} \textstyle\sum_{j \in [m/2]} x_j^{(i)} (x_j^{(i)})^\top \text{ and } \widetilde{U} = \mathrm{QR}(U) \Big\}$$

Theorem 5. *Let there be t linear regression tasks, each with m samples satisfying Assumptions 1, 2. Let $\kappa := \lambda_1^*/\lambda_r^*$ and let,*

$$m \ge \widetilde{\Omega}(r^2 + r \log t + (\sigma/\sqrt{\lambda_r^*})^2 r^2 \log t),\ t \ge \widetilde{\Omega}(\mu^2 r^3), \ \textit{and}$$
$$mt \ge \widetilde{\Omega}(\kappa \cdot \mu d r^2 + \kappa \cdot \mu d r^2 (\sigma/\sqrt{\lambda_r^*})^2).$$

Then AltMin (Algorithm 3), initialized at U_{init} s.t. $\|(\mathbf{I} - U^*(U^*)^\top)U_{\text{init}}\|_F \le \min(21/121, \widetilde{O}(1/\sqrt{\kappa}))$ and run for $K = \Omega(\lceil \log(mt/(\kappa \cdot \mu d r \cdot (\sigma/\sqrt{\lambda_r^*}))\rceil)$ iterations, outputs U so that the following holds (w.p. $\ge 1 - K/(dr)^{10}$):

$$\frac{\|(\mathbf{I} - U^*(U^*)^\top)U\|_F}{\sqrt{r}} \le \widetilde{O}\left(\left(\frac{\sigma}{\sqrt{\lambda_r^*}} \right) \sqrt{\frac{\mu\, r\, d}{m\, t}} \right). \tag{9}$$

Remark 7 (Initialization): Our result holds if the initial point U_{init} is reasonably accurate. One choice of initialization is to use the Method-of-Moments (MoM) [39]. Due to sub-optimality of MoM approach ([39, Theorem 3], also provided in Theorem 12 in Appendix), we get an additional sample complexity requirement of $mt \ge \widetilde{\Omega}(\kappa d r^2 (\mu\kappa + r(\sigma/\sqrt{\lambda_r^*})^4)$. Note that this does not degrade the asymptotic error rate, $\widetilde{O}(\sqrt{dr/mt})$ when $\varepsilon = \widetilde{O}(\sqrt{dr/mt}) \to 0$. Similar to case of AltMinGD, in our experiments, we observed that random initialization works just as well for AltMin too. This is analogous to alternating minimizaition for other problems where it has been widely observed that random initialization works well in practice [22, 29].

Remark 8 (Optimality and comparison to AltMinGD): Similar to Theorem 1, this error rate is nearly optimal in terms of d, r, m, t and $\sigma/\sqrt{\lambda_r^*}$, as it matches best possible rate when V^* is specified a priori (Corollary 2). However, we see that the error rate, the additional sample complexities and required initial error are all tighter than Theorem 1 in terms of condition number κ factors. However, in high-dimensions, one iteration of AltMinGD may be much faster than that of AltMin.

***Task subset selection (AltMin-S)*:** Just as we did in AltMinGD-S (Algorithm 2), we reduce the dependence of the per-task sample complexity $m = \Omega(\log(t))$ of AltMin (Algorithm 3) on the number of tasks t. This achieved through a subset selection-based AltMin-S (Algorithm 4) algorithm, which has the following guarantees for noisy and noiseless ($\sigma = 0$) observations.

Theorem 6. *Let there be t linear regression tasks, each with m samples satisfying Assumptions 1, 2. Let $\kappa := \lambda_1^*/\lambda_r^*$ and let,*

$$m \geq \widetilde{\Omega}(r^2 + (\sigma/\sqrt{\lambda_r^*})^2 r^2 \log t),\ t \geq \widetilde{\Omega}(\mu^2 r^3),\ \text{ and}$$
$$mt \geq \widetilde{\Omega}(\kappa \cdot \mu d r^2 + \kappa \cdot \mu d r^2 (\sigma/\sqrt{\lambda_r^*})^2).$$

Then AltMin-S (Algorithm 4), initialized at U_{init} s.t. $\|(\mathbf{I} - U^*(U^*)^\top)U_{\text{init}}\|_F \leq \min(21/121, \widetilde{O}(1/\sqrt{\kappa}))$ and run for $K = \Omega(\lceil \log(mt/(\kappa \cdot \mu d r \cdot (\sigma/\sqrt{\lambda_r^*}))\rceil)$ iterations, outputs U so that, w.p. $\geq 1 - K/(dr)^{10}$

$$\frac{\|(\mathbf{I} - U^*(U^*)^\top)U\|_F}{\sqrt{r}} \;\leq\; \widetilde{O}\left(\left(\frac{\sigma}{\sqrt{\lambda_r^*}}\right)\sqrt{\frac{\mu\, r\, d}{m\, t}}\right)\ . \tag{10}$$

Corollary 7. *Consider t linear regression tasks, each with m samples satisfying Assumptions 1 and 2 with $\sigma = 0$, and*

$$m \geq \widetilde{\Omega}(r^2),\ t \geq \widetilde{\Omega}(\mu^2 r^3),\ \text{ and }\ mt \geq \widetilde{\Omega}(\kappa \cdot \mu d r^2).$$

Then AltMin-S (Algorithm 4), initialized at U_{init} s.t. $\|(\mathbf{I} - U^*(U^*)^\top)U_{\text{init}}\|_F \leq \min(21/121, \widetilde{O}(1/\sqrt{\kappa}))$, and run for K iterations outputs U so that the following holds (w.p. $\geq 1 - K/(dr)^{10}$):

$$\frac{\|(\mathbf{I} - U^*(U^*)^\top)U\|_F}{\sqrt{r}} \;\leq\; \widetilde{O}\left(\frac{\sqrt{\lambda_r^*/\lambda_1^*}}{\sqrt{r}2^K}\right)\ . \tag{11}$$

Remark 9 (Subset selection): Note that when noise is very small $\sigma \ll O(\sqrt{\lambda_r^*}/\log t)$ or when the observations are noiseless ($\sigma = 0$), AltMin-S only needs $m \geq \widetilde{\Omega}(r^2)$ samples per task. Then, the per-task sample complexity does not grow with the number of tasks t. Again, we see that the sample complexity and iteration complexity of AltMin-S is smaller than AltMinGD-S. However, AltMinGD-S could still be faster than AltMin-S, due to its faster iterations.

Proofs of Theorems 5 & 6 are in the Appendices B.1 & D.1.

A.3 Proof sketch for noiseless case

Here we provide proof sketches of Theorem 5. To highlight the main ideas behind our analysis, we start with the simplest case when there is no noise ($\sigma^2 = 0$) and all the task specific regression parameters lie on a single dimensional subspace ($r = 1$). The analysis gets quite challenging as we go to multi-dimensional shared subspace ($r > 1$), and we illustrate these challenges and how to resolve them in Section A.3.2.

A.3.1 Proof sketch for the one-dimensional case

Let $u^* \in \mathbb{R}^d$ be the unit vector of the one-dimensional true subspace, and $v^* \in \mathbb{R}^t$ the vector of the true regression parameters of the t tasks. In the noiseless setting ($\varepsilon_j^{(i)} = 0$), the k-th step of AltMin can be written as follows.

For all $i \in \mathcal{T}_k$

$$v^{(i)} \leftarrow (u^\top S_1^{(i)} u)^{-1} u^\top S_1^{(i)} (u^*) v^{*(i)}\ ,$$

$$\widehat{u} \leftarrow \Big(\sum_{i \in \mathcal{T}_k} (v^{(i)})^2 S_2^{(i)}\Big)^\dagger \Big(\sum_{i \in \mathcal{T}_k} v^{*(i)} v^{(i)} S_2^{(i)} u^*\Big)\ ,\ u^+ \leftarrow \frac{\widehat{u}}{\|\widehat{u}\|},$$

where $S_\ell^{(i)} = \frac{2}{m}\sum_{j=(\ell-1)m/2+1}^{\ell m/2} x_j^{(i)}(x_j^{(i)})^\top$ is the data covariance matrix of a half of the dataset $[m]$ of task $i \in [t]$. Our incoherence condition for rank-1 case simplifies to $\|v\|_\infty^2 \leq \frac{\mu}{t}\|v\|^2$. The distance between two unit norm vectors u and u^* is commonly measured by the angular distance defined as $\sin\theta(u, u^*) \triangleq \|(\mathbf{I} - u^*(u^*)^\top)u\|^{1/2}$, where $\mathbf{I} - u^*(u^*)^\top$ is the projection operator to the sub-space orthogonal to u^*. In the following we let $q \triangleq \langle u^*, u\rangle$ and use the relation $\sin\theta(u, u^*) = \|u - u^*q\|$ in the analysis. We use the fact that if we have a good previous iterate u close to u^*, i.e. $\sin\theta(u, u^*) \le 3/4$, then $1/2 \le |q| \le 1$.

Our analysis shows that we get geometrically closer to the true subspace u^* at every iteration in this $\sin\theta$ distance, when initialized sufficiently close to u^*.

Our strategy is to show that the v-update achieves $|v^{(i)} - q^{-1}v^{*(i)}| \le C\|v^{*(i)}\|\sin\theta(u, u^*)$ for some constant C, and the u-update achieves $\sin\theta(u^+, u^*) \le (c/\|v^*\|)\|v - q^{-1}v^*\|)$ where the constant c can be made as small as we want in the assumed sample regime. Together, they imply the desired theorem.

v**-update:** We can write $v^{(i)}q^{-1} - v^{*(i)}$ as

$$v^{(i)} - q^{-1}v^{*(i)} = u^\top S_1^{(i)}(qu^* - u)(u^\top S_1^{(i)}u)^{-1}q^{-1}v^{*(i)}\,.$$

In expectation, $\|\mathbb{E}[u^\top S_1^{(i)}(qu^* - u)]\| = \|u^\top(qu^* - u)\| = 1 - q^2 \le (\sin\theta(u, u^*))^2$ and $\mathbb{E}[u^\top S_1^{(i)}u] = \|u\|^2 = 1$. Therefore, by Lemma B.2, if $\sin\theta(u, u^*) \le \frac{1}{32}$ and there is enough samples per task, i.e. $m \ge \Omega(\log(t/K\,\delta))$, we can bound their deviations in terms of $\sin\theta(u, u^*)$. This implies that, with a probability of at least $1 - \delta/2$,

$$\frac{|v^{(i)} - q^{-1}v^{*(i)}|}{|v^{*(i)}|} \le \frac{\sin\theta(u, u^*)}{4}\text{, for all } i \in \mathcal{T}_k, \tag{12}$$

where we used the fact that $|q| \ge 1/2$. This in turn implies that $(1/4)|v^{*(i)}| \le |v^{(i)}|$ and v is incoherent.

u**-update:** We bound the distance between $\widehat{u}$ and u^*:

$$\begin{aligned}&\widehat{u} - u^*q \\ &= \Big(\underbrace{\sum_{i\in\mathcal{T}_k}\frac{(v^{(i)})^2}{\|v\|^2}S_2^{(i)}}_{:=A}\Big)^\dagger\Big(\underbrace{\sum_{i\in\mathcal{T}_k}\frac{v^{(i)}h^{(i)}}{\|v\|^2}S_2^{(i)}}_{:=\widehat{H}}\,u^*q\Big),\end{aligned} \tag{13}$$

where $h^{(i)} = q^{-1}v^{*(i)} - v^{(i)}$. Notice that, in expectation, $\mathbb{E}[A] = \mathbf{I}$ and $\mathbb{E}[\widehat{H}u^*q] = \frac{v^\top h}{\|v\|^2}u^*q \le \frac{\|h\|}{\|v\|}$. Therefore, by Lemma B.3, when there are enough samples, i.e. $mt \ge K\Omega(\mu d\log(\frac{1}{\delta}))$ deviations form these expected values can be bounded using the distance between v and v^*, $\|h\|$. That is with a probability of at least $1 - \frac{\delta}{2}$, A is invertible and well-conditioned,

$$A^{-1} = \mathbf{I} + E_1, \quad \text{and} \quad Hu^*q = \frac{v^\top h}{\|v\|^2}u^*q + e_2,$$

where $\|E_1\| \le \frac{1}{16}$ and $\|e_2\| \le \frac{1}{32}\Big(\frac{\|h\|}{\|v\|} + \sqrt{\frac{t}{\mu}}\frac{\|h\|_\infty}{\|v\|}\Big)$. Note that we had to critically use incoherence of intermediate v to bound e_2. Therefore

$$\widehat{u} - u^*q = \underbrace{\frac{v^\top h}{\|v\|^2}u^*q}_{:=\widehat{u}_\parallel} + \underbrace{q\frac{v^\top h}{\|v\|^2}E_1u^* + (\mathbf{I} + E_1)e_2}_{:=f}\;.$$

Notice that $\widehat{u}_\parallel$ is parallel to u^*. Rest of the terms are grouped together as f. The angle distance $\sin(u^+, u^*)$ only depends on the portion of u^+ which lie in the orthogonal subspace to u^*. Therefore, $\|\widehat{u}_\parallel\|$ does not directly contribute to the distance, and this is formalized below. Clearly, $\|(\mathbf{I} - u^*(u^*)^\top)u^+\| = \min_{q^+}\|u^+ - u^*q^+\|$. This follows from the trivial solution of the scalar quadratic problem $\min_{q^+\in\mathbb{R}} \|u - u^* q^+\|^2$. Thus,

$$\begin{aligned}\sin\theta(u^+, u^*) &= \min_{q^+} \|u^+ - u^* q^+\| \\ &\le \Big\| \frac{\widehat{u}}{\|\widehat{u}\|} - \Big(1 + \frac{h^\top v}{\|v\|^2}\Big) u^* \frac{q}{\|\widehat{u}\|} \Big\| \\ &\le \frac{\|f\|}{\|\widehat{u}\|} \le \frac{\|f\|}{q\|u^*\| - \|f\| - \|h\|/\|v\|} \,. \end{aligned} \tag{14}$$

Putting them together: We bound f using definitions of E_1 and e_2, incoherence, and (12) as

$$\|f\| \le \frac{1}{16}\frac{\|h\|}{\|v\|} + \frac{1}{32}\Big(\frac{\|h\|}{\|v\|} + \sqrt{\frac{t}{\mu}}\frac{\|h\|_\infty}{\|v\|}\Big) \le \frac{1}{8}\sin\theta(u, u^*)\,.$$

Combining this with (14), we see that with a probability of at least $1-\delta$, the angle distance geometrically decreases at each step, i.e.

$$\sin\theta(u^+, u^*) \le \frac{1}{2}\sin\theta(u, u^*). \tag{15}$$

Finally, if the initialization is good, i.e. $\sin\theta(u_{\text{init}}, u^*) \le \frac{1}{16}$, we can unroll the above inequality across iterations. Taking union bound over the iterations we get that, with a probability of at least $1 - K\delta$, the output u after K iterations satisfies

$$\sin\theta(u, u^*) \le \frac{1}{2^K}\sin\theta(u_{\text{init}}, u^*). \tag{16}$$

To achieve this, we need at least $m \ge \Omega(\log(\frac{t}{K\delta}))$ samples per task and at least $mt \ge \Omega(K\mu d\log(\frac{1}{\delta}))$ total samples.

A.3.2 Proof sketch for the r-dimensional case

Here we do not use $\sin\theta_1(U, u^*)$ distance, as the analysis of $\sin\theta_1$ gets more complicated in the general r-dimensional case. Therefore we use ℓ-2 norm based error, $\Delta(U, U^*) := (\sum_{r'=1}^r \sin^2\theta_{r'}(U, U^*))^{1/2} := \|(\mathbf{I} - U^*(U^*)^\top)U\|_F$. Let $Q = (U^*)^\top U$, then $\Delta(U, U^*) = \|U - U^*Q\|_F$, and $1/2 \le \|Q\| \le 1$ if $\Delta(U, U^*) \le 3/4$.

For all $i \in \mathcal{T}_k$

$$\begin{aligned} v^{(i)} &\leftarrow (U^\top S_1^{(i)} U)^\dagger U^\top S_1^{(i)} U^* v^{*(i)} \,, \\ \widehat{U} &\leftarrow \Big(\mathcal{A}^\dagger\Big(\sum_{i\in\mathcal{T}_k} S_2^{(i)} U^* v^{*(i)} (v^{(i)} W^{-\frac{1}{2}})^\top\Big)\Big) W^{-\frac{1}{2}} \,, \\ U &\leftarrow \text{QR}(\widehat{U}) \,, \end{aligned}$$

where $W = V^\top V$, $\mathcal{A} : \mathbb{R}^{d\times r} \to \mathbb{R}^{d\times r}$ is linear operator such that $\mathcal{A}(U) = \sum_{i\in\mathcal{T}_k} S_2^{(i)} U W^{-\frac{1}{2}} v^{(i)} (v^{(i)})^\top W^{-\frac{1}{2}}$, and $S_\ell^{(i)}$ are defined as in the one-dimensional case.

V-update: We will prove that $\|v^{(i)} - Q^{-1}v^{*(i)}\| = O(\Delta(U, U^*))$. Let $h^{(i)} := v^{(i)}Q^{-1} - v^{*(i)}$, then

$$h^{(i)} = (U^\top S_1^{(i)} U)^\dagger \underbrace{U^\top S_1^{(i)} (U^*Q - U) Q^\dagger v^{*(i)}}_{:=G}.$$

Notice that, in expectation, $\|\mathbb{E}[U^\top S_1^{(i)} U]\| = 1$ and $\|\mathbb{E}[G]\| = \|U^\top(U^*Q - U)\| = \|Q^\top Q - \mathbf{I}\| = \Delta^2(U, U^*)$. Therefore, by Lemma B.2, if $\Delta^2(U, U^*) \le \frac{1}{32}$ and there is enough samples per task, i.e. $m \ge \Omega(r\,\log(\frac{t}{K\,\delta}))$, we can bound their deviations in terms of $\sin\theta(u, u^*)$. This implies that, with a probability of at least $1 - \delta/2$,

$$\|h^{(i)}\| \le \frac{\|v^{*(i)}\|\Delta^2(U, U^*)}{4}\,\text{, for all } i \in \mathcal{T}_k. \tag{17}$$

Furthermore, $\|v^{(i)}\| \leq 4\|v^{*(i)}\|$ and V is incoherent.

U**-update:** We bound the distance between $\widehat{U}$ and U^*:

$$(\widehat{U} - U^*Q)W^{\frac{1}{2}} = \mathcal{A}^\dagger\Big(\underbrace{\sum_{i\in\mathcal{T}_k} S_2^{(i)}U^*Qh^{(i)}(v^{(i)})^\top W^{-\frac{1}{2}}}_{:=-\widehat{\mathcal{H}}(U^*Q)} \Big).$$

Notice that, in expectation, $\mathbb{E}[\widehat{\mathcal{H}}(U^*Q)] = \mathcal{H}(U^*Q) := U^*Q\sum_{i\in\mathcal{T}_k} h^{(i)}(v^{(i)})^\top W^{-\frac{1}{2}}$ and $\mathcal{H}(U^*Q) \leq \|H\|_F$ and $\mathbb{E}[\mathcal{A}]$ is the identity map $\mathcal{I}$. Like in the 1-dimensional case, by Lemma B.3, when there are enough samples, i.e. $mt \geq K\Omega(\mu dr^2\log(\frac{1}{\delta}))$ deviations from these expected values can be bounded using the distance between V and V^*, $\|H\|$. That is, with a probability of at least $1-\delta/2$, $\mathcal{A}$ is invertible and well-conditioned in Frobenius operator norm,

$$\mathcal{A}^{-1} = \mathcal{I} + \mathcal{E}_1, \quad \text{and} \quad \widehat{\mathcal{H}}(U^*Q) = \mathcal{H}(U^*Q) - E_2,$$

where $\|\mathcal{E}_1\|_F \leq 1/16$ and $\|E_2\|_F \leq 1/32(\|H\|_F + \sqrt{t/\mu}\|H\|_{\infty,2})$. Note that we had to critically use incoherence of intermediate V to bound E_2. Therefore,

$$(\widehat{U} - U^*Q)W^{\frac{1}{2}} = -\mathcal{H}(U^*Q) - \underbrace{cE_1\mathcal{H}(U^*Q) + (\mathcal{I}+\mathcal{E}_1)E_2}_{:=F}.$$

Now, using similar arguments as in the one-dimensional case, we get

$$\begin{aligned}\Delta(U^+, U^*) &\leq \left\|\widehat{U}R^{-1} - U^*Q + \mathcal{H}(U^*Q)\right\|_F \|W^{-\frac{1}{2}}\| \\ &\leq \frac{\|F\|_F}{\|R^{-1}\|} \leq \frac{\|F\|_F\lambda_r^{-\frac{1}{2}}}{\|QU^*\| - (\|F\|_F + \|H\|_F)\lambda_r^{-\frac{1}{2}}}.\end{aligned}$$

Putting them together: Using similar arguments as in one-dimensional case, if the initialization is good, i.e. $\Delta(U_{\text{init}}, U^*) \leq 1/16$, we can show that with a probability of at least $1-\delta$, the next iterate U^+ satisfies: $\Delta(U^+, U^*) \leq \frac{1}{2}\Delta(U, U^*)$. To achieve this, we need at least $\Omega(r\log(\frac{t}{K\delta}))$ samples per task (m) and at least $\Omega(K\mu dr^2\log(\frac{1}{\delta}))$ total samples (mt). Result now follows by applying the above result K times.

B Analysis of AltMin (Algorithm 3)

Initialized at U, the k-the step of alternating minimization-based AltMin (Algorithm 3) is:

$$v^{(i)} \leftarrow (U^\top S_1^{(i)}U)^\dagger((U^\top S_1^{(i)}U^*)v^{*(i)} + U^\top z^{(i)}), \quad \text{for } i \in \mathcal{T}_k = [1 + \frac{(k-1)t}{K}, \frac{tk}{K}] \quad (18)$$

$$\widehat{U} \leftarrow \mathcal{A}^\dagger\Big(\sum_{i\in[t]} S_2^{(i)}U^*v^{*(i)}(v^{(i)})^\top + z^{(i)}(v^{(i)})^\top\Big), \quad (19)$$

$$U^+ \leftarrow \mathrm{QR}(\widehat{U}), \quad (20)$$

where U^+ is the next iterate, $S_1^{(i)} = \frac{2}{m}\sum_{j\in[1,m/2]} x_j^{(i)}(x_j^{(i)})^\top$, $S_2^{(i)} = \frac{2}{m}\sum_{j\in[1+m/2,m]} x_j^{(i)}(x_j^{(i)})^\top$, $z^{(i)} \triangleq (1/m)\sum_{j\in[m]}\varepsilon_j^{(i)}x_j^{(i)}$ and $\mathcal{A} : \mathbb{R}^{d\times r} \to \mathbb{R}^{d\times r}$ is a self-adjoint linear operator such that $\mathcal{A}(U) = \sum_{i\in T} S^{(i)}Uv^{(i)}(v^{(i)})^\top$. The self-adjointness of $\mathcal{A}$ follows from the symmetry of $S^{(i)}$ when using cyclic property of trace as follows

$$\begin{aligned}\langle U_2, \mathcal{A}(U_1)\rangle = \sum_{i\in T}\left\langle U_2, S^{(i)}U_1v^{(i)}(v^{(i)})^\top)\right\rangle &= \sum_{i\in T}\mathrm{tr}(U_2^\top S^{(i)}U_1v^{(i)}(v^{(i)})^\top) \\ &= \sum_{i\in T}\mathrm{tr}(v^{(i)}(v^{(i)})^\top U_2^\top S^{(i)}U_1) = \langle\mathcal{A}(U_2), U_1\rangle\end{aligned} \quad (21)$$

Incoherence. $\max_i \|v^{*(i)}\|^2 \leq (\mu\, r/t)\lambda_r(\sum_{i\in[t]} v^{*(i)}(v^{*(i)})^\top)$, and we define $\nu = (1/t)\lambda_r(\sum_{i\in[t]} v^{*(i)}(v^{*(i)})^\top)$. Notice that, this non-standard definition of incoherence is related to the standard definition: $W^* = (V^*)^\top V^* = \sum_{i\in[t]} v^{*(i)}(v^{*(i)})^\top$, $V^* = \tilde{V}^* R^*$ (QR-decomposition), $\max_i \|\widetilde{v}^{*(i)}\|^2 \leq \widetilde{\mu}\, r/t$, as follows $\mu = \widehat{\mu}(\sigma_1^2(R^*)/\sigma_r^2(R^*))$.

Theorem 8. *Let there be t linear regression tasks, each with m samples satisfying Assumptions 1 and 2, and $K = \lceil \log_2(\frac{mt}{(\lambda_1^*/\lambda_r^*)(\sigma/\sqrt{\lambda_r^*})\mu dr})\rceil$, $\|(\mathbf{I} - U^*(U^*)^\top)U_{\text{init}}\|_F \leq \min\Big(\frac{21}{121}, O\Big(\sqrt{\frac{\lambda_r^*}{\lambda_1^*}\frac{1}{\log(t/K)}}\Big)\Big)$, $m \geq \Omega\Big((1 + r\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^2)r\log(\frac{t}{\delta}) + r^2\log(\frac{K}{\delta})\Big)$, $t \geq \Omega(\mu^2 r^3 K \log(\frac{K}{\delta}))$, and $mt \geq \Omega\Big(\mu dr^2 K \frac{\lambda_1^*}{\lambda_r^*}\Big(\log(\frac{t}{\delta}) + \Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^2 \log^2(\frac{t}{\delta})\log(\frac{rK}{\delta})\Big)\Big)$. Then, for any $0 < \delta < 1$, after K iterations, AltMin (Algorithm 3) returns an orthonormal matrix $U \in \mathbb{R}^{d\times r}$, such that with a probability of at least $1-\delta$*

$$\frac{1}{\sqrt{r}}\|(\mathbf{I} - U^*(U^*)^\top)U\|_F \leq O\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\mu dr K \log(\frac{t}{\delta})\log(\frac{rK}{\delta})}{mt}}\Big) \tag{22}$$

and the algorithm uses an additional memory of size $O(d^2r^2)$.

A proof is in Section B.1.

Initialization. If we initialize AltMin (Algorithm 3) with Method-of-Moments (Theorem 12), we need at least

$$mt \geq \widetilde{\Omega}\Big(\frac{\lambda_1^{*2}}{\lambda_r^{*2}}\mu dr^2 + \Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^4 \frac{\lambda_1^*}{\lambda_r^*} dr^3\Big) \tag{23}$$

initial number of samples, where $\widetilde{\Omega}$ hides polylog factors.

B.1 Proof of Theorem 8

Proof sketch: We first prove that distance between U^* and U decreases at each iteration up to some additional noise terms. Then this per iterate result is unrolled to obtained the final guarantees.

First we focus on the k-th iterate. In this analysis, unless specified $[t]$, represents the k-th K-way partition used for the k-th iterate. In the following lemma we prove that tasks subset used for each iteration, satisfy approximate incoherence.

Lemma B.1 (Shuffling and partition of tasks)**.** *Let $\mathcal{T}_k$ be the k-th subset ($k \in [K]$) of the K-way partition of the shuffled set of all t tasks. If $t \geq \Omega(\mu^2 r^3 K \log(1/\delta))$, then with a probability of at least $1 - \delta/3$,*

$$\lambda_1(\sum_{i\in\mathcal{T}_k} v^{*(i)}(v^{*(i)})^\top) = \frac{1}{K}\Theta(\lambda_1((V^*)^\top V^*)) \quad \textit{and} \tag{24}$$

$$\lambda_r(\sum_{i\in\mathcal{T}_k} v^{*(i)}(v^{*(i)})^\top) = \frac{1}{K}\Theta(\lambda_r((V^*)^\top V^*))\ , \quad \textit{for all } r' \in [r] \tag{25}$$

where are $\lambda_1(\cdot)$ and $\lambda_r(\cdot)$ are the largest and smallest, respectively, eigenvalue operators of real-symmetric $r\times r$ matrix.

A proof is in Section B.5.

In the analysis of an iterate we denote the current iterate using U and the next iterate using U^+. First we prove that the distance between the true $v^{*(i)}$ and the current $v^{(i)}$ is approximately upper-bounded by multiple of distance between U and U^*. Next we prove that distance between U^+ and U^* is approximately a fraction of the distance between $v^{*(i)}$ and $v^{(i)}$. Finally, combining the above two results gives us desired result.

Preliminaries: Let $Q = (U^*)^\top U$. Using Lemma F.4, if $\|U - U^*(U^*)^\top U\|_F < 1$, Q is invertible. Let Q^{-1} be the right inverse of Q, i.e. $QQ^{-1} = \mathbf{I}$. Let $W = (V^*)^\top V^* = \sum_{i\in[t]} v^{*(i)}(v^{*(i)})^\top$, then using Assumption 2 we have that $\lambda_1^* = (r/t)\max_{\|z\|=1} z^\top W^* z$ and $\lambda_r^* = (r/t)\min_{\|z\|=1} z^\top W^* z$.

Update on V: Let $h^{(i)} = v^{(i)} - Q^{-1}v^{*(i)}$ and $H^T = [h^{(1)}h^{(2)}\dots h^{(t)}]$. Let $\|H\|_F \triangleq \sqrt{\sum_{i\in[t]}\|h^{(i)}\|^2}$ and $\|H\|_{\infty,2} \triangleq \max_{i\in[t]}\|h^{(i)}\|$. Let $W = V^\top V = \sum_{i\in[t]} v^{(i)}(v^{(i)})^\top$, and $\lambda_1 = (r/t)\max_{\|z\|=1} z^\top W z$ and $\lambda_r = (r/t)\min_{\|z\|=1} z^\top W z$.

Lemma B.2. *Assume that all conditions and the large probability event in Lemma B.1 holds true. If* $\|(\mathbf{I} - U^*(U^*)^\top)U\|_F \leq \min\left(\frac{21}{121}, O\left(\sqrt{\frac{\lambda_r^*}{\lambda_1^*}\frac{1}{\log(t/K)}}\right)\right)$ *and* $m \geq \Omega\Big(\left(\frac{\sigma}{\sqrt{\lambda_r^*}}\right)^2 r^2\log(\frac{t}{K\delta}) + r\log(\frac{t}{K\delta})\Big)$*, then with a probability of at least* $1-\delta/3$*,*

$$\|v^{(i)}\| \leq O\Big(\mu\,\lambda_r\Big)\,,\ \ \lambda_1 \leq 2\lambda_1^*\,,\ \text{and}\ \lambda_r^*/2 \leq \lambda_r \leq 2\lambda_r^* \tag{26}$$

and

$$\sqrt{\frac{rK}{t}}\frac{\|H\|_F}{\sqrt{\lambda_r}} \leq O\Big(\sqrt{\frac{\log(\frac{t}{K\delta})}{\log(\frac{1}{\delta})}}\sqrt{\frac{\lambda_1^*}{\lambda_r^*}}\|(\mathbf{I} - U^*(U^*)^\top)U\|_F + \frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2\log(\frac{t}{K\delta})}{m}}\Big) \tag{27}$$

$$\sqrt{\frac{rK}{t}}\frac{\|H\|_{\infty,2}}{\sqrt{\lambda_r}} \leq O\Big(\sqrt{\frac{\log(\frac{t}{K\delta})}{\log(\frac{1}{\delta})}}\|(\mathbf{I} - U^*(U^*)^\top)U\|\sqrt{\frac{\mu rK}{t}} + \frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2K\log(\frac{t}{K\delta})}{mt}}\Big) \tag{28}$$

A proof is in Section B.2.1.

Update on U: Let $W, \mathcal{H}, \widehat{\mathcal{H}} : \mathbb{R}^{d\times r} \to \mathbb{R}^{d\times r}$ be three linear operators, such that $\mathcal{W}(U) = U\sum_{i\in\mathcal{T}_k} v^{(i)}(v^{(i)})^\top = UW$, $\mathcal{H}(U) = U\sum_{i\in\mathcal{T}_k} h^{(i)}(v^{(i)})^\top$ and $\widehat{\mathcal{H}}(U) = \sum_{i\in\mathcal{T}_k} S_2^{(i)}Uh^{(i)}(v^{(i)})^\top$, where $h^{(i)} = v^{(i)} - Q^{-1}v^{*(i)}$. $\mathcal{W}$ is invertible and self-adjoint. Therefore $\mathcal{W}^{-\frac{1}{2}}$ and $\mathcal{W}^{\frac{1}{2}}$ exist. Let $\mathcal{I} : \mathbb{R}^{d\times r} \to \mathbb{R}^{d\times r}$ be the identity mapping, such that $\mathcal{I}(U) = U$.

$$\widehat{U} - U^*Q = \mathcal{A}^\dagger\big(\sum_{i\in\mathcal{T}_k} S_2^{(i)}U^*Q(Q^{-1}v^{*(i)} - v^{(i)})(v^{(i)})^\top + z^{(i)}(v^{(i)})^\top\big) \tag{29}$$

$$= \mathcal{A}^\dagger\big(-\widehat{\mathcal{H}}(U^*Q) + \sum_{i\in\mathcal{T}_k} z^{(i)}(v^{(i)})^\top\big) \tag{30}$$

$$= \mathcal{W}^{-\frac{1}{2}}(\mathcal{W}^{\frac{1}{2}}\mathcal{A}^\dagger\mathcal{W}^{\frac{1}{2}})\mathcal{W}^{-\frac{1}{2}}\big(-\widehat{\mathcal{H}}(U^*Q) + \sum_{i\in\mathcal{T}_k} z^{(i)}(v^{(i)})^\top\big) \tag{31}$$

$$= \mathcal{W}^{-\frac{1}{2}}(\mathcal{I} + \mathcal{E}_1)\big(-(\mathcal{W}^{-\frac{1}{2}}\mathcal{H} + \mathcal{E}_2)(U^*Q) + \mathcal{W}^{-\frac{1}{2}}\big(\sum_{i\in\mathcal{T}_k} z^{(i)}(v^{(i)})^\top\big)\big) \tag{32}$$

where $\mathcal{E}_1 = (\mathcal{W}^{-\frac{1}{2}}\mathcal{A}\mathcal{W}^{-\frac{1}{2}})^\dagger - \mathcal{I}$ and $\mathcal{E}_2 = \mathcal{W}^{-\frac{1}{2}}\widehat{\mathcal{H}} - \mathcal{W}^{-\frac{1}{2}}\mathcal{H}$, and $F = \widehat{U} - U^*Q + \mathcal{W}^{-1}(\mathcal{H}(U^*Q))$. Let $F = \widehat{U} - U^*Q + \mathcal{W}^{-1}(\mathcal{H}(U^*Q))$

Lemma B.3. *Assume that all conditions and the large probability event in Lemma B.2 holds true. Then,*

$$\|\mathcal{W}^{-1}\mathcal{H}(U^*Q)\|_F \leq O\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\log(\frac{t}{K})}\|(\mathbf{I} - U^*(U^*)^\top)U\|_F + \frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2\log(\frac{t}{K\delta})}{m}}\Big) \tag{33}$$

and if $mt \geq \Omega(\mu d r^2 K\log(t/K\delta))$*, then with probability at least* $1-\delta/3$

$$\|F\|_F \leq O\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{\mu d r^2 K\log(\frac{t}{K\delta})}{mt}}\|(\mathbf{I} - U^*(U^*)^\top)U\|_F + \frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\mu d r^2 K\log(\frac{t}{K\delta})\log(\frac{r}{\delta})}{mt}}\Big) \tag{34}$$

A proof is in Section B.3.1.

Lemma B.4. *If* $\frac{1}{2} \leq \sigma_{\min}(Q)$*,* $\|F\|_F \leq \frac{1}{8}$ *and* $\|\mathcal{W}^{-1}(\mathcal{H}(U^*Q))\|_F \leq \frac{1}{8}$*, then* R *is invertible and* $\|R^{-1}\| \leq 4$.

A proof is in Section B.4. Clearly, from (33) and (34), a sufficient condition for the above lemma is

$$O\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\log(\frac{t}{K})}\|(\mathbf{I}-U^*(U^*)^\top)U\|_F + \frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2\log(\frac{t}{K\delta})}{m}}\Big) \leq \frac{1}{8} \text{ , and} \tag{35}$$

$$O\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{\mu d r^2 K\log(\frac{t}{K\delta})}{mt}}\|(\mathbf{I}-U^*(U^*)^\top)U\|_F + \frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\mu d r^2 K\log(\frac{t}{K\delta})\log(\frac{r}{\delta})}{mt}}\Big) \leq \frac{1}{8} \tag{36}$$

which can be satisfied with

$$\|(\mathbf{I}-U^*(U^*)^\top)U\|_F \leq O\Big(\sqrt{\frac{\lambda_r^*}{\lambda_1^*}\frac{1}{\log(t/K)}}\Big), \quad m \geq \Omega\Big(\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^2 r^2\log(\frac{t}{K\delta}) + r^2\log(\frac{1}{\delta})\Big) \text{ , and} \tag{37}$$

$$mt \geq \Omega\Big(\mu d r^2 K\Big(1+\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^2\log(\frac{t}{K\delta})\log(\frac{r}{\delta})\Big)\Big) \tag{38}$$

Finally, we bound the Frobenius norm distance of the next iterate U^+ from the optimal U^*.

$$\begin{aligned}
&\|(\mathbf{I}-U^*(U^*)^\top)U^+\|_F && (39)\\
&\quad= \min_{Q^+}\|U^+ - U^*Q^+\|_F && (40)\\
&\quad\leq \|\widehat{U}R^{-1} - U^*QR^{-1} + (\mathcal{W}^{-1}\mathcal{H}(U^*Q))R^{-1}\| && (41)\\
&\quad\leq \|\widehat{U} - U^*Q + \mathcal{W}^{-1}\mathcal{H}(U^*Q)\|_F\|R^{-1}\| && (42)\\
&\quad= \|F\|_F\|R^{-1}\| && (43)\\
&\quad\leq O\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{\mu d r^2 K\log(\frac{t}{K\delta})}{mt}}\|(\mathbf{I}-U^*(U^*)^\top)U\|_F + \frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\mu d r^2 K\log(\frac{t}{K\delta})\log(\frac{r}{\delta})}{mt}}\Big) && (44)
\end{aligned}$$

If

$$mt \geq \Omega\Big(\mu d r^2 K\frac{\lambda_1^*}{\lambda_r^*}\Big(\log(\frac{t}{K\delta}) + \Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^2\log^2(\frac{t}{K\delta})\log(\frac{r}{\delta})\Big)\Big) \tag{45}$$

then,

$$\|(\mathbf{I}-U^*(U^*)^\top)U^+\|_F \leq \frac{1}{2}\|(\mathbf{I}-U^*(U^*)^\top)U\|_F + \frac{1}{2}\min\Big(\frac{21}{121}, O\Big(\sqrt{\frac{\lambda_r^*}{\lambda_1^*}\frac{1}{\log(t/K)}}\Big)\Big) \tag{46}$$

Thus if $\|(\mathbf{I}-U^*(U^*)^\top)U\|_F \leq \min\Big(\frac{21}{121}, O\Big(\sqrt{\frac{\lambda_r^*}{\lambda_1^*}\frac{1}{\log(t/K)}}\Big)\Big)$, then $\|(\mathbf{I}-U^*(U^*)^\top)U^+\|_F \leq \min\Big(\frac{21}{121}, O\Big(\sqrt{\frac{\lambda_r^*}{\lambda_1^*}\frac{1}{\log(t/K)}}\Big)\Big)$.

Therefore, using union-bound, we can un-roll the relation, between current iterate U and the next iterate U^+, over K iterations, starting from U_{init} and ending at some U iterations, to get

$$\|(\mathbf{I}-U^*(U^*)^\top)U\|_F \leq \frac{1}{2^K}\|(\mathbf{I}-U^*(U^*)^\top)U_{\text{init}}\|_F + O\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\mu d r^2 K\log(\frac{t}{K\delta})\log(\frac{r}{\delta})}{mt}}\Big) \tag{47}$$

with probability at least $1-K\delta$. Finally setting $K = \lceil\log_2(\frac{mt}{(\lambda_1^*/\lambda_r^*)(\sigma/\sqrt{\lambda_r^*})\mu d r})\rceil$ we get that, with a probability of at least $1-K\delta$

$$\|(\mathbf{I}-U^*(U^*)^\top)U\|_F \leq O\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\mu d r^2 K\log(\frac{t}{K\delta})\log(\frac{r}{\delta})}{mt}}\Big) \tag{48}$$

B.2 Analysis of update on V

B.2.1 Proof of Lemma B.2

Proof of Lemma B.2. In this proof for brevity, we will first set that $\mathcal{T}_k \leftarrow [t]$, $|\mathcal{T}_k| = t/K \leftarrow t$, $S_1^{(i)} \leftarrow S^{(i)} = \frac{1}{m}\sum_{j\in[m]} x_j^{(i)}(x_j^{(i)})^\top$. This can be done due to the approximate equivalence of the subset $\mathcal{T}_k$ and the set of all tasks $[t]$ by Lemma B.1 which requires that $t \geq \Omega(\mu^2 r^3 K \log(\frac{K}{\delta}))$. Finally at the end of the analysis we will reset $\mathcal{T}_k \leftarrow \mathcal{T}_k$, $|\mathcal{T}_k| = t/K \leftarrow t/K$, $S_1^{(i)} \leftarrow S_1^{(i)} = \frac{2}{m}\sum_{j\in[1,m/2]} x_j^{(i)}(x_j^{(i)})^\top$.

Recall the definition of $v^{(i)}$ from the update (18), and that Q^{-1} is right inverse of Q, i.e. $QQ^{-1} = \mathbf{I}$.

$$v^{(i)} - Q^{-1}v^{*(i)} = (U^\top S^{(i)}U)^\dagger(U^\top S^{(i)}(U^*Q - U))Q^{-1}v^{*(i)} + (U^\top S^{(i)}U)^\dagger U^\top z^{(i)} \tag{49}$$

We can use re-write the first term as,

$$\begin{aligned}
&(U^\top S^{(i)}U)^\dagger U^\top S^{(i)}(U^*Q - U)Q^{-1} &(50)\\
&= (U^\top S^{(i)}U)^\dagger U^\top S^{(i)}(UU^\top + U_\perp U_\perp^\top)(U^*Q - U)Q^{-1} &(51)\\
&= U^\top(U^*Q - U)Q^{-1} + (U^\top S^{(i)}U)^\dagger U^\top S^{(i)}U_\perp U_\perp^\top(U^*Q - U)Q^{-1} &(52)\\
&= -U^\top(\mathbf{I} - U^*(U^*)^\top)^2 UQ^{-1} + (U^\top S^{(i)}U)^\dagger U^\top S^{(i)}U_\perp U_\perp^\top U^* &(53)\\
&= -(U - U^*Q)^\top(U - U^*Q)Q^{-1} + (U^\top S^{(i)}U)^\dagger U^\top S^{(i)}U_\perp U_\perp^\top U^* &(54)
\end{aligned}$$

where we used the fact that $Q = (U^*)^\top U$. Therefore

$$\begin{aligned}
&\|v^{(i)} - Q^{-1}v^{*(i)}\| \leq\\
&\|U - U^*Q\|\|(U - U^*Q)Q^{-1}v^{*(i)}\| + \|(U^\top S^{(i)}U)^\dagger\|(\|U^\top S^{(i)}U_\perp U_\perp^\top U^* v^{*(i)}\| + \|U^\top z^{(i)}\|)
\end{aligned} \tag{55}$$

If $m \geq \Omega(r\log(t/\delta))$, then $\alpha = c\sqrt{\frac{r\log(27t/\delta)}{m}} \leq 1/2$ and by Lemma B.5, with a probability of at least $1-\delta$,

$$\left.\begin{aligned}
\|(U^\top S^{(i)}U)^\dagger\| &\leq (1+2\alpha),\\
\left\|U^\top S^{(i)}U_\perp U_\perp^\top U^* v^{*(i)}\right\| &\leq \alpha\|U_\perp^\top U^* v^{*(i)}\|, \quad \text{and}\\
\left\|U^\top z^{(i)}\right\| &\leq \sigma\alpha\,,
\end{aligned}\right\} \text{ for all } i \in [t] \tag{56}$$

Now if $m \geq \Omega(r\log(1/\delta))$ and $\|U^*Q - U\| \leq O\Big(\sqrt{\frac{\log(\frac{t}{\delta})}{\log(\frac{1}{\delta})}}\Big)$, then

$$\|v^{(i)} - Q^{-1}v^{*(i)} \leq O(\sqrt{\frac{\log(\frac{t}{\delta})}{\log(\frac{1}{\delta})}}(\|(U^*Q - U)Q^{-1}v^{*(i)}\| + \|U_\perp^\top U^* v^{*(i)}\|) + \sigma\sqrt{\frac{r\log(\frac{t}{\delta})}{m}}) \tag{57}$$

Next we bound $\|H\|_F$, which by definition is $\|H\|_F = \sqrt{\sum_{i\in[t]}\|h^{(i)}\|^2} = \sqrt{\sum_{i\in[t]}\|v^{(i)} - Q^{-1}v^{*(i)}\|^2}$. Using (57) and the fact that $(a^2+b^2) \leq 2(a^2+b^2)$ we get

$$\|H\|_F^2 \leq \frac{\log(\frac{t}{\delta})}{\log(\frac{1}{\delta})}[\sum_{i\in\mathcal{T}} O(\|(U^*Q - U)Q^{-1}v^{*(i)}\|^2 + \|U_\perp^\top U^* v^{*(i)}\|^2)] + t(\sigma\sqrt{\frac{r\log(\frac{t}{\delta})}{m}})^2) \tag{58}$$

Clearly $\|Q\| = \|(U^*)^\top U\| \leq \|U^*\|\|U\| \leq 1$. If $\|(\mathbf{I} - U^*(U^*)^\top)U\| \leq \|(\mathbf{I} - U^*(U^*)^\top)U\|_F \leq \frac{3}{4}$, then by using Lemma F.4, $\|Q^{-1}\| \leq 2$.

$$\sum_{i\in[t]} \|(U^*Q - U)Q^{-1}v^{*(i)}\|^2 = \sum_{i\in[t]} \operatorname{tr}((v^{*(i)})^\top((U^*Q-U)Q^{-1})^\top(U^*Q-U)Q^{-1}v^{*(i)}) \quad (59)$$
$$= \operatorname{tr}(((U^*Q-U)Q^{-1})^\top(U^*Q-U)Q^{-1})\sum_{i\in[t]} v^{*(i)}(v^{*(i)})^\top) \quad (60)$$
$$\leq \|(U^*Q-U)\|_F^2\|Q^{-1}\|^2 O(\lambda_1^*)(t/r) \quad (61)$$
$$\leq 4\|(U^*Q-U)\|_F^2 O(\lambda_1^*)(t/r) \quad (62)$$

Similarly we can use Lemma F.4, to get

$$\sum_{i\in[t]} \|U_\perp^\top U^* v^{*(i)}\|^2 = \sum_{i\in[t]} \operatorname{tr}((v^{*(i)})^\top(U_\perp^\top U^*)^\top U_\perp^\top U^* v^{*(i)}) \quad (63)$$
$$= \operatorname{tr}((U_\perp^\top U^*)^\top(U_\perp^\top U^*)\sum_{i\in[t]} v^{*(i)}(v^{*(i)})^\top) \quad (64)$$
$$\leq \|U_\perp^\top U^*\|_F^2 O(\lambda_1^*)(t/r) \quad (65)$$
$$\leq \|(U^*Q-U)\|_F^2 O(\lambda_1^*)(t/r) \quad (66)$$

Therefore substituting the above two inequalities into (58) and using the fact that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for all $0 \leq a, b$ we get

$$\|H\|_F \leq O(\sqrt{\frac{\log(\frac{t}{\delta})}{\log(\frac{1}{\delta})}}\|U^*Q-U\|_F\sqrt{\lambda_1^*(t/r)} + \sqrt{t}\sigma\sqrt{\frac{r\log(\frac{t}{\delta})}{m}}) \quad (67)$$

Then as $\|(\mathbf{I} - U^*(U^*)^\top)U\|_F \leq O\Big(\sqrt{\frac{\lambda_r^*}{\lambda_1^*}\frac{1}{\log(t)}}\Big)$ and $m \geq \Omega\Big(\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^2 r^2\log(\frac{t}{\delta})\Big)$, $\|H\|_F \leq (1 - \frac{1}{\sqrt{2}})\sqrt{(t/r)\lambda_r^*}$. Using $\|Q^{-1}\| \leq 2$ in (57) we also get that

$$\|h^{(i)}\| = \|v^{(i)} - Q^{-1}v^{*(i)}\| \leq O(\sqrt{\frac{\log(\frac{t}{\delta})}{\log(\frac{1}{\delta})}}\|(U^*Q-U)\|\|v^{*(i)}\| + \sigma\sqrt{\frac{r\log(\frac{t}{\delta})}{m}}) \quad (68)$$

By definition is $\|H\|_{\infty,2} = \max_{i\in[t]}\|h^{(i)}\| = \max_{i\in[t]}\|v^{(i)} - Q^{-1}v^{*(i)}\|$. Then as $\|(\mathbf{I} - U^*(U^*)^\top)U\| \leq \|(\mathbf{I} - U^*(U^*)^\top)U\|_F \leq O\Big(\sqrt{\frac{\lambda_r^*}{\lambda_1^*}\frac{1}{\log(t)}}\Big) \leq O(1)$, $m \geq \Omega\Big(\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^2 r^2\log(\frac{t}{\delta})\Big) \geq \Omega\Big(\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^2 r\log(\frac{t}{\delta})\Big)$, $\|H\|_{\infty,2} \leq O(\mu\lambda_r^*)$. Now, using $\|H\|_F \leq (1 - \frac{1}{\sqrt{2}})\sqrt{(t/r)\lambda_r^*}$, $\|H\|_{\infty,2} \leq O(\mu\lambda_r^*)$, $\|Q\| \leq 1$ and $\frac{10}{11} \leq \sigma_{\min}(Q)$, by Lemma B.6, we get the approximate incoherence relation for the intermediate V

$$\|v^{(i)}\| \leq O\Big(\mu\,\lambda_r\Big)\ ,\ \lambda_1 \leq 2\lambda_1^*\ \text{, and}\ \ \lambda_r^* \leq 2\lambda_r \quad (69)$$

Using this we bound $\|H\|_{\infty,2}$. Using the above incoherence relation and (68), we get

$$\sqrt{\frac{r}{t}}\frac{\|H\|_{\infty,2}}{\sqrt{\lambda_r}} \leq 2\sqrt{\frac{r}{t}}\frac{\|H\|_{\infty,2}}{\sqrt{\lambda_r^*}}$$
$$\leq O\Big(\sqrt{\frac{r}{t}}\sqrt{\frac{\log(\frac{t}{\delta})}{\log(\frac{1}{\delta})}}\|U^*Q-U\|\max_{i\in[t]}\frac{\|v^{*(i)}\|}{\sqrt{\lambda_r^*}} + 2\sqrt{\frac{r}{t}}\frac{2c\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r\log(\frac{27t}{\delta})}{m}}$$
$$\leq O\Big(\sqrt{\frac{\log(\frac{t}{\delta})}{\log(\frac{1}{\delta})}}\sqrt{\frac{\mu r}{t}}\|U^*Q-U\| + \frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2\log(\frac{t}{\delta})}{mt}}\Big) \quad (70)$$

Using (69) in (67), we get

$$\sqrt{\frac{r}{t}}\frac{\|H\|_F}{\sqrt{\lambda_r}} \leq 2\sqrt{\frac{r}{t}}\frac{\|H\|_F}{\sqrt{\lambda_r^*}} \leq O\Big(\sqrt{\frac{\log(\frac{t}{\delta})}{\log(\frac{1}{\delta})}}\sqrt{\frac{\lambda_1^*}{\lambda_r^*}}\|(\mathbf{I} - U^*(U^*)^\top)U\|_F + \frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2\log(\frac{t}{\delta})}{m}}\Big) \quad (71)$$

Finally, by resetting $\mathcal{T}_k \leftarrow \mathcal{T}_k$, $|\mathcal{T}_k| = t/K \leftarrow t/K$, $S_1^{(i)} \leftarrow S_1^{(i)} = \frac{2}{m}\sum_{j\in[1,m/2]} x_j^{(i)}(x_j^{(i)})^\top$, we obtain the desired result. $\square$

B.2.2 Supporting lemmas for the analysis of update on V

Here we bound the linear operators in the $v^{(i)}$ update.

Lemma B.5. *Let $\alpha = c\sqrt{\frac{r\log(27t/\delta)}{m}}$. With a probability of at least $1-\delta$, the following are true for all $i \in [t]$*

$$\|(U^\top S^{(i)} U)^\dagger\| \leq (1+2\alpha), \tag{72}$$

$$\left\|(U^\top S^{(i)}(U^*Q - U)Q^{-1}v^{*(i)}\right\| \leq (\|(\mathbf{I} - U^*(U^*)^\top)U\| + \alpha)\|(U^*Q-U)Q^{-1}v^{*(i)}\| \tag{73}$$

$$\leq (1+\alpha)\|(U^*Q-U)Q^{-1}v^{*(i)}\| \tag{74}$$

$$\left\|U^\top S^{(i)} U_\perp U_\perp^\top U^* v^{*(i)}\right\| \leq \alpha\left\|U_\perp^\top U^* v^{*(i)}\right\|, \text{ and} \tag{75}$$

$$\left\|U^\top z^{(i)}\right\| \leq \sigma\alpha \tag{76}$$

Proof of Lemma B.5. Let $i \in [t]$.

Let $\mathcal{S} = \{v \in \mathbb{R}^r \mid \|v\| = 1\}$ be the set of all real vectors of dimension r with unit Euclidean norm. For $\epsilon \leq 1$, there exists an ϵ-net, $N_\epsilon \subset \mathcal{S}$, of size $(1+2/\epsilon)^r$ with respect to the Euclidean norm [40, Lemma 5.2]. That is for any $v' \in \mathcal{S}$, there exists some $v \in N_\epsilon$ such that $\|v' - v\|_F \leq \epsilon$.

Consider a $v \in N_\epsilon$, such that $\|v\|_F = 1$. Now we will prove with high-probability that $\left\langle((U^\top S^{(i)}U) - \mathbf{I})v, v\right\rangle$ is small. Consider the the following quadratic form

$$v^\top(U^\top S^{(i)}U)v = \frac{1}{m}\sum_{j\in[m]} \mathrm{tr}(v^\top(U^\top x_j^{(i)}(x_j^{(i)})^\top U)v) = \frac{1}{m}\sum_{j\in[m]} \mathrm{tr}((x_j^{(i)})^\top U v v^\top U^\top x_j^{(i)}) \tag{77}$$

$x_j^{(i)} \sim \mathcal{N}(0, \mathbf{I}_{d\times d})$ are i.i.d. standard Gaussian random vectors. We will use Hanson-Wright inequality (Lemma F.5) to prove that the above quadratic form concentrates around its mean. In Lemma F.6 (which is a straightforward Corollary of Hanson-Wright inequality), by setting $a \leftarrow Uv, b \leftarrow Uv$, we get that with a probability of at least $1-\delta$

$$\left|v^\top((U^\top S^{(i)}U) - \mathbf{I})v\right| \leq c\max\left(\sqrt{\frac{\log(1/\delta)}{m}}, \frac{\log(1/\delta)}{m}\right) := \Delta_\epsilon \tag{78}$$

For brevity, let $E = (U^\top S^{(i)}U) - \mathbf{I}$. Notice that E is a real symmetric matrix, therefore it has an eigen decomposition. Then, let $v' \in \mathcal{S} \subset \mathbb{R}^r$ be the largest "eigenvector" of E, such that $(v')^\top E v' = \|E\| = \max_{\|\widetilde{v}\|=1} \widetilde{v}^\top E\widetilde{v} = \max_{\|\widetilde{v}\|=\|\widetilde{v}'\|_F=1} \widetilde{v}^\top E\widetilde{v}'$. Then there exists some $v \in N_\epsilon$ such that $\|v' - v\| \leq \epsilon$.

$$\|E\|_F = (v')^\top E v = v^\top E v + (v'-v)^\top E v + (v')^\top E(v'-v) \tag{79}$$

$$\leq v^\top E v + \|v'-v\|\|E\|\|v\| + \|v'\|\|E\|\|v'-v\| \tag{80}$$

$$\leq v^\top E v + 2\epsilon\|E\| \tag{81}$$

Re-arranging and setting $\epsilon = 1/4$, and $c \leftarrow 2c$, we get

$$\|(U^\top S^{(i)}U) - \mathbf{I}\| = \|E\| \leq \Delta_{\frac{1}{4}} = \Delta. \tag{82}$$

where $\Delta = c\max\left(\sqrt{\frac{r\ \log(9/\delta)}{m}}, \frac{r\ \log(9/\delta)}{m}\right)$. If $m \geq \max(1, 4c^2)r\log(27t/\delta)$, then $\Delta \leq \alpha \leq 1/2$.

Thus with a probability of at least is is also implies that

$$\|(U^\top S^{(i)}U)^\dagger\| = (\sigma_{\min}(U^\top S^{(i)}U))^{-1} \leq \frac{1}{1-\alpha} \leq 2. \tag{83}$$

Using similar arguments we can also prove that with a probability of at least $1-\delta$

$$\begin{aligned}
&\left\|(U^\top S^{(i)}(U^*Q-U)Q^{-1}v^{*(i)}\right\| \\
&\le \left\|U^\top(U^*Q-U)Q^{-1}v^{*(i)}\right\| + \alpha\left\|(U^*Q-U)Q^{-1}v^{*(i)}\right\| && (84)\\
&\le \left\|U^\top(\mathbf{I}-U^*(U^*)^\top)UQ^{-1}v^{*(i)}\right\| + \alpha\left\|(U^*Q-U)Q^{-1}v^{*(i)}\right\| && (85)\\
&\le \left\|U^\top(\mathbf{I}-U^*(U^*)^\top)^2UQ^{-1}v^{*(i)}\right\| + \alpha\left\|(U^*Q-U)Q^{-1}v^{*(i)}\right\| && (86)\\
&\le \left\|U^\top(\mathbf{I}-U^*(U^*)^\top)(U^*Q-U)Q^{-1}v^{*(i)}\right\| + \alpha\left\|(U^*Q-U)Q^{-1}v^{*(i)}\right\| && (87)\\
&\le \left\|(\mathbf{I}-U^*(U^*)^\top)U\right\|\left\|(U^*Q-U)Q^{-1}v^{*(i)}\right\| + \alpha\left\|(U^*Q-U)Q^{-1}v^{*(i)}\right\| && (88)\\
&\le (\left\|(\mathbf{I}-U^*(U^*)^\top)U\right\| + \alpha)\left\|(U^*Q-U)Q^{-1}v^{*(i)}\right\| && (89)\\
&\le (1+\alpha)\left\|(U^*Q-U)Q^{-1}v^{*(i)}\right\|, && (90)
\end{aligned}$$

Using similar arguments we can also prove that with a probability of at least $1-\delta$

$$\left\|U^\top S^{(i)}U_\perp U_\perp^\top U^*v^{*(i)}\right\| \le \alpha\left\|U_\perp^\top U^*v^{*(i)}\right\| \tag{91}$$

and with a probability of at least $1-\delta$

$$\left\|U^\top z^{(i)}\right\| \le \sigma\alpha \tag{92}$$

Finally setting $\delta \leftarrow \delta/3/t$ and taking the union bound over three bounds over all the tasks in $[t]$ gets us the desired result. ☐

Here we prove the approximate incoherence of the intermediate V and the spectrum of intermediate W.

Lemma B.6 (Incoherence of intermediate $v^{(i)}$)**.** *If $\|H\|_F \le (1-\frac{1}{\sqrt{2}})\sqrt{(t/r)\lambda_r((r/t)W^*)}$, $\|H\|_{\infty,2}^2 \le O(\mu\lambda_r((r/t)W^*))$, $\|Q\| \le 1$ and $\frac{10}{11} \le \sigma_{\min}(Q)$, and* (67) *and* (68) *are true, then*

$$\|v^{(i)}\| \le O\Big(\mu\,\lambda_r((r/t)W)\Big)\,,\ \lambda_1((r/t)W) \le 2\lambda_1((r/t)W^*)\text{ , and} \tag{93}$$

$$(1/2)\lambda_r((r/t)W^*) \le \lambda_r((r/t)W) \le (4/3)\lambda_r((r/t)W^*) \tag{94}$$

Proof of Lemma B.6.

$$\|v^{(i)}\| \le \|Q^{-1}v^{*(i)}\| + \|v^{(i)} - Q^{-1}v^{*(i)}\| \le 2\|v^{*(i)}\| + \|h^{(i)}\| \tag{95}$$

$$\implies \|v^{(i)}\|^2 \le O(\|V^*\|_{\infty,2}^2) + O(\|H\|_{\infty,2}^2) \le O\Big(\mu\lambda_r((r/t)W^*)\Big) \tag{96}$$

where the second inequality use the definition $h^{(i)} = v^{(i)} - Q^{-1}v^{*(i)}$ and $\|Q^{-1}\| \le 2$ (as $\sigma_{\min}(Q) \ge \frac{1}{2}$), the third inequality use the fact that $a+b \le 2a^2+2b^2$ an (68), and the final inequality uses $\|H\|_{\infty,2} \le \|V\|_{\infty,2}$.

Notice that $W = V^\top V$ and $W^* = (V^*)^\top V^*$. Thus $\sqrt{\lambda_r((r/t)W)} = \sqrt{(r/t)}\sigma_r(V)$ and $\sqrt{\lambda_r((r/t)W^*)} = \sqrt{(r/t)}\sigma_r(W^*)$, and both W and W^* are positive semi-definite (PSD). Similarly, using $\sigma_{\min}(Q^{-1}) = \sigma_{\min}(((U^*)^\top U)^{-1}) \ge 1$ and Lemma F.1 we can get that

$$\begin{aligned}
\sqrt{\lambda_r((r/t)W^*)} \le \sqrt{\sigma_{\min}^2(Q^{-1})\lambda_r((r/t)W^*)} &\le \sqrt{(r/t)\lambda_r(Q^{-1}(V^*)^\top V^*Q^{-\top})} \\
&\le \sqrt{(r/t)}\sigma_r(V^*Q^{-\top}) && (97)
\end{aligned}$$

Therefore, instead of analyzing the relation between $\lambda_r(W)$ and $\lambda_r(W^*)$, we can analyze the relation between $\sigma_r(V)$ and $\sigma_r(V^*)$. Notice that $V^*Q^{-\top} = V + V^*Q^{-\top} - V$. Then by Weyl's inequality (Lemma F.2, by setting $A \leftarrow V^*Q^{-\top}$, $B \leftarrow V$, and $C \leftarrow V^*Q^{-\top} - V$) we get that

$$\begin{aligned}
\sqrt{\lambda_r((r/t)W^*)} \le \sqrt{(r/t)}\sigma_r(V^*Q^{-\top}) &\le \sqrt{(r/t)}\sigma_r(V) + \sqrt{(r/t)}\|V - V^*Q^{-\top}\| && (98)\\
&\le \sqrt{\lambda_r((r/t)W)} + \sqrt{(r/t)}\|H\| && (99)\\
&\le \sqrt{\lambda_r((r/t)W)} + \sqrt{(r/t)}\|H\|_F && (100)\\
&\le \sqrt{\lambda_r((r/t)W)} + (1-\frac{1}{\sqrt{2}})\sqrt{\lambda_r((r/t)W^*)} && (101)
\end{aligned}$$

where the last inequality uses $\|H\|_F \leq (1-\frac{1}{\sqrt{2}})\sqrt{(t/r)\lambda_r((r/t)W^*)}$. Finally we get the desired result: $\lambda_r((r/t)W^*) \leq 2\lambda_r((r/t)W^*)$ by re-arranging the terms. Using similar arguments we can show that $\lambda_r((r/t)W^*) \leq 2\lambda_r((r/t)W)$ as follows.

$$\sqrt{\lambda_r((r/t)W)} \leq \sqrt{(r/t)}\sigma_r(V) \leq \sqrt{(r/t)}\sigma_r(V^*Q^{-\top}) + \sqrt{(r/t)}\|V - V^*Q^{-\top}\| \tag{102}$$
$$\leq \|Q^{-\top}\|\sqrt{(r/t)}\sigma_r(V^*) + \sqrt{(r/t)}\|H\| \tag{103}$$
$$\leq (1+\frac{1}{10})\sqrt{\lambda_r((r/t)W^*)} + (1-\frac{1}{\sqrt{2}})\sqrt{\lambda_r((r/t)W^*)} \tag{104}$$
$$\leq \sqrt{2\lambda_r((r/t)W^*)} \tag{105}$$

Similarly, we derive a relation between $\lambda_r(W)$ and $\lambda_r(W^*)$. Notice that $V^*Q^{-\top} = V + V^*Q^{-\top} - V$. Then by Weyl's inequality (Lemma F.2, by setting $B \leftarrow V^*Q^{-\top}$, $A \leftarrow V$, and $C \leftarrow V^*Q^{-\top} - V$) we get that

$$\sqrt{\lambda_1((r/t)W)} \leq \sqrt{(r/t)}\sigma_1(V) \leq \sqrt{(r/t)}\sigma_1(V^*Q^{-\top}) + \sqrt{(r/t)}\|V - V^*Q^{-T}\| \tag{106}$$
$$\leq \|Q^{-1}\|\sqrt{\lambda_1((r/t)W^*)} + \sqrt{(r/t)}\|H\| \tag{107}$$
$$\leq (1+\frac{1}{10})\sqrt{\lambda_1((r/t)W^*)} + \sqrt{(r/t)}\|H\|_F \tag{108}$$
$$\leq (1+\frac{1}{10})\sqrt{\lambda_1((r/t)W^*)} + (1-\frac{1}{\sqrt{2}})\sqrt{\lambda_r((r/t)W^*)} \tag{109}$$
$$\leq \sqrt{2}\sqrt{\lambda_1((r/t)W^*)} \tag{110}$$

where the last second inequality uses $\|H\|_F \leq (1-\frac{1}{\sqrt{2}})\sqrt{(t/r)\lambda_r((r/t)W^*)}$, and the last inequality uses $\lambda_r(\cdot) \leq \lambda_1(\cdot)$. Finally we get the desired result by re-arranging the terms. ☐

B.3 Analysis of update on U

B.3.1 Proof of Lemma B.3

Proof of Lemma B.3. In this proof for brevity, we will first set that $\mathcal{T}_k \leftarrow [t]$, $|\mathcal{T}_k| = t/K \leftarrow t$, $S_2^{(i)} \leftarrow S^{(i)} = \frac{1}{m}\sum_{j\in[m]} x_j^{(i)}(x_j^{(i)})^\top$. This can be done due to the approximate equivalence of the subset $\mathcal{T}_k$ and the set of all task $[t]$ by Lemma B.1, which requires that $t \geq \Omega(\mu^2 r^3 K \log(\frac{K}{\delta}))$. Finally at the end of the analysis we will reset $\mathcal{T}_k \leftarrow \mathcal{T}_k$, $|\mathcal{T}_k| = t/K \leftarrow t/K$, $S_2^{(i)} \leftarrow S_2^{(i)} = \frac{2}{m}\sum_{j\in[m/2+1,m]} x_j^{(i)}(x_j^{(i)})^\top$.

Recall that

$$\widehat{U} - U^*Q = \mathcal{W}^{-\frac{1}{2}}(\mathcal{I}+\mathcal{E}_1)(-(\mathcal{W}^{-\frac{1}{2}}\mathcal{H}+\mathcal{E}_2)(U^*Q) + \mathcal{W}^{-\frac{1}{2}}(\sum_{i\in[t]} z^{(i)}(v^{(i)})^\top)) \tag{111}$$

where $\mathcal{E}_1 = (\mathcal{W}^{-\frac{1}{2}}\mathcal{A}\mathcal{W}^{-\frac{1}{2}})^\dagger - \mathcal{I}$ and $\mathcal{E}_2 = \mathcal{W}^{-\frac{1}{2}}\widehat{\mathcal{H}} - \mathcal{W}^{-\frac{1}{2}}\mathcal{H}$, and $F = \widehat{U} - U^*Q + \mathcal{W}^{-1}(\mathcal{H}(U^*Q))$. Therefore

$$\|F\|_F \leq \|\mathcal{W}^{-\frac{1}{2}}\|_F(\|\mathcal{E}_1\|_F\|\mathcal{W}^{-\frac{1}{2}}\mathcal{H}(U^*Q)\|_F + \|\mathcal{I} + \mathcal{E}_1\|_F(\|\mathcal{E}_2(U^*Q)\|_F + \|\mathcal{W}^{-\frac{1}{2}}(\sum_{i\in[t]} z^{(i)}(v^{(i)})^\top))\|_F)) \tag{112}$$

We can trivially bound $\|\mathcal{W}^{-\frac{1}{2}}\|_F$ as follows. For all $\|U\|_F = 1$, the following is true.

$$\|\mathcal{W}^{-\frac{1}{2}}(U)\|_F = \|UW^{-\frac{1}{2}}\|_F \leq \|U\|_F\|W^{-\frac{1}{2}}\| \leq \sqrt{\frac{r/t}{\lambda_r}} \tag{113}$$

$\Omega(\mu d r^2 \log(1/\delta)) \leq mt$ and approximate incoherence of intermediate V (26) implies that $\Omega(dr \frac{\|V\|_{\infty,2}^2}{\lambda_r(W)/t} \log(1/\delta)) \leq \Omega(\mu d r^2 \log(1/\delta)) \leq mt$, then by Lemma B.7 we have that, with a probability of at least $1 - \delta/3$

$$\|\mathcal{E}_1\|_F \leq 3c\sqrt{\frac{dr\,\|V\|_{\infty,2}^2 \log(27/\delta)}{m\,\lambda_r(W)}} \leq 3c\sqrt{\frac{\mu d r^2 \log(27/\delta)}{mt}} \leq \frac{1}{2} \tag{114}$$

This also implies that

$$\|\mathcal{I} + \mathcal{E}_1\|_F \leq \|\mathcal{I}\| + \|\mathcal{E}_1\|_F \leq 1 + \Delta \leq \frac{3}{2} \tag{115}$$

By Lemma B.8,

$$\|(\mathcal{W}^{-\frac{1}{2}}\mathcal{H})(U^*Q)\|_F \leq \|H\|_F \tag{116}$$

and with a probability of at least $1 - \delta/3$

$$\begin{aligned}\|\mathcal{E}_2(U^*Q)\|_F \leq c(\min(\|H\|_F \frac{\|V\|_{\infty,2}}{\sqrt{\lambda_r(W)}}, \|H\|_{\infty,2})\sqrt{\frac{dr\ \log(15/\delta)}{m}} +\\ \|H\|_{\infty,2}\frac{\|V\|_{\infty,2}}{\sqrt{\lambda_r(W)}}\frac{dr\ \log(15/\delta)}{m})\end{aligned} \tag{117}$$

Using the approximate incoherence of V (26) in the above inequality, we get that

$$\|\mathcal{E}_2(U^*Q)\|_F \leq c(\min(\|H\|_F\sqrt{\frac{\mu r}{t}}, \|H\|_{\infty,2})\sqrt{\frac{dr\ \log(15/\delta)}{m}} + \|H\|_{\infty,2}\sqrt{\frac{\mu r}{t}} \cdot \frac{dr\ \log(15/\delta)}{m}) \tag{118}$$

By Lemma B.9 with a probability of at least $1 - \delta/3$

$$\|\sum_{i\in[t]} \mathcal{W}^{-\frac{1}{2}}(z^{(i)}(v^{(i)})^\top))\|_F \leq O\Big(\sigma\sqrt{\frac{dr}{m}\log\Big(\frac{t}{\delta}\Big)\log\Big(\frac{r}{\delta}\Big)}\Big) \tag{119}$$

Finally taking union bound over the above results and using Lemma B.2, we can bound each of the terms constituting F. Using (113), (116) and (27) (recall that we set $t \leftarrow t/K$) we get

$$\|\mathcal{W}^{-1}\mathcal{H}(U^*Q)\|_F \leq \|\mathcal{W}^{-\frac{1}{2}}\|_F\|\mathcal{W}^{-\frac{1}{2}}\mathcal{H}(U^*Q)\|_F \tag{120}$$

$$\leq \sqrt{\frac{r}{t}}\frac{\|H\|_F}{\sqrt{\lambda_r}} \leq O\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}}\sqrt{\frac{\log(\frac{t}{\delta})}{\log(\frac{1}{\delta})}}\|(\mathbf{I} - U^*(U^*)^\top)U\|_F + \frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2\log(\frac{t}{\delta})}{m}}\Big) \tag{121}$$

Using (113), (115), (116), and (27) we get

$$\|\mathcal{W}^{-\frac{1}{2}}\|_F\|\mathcal{E}_1\|_F\|\mathcal{W}^{-\frac{1}{2}}\mathcal{H}(U^*Q)\|_F \tag{122}$$

$$\leq O\Big(\sqrt{\frac{\mu d r^2\log(\frac{1}{\delta})}{mt}}\sqrt{\frac{r}{t}}\frac{\|H\|_F}{\sqrt{\lambda_r}}\Big) \tag{123}$$

$$\leq O\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{\mu d r^2\log(\frac{t}{\delta})}{mt}}\|(\mathbf{I} - U^*(U^*)^\top)U\|_F + \sqrt{\frac{\mu d r^2\log(\frac{1}{\delta})}{mt}}\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2\log(\frac{t}{\delta})}{m}}\Big) \tag{124}$$

Using (113), (115), (118), (27) and (28) we get

$$\|\mathcal{W}^{-\frac{1}{2}}\|_F\|\mathcal{I}+\mathcal{E}_1\|_F\|\mathcal{E}_2(U^*Q)\|_F \tag{125}$$

$$\leq O\Big(\sqrt{\frac{r}{t}}\min\Big(\frac{\|H\|_F}{\sqrt{\lambda_r}}\sqrt{\frac{\mu r}{t}}, \frac{\|H\|_{\infty,2}}{\sqrt{\lambda_r}}\Big)\sqrt{\frac{dr\ \log(\frac{1}{\delta})}{m}}+\sqrt{\frac{r}{t}}\frac{\|H\|_{\infty,2}}{\sqrt{\lambda_r}}\sqrt{\frac{\mu r}{t}}\frac{dr\ \log(\frac{1}{\delta})}{m}\Big) \tag{126}$$

$$\leq O\Big(\min\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{\mu dr^2\log(\frac{t}{\delta})}{mt}}\|(\mathbf{I}-U^*(U^*)^\top)U\|_F+\sqrt{\frac{\mu dr^2\log(\frac{1}{\delta})}{mt}}\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2\log(\frac{t}{\delta})}{m}},$$
$$\sqrt{\frac{\mu dr^2\log(\frac{t}{\delta})}{mt}}\|(\mathbf{I}-U^*(U^*)^\top)U\|+\sqrt{\frac{dr\log(\frac{1}{\delta})}{m}}\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2\log(\frac{t}{\delta})}{mt}}\Big)+$$
$$\frac{\mu dr^2\log(\frac{t}{\delta})}{mt}\|(\mathbf{I}-U^*(U^*)^\top)U\|+\frac{\sqrt{\mu}dr\sqrt{r}\log(\frac{1}{\delta})}{m\sqrt{t}}\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2\log(\frac{t}{\delta})}{mt}}\Big) \tag{127}$$

$$\leq O\Big(\sqrt{\frac{\mu dr^2\log(\frac{t}{\delta})}{mt}}\|(\mathbf{I}-U^*(U^*)^\top)U\|+\sqrt{\frac{dr\log(\frac{1}{\delta})}{m}}\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2\log(\frac{t}{\delta})}{mt}}\Big)+ \tag{128}$$

$$\frac{\mu dr^2\log(\frac{t}{\delta})}{mt}\|(\mathbf{I}-U^*(U^*)^\top)U\|+\frac{\sqrt{\mu}dr\sqrt{r}\log(\frac{1}{\delta})}{m\sqrt{t}}\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2\log(\frac{t}{\delta})}{mt}}\Big) \tag{129}$$

Using (113), (115), (119), and (26) we get

$$\|\mathcal{W}^{-\frac{1}{2}}\|_F\|\mathcal{I}+\mathcal{E}_1\|_F\|\sum_{i\in[t]}\mathcal{W}^{-\frac{1}{2}}(z^{(i)}(v^{(i)})^\top))\|_F\leq O\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{dr^2\log(\frac{t}{\delta})\log(\frac{r}{\delta})}{mt}}\Big) \tag{130}$$

Substituting (121), (124), (129), and (130) in (112) we get

$$\|F\|_F\leq\|\mathcal{W}^{-\frac{1}{2}}\|_F(\|\mathcal{E}_1\|_F\|\mathcal{W}^{-\frac{1}{2}}\mathcal{H}(U^*Q)\|_F+\|\mathcal{I}+\mathcal{E}_1\|_F(\|\mathcal{E}_2(U^*Q)\|_F+\|\sum_{i\in[t]}\mathcal{W}^{-\frac{1}{2}}(z^{(i)}(v^{(i)})^\top)\|_F)) \tag{131}$$

$$\leq O\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{\mu dr^2\log(\frac{t}{\delta})}{mt}}\|(\mathbf{I}-U^*(U^*)^\top)U\|_F+\sqrt{\frac{\mu dr^2\log(\frac{1}{\delta})}{mt}}\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2\log(\frac{t}{\delta})}{m}}\Big)+ \tag{132}$$

$$O\Big(\frac{\mu dr^2\log(\frac{t}{\delta})}{mt}\|(\mathbf{I}-U^*(U^*)^\top)U\|+\frac{\sqrt{\mu}dr\sqrt{r}\log(\frac{1}{\delta})}{mt}\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2\log(\frac{t}{\delta})}{m}}\Big)+$$
$$O\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{dr^2\log(\frac{t}{\delta})\log(\frac{r}{\delta})}{mt}}\Big) \tag{133}$$

$$\leq O\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{\mu dr^2\log(\frac{t}{\delta})}{mt}}\|(\mathbf{I}-U^*(U^*)^\top)U\|_F+\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\mu dr^2\log(\frac{t}{\delta})\log(\frac{r}{\delta})}{mt}}\Big) \tag{134}$$

where the second-last inequality used the fact that $mt\geq\Omega(\mu dr^2\log(\frac{t}{\delta}))$. Finally, by resetting $\mathcal{T}_k\leftarrow\mathcal{T}_k$, $|\mathcal{T}_k|=t/K\leftarrow t/K$, $S_2^{(i)}\leftarrow S_2^{(i)}=\frac{2}{m}\sum_{j\in[m/2+1,m]}x_j^{(i)}(x_j^{(i)})^\top$, we obtain the desired result. □

B.3.2 Supporting lemmas for the analysis of update on U

Lemma B.7. *If* $\max(1,4c^2)dr\frac{\|V\|_{\infty,2}^2}{\lambda_r(W)/t}\log(27/\delta)\leq mt$*, then with a probability of at least* $1-\delta/3$*,*

$$\|\mathcal{E}_1\|_F\leq 3c\sqrt{\frac{dr\ \|V\|_{\infty,2}^2\log(27/\delta)}{m\,\lambda_r(W)}} \tag{135}$$

Proof of Lemma B.7. Let $\mathcal{S}_F = \{U \in \mathbb{R}^{d\times r} \mid \|U\|_F = 1\}$ be the set of all real matrices of dimensions $d \times r$ with unit Frobenius norm. For $\epsilon \le 1$, there exists an ϵ-net, $N_\epsilon \subset \mathcal{S}_F$, of size $(1 + 2/\epsilon)^{dr}$ with respect to the Frobenius norm [40, Lemma 5.2]. That is for any $U' \in \mathcal{S}_F$, there exists some $U \in N_\epsilon$ such that $\|U' - U\|_F \le \epsilon$.

Consider a $U \in N_\epsilon$, such that $\|U\|_F = 1$. Now we will prove with high-probability that $\left\langle(\mathcal{W}^{-\frac{1}{2}}\mathcal{A}\mathcal{W}^{-\frac{1}{2}} - \mathcal{I})(U), U\right\rangle$ is small. Consider the the following quadratic form

$$\left\langle(\mathcal{W}^{-\frac{1}{2}}\mathcal{A}\mathcal{W}^{-\frac{1}{2}})(U), U\right\rangle = \left\langle \sum_{i\in[t]} S^{(i)} U W^{-\frac{1}{2}} v^{(i)} (v^{(i)})^\top W^{-\frac{1}{2}}, U \right\rangle \tag{136}$$

$$= \sum_{i\in[t]} \frac{1}{m} \sum_{j\in[m]} (x_j^{(i)})^\top (U W^{-\frac{1}{2}} v^{(i)} (v^{(i)})^\top W^{-\frac{1}{2}} U^\top) x_j^{(i)} \tag{137}$$

where $S^{(i)} = \frac{1}{m}\sum_{j\in[m]} x_j^{(i)} (x_j^{(i)})^\top$ and $x_j^{(i)} \sim \mathcal{N}(0, \mathbf{I}_{d\times d})$ are i.i.d. standard Gaussian random vectors and $W = \sum_{i\in[t]} v^{(i)}(v^{(i)})^\top$ is rank-r matrix. We will use Hanson-Wright inequality (Lemma F.5) to prove that the above quadratic form concentrates around its mean. Notice that the the expectation of $\left\langle(\mathcal{W}^{-\frac{1}{2}}\mathcal{A}\mathcal{W}^{-\frac{1}{2}})(U), U\right\rangle$ is $\langle \mathcal{I}(U), U\rangle$.

$$\sum_{i\in[t]} \mathbb{E}\left[\left\langle S^{(i)} U W^{-\frac{1}{2}} v^{(i)} (v^{(i)})^\top W^{-\frac{1}{2}}, U \right\rangle\right] = \left\langle U W^{-\frac{1}{2}} \sum_{i\in[t]} v^{(i)} (v^{(i)})^\top W^{-\frac{1}{2}}, U \right\rangle \tag{138}$$

$$= \langle U, U\rangle = \|U\|_F^2 = 1\,. \tag{139}$$

We will also need the following bounds to apply the Hanson-Wright inequality. Recall that $\|V\|_{\infty,2} = \max_{i\in[t]} \|v^{(i)}\|$. Then,

$$\max_{i\in[t]} \|U W^{-\frac{1}{2}} v^{(i)} (v^{(i)})^\top W^{-\frac{1}{2}} U^\top\| = \max_{i\in[t]} \|U W^{-\frac{1}{2}} v^{(i)}\|^2 \le \max_{i\in[t]} \|U\|^2 \|W^{-1}\|^2 \|v^{(i)}\|^2 \tag{140}$$

$$\le \frac{\|V\|_{\infty,2}^2}{\lambda_r(W)} \tag{141}$$

Also note that,

$$\sum_{i\in[t]} \|U W^{-\frac{1}{2}} v^{(i)} (v^{(i)})^\top W^{-\frac{1}{2}} U^\top\|_F^2 = \sum_{i\in[t]} \|U W^{-\frac{1}{2}} v^{(i)}\|^4 \tag{142}$$

$$= \max_{i\in[t]} \|U W^{-\frac{1}{2}} v^{(i)}\|^2 \sum_{i\in[t]} \left\langle U W^{-\frac{1}{2}} v^{(i)}, U W^{-\frac{1}{2}} v^{(i)} \right\rangle \tag{143}$$

$$\le \frac{\|V\|_{\infty,2}^2}{\lambda_r(W)} \tag{144}$$

where the last inequality used (138) and (141). Then by Hanson-Wright inequality (Lemma F.5), with probability at least $1 - \delta/|N_\epsilon|$

$$\left|\left\langle(\mathcal{W}^{-\frac{1}{2}}\mathcal{A}\mathcal{W}^{-\frac{1}{2}} - \mathcal{I})(U), U\right\rangle\right| \tag{145}$$

$$= \left|\left\langle \sum_{i\in[t]} \frac{1}{m} \sum_{j\in[m]} x_j^{(i)} (x_j^{(i)})^\top U W^{-\frac{1}{2}} v^{(i)} (v^{(i)})^\top W^{-\frac{1}{2}}, U \right\rangle - \langle U, U\rangle\right| \le \Delta_\epsilon \tag{146}$$

where $\Delta_\epsilon = c \max\left(\sqrt{\frac{\|V\|_{\infty,2}^2 \log(|N_\epsilon|/\delta)}{m\,\lambda_r(W)}}, \frac{\|V\|_{\infty,2}^2 \log(|N_\epsilon|/\delta)}{m\,\lambda_r(W)}\right)$. Taking union bound over all $U \in N_\epsilon$ implies that with probability at least $1 - \delta$

$$\left|\left\langle(\mathcal{W}^{-\frac{1}{2}}\mathcal{A}\mathcal{W}^{-\frac{1}{2}} - \mathcal{I})(U), U\right\rangle\right| \le \Delta_\epsilon\ , \ \text{for all } U \in N_\epsilon\,. \tag{147}$$

For brevity, let $\mathcal{E}_1'(U) = (\mathcal{W}^{-\frac{1}{2}}\mathcal{A}\mathcal{W}^{-\frac{1}{2}} - \mathcal{I})(U)$. Notice that $\mathcal{E}_1'$ is self-adjoint, therefore it has an eigen decomposition with respect to the Frobenius norm. Then, let $U' \in \mathcal{S}_F \subset \mathbb{R}^{d\times r}$ be the largest "eigenmatrix" of $\mathcal{E}_1$, such that $\langle \mathcal{E}_1'(U), U\rangle = \|\mathcal{E}_1'\|_F = \max_{\|\widetilde{U}\|_F=1} \left\langle \mathcal{E}_1'(\widetilde{U}), \widetilde{U}\right\rangle = \max_{\|\widetilde{U}\|_F=\|\widetilde{U}'\|_F=1} \left\langle \mathcal{E}_1'(\widetilde{U}), \widetilde{U}'\right\rangle$. Then there exists some $U \in N_\epsilon$ such that $\|U' - U\|_F \leq \epsilon$.

$$\begin{aligned}\|\mathcal{E}_1'\|_F = \langle \mathcal{E}_1'(U'), U'\rangle &= \langle \mathcal{E}_1'(U), U\rangle + \langle \mathcal{E}_1'(U'-U), U\rangle + \langle \mathcal{E}_1'(U'), U'-U\rangle && (148)\\ &\leq \langle \mathcal{E}_1'(U), U\rangle + \|\mathcal{E}_1'\|_F \|U'-U\|_F(\|U\|_F + \|U'\|_F) && (149)\\ &\leq \langle \mathcal{E}_1'(U), U\rangle + 2\epsilon\|\mathcal{E}_1'\|_F && (150)\end{aligned}$$

Re-arranging and setting $\epsilon = 1/4$, and $c \leftarrow 2c$, we get

$$\|\mathcal{W}^{-\frac{1}{2}}\mathcal{A}\mathcal{W}^{-\frac{1}{2}} - \mathcal{I}\|_F = \|\mathcal{E}_1'\|_F \leq \Delta_{\frac{1}{4}} = \Delta. \tag{151}$$

where $\Delta = c\max\left(\sqrt{\frac{dr\,\|V\|_{\infty,2}^2\log(9/\delta)}{m\,\lambda_r(W)}}, \frac{dr\,\|V\|_{\infty,2}^2\log(9/\delta)}{m\,\lambda_r(W)}\right)$.

For brevity, let $\widehat{\mathcal{A}}(U) = (\mathcal{W}^{-\frac{1}{2}}\mathcal{A}\mathcal{W}^{-\frac{1}{2}})(U)$. Notice that $\widehat{\mathcal{A}}$ is self-adjoint, therefore it has an eigen decomposition with respect to the Frobenius norm. Then, let $U' \in \mathcal{S}_F \subset \mathbb{R}^{d\times r}$ be the smallest "eigenmatrix" of $\widehat{\mathcal{A}}$, such that $\left\langle \widehat{\mathcal{A}}(U), U\right\rangle = \lambda_{\min}(\widehat{\mathcal{A}}) = \min_{\|\widetilde{U}\|_F=1}\left\langle \widehat{\mathcal{A}}(\widetilde{U}), \widetilde{U}\right\rangle = \min_{\|\widetilde{U}\|_F=\|\widetilde{U}'\|_F=1}\left\langle \widehat{\mathcal{A}}(\widetilde{U}), \widetilde{U}'\right\rangle$. Then there exists some $U \in N_\epsilon$ such that $\|U' - U\|_F \leq \epsilon$.

$$\begin{aligned}\lambda_{\min}(\widehat{\mathcal{A}}) = \left\langle \widehat{\mathcal{A}}(U'), U'\right\rangle &= \langle \mathcal{I}(U), U\rangle + \left\langle (\widehat{\mathcal{A}} - \mathcal{I})(U), U\right\rangle + \left\langle \widehat{\mathcal{A}}(U'-U), U\right\rangle + \left\langle \widehat{\mathcal{A}}(U'), U'-U\right\rangle && (152)\\ &\geq 1 - \left|\left\langle (\widehat{\mathcal{A}} - \mathcal{I})(U), U\right\rangle\right| - \lambda_{\min}(\widehat{\mathcal{A}})\|U'-U\|_F(\|U\|_F + \|U'\|_F) && (153)\\ &\geq 1 - \Delta_\epsilon - 2\epsilon\lambda_{\min}(\widehat{\mathcal{A}}) && (154)\end{aligned}$$

Re-arranging and setting $\epsilon = 1/4$, and $c \leftarrow 2c$, we get that $\lambda_{\min}(\widehat{\mathcal{A}}) \geq \frac{2}{3}(1-\Delta)$. Therefore,

$$\|(\mathcal{W}^{-\frac{1}{2}}\mathcal{A}\mathcal{W}^{-\frac{1}{2}})^\dagger\|_F = \frac{1}{\lambda_{\min}(\widehat{\mathcal{A}})} \leq \frac{3}{2(1-\Delta)}. \tag{155}$$

where $\Delta = c\max\left(\sqrt{\frac{dr\,\|V\|_{\infty,2}^2\log(9/\delta)}{m\,\lambda_r(W)}}, \frac{dr\,\|V\|_{\infty,2}^2\log(9/\delta)}{m\,\lambda_r(W)}\right)$. If $\max(1, 4c^2)dr\frac{\|V\|_{\infty,2}^2}{\lambda_r(W)/t}\log(27/\delta) \leq mt$, we get that $\Delta \leq c\sqrt{\frac{dr\,\|V\|_{\infty,2}^2\log(9/\delta)}{m\,\lambda_r(W)}} \leq \frac{1}{2}$.

By setting $A + B = \mathcal{W}^{-\frac{1}{2}}\mathcal{A}\mathcal{W}^{-\frac{1}{2}}$ and $A = \mathcal{I}$ such that $\mathcal{E}_1 = (A+B)^{-1} - B^{-1}$, in the Woodburry matrix inverse identity (359) (Lemma F.3) we get that, with a probability of at least $1-\delta$

$$\begin{aligned}&\|(A+B)^{-1} - A^{-1}\|_F \leq \|A^{-1}\|_F\|B\|_F\|(A+B)^{-1}\|_F && (156)\\ \implies &\|\mathcal{E}_1\|_F \leq \left\|(\mathcal{W}^{-\frac{1}{2}}\mathcal{A}\mathcal{W}^{-\frac{1}{2}})^\dagger - \mathcal{I}\right\|_F \leq \|\mathcal{I}^\dagger\|_F\|\mathcal{W}^{-\frac{1}{2}}\mathcal{A}\mathcal{W}^{-\frac{1}{2}} - \mathcal{I}\|_F\|(\mathcal{W}^{-\frac{1}{2}}\mathcal{A}\mathcal{W}^{-\frac{1}{2}})^\dagger\|_F && (157)\\ &\leq 1\cdot\Delta\cdot\frac{3}{2(1-\Delta)} \leq 3\Delta \leq 3c\sqrt{\frac{dr\,\|V\|_{\infty,2}^2\log(9/\delta)}{m\,\lambda_r(W)}} && (158)\end{aligned}$$

Finally, setting $\delta \leftarrow \delta/3$ get us the desired result. □

Lemma B.8. *$\|(\mathcal{W}^{-\frac{1}{2}}\mathcal{H})(U^*Q)\|_F \leq \|H\|_F$ and with a probability of at least $1 - \delta/3$*

$$\|\mathcal{E}_2(U^*Q)\|_F \leq c(\min(\|H\|_F\frac{\|V\|_{\infty,2}}{\sqrt{\lambda_r(W)}}, \|H\|_{\infty,2})\sqrt{\frac{dr\,\log(15/\delta)}{m}} + \|H\|_{\infty,2}\frac{\|V\|_{\infty,2}}{\sqrt{\lambda_r(W)}}\frac{dr\,\log(15/\delta)}{m}) \tag{159}$$

Proof of Lemma B.8. First we prove that the expected value $\mathbb{E}[(\mathcal{W}^{-\frac{1}{2}}\widehat{\mathcal{H}})(U^*Q)] = (\mathcal{W}^{-\frac{1}{2}}\mathcal{H})(U^*Q)$ is bounded.

$$\|(\mathcal{W}^{-\frac{1}{2}}\mathcal{H})(U^*Q)\|_F = \max_{\|U\|_F=1} \left\langle (\mathcal{W}^{-\frac{1}{2}}\mathcal{H})(U^*Q), U \right\rangle \tag{160}$$

$$= \max_{\|U\|_F=1} \sum_{i\in[t]} \left\langle U^*Qh^{(i)}(v^{(i)})^\top W^{-\frac{1}{2}}, U \right\rangle \tag{161}$$

$$= \max_{\|U\|_F=1} \sum_{i\in[t]} \left\langle U^*Qh^{(i)}, UW^{-\frac{1}{2}}v^{(i)} \right\rangle \tag{162}$$

$$\leq \max_{\|U\|_F=1} \sqrt{\sum_{i\in[t]} \|U^*Qh^{(i)}\|^2} \sqrt{\sum_{i\in[t]} \left\langle UW^{-\frac{1}{2}}v^{(i)}, UW^{-\frac{1}{2}}v^{(i)} \right\rangle} \tag{163}$$

$$\leq \max_{\|U\|_F=1} \|Q\| \sqrt{\sum_{i\in[t]} \|h^{(i)}\|^2} \sqrt{\left\langle U \sum_{i\in[t]} W^{-\frac{1}{2}}v^{(i)}(v^{(i)})^\top W^{-\frac{1}{2}}, U \right\rangle} \tag{164}$$

$$\leq \max_{\|U\|_F=1} \|H\|_F \|U\|_F = \|H\|_F \tag{165}$$

where used the fact that $\langle AB, C\rangle = \langle A, CB^\top \rangle$ and $(U^*)^\top U^* = \mathbf{I}$.

Let $\mathcal{S}_F = \{U \in \mathbb{R}^{d\times r} \,|\, \|U\|_F = 1\}$ be the set of all real matrices of dimensions $d \times r$ with unit Frobenius norm. For $\epsilon \leq 1$, there exists an ϵ-net, $N_\epsilon \subset \mathcal{S}_F$, of size $(1 + 2/\epsilon)^{dr}$ with respect to the Frobenius norm [40, Lemma 5.2]. That is for any $U' \in \mathcal{S}_F$, there exists some $U \in N_\epsilon$ such that $\|U' - U\|_F \leq \epsilon$.

Consider a $U \in N_\epsilon$, such that $\|U\|_F = 1$. Now we will prove with high-probability that $\left\langle (\mathcal{W}^{-\frac{1}{2}}\mathcal{H})(U^*Q)(U) - \mathcal{W}^{-\frac{1}{2}}(\sum_{i\in[t]} S^{(i)}U^*Qh^{(i)}(v^{(i)})^\top), U \right\rangle$ is small. Consider the the following quadratic form

$$\left\langle \mathcal{W}^{-\frac{1}{2}}(\sum_{i\in[t]} S^{(i)}U^*Qh^{(i)}(v^{(i)})^\top), U \right\rangle = \left\langle \sum_{i\in[t]} S^{(i)}U^*Qh^{(i)}(v^{(i)})^\top W^{-\frac{1}{2}}, U \right\rangle \tag{166}$$

$$= \sum_{i\in[t]} \frac{1}{m} \sum_{j\in[m]} (x_j^{(i)})^\top (U^*Qh^{(i)}(v^{(i)})^\top W^{-\frac{1}{2}}U^\top) x_j^{(i)} \tag{167}$$

where $S^{(i)} = \frac{1}{m}\sum_{j\in[m]} x_j^{(i)}(x_j^{(i)})^\top$ and $x_j^{(i)} \sim \mathcal{N}(0, \mathbf{I}_{d\times d})$ are i.i.d. standard Gaussian random vectors and $W = \sum_{i\in[t]} v^{(i)}(v^{(i)})^\top$ is rank-r matrix. We will use Hanson-Wright inequality (Lemma F.5) to prove that the above quadratic form concentrates around its mean. Notice that the the expectation of $\left\langle \mathcal{W}^{-\frac{1}{2}}(\sum_{i\in[t]} S^{(i)}U^*Qh^{(i)}(v^{(i)})^\top), U \right\rangle$ is $\left\langle W^{-\frac{1}{2}}\mathcal{H}(U), U \right\rangle$.

$$\mathbb{E}[\mathcal{W}^{-\frac{1}{2}}(\sum_{i\in[t]} S^{(i)}U^*Qh^{(i)}(v^{(i)})^\top)] = \mathcal{W}^{-\frac{1}{2}}(\sum_{i\in[t]} U^*Qh^{(i)}(v^{(i)})^\top) = (\mathcal{W}^{-\frac{1}{2}}\mathcal{H})(U^*Q)\,. \tag{168}$$

We will also need the following bounds to apply the Hanson-Wright inequality. Recall that $\|H\|_{\infty,2} = \max_{i\in[t]} \|h^{(i)}\|$ and $\|V\|_{\infty,2} = \max_{i\in[t]} \|v^{(i)}\|$. Then,

$$\max_{i\in[t]} \|U^*Qh^{(i)}(v^{(i)})^\top W^{-\frac{1}{2}}U^\top\| \leq \max_{i\in[t]} \|U^*\|\|Q\|\|h^{(i)}\| \max_{i\in[t]} \frac{\|v^{(i)}\|}{\sqrt{\lambda_r(W)}}\|U\| \leq \|H\|_{\infty,2}\frac{\|V\|_{\infty,2}}{\sqrt{\lambda_r(W)}} \tag{169}$$

Also note that

$$\sum_{i\in[t]} \|U^*Qh^{(i)}(v^{(i)})^\top W^{-\frac{1}{2}}U^\top\|_F^2 = \sum_{i\in[t]} \|U^*Qh^{(i)}\|^2 \|UW^{-\frac{1}{2}}v^{(i)}\|^2 \tag{170}$$

$$\le (\sum_{i\in[t]} \|U^*Qh^{(i)}\|^2)(\max_{i\in[t]} \|UW^{-\frac{1}{2}}v^{(i)}\|^2) \tag{171}$$

$$\le (\|Q\|^2 \sum_{i\in[t]} \|h^{(i)}\|^2)(\max_{i\in[t]} \|U\|^2 \|W^{-\frac{1}{2}}\|^2 \|v^{(i)}\|^2) \tag{172}$$

$$\le \|H\|_F^2 \frac{\|V\|_{\infty,2}^2}{\lambda_r(W)} \tag{173}$$

and

$$\sum_{i\in[t]} \|U^*Qh^{(i)}(v^{(i)})^\top W^{-\frac{1}{2}}U^\top\|_F^2 = \sum_{i\in[t]} \|U^*Qh^{(i)}\|^2 \|UW^{-\frac{1}{2}}v^{(i)}\|^2 \tag{174}$$

$$\le (\max_{i\in[t]} \|U^*Qh^{(i)}\|^2) \mathrm{tr}(UW^{-\frac{1}{2}} \sum_{i\in[t]} v^{(i)}(v^{(i)})^\top W^{-\frac{1}{2}}U^\top) \tag{175}$$

$$\le \|Q\| \max_{i\in[t]} \|h^{(i)}\|^2 \|U\|_F^2 \tag{176}$$

$$= \|H\|_{\infty,2}^2 \,. \tag{177}$$

Therefore, $\sum_{i\in[t]} \|U^*Qh^{(i)}(v^{(i)})^\top W^{-\frac{1}{2}}U^\top\|_F^2 \le \min\{\|H\|_F^2 \frac{\|V\|_{\infty,2}^2}{\lambda_r(W)}, \|H\|_{\infty,2}^2\}$. For brevity, let $\mathcal{E}_2(U) =$
$\mathcal{W}^{-\frac{1}{2}}(\sum_{i\in[t]} S^{(i)}Uh^{(i)}(v^{(i)})^\top) - (\mathcal{W}^{-\frac{1}{2}}\mathcal{H})(U)$. Then by Hanson-Wright inequality (Lemma F.5), with probability at least $1-\delta/|N_\epsilon|$

$$\left|\left\langle \mathcal{E}_2(U^*Q), U\right\rangle\right| \tag{178}$$

$$= \left|\left\langle \sum_{i\in[t]} \frac{1}{m} \sum_{j\in[m]} x_j^{(i)}(x_j^{(i)})^\top U^*Qh^{(i)}(v^{(i)})^\top W^{-\frac{1}{2}}, U\right\rangle - \left\langle (\mathcal{W}^{-\frac{1}{2}}\mathcal{H})(U^*Q), U\right\rangle\right| \le \Delta_\epsilon \tag{179}$$

where $\Delta_\epsilon = c(\min(\|H\|_F \frac{\|V\|_{\infty,2}}{\sqrt{\lambda_r(W)}}, \|H\|_{\infty,2})\sqrt{\frac{\log(|N_\epsilon|/\delta)}{m}} + \|H\|_{\infty,2} \frac{\|V\|_{\infty,2}}{\sqrt{\lambda_r(W)}} \frac{\log(|N_\epsilon|/\delta)}{m})$. Taking union bound over all $U \in N_\epsilon$ implies that with probability at least $1-\delta$

$$\left|\left\langle \mathcal{E}_2(U), U\right\rangle\right| \le \Delta_\epsilon \;,\; \text{ for all } U \in N_\epsilon \,. \tag{180}$$

Let $U' \in \mathcal{S}_F \subset \mathbb{R}^{d\times r}$ be the matrix "parallel" to $\mathcal{E}_2(U^*Q)$, that is $\|\mathcal{E}_2(U^*Q)\|_F = \max_{\|\widetilde{U}\|_F=1} \left\langle \mathcal{E}_1(U^*Q), \widetilde{U}\right\rangle = \langle \mathcal{E}_2(U^*Q), U'\rangle$. Then there exists some $U \in N_\epsilon$ such that $\|U'-U\|_F \le \epsilon$.

$$\|\mathcal{E}_2(U^*Q)\|_F = \langle \mathcal{E}_2(U^*Q), U'\rangle = \langle \mathcal{E}_2(U^*Q), U\rangle + \langle \mathcal{E}_2(U^*Q), U'-U\rangle \tag{181}$$

$$\le \langle \mathcal{E}_1(U), U\rangle + \|\mathcal{E}_2(U^*Q)\|_F \|U'-U\|_F \tag{182}$$

$$\le \langle \mathcal{E}_1(U), U\rangle + \epsilon \|\mathcal{E}_2(U^*Q)\|_F \tag{183}$$

Re-arranging and setting $\epsilon = 1/2$, and $c \leftarrow 2c$, we get

$$\|\mathcal{E}_2(U^*Q)\|_F \le \Delta_{\frac{1}{2}} = c(\min(\|H\|_F \frac{\|V\|_{\infty,2}}{\sqrt{\lambda_r(W)}}, \|H\|_{\infty,2})\sqrt{\frac{dr\,\log(5/\delta)}{m}} + \|H\|_{\infty,2} \frac{\|V\|_{\infty,2}}{\sqrt{\lambda_r(W)}} \frac{dr\,\log(5/\delta)}{m}) \tag{184}$$

Finally setting $\delta \leftarrow \delta/3$ get us the desired result.

□

Lemma B.9. *With a probability of at least* $1 - \delta/3$

$$\| \sum_{i \in [t]} \mathcal{W}^{-\frac{1}{2}}(z^{(i)}(v^{(i)})^\top))\|_F \le O\Big(\sigma\sqrt{\frac{dr}{m}\log\Big(\frac{t}{\delta}\Big)\log\Big(\frac{r}{\delta}\Big)}\Big) \tag{185}$$

Proof of Lemma B.9. Notice that $z^{(i)}$ (defined in Appendix B) is a Gaussian random vector of the following form

$$z^{(i)} = \frac{1}{m}\sum_{j \in [m]} \varepsilon_j^{(i)} x_j^{(i)} = \frac{1}{m}\|\varepsilon^{(i)}\| g^{(i)}, g^{(i)} \sim \mathcal{N}(0, \mathbf{I}_{d \times d}) \tag{186}$$

Using Hanson-Wright inequality (Lemma F.5, by setting $m \leftarrow 1$, $x_1 \leftarrow \varepsilon^{(i)}$, and $A_1 \leftarrow \mathbf{I}_{m \times m}$) and taking union bound over all tasks, we get that, with probability of at least $1 - \frac{\delta}{2}$

$$\|\varepsilon^{(i)}\|^2 \le \sigma^2 m(1 + c\sqrt{\frac{\log(\frac{2t}{\delta})}{m}} + c\frac{\log(\frac{2t}{\delta})}{m}) \le 2c\,\sigma^2 m \log\Big(\frac{2t}{\delta}\Big), \text{ for all } i \in [t] \tag{187}$$

where used the fact that $m \ge 1$ and $\log\Big(\frac{2t}{\delta}\Big) \ge 1$.

Let $\widehat{v}^{(i)} = W^{-\frac{1}{2}} v^{(i)}$, then

$$\sum_{i \in [t]} \|\widehat{v}^{(i)}\|^2 = \sum_{i \in [t]} \operatorname{tr}((v^{(i)})^\top W^{-1} v^{(i)}) = \sum_{i \in [t]} \operatorname{tr}(W^{-1} v^{(i)}(v^{(i)})^\top) = r \tag{188}$$

Notice that $\sum_{i \in [t]} \frac{1}{m}\|\varepsilon^{(i)}\| g^{(i)} \widehat{v}_j^{(i)}$ is a Gaussian random vector of the following form

$$\sum_{i \in [t]} \frac{1}{m}\|\varepsilon^{(i)}\| g^{(i)} \widehat{v}_j^{(i)} = \frac{1}{m}\sqrt{\sum_{i \in [t]} \|\varepsilon^{(i)}\|^2 (\widehat{v}_j^{(i)})^2}\; \widehat{g}_j\,, \widehat{g}_j \sim \mathcal{N}(0, \mathbf{I}_{d \times d}) \tag{189}$$

Using Hanson-Wright inequality (Lemma F.5, by setting $m \leftarrow 1$, $x_1 \leftarrow \widehat{g}_j$, and $A_1 \leftarrow \mathbf{I}_{d \times d}$) and taking union bound over all $j \in [r]$, we get that, with probability of at least $1 - \frac{\delta}{2}$

$$\|\widehat{g}_j\|^2 \le d(1 + c\sqrt{\frac{\log(\frac{2r}{\delta})}{d}} + c\frac{\log(\frac{2r}{\delta})}{d}) \le 2cd\log\Big(\frac{2r}{\delta}\Big), \text{ for all } j \in [r] \tag{190}$$

where used the fact that $d \ge 1$ and $\log\Big(\frac{2r}{\delta}\Big) \ge 1$.

Combining the above results and using union bound, we get that, with a probability of at least $1 - \delta$,

$$\begin{aligned}
\Big\|\sum_{i \in [t]} \mathcal{W}^{-\frac{1}{2}}(z^{(i)}(v^{(i)})^\top))\Big\|_F^2 &= \Big\|\sum_{i \in [t]} z^{(i)}(v^{(i)})^\top W^{-\frac{1}{2}}\Big\|_F^2 && (191)\\
&= \Big\|\sum_{i \in [t]} \frac{1}{m}\|\varepsilon^{(i)}\| g^{(i)}(\widehat{v}^{(i)})^\top\Big\|_F^2 && (192)\\
&= \sum_{j \in [r]} \Big\|\sum_{i \in [t]} \frac{1}{m}\|\varepsilon^{(i)}\| g^{(i)} \widehat{v}_j^{(i)}\Big\|^2 && (193)\\
&\le \sum_{j \in [r]}\sum_{i \in [t]} \frac{\|\varepsilon^{(i)}\|^2}{m^2}(\widehat{v}_j^{(i)})^2 \|\widehat{g}_j\|^2 && (194)\\
&\le \sum_{j \in [r]}\sum_{i \in [t]} O\Big(\frac{m\sigma^2}{m^2}\log\Big(\frac{t}{\delta}\Big)\Big)(\widehat{v}_j^{(i)})^2 O\Big(d\log\Big(\frac{r}{\delta}\Big)\Big) && (195)\\
&\le O\Big(\frac{d\sigma^2}{m}\log\Big(\frac{t}{\delta}\Big)\log\Big(\frac{r}{\delta}\Big)\Big)\sum_{i \in [t]} \|\widehat{v}^{(i)}\|^2 && (196)\\
&\le O\Big(\frac{\sigma^2 dr}{m}\log\Big(\frac{t}{\delta}\Big)\log\Big(\frac{r}{\delta}\Big)\Big). && (197)
\end{aligned}$$

Finally, we get the desired result by setting $\delta \leftarrow \delta/3$. □

B.4 Analysis of QR decomposition

Proof of Lemma B.4.

$$\sigma_{\min}(R) \geq \min_{\|z\|=1} \|Rz\| = \min_{\|z\|=1} \|U^{+}Rz\| = \min_{\|z\|=1} \|\widehat{U}z\| \tag{198}$$

$$\geq \min_{\|z\|=1} \|(U^{*}Q - \mathcal{W}^{\dagger}\mathcal{H}(U^{*}Q) + F)z\| \tag{199}$$

$$\geq \min_{\|z\|=1} \sqrt{z^{\top}Q^{\top}Qz} - \|\mathcal{W}^{\dagger}\mathcal{H}(U^{*}Q)\| - \|F\| \tag{200}$$

$$\geq \min_{\|z\|=1} \sigma_{\min}(Q) - \|\mathcal{W}^{\dagger}\mathcal{H}(U^{*}Q)\| - \|F\| \tag{201}$$

$$\geq \frac{1}{2} - \frac{1}{8} - \frac{1}{8} \geq \frac{1}{4} \tag{202}$$

There fore R is invertible and $\|R^{-1}\| = (\sigma_{\min}(R))^{-1} \leq 4$ □

B.5 Analysis of shuffling and partitioning

Proof of Lemma B.1. We will assume that the set of tasks $[t]$ is shuffled. We will prove that incoherence holds for the all subset $\mathcal{T}_k = [1 + \frac{t(k-1)}{K}, \frac{tk}{K}]$ of size t/K. Shuffling and K-way partitioning to get $\mathcal{T}_k$ is equivalent to uniformly sampling without replacement t/K elements from $[t]$. We prove that incoherrence holds for the first subset $\mathcal{T}_1$, then this is equivalent to proving that incoherence holds for the k-th partition $\mathcal{T}_k$ by symmetry. Let the tasks sampled for $\mathcal{T}_1$ without replacement be $\{i_l\}_{l=1}^{t/k}$, where i_l is the l-th sample.

Let $\mathcal{S}_F = \{z \in \mathbb{R}^r \mid \|z\| = 1\}$ be the set of all real vectors of dimensions r with unit Euclidean norm. For $\epsilon \leq 1$, there exists an ϵ-net, $N_\epsilon \subset \mathcal{S}_F$, of size $(1 + 2/\epsilon)^r$ with respect to the Euclidean norm [40, Lemma 5.2]. That is for any $z' \in \mathcal{S}_F$, there exists some $z \in N_\epsilon$ such that $\|z' - z\| \leq \epsilon$.

Consider a $z \in N_\epsilon$, such that $\|z\| = 1$. Now we will prove with high-probability that $z^{\top}(\sum_{l=1}^{t/K} v^{*(i_l)}(v^{*(i_l)})^{\top})z$ is approximately equal to $z^{\top}\mathbb{E}[\sum_{l=1}^{t/K} v^{*(i_l)}(v^{*(i_l)})^{\top}]z$. Now consider the martingale X_l, such that $X_0 = 0$ and $X_l = X_{l-1} + z^{\top}(v^{*(i_l)}(v^{*(i_l)})^{\top} - \mathbb{E}[v^{*(i_l)}(v^{*(i_l)})^{\top}|X_0, \ldots, X_{l-1}])z$, for all $l \in [t/K]$. Clearly this is a martginagle as $\mathbb{E}[X_l|X_0, \ldots, X_{l-1}] = 0$, for all $l \in [t/K]$. The maximum difference two consecutive steps is $\max_l |X_l - X_{l-1}| \leq 2\|v^{*(i_l)}\|^2 \leq 2\|V^{*}\|_{\infty,2}^2$. Therefore by Azuma-Hoeffding martingale inequality,

$$|\sum_{l=1}^{t/K} z^{\top}v^{*(i_l)}(v^{*(i_l)})^{\top}z - z^{\top}\mathbb{E}[\sum_{l=1}^{t/K} v^{*(i_l)}(v^{*(i_l)})^{\top}]z| = |X_{t/K}| \leq \sqrt{\frac{2t}{K}\|V\|_{\infty,2}^4 \log(\frac{2|N_\epsilon|}{\delta})} \tag{203}$$

with a probability of at least $1 - \delta/|N_\epsilon|$.

For brevity, let $E = \sum_{l=1}^{t/K} v^{*(i_l)}(v^{*(i_l)})^{\top} - \mathbb{E}[\sum_{l=1}^{t/K} v^{*(i_l)}(v^{*(i_l)})^{\top}]$. Notice that E is a real symmetric matrix, therefore it has an eigen decomposition. Then, let $v' \in \mathcal{S} \subset \mathbb{R}^r$ be the largest "eigenvector" of E, such that $(v')^{\top}Ev' = \|E\| = \max_{\|\widetilde{v}\|=1} \widetilde{v}^{\top}E\widetilde{v} = \max_{\|\widetilde{v}\|=\|\widetilde{v}'\|_F=1} \widetilde{v}^{\top}E\widetilde{v}'$. Then there exists some $v \in N_\epsilon$ such that $\|v' - v\| \leq \epsilon$.

$$\|E\|_F = (v')^{\top}Ev = v^{\top}Ev + (v' - v)^{\top}Ev + (v')^{\top}E(v' - v) \tag{204}$$

$$\leq v^{\top}Ev + \|v' - v\|\|E\|\|v\| + \|v'\|\|E\|\|v' - v\| \tag{205}$$

$$\leq v^{\top}Ev + 2\epsilon\|E\| \tag{206}$$

Re-arranging and setting $\epsilon = 1/4$, and $c \leftarrow 2c$, we get

$$\|\sum_{l=1}^{t/K} v^{*(i_l)}(v^{*(i_l)})^{\top} - \mathbb{E}[\sum_{l=1}^{t/K} v^{*(i_l)}(v^{*(i_l)})^{\top}]\| = \|E\| \leq \sqrt{\frac{2tr}{K}\|V\|_{\infty,2}^4 \log(\frac{18}{\delta})} \tag{207}$$

$$\leq \frac{1}{2}\lambda_r(\mathbb{E}[\sum_{l=1}^{t/K} v^{*(i_l)}(v^{*(i_l)})^{\top}]). \tag{208}$$

with probability at least $1-\delta/k$, where the last inequality used the fact that $t \geq \Omega(\mu^2 r^3 K \log(1/\delta))$. Additionally note that $\mathbb{E}[\sum_{l=1}^{t/k} v^{*(i_l)}(v^{*(i_l)})^\top] = \frac{1}{K}\sum_{i=1}^{t} v^{*(i)}(v^{*(i)})^\top = \frac{1}{K}(V^*)^\top V^*$, Therefore

$$\lambda_{r'}(\sum_{i\in\mathcal{T}_k} v^{*(i)}(v^{*(i)})^\top) = \frac{1}{K}\Theta(\lambda_{r'}((V^*)^\top V^*)) \text{ for all } r' \in [r] \tag{209}$$

where $\lambda_i(\cdot)$ is the r'-th largest eigenvalue matrix operator.

□

C Analysis of AltMinGD (Algorithm 1)

Initialized at U, the k-the step of alternating minimization-based AltMin (Algorithm 1) is:

$$v^{(i)} \leftarrow (U^\top S_1^{(i)} U)^\dagger((U^\top S_1^{(i)} U^*)v^{*(i)} + U^\top z^{(i)}) \;, \text{ for } i \in \mathcal{T}_k = [1+\frac{(k-1)t}{K}, \frac{tk}{K}] \tag{210}$$

$$\widetilde{U} \leftarrow U - \eta\Big(\sum_{i\in[t]} S_2^{(i)}(Uv^{(i)} - U^* v^{*(i)})(v^{(i)})^\top + z^{(i)}(v^{(i)})^\top\Big) \;, \tag{211}$$

$$U^+ \leftarrow \mathrm{QR}(\widetilde{U}) \;, \tag{212}$$

where U^+ is the next iterate, and $S_1^{(i)} = \frac{2}{m}\sum_{j\in[1,m/2]} x_j^{(i)}(x_j^{(i)})^\top$, $S_2^{(i)} = \frac{2}{m}\sum_{j\in[1+m/2,m]} x_j^{(i)}(x_j^{(i)})^\top$, and $z^{(i)} \triangleq (1/m)\sum_{j\in[m]} \varepsilon_j^{(i)} x_j^{(i)}$ and $\mathcal{A} : \mathbb{R}^{d\times r} \to \mathbb{R}^{d\times r}$ is a self-adjoint linear operator such that $\mathcal{A}(U) = \sum_{i\in T} S^{(i)} U v^{(i)}(v^{(i)})^\top$. The self-adjointness of $\mathcal{A}$ follows from the symmetry of $S^{(i)}$ when using cyclic property of trace as follows

$$\begin{aligned}\langle U_2, \mathcal{A}(U_1)\rangle = \sum_{i\in T}\Big\langle U_2, S^{(i)}U_1 v^{(i)}(v^{(i)})^\top)\Big\rangle &= \sum_{i\in T}\mathrm{tr}(U_2^\top S^{(i)} U_1 v^{(i)}(v^{(i)})^\top) \\ &= \sum_{i\in T}\mathrm{tr}(v^{(i)}(v^{(i)})^\top U_2^\top S^{(i)} U_1) = \langle \mathcal{A}(U_2), U_1\rangle\end{aligned} \tag{213}$$

QR-decomposition after every update is required to ensure that magnitude of U and V does not stray far away from that of true U^* and V^*, respectively. Otherwise, the sample complexity requirements of our algorithm increase in the condition number factors.

Incoherence. $\max_i \|v^{*(i)}\|^2 \leq (\mu\, r/t)\lambda_r(\sum_{i\in[t]} v^{*(i)}(v^{*(i)})^\top)$, and we define $\nu = (1/t)\lambda_r(\sum_{i\in[t]} v^{*(i)}(v^{*(i)})^\top)$. Notice that, this non-standard definition of incoherence is related to the standard definition: $W^* = (V^*)^\top V^* = \sum_{i\in[t]} v^{*(i)}(v^{*(i)})^\top$, $V^* = \tilde{V}^* R^*$ (QR-decomposition), $\max_i \|\widetilde{v}^{*(i)}\|^2 \leq \widetilde{\mu}\, r/t$, as follows $\mu = \widehat{\mu}(\sigma_1^2(R^*)/\sigma_r^2(R^*))$.

Theorem 9. *Let there be t linear regression tasks, each with m samples satisfying Assumptions 1 and 2, and number of iterations $K = \Omega(\lceil \frac{\lambda_1^*}{\lambda_r^*}\log(\frac{mt}{(\lambda_1^*/\lambda_r^*)(\sigma/\sqrt{\lambda_r^*})\mu dr})\rceil)$, $\|(\mathbf{I} - U^*(U^*)^\top)U_{\text{init}}\|_F \leq \min\Big(\frac{21}{121}, O\Big(\frac{\lambda_r^*}{\lambda_1^*}\sqrt{\frac{1}{\log(t/K)}}\Big)\Big)$, $m \geq \Omega\Big((1 + r\frac{\lambda_1^*}{\lambda_r^*}\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^2)r\log(\frac{t}{\delta}) + r^2\log(\frac{K}{\delta})\Big)$, $t \geq \Omega(\mu^2 r^3 K\log(\frac{K}{\delta}))$, and $mt \geq \Omega\Big(\mu dr^2 K\log(\frac{t}{\delta})\Big(1 + \Big(\frac{\lambda_1^*}{\lambda_r^*}\Big)^2\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^2\log(\frac{t}{\delta})\log(\frac{rK}{\delta})\Big)\Big)$. Then, for any $0<\delta<1$, after K iterations and using the stepsize $\eta = (r/t)/2\lambda_1^*$, AltMinGD (Algorithm 1) returns an orthonormal matrix $U \in \mathbb{R}^{d\times r}$, such that with a probability of at least $1-\delta$*

$$\frac{1}{\sqrt{r}}\|(\mathbf{I} - U^*(U^*)^\top)U\|_F \leq O\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\mu dr^2 K\log(\frac{t}{\delta})\log(\frac{rK}{\delta})}{mt}}\Big) \tag{214}$$

A proof is in Section C.1.

Initialization. If we initialize AltMinGD (Algorithm 1) with Method-of-Moments (Theorem 12), we need at least

$$mt \geq \widetilde{\Omega}\Big(\Big(\frac{\lambda_1^*}{\lambda_r^*}\Big)^3 \mu d r^2 + \Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^4 \Big(\frac{\lambda_1^*}{\lambda_r^*}\Big)^2 d r^3\Big) \tag{215}$$

initial number of samples, where $\widetilde{\Omega}$ hides polylog factors.

C.1 Proof of Theorem 9

Proof sketch: We first prove that distance between U^* and U decreases at each iteration up to some additional noise terms. Then this per iterate result is unrolled to obtained the final guarantees.

First we focus on the k-th iterate. In this analysis, unless specified $[t]$, represents the k-th K-way partition used for the k-th iterate. Same result as Lemma B.1 for AltMin (Algorithm 3), holds for AltMinGD too. Therefore the tasks subset used for each iteration, satisfy approximate incoherence.

In the analysis of an iterate we denote the current iterate using U and the next iterate using U^+. First we prove that the distance between the true $v^{*(i)}$ and the current $v^{(i)}$ is approximately upper-bounded by multiple of distance between U and U^*. Next we prove that distance between U^+ and U^* is approximately a fraction of the distance between $v^{*(i)}$ and $v^{(i)}$. Finally, combining the above two results gives us desired result.

Preliminaries: Let $Q = (U^*)^\top U$. Using Lemma F.4, if $\|U - U^*(U^*)^\top U\|_F < 1$, Q is invertible. Let Q^{-1} be the right inverse of Q, i.e. $QQ^{-1} = \mathbf{I}$. Let $W = (V^*)^\top V^* = \sum_{i\in[t]} v^{*(i)}(v^{*(i)})^\top$, then using Assumption 2 we have that $\lambda_1^* = (r/t)\max_{\|z\|=1} z^\top W^* z$ and $\lambda_r^* = (r/t)\min_{\|z\|=1} z^\top W^* z$.

Update on V: Let $h^{(i)} = v^{(i)} - Q^{-1}v^{*(i)}$ and $H^T = [h^{(1)}h^{(2)}\dots h^{(t)}]$. Let $\|H\|_F \triangleq \sqrt{\sum_{i\in[t]}\|h^{(i)}\|^2}$ and $\|H\|_{\infty,2} \triangleq \max_{i\in[t]}\|h^{(i)}\|$. Let $W = V^\top V = \sum_{i\in[t]} v^{(i)}(v^{(i)})^\top$, and $\lambda_1 = (r/t)\max_{\|z\|=1} z^\top W z$ and $\lambda_r = (r/t)\min_{\|z\|=1} z^\top W z$.

Same result as Lemma B.2 for AltMin (Algorithm 3), holds for AltMinGD too. Therefore V update of AltMinGD satisfies Lemma B.2.

Update on U: Let $\mathcal{W}, \mathcal{H}, \widehat{\mathcal{H}} : \mathbb{R}^{d\times r} \to \mathbb{R}^{d\times r}$ be three linear operators, such that $\mathcal{W}(U) = U\sum_{i\in\mathcal{T}_k} v^{(i)}(v^{(i)})^\top = UW$, $\mathcal{H}(U) = U\sum_{i\in\mathcal{T}_k} h^{(i)}(v^{(i)})^\top$ and $\widehat{\mathcal{H}}(U) = \sum_{i\in\mathcal{T}_k} S_2^{(i)} U h^{(i)}(v^{(i)})^\top$, where $h^{(i)} = v^{(i)} - Q^{-1}v^{*(i)}$. $\mathcal{W}$ is invertible and self-adjoint. Therefore $\mathcal{W}^{-\frac{1}{2}}$ and $\mathcal{W}^{\frac{1}{2}}$ exist. Let $\mathcal{I} : \mathbb{R}^{d\times r} \to \mathbb{R}^{d\times r}$ be the identity mapping, such that $\mathcal{I}(U) = U$.

$$\begin{aligned}
&\widetilde{U} - U^*Q \\
&= U - U^*Q - \eta\Big(\sum_{i\in\mathcal{T}_k} S_2^{(i)}(Uv^{(i)} - U^*v^{*(i)})(v^{(i)})^\top + z^{(i)}(v^{(i)})^\top\Big) \\
&= U - U^*Q - \eta(\sum_{i\in\mathcal{T}_k} S_2^{(i)}(U - U^*Q)v^{(i)}(v^{(i)})^\top - S_2^{(i)}U^*Q(Q^{-1}v^{*(i)} - v^{(i)})(v^{(i)})^\top + z^{(i)}(v^{(i)})^\top) \\
&= (\mathcal{I} - \eta\mathcal{A})(U - U^*Q) + \eta(-\widehat{\mathcal{H}}(U^*Q) + \sum_{i\in\mathcal{T}_k} z^{(i)}(v^{(i)})^\top) \\
&= (\mathcal{I} - \eta\mathcal{W})(U - U^*Q) + \eta\mathcal{E}_1(U - U^*Q) + \eta(-\mathcal{H} + \mathcal{E}_2)(U^*Q) + \eta\sum_{i\in\mathcal{T}_k} z^{(i)}(v^{(i)})^\top
\end{aligned} \tag{216}$$

where $\mathcal{E}_1 = \mathcal{A} - \mathcal{W}$ and $\mathcal{E}_2 = \widehat{\mathcal{H}} - \mathcal{H}$. Let $F = \widetilde{U} - U^*Q + \eta\mathcal{H}(U^*Q)$

Lemma C.1. *Assume that all conditions and the large probability event in Lemma B.2 holds true. Then,*

$$\|\mathcal{H}(U^*Q)\|_F \leq \lambda_r(W) O\Big(\frac{\lambda_1^*}{\lambda_r^*}\sqrt{\frac{\log(\frac{t}{K\delta})}{\log(\frac{1}{\delta})}}\|(\mathbf{I} - U^*(U^*)^\top)U\|_F + \frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{r^2\log(\frac{t}{K\delta})}{m}}\Big) \tag{217}$$

and if $mt \geq \Omega(\mu d r^2 K \log(t/K\delta))$, $m \geq \Omega(r^2 \log(1/\delta))$, *and* $\eta \leq \frac{1}{\lambda_1(W)}$, *then with probability at least* $1 - \delta/3$

$$\|F\|_F \leq (1 - \frac{\eta}{2}\lambda_r(W))\|(\mathbf{I} - U^*(U^*)^\top)U\|_F \ + \ \eta\lambda_r(W) O\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\mu d r^2 K \log(\frac{t}{K\delta})\log(\frac{r}{\delta})}{mt}}\Big) \tag{218}$$

A proof is in Section C.2.1.

Lemma C.2. *If* $\eta \leq \frac{1}{\lambda_1(W)}$, $1 - \frac{1}{21}\eta\lambda_r(W) \leq \sigma_{\min}(Q)$, $\|F\|_F \leq \frac{1}{21}\eta\lambda_r(W)$ *and* $\eta\|\mathcal{H}(U^*Q)\|_F \leq \frac{1}{21}\eta\lambda_r(W)$*, then* R *is invertible and* $\|R^{-1}\| \leq (1 + \frac{1}{6}\eta\lambda_r(W)) \leq \frac{7}{6}$.

A proof is in Section C.3. Clearly, from (217) and (218), a sufficient condition for the above lemma is

$$1 - \frac{1}{21}\eta\lambda_r(W) \leq \sigma_{\min}(Q) \tag{219}$$

$$\eta\lambda_r(W) O\Big(\frac{\lambda_1^*}{\lambda_r^*}\sqrt{\frac{\log(\frac{t}{K\delta})}{\log(\frac{1}{\delta})}}\|(\mathbf{I} - U^*(U^*)^\top)U\|_F + \frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{r^2\log(\frac{t}{K\delta})}{m}}\Big) \leq \frac{1}{21}\eta\lambda_r(W) \text{ , and} \tag{220}$$

$$(1 - \frac{\eta}{2}\lambda_r(W))\|(\mathbf{I} - U^*(U^*)^\top)U\|_F + \eta\lambda_r(W) O\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\mu d r^2 K \log(\frac{t}{K\delta})\log(\frac{r}{\delta})}{mt}}\Big) \leq \frac{1}{21}\eta\lambda_r(W) \tag{221}$$

which can be satisfied with

$$\|(\mathbf{I} - U^*(U^*)^\top)U\|_F \leq \min\Big(O(\eta\lambda_r(W)), O\Big(\frac{\lambda_r^*}{\lambda_1^*}\sqrt{\frac{1}{\log(t/K)}}\Big)\Big), \tag{222}$$

$$m \geq \Omega(\frac{\lambda_1^*}{\lambda_r^*}\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^2 r^2 \log(\frac{t}{K\delta})\Big) \text{ , and} \tag{223}$$

$$mt \geq \Omega\Big(\mu d r^2 K\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^2 \log(\frac{t}{K\delta})\log(\frac{r}{\delta})\Big) \tag{224}$$

Finally, we bound the Frobenius norm distance of the next iterate U^+ from the optimal U^*.

$$\begin{aligned}
\|(\mathbf{I} - U^*(U^*)^\top)U^+\|_F &= \min_{Q^+}\|U^+ - U^*Q^+\|_F \\
&\leq \|\widetilde{U}R^{-1} - U^*QR^{-1} + \eta(\mathcal{H}(U^*Q))R^{-1}\|_F \\
&\leq \|\widetilde{U} - U^*Q + \eta\mathcal{H}(U^*Q)\|_F\|R^{-1}\| \\
&= \|F\|_F\|R^{-1}\| \\
&\leq (1 - \frac{\eta}{2}\lambda_r(W))(1 + \frac{\eta}{6}\lambda_r(W))\|(\mathbf{I} - U^*(U^*)^\top)U\|_F + \\
&\quad \frac{7}{6}\eta\lambda_r(W) O\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\mu d r^2 K \log(\frac{t}{K\delta})\log(\frac{r}{\delta})}{mt}}\Big) \\
&\leq (1 - \frac{\eta}{3}\lambda_r(W))\|(\mathbf{I} - U^*(U^*)^\top)U\|_F + \\
&\quad \eta\lambda_r(W) O\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\mu d r^2 K \log(\frac{t}{K\delta})\log(\frac{r}{\delta})}{mt}}\Big)
\end{aligned} \tag{225}$$

Finally setting $\eta = \frac{1}{2\lambda_1(W^*)} \leq \frac{1}{\lambda_1(W)}$, we get

$$\begin{aligned}
\|(\mathbf{I} - U^*(U^*)^\top)U^+\|_F &\leq (1 - \frac{\lambda_r}{6\lambda_1^*})\|(\mathbf{I} - U^*(U^*)^\top)U\|_F + \\
&\quad \frac{\lambda_r}{6\lambda_1^*} O\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\mu d r^2 K \log(\frac{t}{K\delta})\log(\frac{r}{\delta})}{mt}}\Big)
\end{aligned} \tag{226}$$

If

$$mt \geq \Omega\Big(\mu d r^2 K \Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^2 \log(\frac{t}{K\delta}) \log(\frac{r}{\delta}) \Big(1 + \Big(\frac{\lambda_1^*}{\lambda_r^*}\Big)^2 \log(\frac{t}{K\delta})\Big)\Big) \tag{227}$$

then,

$$\|(\mathbf{I} - U^*(U^*)^\top)U^+\|_F \leq (1 - \frac{1}{6}\frac{\lambda_r}{\lambda_1^*})\|(\mathbf{I} - U^*(U^*)^\top)U\|_F + \frac{1}{6}\frac{\lambda_r}{\lambda_1^*}\min\Big(\frac{21}{121}, O\Big(\frac{\lambda_r^*}{\lambda_1^*}\sqrt{\frac{1}{\log(t/K)}}\Big)\Big) \tag{228}$$

Thus if $\|(\mathbf{I} - U^*(U^*)^\top)U\|_F \leq \min\Big(\frac{21}{121}, O\Big(\frac{\lambda_r^*}{\lambda_1^*}\sqrt{\frac{1}{\log(t/K)}}\Big)\Big)$, then $\|(\mathbf{I} - U^*(U^*)^\top)U^+\|_F \leq \min\Big(\frac{21}{121}, O\Big(\frac{\lambda_r^*}{\lambda_1^*}\sqrt{\frac{1}{\log(t/K)}}\Big)\Big)$.

Therefore, using Lemma B.1 (which requires that $t \geq \Omega(\mu^2 r^3 K \log(\frac{K}{\delta})))$ and union-bound and $\lambda_r^*/2 \leq \lambda_r \leq 2\lambda_r^*$ (Lemma B.2), we can un-roll the relation, between current iterate U and the next iterate U^+, over K iterations, starting from U_{init} and ending at some U iterations, to get

$$\|(\mathbf{I} - U^*(U^*)^\top)U\|_F \leq (1 - \frac{1}{12}\frac{\lambda_r^*}{\lambda_1^*})^K \|(\mathbf{I} - U^*(U^*)^\top)U_{\text{init}}\|_F + O\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\mu d r^2 K \log(\frac{t}{K\delta})\log(\frac{r}{\delta})}{mt}}\Big) \tag{229}$$

with probability at least $1 - K\delta$. Finally setting the number of iterations as $K := \Theta(\lceil \frac{\lambda_1^*}{\lambda_r^*} \log(\frac{mt}{(\lambda_1^*/\lambda_r^*)(\sigma/\sqrt{\lambda_r^*})\mu d r})\rceil)$ we get that, with a probability of at least $1 - K\delta$

$$\|(\mathbf{I} - U^*(U^*)^\top)U\|_F \leq O\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\mu d r^2 K \log(\frac{t}{K\delta})\log(\frac{r}{\delta})}{mt}}\Big) \tag{230}$$

C.2 Analysis of update on U

C.2.1 Proof of Lemma C.1

Proof of Lemma C.1. In this proof for brevity, we will first set that $\mathcal{T}_k \leftarrow [t]$, $|\mathcal{T}_k| = t/K \leftarrow t$, $S_2^{(i)} \leftarrow S^{(i)} = \frac{1}{m}\sum_{j\in[m]} x_j^{(i)}(x_j^{(i)})^\top$. This can be done due to the approximate equivalence of the subset $\mathcal{T}_k$ and the set of all tasks $[t]$ by Lemma B.1, which requires that $t \geq \Omega(\mu^2 r^3 K \log(\frac{K}{\delta}))$. Finally at the end of the analysis we will reset $\mathcal{T}_k \leftarrow \mathcal{T}_k$, $|\mathcal{T}_k| = t/K \leftarrow t/K$, $S_2^{(i)} \leftarrow S_2^{(i)} = \frac{2}{m}\sum_{j\in[m/2+1,m]} x_j^{(i)}(x_j^{(i)})^\top$.

Recall that

$$\widetilde{U} - U^*Q = (\mathcal{I} - \eta\mathcal{W})(U - U^*Q) + \eta\mathcal{E}_1(U - U^*Q) + \eta(-\mathcal{H} + \mathcal{E}_2)(U^*Q) + \eta\sum_{i\in\mathcal{T}_k} z^{(i)}(v^{(i)})^\top \tag{231}$$

where $\mathcal{E}_1 = \mathcal{A} - \mathcal{W}$ and $\mathcal{E}_2 = \mathcal{W}^{-\frac{1}{2}}\widehat{\mathcal{H}} - \mathcal{W}^{-\frac{1}{2}}\mathcal{H}$, and $F = \widehat{U} - U^*Q + \eta\mathcal{H}(U^*Q)$. Assume that $0 < 1 - \eta\lambda_1(W) < 1 - \eta\lambda_r(W)$. Therefore

$$\|F\|_F \leq (1 - \eta\lambda_r(W) + \eta\|\mathcal{E}_1\|_F)\|U - U^*Q\|_F + \eta(\|\mathcal{E}_2(U^*Q)\|_F + \|\sum_{i\in[t]} z^{(i)}(v^{(i)})^\top\|_F) \tag{232}$$

$\Omega(\mu d r^2 \log(1/\delta)) \leq mt$ and approximate incoherence of intermediate V (26) implies that $\Omega(dr\frac{\|V\|_{\infty,2}^2}{\lambda_r(W)/t}\log(1/\delta)) \leq \Omega(\mu d r^2 \log(1/\delta)) \leq mt$, then by Lemma C.3 we have that, with a probability of at least $1 - \delta/3$

$$\|\mathcal{E}_1\|_F \leq \lambda_r(W) O\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{\mu d r^2 \log(27/\delta)}{m\,t}}\Big) \tag{233}$$

By Lemma C.4,

$$\|\mathcal{H}(U^*Q)\|_F \le \sqrt{\lambda_1(W)}\|H\|_F \tag{234}$$

and with a probability of at least $1-\delta/3$

$$\begin{aligned}&\|\mathcal{E}_2(U^*Q)\|_F\\ &\le c(\min(\|H\|_F\|V\|_{\infty,2}, \|H\|_{\infty,2}\sqrt{\lambda_1(W)})\sqrt{\frac{dr\ \log(5/\delta)}{m}} + \|H\|_{\infty,2}\|V\|_{\infty,2}\frac{dr\ \log(5/\delta)}{m})\end{aligned} \tag{235}$$

Using the approximate incoherence of V (26) in the above inequality, we get that

$$\begin{aligned}\|\mathcal{E}_2(U^*Q)\|_F \le \sqrt{\lambda_r(W)}O(\,&\min(\|H\|_F\sqrt{\frac{\mu r}{t}}, \|H\|_{\infty,2}\sqrt{\frac{\lambda_1^*}{\lambda_r^*}})\sqrt{\frac{dr\ \log(15/\delta)}{m}} +\\ &\|H\|_{\infty,2}\sqrt{\frac{\mu r}{t}}\cdot\frac{dr\ \log(15/\delta)}{m})\end{aligned} \tag{236}$$

By Lemma C.5 with a probability of at least $1-\delta/3$

$$\|\sum_{i\in[t]} z^{(i)}(v^{(i)})^\top\|_F \le O\Big(\sigma\sqrt{\frac{d\mathrm{tr}(W)}{m}\log\Big(\frac{t}{\delta}\Big)\log\Big(\frac{r}{\delta}\Big)}\Big) \tag{237}$$

Finally taking union bound over the above results and using Lemma B.2, we can bound each of the terms constituting F. Using the definitions of $\lambda_1 = (r/t)\lambda_1(W)$ and $\lambda_r = (r/t)\lambda_r(W)$, (26), and (27) (recall that we set $t \leftarrow t/K$) in (234) we get

$$\begin{aligned}\|\mathcal{H}(U^*Q)\|_F &\le \sqrt{\lambda_1(W)}\|H\|_F &(238)\\ &\le \lambda_r(W)\sqrt{\frac{\lambda_1}{\lambda_r}}\sqrt{\frac{r}{t}}\frac{\|H\|_F}{\sqrt{\lambda_r}} &(239)\\ &\le \lambda_r(W)O\Big(\frac{\lambda_1^*}{\lambda_r^*}\sqrt{\frac{\log(\frac{t}{\delta})}{\log(\frac{1}{\delta})}}\|(\mathbf{I}-U^*(U^*)^\top)U\|_F + \frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{r^2\log(\frac{t}{\delta})}{m}}\Big) &(240)\end{aligned}$$

Using the definitions of $\lambda_1 = (r/t)\lambda_1(W)$ and $\lambda_r = (r/t)\lambda_r(W)$, (27) and (28) in (236) we get

$$\begin{aligned}
&\|\mathcal{E}_2(U^*Q)\|_F \\
&\leq \sqrt{\lambda_r(W)}O(\min(\|H\|_F\sqrt{\frac{\mu r}{t}}, \|H\|_{\infty,2}\sqrt{\frac{\lambda_1^*}{\lambda_r^*}})\sqrt{\frac{dr\ \log(15/\delta)}{m}} + \|H\|_{\infty,2}\sqrt{\frac{\mu r}{t}}\cdot\frac{dr\ \log(15/\delta)}{m}) \\
&\leq \lambda_r(W)O(\sqrt{\frac{r}{t}}\min(\frac{\|H\|_F}{\sqrt{\lambda_r^*}}\sqrt{\frac{\mu r}{t}}, \frac{\|H\|_{\infty,2}}{\sqrt{\lambda_r^*}}\sqrt{\frac{\lambda_1^*}{\lambda_r^*}})\sqrt{\frac{dr\ \log(15/\delta)}{m}} + \\
&\qquad \sqrt{\frac{r}{t}}\frac{\|H\|_{\infty,2}}{\sqrt{\lambda_r^*}}\sqrt{\frac{\mu r}{t}}\cdot\frac{dr\ \log(15/\delta)}{m}) \\
&\leq \lambda_r(W)O\Big(\min\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{\mu dr^2\log(\frac{t}{\delta})}{mt}}\|(\mathbf{I}-U^*(U^*)^\top)U\|_F + \sqrt{\frac{\mu dr^2\log(\frac{1}{\delta})}{mt}}\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2\log(\frac{t}{\delta})}{m}}, \\
&\qquad \sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{\mu dr^2\log(\frac{t}{\delta})}{mt}}\|(\mathbf{I}-U^*(U^*)^\top)U\| + \sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{dr\log(\frac{1}{\delta})}{m}}\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2\log(\frac{t}{\delta})}{mt}}\Big) + \\
&\qquad \frac{\mu dr^2\log(\frac{t}{\delta})}{mt}\|(\mathbf{I}-U^*(U^*)^\top)U\| + \frac{\sqrt{\mu}dr\sqrt{r}\log(\frac{1}{\delta})}{m\sqrt{t}}\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2\log(\frac{t}{\delta})}{mt}}\Big) \\
&\leq \lambda_r(W)O\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{\mu dr^2\log(\frac{t}{\delta})}{mt}}\|(\mathbf{I}-U^*(U^*)^\top)U\| + \sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{dr\log(\frac{1}{\delta})}{m}}\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2\log(\frac{t}{\delta})}{mt}} + \\
&\qquad \frac{\mu dr^2\log(\frac{t}{\delta})}{mt}\|(\mathbf{I}-U^*(U^*)^\top)U\| + \frac{\sqrt{\mu}dr\sqrt{r}\log(\frac{1}{\delta})}{m\sqrt{t}}\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2\log(\frac{t}{\delta})}{mt}}\Big) \\
&\leq \lambda_r(W)O\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{\mu dr^2\log(\frac{t}{\delta})}{mt}}\|(\mathbf{I}-U^*(U^*)^\top)U\| + \sqrt{\frac{\mu dr^2\log(\frac{t}{\delta})}{mt}}\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2\log(\frac{1}{\delta})}{m}}\Big) \qquad (241)
\end{aligned}$$

where the second-last inequality used the fact that $mt \geq \Omega(\mu dr^2\log(\frac{t}{\delta}))$ and last inequality uses $\lambda_1^*/\lambda_r^* \leq \mu r$ (which follows from Assumption 2). Using the definitions of $\lambda_1 = (r/t)\lambda_1(W)$ and $\lambda_r = (r/t)\lambda_r(W)$, and (26) in (237) we get

$$\|\sum_{i\in[t]} z^{(i)}(v^{(i)})^\top)\|_F \leq \lambda_r(W)O\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\mathrm{tr}(W)}{\lambda_r(W)}\frac{dr\log(\frac{t}{\delta})\log(\frac{r}{\delta})}{mt}}\Big) \qquad (242)$$

$$\leq \lambda_r(W)O\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\mu dr^2\log(\frac{t}{\delta})\log(\frac{r}{\delta})}{mt}}\Big) \qquad (243)$$

where the last inequality uses $\mathrm{tr}(W)/\lambda_r(W) \leq \mu r$ (which follows from Assumption 2 and (26)) Substituting (233), (241), (243), and (26) in (232) and using $m \geq \Omega(r^2\log(1/\delta))$we get

$$\begin{aligned}
\|F\|_F &\leq (1-\eta\lambda_r(W)+\eta\|\mathcal{E}_1\|_F)\|U-U^*Q\|_F + \\
&\qquad \eta(\|\mathcal{E}_2(U^*Q)\|_F + \|\mathcal{W}^{\frac{1}{2}}\|\|\mathcal{W}^{-\frac{1}{2}}(\sum_{i\in[t]} z^{(i)}(v^{(i)})^\top)\|_F) \qquad (244) \\
&\leq \Big(1-\eta\lambda_r(W)\Big(1-O\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{\mu dr^2\log(t/\delta)}{mt}}\Big)\Big)\Big)\|(\mathbf{I}-U^*(U^*)^\top)U\|_F + \\
&\qquad \eta\lambda_r(W)O\Big(\sqrt{\frac{\mu dr^2\log(\frac{t}{\delta})}{mt}}\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2\log(\frac{1}{\delta})}{m}} + \frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\mu dr^2\log(\frac{t}{\delta})\log(\frac{r}{\delta})}{mt}}\Big) \qquad (245) \\
&\leq (1-\frac{\eta}{2}\lambda_r(W))\|(\mathbf{I}-U^*(U^*)^\top)U\|_F \ + \ \eta\lambda_r(W)O\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\mu dr^2\log(\frac{t}{\delta})\log(\frac{r}{\delta})}{mt}}\Big) \qquad (246)
\end{aligned}$$

Finally, by resetting $\mathcal{T}_k \leftarrow \mathcal{T}_k$, $|\mathcal{T}_k| = t/K \leftarrow t/K$, $S_2^{(i)} \leftarrow S_2^{(i)} = \frac{2}{m}\sum_{j\in[m/2+1,m]} x_j^{(i)}(x_j^{(i)})^\top$, we obtain the desired result. $\square$

C.2.2 Supporting lemmas for the analysis of update on U

Lemma C.3. *If $\Omega(\mu d r^2 \log(27/\delta)) \le mt$, then with a probability of at least $1-\delta/3$,*

$$\|\mathcal{E}_1\|_F \le \lambda_r(W) O\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{\mu d r^2 \log(27/\delta)}{m\, t}}\Big) \tag{247}$$

Proof of Lemma C.3. Let $\mathcal{S}_F = \{U \in \mathbb{R}^{d\times r} \mid \|U\|_F = 1\}$ be the set of all real matrices of dimensions $d\times r$ with unit Frobenius norm. For $\epsilon \le 1$, there exists an ϵ-net, $N_\epsilon \subset \mathcal{S}_F$, of size $(1+2/\epsilon)^{dr}$ with respect to the Frobenius norm [40, Lemma 5.2]. That is for any $U' \in \mathcal{S}_F$, there exists some $U \in N_\epsilon$ such that $\|U'-U\|_F \le \epsilon$.

Consider a $U \in N_\epsilon$, such that $\|U\|_F = 1$. Now we will prove with high-probability that $\big\langle(\mathcal{A}-\mathcal{W})(U), U\big\rangle$ is small. Consider the the following quadratic form

$$\big\langle(\mathcal{A})(U), U\big\rangle = \Big\langle \sum_{i\in[t]} S^{(i)} U v^{(i)}(v^{(i)})^\top, U\Big\rangle \tag{248}$$

$$= \sum_{i\in[t]} \frac{1}{m}\sum_{j\in[m]} (x_j^{(i)})^\top (U v^{(i)}(v^{(i)})^\top U^\top) x_j^{(i)} \tag{249}$$

where $S^{(i)} = \frac{1}{m}\sum_{j\in[m]} x_j^{(i)}(x_j^{(i)})^\top$ and $x_j^{(i)} \sim \mathcal{N}(0, \mathbf{I}_{d\times d})$ are i.i.d. standard Gaussian random vectors and $W = \sum_{i\in[t]} v^{(i)}(v^{(i)})^\top$ is rank-r matrix. We will use Hanson-Wright inequality (Lemma F.5) to prove that the above quadratic form concentrates around its mean. Notice that the the expectation of $\big\langle \mathcal{A}(U), U\big\rangle$ is $\langle \mathcal{W}(U), U\rangle$.

$$\sum_{i\in[t]} \mathbb{E}\Big[\Big\langle S^{(i)} U v^{(i)}(v^{(i)})^\top, U\Big\rangle\Big] = \Big\langle U \sum_{i\in[t]} v^{(i)}(v^{(i)})^\top, U\Big\rangle = \langle UW, U\rangle = \langle \mathcal{W}(U), U\rangle\ . \tag{250}$$

We will also need the following bounds to apply the Hanson-Wright inequality. Recall that $\|V\|_{\infty,2} = \max_{i\in[t]} \|v^{(i)}\|$. Then,

$$\max_{i\in[t]} \|U v^{(i)}(v^{(i)})^\top U^\top\| = \max_{i\in[t]} \|U v^{(i)}\|^2 \le \max_{i\in[t]} \|U\|^2 \|v^{(i)}\|^2 \le \|V\|_{\infty,2}^2 \tag{251}$$

Also note that,

$$\sum_{i\in[t]} \|U v^{(i)}(v^{(i)})^\top U^\top\|_F^2 = \sum_{i\in[t]} \|U v^{(i)}\|^4 = \max_{i\in[t]} \|U v^{(i)}\|^2 \sum_{i\in[t]} \Big\langle U v^{(i)}, U v^{(i)}\Big\rangle \tag{252}$$

$$= \max_{i\in[t]} \|U\|^2 \|v^{(i)}\|^2 \sum_{i\in[t]} \Big\langle U U^\top, \sum_{i\in[t]} v^{(i)}(v^{(i)})^\top \Big\rangle \tag{253}$$

$$\le \|V\|_{\infty,2}^2 \lambda_1(W) \tag{254}$$

where the last inequality used (250) and (251). Then by Hanson-Wright inequality (Lemma F.5), with probability at least $1-\delta/|N_\epsilon|$

$$\big|\big\langle(\mathcal{A}-\mathcal{W})(U), U\big\rangle\big| = \big|\Big\langle \sum_{i\in[t]} \frac{1}{m}\sum_{j\in[m]} x_j^{(i)}(x_j^{(i)})^\top U v^{(i)}(v^{(i)})^\top, U\Big\rangle - \langle \mathcal{W}(U), U\rangle\,\big| \le \Delta_\epsilon \tag{255}$$

where $\Delta_\epsilon = c\max(\sqrt{\frac{\|V\|_{\infty,2}^2 \lambda_1(W)\log(|N_\epsilon|/\delta)}{m}}, \frac{\|V\|_{\infty,2}^2 \log(|N_\epsilon|/\delta)}{m})$. Taking union bound over all $U \in N_\epsilon$ implies that with probability at least $1-\delta$

$$\big|\big\langle(\mathcal{A}-\mathcal{W})(U), U\big\rangle\big| \le \Delta_\epsilon\ ,\ \text{ for all } U \in N_\epsilon\,. \tag{256}$$

For brevity, let $\mathcal{E}_1(U) = (\mathcal{A} - \mathcal{W})(U)$. Notice that $\mathcal{E}_1$ is self-adjoint, therefore it has an eigen decomposition with respect to the Frobenius norm. Then, let $U' \in \mathcal{S}_F \subset \mathbb{R}^{d\times r}$ be the largest "eigenmatrix" of $\mathcal{E}_1$, such that $\langle \mathcal{E}_1(U), U\rangle = \|\mathcal{E}_1\|_F = \max_{\|\widetilde{U}\|_F=1} \left\langle \mathcal{E}_1(\widetilde{U}), \widetilde{U}\right\rangle = \max_{\|\widetilde{U}\|_F=\|\widetilde{U}'\|_F=1} \left\langle \mathcal{E}_1(\widetilde{U}), \widetilde{U}'\right\rangle$. Then there exists some $U \in N_\epsilon$ such that $\|U' - U\|_F \le \epsilon$.

$$\|\mathcal{E}_1\|_F = \langle \mathcal{E}_1(U'), U'\rangle = \langle \mathcal{E}_1(U), U\rangle + \langle \mathcal{E}_1(U'-U), U\rangle + \langle \mathcal{E}_1(U'), U'-U\rangle \tag{257}$$
$$\le \langle \mathcal{E}_1(U), U\rangle + \|\mathcal{E}_1\|_F\|U'-U\|_F(\|U\|_F + \|U'\|_F) \tag{258}$$
$$\le \langle \mathcal{E}_1(U), U\rangle + 2\epsilon\|\mathcal{E}_1\|_F \tag{259}$$

Re-arranging and setting $\epsilon = 1/4$, and $c \leftarrow 2c$, we get

$$\|\mathcal{A} - \mathcal{W}\|_F = \|\mathcal{E}_1'\|_F \le \Delta_{\frac{1}{4}} \le O\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{\mu d r^2\log(9/\delta)}{m\,t}}\Big). \tag{260}$$

where we use the approximate incoherence of intermediate variable V Lemma B.2 and the fact that $\Omega(\mu d r^2\log(9/\delta)) \le mt$, which implies that $\Delta_{\frac{1}{4}} = c\max\Big(\sqrt{\frac{dr\,\|V\|_{\infty,2}^2\lambda_1(W)\log(9/\delta)}{m}}, \frac{dr\,\|V\|_{\infty,2}^2\log(9/\delta)}{m}\Big) \le \lambda_r(W)O\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}}\max\Big(\sqrt{\frac{\mu dr^2\,\log(9/\delta)}{mt}}, \frac{\mu dr^2\,\log(9/\delta)}{mt}\Big)\Big) \le \lambda_r(W)O\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{\mu dr^2\log(9/\delta)}{m\,t}}\Big)$. Finally, setting $\delta \leftarrow \delta/3$ get us the desired result. □

Lemma C.4. *$\|(\mathcal{W}^{-\frac{1}{2}}\mathcal{H})(U^*Q)\|_F \le \sqrt{\lambda_1(W)}\|H\|_F$ and with a probability of at least $1-\delta/3$*

$$\|\mathcal{E}_2(U^*Q)\|_F$$
$$\le c(\min(\|H\|_F\|V\|_{\infty,2}, \|H\|_{\infty,2}\sqrt{\lambda_1(W)})\sqrt{\frac{dr\,\log(5/\delta)}{m}} + \|H\|_{\infty,2}\|V\|_{\infty,2}\frac{dr\,\log(5/\delta)}{m}) \tag{261}$$

Proof of Lemma C.4. First we prove that the expected value $\mathbb{E}[(\widehat{\mathcal{H}})(U^*Q)] = (\mathcal{H})(U^*Q)$ is bounded.

$$\|\mathcal{H}(U^*Q)\|_F = \max_{\|U\|_F=1}\langle \mathcal{H}(U^*Q), U\rangle \tag{262}$$
$$= \max_{\|U\|_F=1}\sum_{i\in[t]}\left\langle U^*Qh^{(i)}(v^{(i)})^\top, U\right\rangle \tag{263}$$
$$= \max_{\|U\|_F=1}\sum_{i\in[t]}\left\langle U^*Qh^{(i)}, Uv^{(i)}\right\rangle \tag{264}$$
$$\le \max_{\|U\|_F=1}\sqrt{\sum_{i\in[t]}\|U^*Qh^{(i)}\|^2}\sqrt{\sum_{i\in[t]}\langle Uv^{(i)}, Uv^{(i)}\rangle} \tag{265}$$
$$\le \max_{\|U\|_F=1}\|Q\|\sqrt{\sum_{i\in[t]}\|h^{(i)}\|^2}\sqrt{\left\langle U\sum_{i\in[t]}v^{(i)}(v^{(i)})^\top, U\right\rangle} \tag{266}$$
$$\le \max_{\|U\|_F=1}\|H\|_F\|U\|_F\sqrt{\lambda_1(W)} = \sqrt{\lambda_1(W)}\|H\|_F \tag{267}$$

where used the fact that $\langle AB, C\rangle = \langle A, CB^\top\rangle$ and $(U^*)^\top U^* = \mathbf{I}$.

Let $\mathcal{S}_F = \{U \in \mathbb{R}^{d\times r} \mid \|U\|_F = 1\}$ be the set of all real matrices of dimensions $d\times r$ with unit Frobenius norm. For $\epsilon \le 1$, there exists an ϵ-net, $N_\epsilon \subset \mathcal{S}_F$, of size $(1+2/\epsilon)^{dr}$ with respect to the Frobenius norm [40, Lemma 5.2]. That is for any $U' \in \mathcal{S}_F$, there exists some $U \in N_\epsilon$ such that $\|U' - U\|_F \le \epsilon$.

Consider a $U \in N_\epsilon$, such that $\|U\|_F = 1$. Now we will prove with high-probability that $\left\langle \mathcal{H}(U^*Q)(U) - \sum_{i\in[t]} S^{(i)}U^*Qh^{(i)}(v^{(i)})^\top, U\right\rangle$ is small. Consider the the following quadratic form

$$\Big\langle \sum_{i\in[t]} S^{(i)}U^*Qh^{(i)}(v^{(i)})^\top, U\Big\rangle = \Big\langle \sum_{i\in[t]} S^{(i)}U^*Qh^{(i)}(v^{(i)})^\top, U\Big\rangle \tag{268}$$

$$= \sum_{i\in[t]} \frac{1}{m} \sum_{j\in[m]} (x_j^{(i)})^\top (U^*Qh^{(i)}(v^{(i)})^\top U^\top) x_j^{(i)} \tag{269}$$

where $S^{(i)} = \frac{1}{m}\sum_{j\in[m]} x_j^{(i)}(x_j^{(i)})^\top$ and $x_j^{(i)} \sim \mathcal{N}(0, \mathbf{I}_{d\times d})$ are i.i.d. standard Gaussian random vectors. We will use Hanson-Wright inequality (Lemma F.5) to prove that the above quadratic form concentrates around its mean. Notice that the the expectation of $\big\langle \sum_{i\in[t]} S^{(i)}U^*Qh^{(i)}(v^{(i)})^\top, U\big\rangle$ is $\big\langle \mathcal{H}(U), U\big\rangle$.

$$\mathbb{E}[\sum_{i\in[t]} S^{(i)}U^*Qh^{(i)}(v^{(i)})^\top] = \sum_{i\in[t]} U^*Qh^{(i)}(v^{(i)})^\top = \mathcal{H}(U^*Q)\,. \tag{270}$$

We will also need the following bounds to apply the Hanson-Wright inequality. Recall that $\|H\|_{\infty,2} = \max_{i\in[t]}\|h^{(i)}\|$ and $\|V\|_{\infty,2} = \max_{i\in[t]}\|v^{(i)}\|$. Then,

$$\max_{i\in[t]}\|U^*Qh^{(i)}(v^{(i)})^\top U^\top\| \le \max_{i\in[t]}\|U^*\|\|Q\|\|h^{(i)}\| \max_{i\in[t]}\|v^{(i)}\|\|U\| \le \|H\|_{\infty,2}\|V\|_{\infty,2} \tag{271}$$

Also note that

$$\sum_{i\in[t]}\|U^*Qh^{(i)}(v^{(i)})^\top U^\top\|_F^2 = \sum_{i\in[t]}\|U^*Qh^{(i)}\|^2\|Uv^{(i)}\|^2 \tag{272}$$

$$\le (\sum_{i\in[t]}\|U^*Qh^{(i)}\|^2)(\max_{i\in[t]}\|Uv^{(i)}\|^2) \tag{273}$$

$$\le (\|Q\|^2\sum_{i\in[t]}\|h^{(i)}\|^2)(\max_{i\in[t]}\|U\|^2\|v^{(i)}\|^2) \tag{274}$$

$$\le \|H\|_F^2\|V\|_{\infty,2}^2 \tag{275}$$

and

$$\sum_{i\in[t]}\|U^*Qh^{(i)}(v^{(i)})^\top U^\top\|_F^2 = \sum_{i\in[t]}\|U^*Qh^{(i)}\|^2\|Uv^{(i)}\|^2 \tag{276}$$

$$\le (\max_{i\in[t]}\|U^*Qh^{(i)}\|^2)\mathrm{tr}(U\sum_{i\in[t]}v^{(i)}(v^{(i)})^\top U^\top) \tag{277}$$

$$\le (\max_{i\in[t]}\|U^*Qh^{(i)}\|^2)\left\langle UU^\top, W\right\rangle \tag{278}$$

$$\le \|Q\|\max_{i\in[t]}\|h^{(i)}\|^2\|U\|_F^2\lambda_1(W) \tag{279}$$

$$= \|H\|_{\infty,2}^2\lambda_1(W)\,. \tag{280}$$

Therefore, $\sum_{i\in[t]}\|U^*Qh^{(i)}(v^{(i)})^\top U^\top\|_F^2 \le \min\{\|H\|_F^2\|V\|_{\infty,2}^2, \|H\|_{\infty,2}^2\lambda_1(W)\}$. For brevity, let $\mathcal{E}_2(U) = \sum_{i\in[t]} S^{(i)}Uh^{(i)}(v^{(i)})^\top - \mathcal{H}(U)$. Then by Hanson-Wright inequality (Lemma F.5), with probability at least $1-\delta/|N_\epsilon|$

$$\big|\big\langle \mathcal{E}_2(U^*Q), U\big\rangle\big| = \big|\Big\langle \sum_{i\in[t]}\frac{1}{m}\sum_{j\in[m]} x_j^{(i)}(x_j^{(i)})^\top U^*Qh^{(i)}(v^{(i)})^\top, U\Big\rangle - \langle \mathcal{H}(U^*Q), U\rangle\,\big| \le \Delta_\epsilon \tag{281}$$

where $\Delta_\epsilon = c(\min(\|H\|_F\|V\|_{\infty,2}, \|H\|_{\infty,2}\sqrt{\lambda_1(W)})\sqrt{\frac{\log(|N_\epsilon|/\delta)}{m}} + \|H\|_{\infty,2}\|V\|_{\infty,2}\frac{\log(|N_\epsilon|/\delta)}{m})$. Taking union bound over all $U \in N_\epsilon$ implies that with probability at least $1-\delta$

$$\big|\big\langle \mathcal{E}_2(U), U\big\rangle\big| \le \Delta_\epsilon\ ,\ \text{for all } U \in N_\epsilon\,. \tag{282}$$

Let $U' \in \mathcal{S}_F \subset \mathbb{R}^{d\times r}$ be the matrix "parallel" to $\mathcal{E}_2(U^*Q)$, that is $\|\mathcal{E}_2(U^*Q)\|_F = \max_{\|\widetilde{U}\|_F=1} \left\langle \mathcal{E}_1(U^*Q), \widetilde{U} \right\rangle = \langle \mathcal{E}_2(U^*Q), U' \rangle$. Then there exists some $U \in N_\epsilon$ such that $\|U' - U\|_F \leq \epsilon$.

$$\|\mathcal{E}_2(U^*Q)\|_F = \langle \mathcal{E}_2(U^*Q), U' \rangle = \langle \mathcal{E}_2(U^*Q), U \rangle + \langle \mathcal{E}_2(U^*Q), U' - U \rangle \tag{283}$$
$$\leq \langle \mathcal{E}_1(U), U \rangle + \|\mathcal{E}_2(U^*Q)\|_F \|U' - U\|_F \tag{284}$$
$$\leq \langle \mathcal{E}_1(U), U \rangle + \epsilon \|\mathcal{E}_2(U^*Q)\|_F \tag{285}$$

Re-arranging and setting $\epsilon = 1/2$, and $c \leftarrow 2c$, we get

$$\|\mathcal{E}_2(U^*Q)\|_F \leq \Delta_{\frac{1}{2}}$$
$$= c(\min(\|H\|_F \|V\|_{\infty,2}, \|H\|_{\infty,2}\sqrt{\lambda_1(W)})\sqrt{\frac{dr\ \log(5/\delta)}{m}} + \|H\|_{\infty,2}\|V\|_{\infty,2}\frac{dr\ \log(5/\delta)}{m}) \tag{286}$$

Finally setting $\delta \leftarrow \delta/3$ get us the desired result.

□

Lemma C.5. *With a probability of at least* $1 - \delta/3$

$$\|\sum_{i\in[t]} z^{(i)}(v^{(i)})^\top)\|_F \leq O\Big(\sigma\sqrt{\frac{d\mathrm{tr}(W)}{m}\log\Big(\frac{t}{\delta}\Big)\log\Big(\frac{r}{\delta}\Big)}\Big) \tag{287}$$

Proof of Lemma C.5. Notice that $z^{(i)}$ (defined in Appendix C) is a Gaussian random vector of the following form

$$z^{(i)} = \frac{1}{m}\sum_{j\in[m]} \varepsilon_j^{(i)} x_j^{(i)} = \frac{1}{m}\|\varepsilon^{(i)}\| g^{(i)}, g^{(i)} \sim \mathcal{N}(0, \mathbf{I}_{d\times d}) \tag{288}$$

Using Hanson-Wright inequality (Lemma F.5, by setting $m \leftarrow 1$, $x_1 \leftarrow \varepsilon^{(i)}$, and $A_1 \leftarrow \mathbf{I}_{m\times m}$) and taking union bound over all tasks, we get that, with probability of at least $1 - \frac{\delta}{2}$

$$\|\varepsilon^{(i)}\|^2 \leq \sigma^2 m(1 + c\sqrt{\frac{\log(\frac{2t}{\delta})}{m}} + c\frac{\log(\frac{2t}{\delta})}{m}) \leq 2c\,\sigma^2 m \log\Big(\frac{2t}{\delta}\Big),\ \text{ for all } i \in [t] \tag{289}$$

where used the fact that $m \geq 1$ and $\log\Big(\frac{2t}{\delta}\Big) \geq 1$.

Now it is easy check that

$$\sum_{i\in[t]} \|v^{(i)}\|^2 = \sum_{i\in[t]} \mathrm{tr}((v^{(i)})^\top v^{(i)}) = \sum_{i\in[t]} \mathrm{tr}(v^{(i)}(v^{(i)})^\top) = \mathrm{tr}(W) \leq \mu r \lambda_r(W) \tag{290}$$

Notice that $\sum_{i\in[t]} \frac{1}{m}\|\varepsilon^{(i)}\| g^{(i)} v_j^{(i)}$ is a Gaussian random vector of the following form

$$\sum_{i\in[t]} \frac{1}{m}\|\varepsilon^{(i)}\| g^{(i)} v_j^{(i)} = \frac{1}{m}\sqrt{\sum_{i\in[t]} \|\varepsilon^{(i)}\|^2 (v_j^{(i)})^2}\ \widehat{g}_j\ , \widehat{g}_j \sim \mathcal{N}(0, \mathbf{I}_{d\times d}) \tag{291}$$

Using Hanson-Wright inequality (Lemma F.5, by setting $m \leftarrow 1$, $x_1 \leftarrow \widehat{g}_j$, and $A_1 \leftarrow \mathbf{I}_{d\times d}$) and taking union bound over all $j \in [r]$, we get that, with probability of at least $1 - \frac{\delta}{2}$

$$\|\widehat{g}_j\|^2 \leq d(1 + c\sqrt{\frac{\log(\frac{2r}{\delta})}{d}} + c\frac{\log(\frac{2r}{\delta})}{d}) \leq 2cd \log\Big(\frac{2r}{\delta}\Big),\ \text{ for all } j \in [r] \tag{292}$$

where used the fact that $d \geq 1$ and $\log\Big(\frac{2r}{\delta}\Big) \geq 1$.

Combining the above results and using union bound, we get that, with a probability of at least $1-\delta$,

$$\Big\|\sum_{i\in[t]} z^{(i)}(v^{(i)})^\top)\Big\|_F^2 = \Big\|\sum_{i\in[t]} z^{(i)}(v^{(i)})^\top\Big\|_F^2 \tag{293}$$

$$= \Big\|\sum_{i\in[t]} \frac{1}{m}\|\varepsilon^{(i)}\| g^{(i)}(v^{(i)})^\top\Big\|_F^2 \tag{294}$$

$$= \sum_{j\in[r]}\Big\|\sum_{i\in[t]} \frac{1}{m}\|\varepsilon^{(i)}\| g^{(i)} v_j^{(i)}\Big\|^2 \tag{295}$$

$$\le \sum_{j\in[r]}\sum_{i\in[t]} \frac{\|\varepsilon^{(i)}\|^2}{m^2}(v_j^{(i)})^2\|\widehat{g}_j\|^2 \tag{296}$$

$$\le \sum_{j\in[r]}\sum_{i\in[t]} O\Big(\frac{m\sigma^2}{m^2}\log\Big(\frac{t}{\delta}\Big)\Big)(v_j^{(i)})^2 O\Big(d\log\Big(\frac{r}{\delta}\Big)\Big) \tag{297}$$

$$\le O\Big(\frac{d\sigma^2}{m}\log\Big(\frac{t}{\delta}\Big)\log\Big(\frac{r}{\delta}\Big)\Big)\sum_{i\in[t]}\|v^{(i)}\|^2 \tag{298}$$

$$\le O\Big(\frac{\sigma^2 d\,\mathrm{tr}(W)}{m}\log\Big(\frac{t}{\delta}\Big)\log\Big(\frac{r}{\delta}\Big)\Big)\,. \tag{299}$$

Finally, we get the desired result by setting $\delta \leftarrow \delta/3$. □

C.3 Analysis of QR decomposition

Proof of Lemma C.2.

$$\sigma_{\min}(R) \ge \min_{\|z\|=1}\|Rz\| = \min_{\|z\|=1}\|U^+Rz\| = \min_{\|z\|=1}\|\widehat{U}z\| \tag{300}$$

$$\ge \min_{\|z\|=1}\|(U^*Q - \eta\mathcal{H}(U^*Q) + F)z\| \tag{301}$$

$$\ge \min_{\|z\|=1}\sqrt{z^\top Q^\top Q z} - \eta\|\mathcal{H}(U^*Q)\| - \|F\| \tag{302}$$

$$\ge \min_{\|z\|=1}\sigma_{\min}(Q) - \eta\|\mathcal{H}(U^*Q)\| - \|F\| \tag{303}$$

$$\ge 1 - \frac{1}{21}\frac{\lambda_r^*}{\lambda_1^*} - \frac{1}{21}\frac{\lambda_r^*}{\lambda_1^*} - \frac{1}{21}\frac{\lambda_r^*}{\lambda_1^*} \ge 1 - \frac{1}{7}\frac{\lambda_r^*}{\lambda_1^*} \tag{304}$$

There fore R is invertible and $\|R^{-1}\| = (\sigma_{\min}(R))^{-1} \le \frac{1}{1-\frac{1}{7}\frac{\lambda_r^*}{\lambda_1^*}} \le 1 + \frac{1}{6}\frac{\lambda_r^*}{\lambda_1^*}$ □

D Analysis of AltMinGD-S (Algorithm 2) and AltMin-S (Algorithm 4) with subset selection

In this section analyze the task subset selection-based algorithms: AltMinGD-S (Algorithm 2) and AltMin-S (Algorithm 4).

AltMin-S: Initialized at U, the k-the step of alternating minimization-based AltMin-S (Algorithm 4) is:

$$\mathcal{T}_k \;=\; \Big\{\, i \in [1 + \frac{(k-1)t}{K}, \frac{tk}{K}] \mid \sigma_{\min}(U^\top S^{(i)}U) \ge 1/2 \text{ and } \sigma_{\max}(U^\top S^{(i)}U) \le 2 \,\Big\} \tag{305}$$

$$v^{(i)} \;\leftarrow\; (U^\top S^{(i)}U)^\dagger((U^\top S^{(i)}U^*)v^{*(i)} + U^\top z^{(i)})\,, \qquad \text{for } i \in \mathcal{T}_k \tag{306}$$

$$\widehat{U} \;\leftarrow\; \mathcal{A}^\dagger\Big(\sum_{i\in\mathcal{T}} S^{(i)}U^*v^{*(i)}(v^{(i)})^\top + z^{(i)}(v^{(i)})^\top\Big)\,, \tag{307}$$

$$U^+ \;\leftarrow\; \mathrm{QR}(\widehat{U})\,, \tag{308}$$

where U^+ is the next iterate, $S_1^{(i)} = \frac{2}{m}\sum_{j\in[1,m/2]} x_j^{(i)}(x_j^{(i)})^\top$, $S_2^{(i)} = \frac{2}{m}\sum_{j\in[1+m/2,m]} x_j^{(i)}(x_j^{(i)})^\top$, $z^{(i)} \triangleq (1/m)\sum_{j\in[m]}\varepsilon_j^{(i)}x_j^{(i)}$ and $\mathcal{A} : \mathbb{R}^{d\times r} \to \mathbb{R}^{d\times r}$ is a self-adjoint linear operator such that $\mathcal{A}(U) = \sum_{i\in T} S^{(i)}Uv^{(i)}(v^{(i)})^\top$.

Theorem 10. *Let there be t linear regression tasks, each with m samples satisfying Assumptions 1 and 2, and $K = \lceil\log_2(\frac{(\lambda_r^*/\lambda_1^*)mt}{\mu dr^2})\rceil$, $\|(\mathbf{I} - U^*(U^*)^\top)U_{\rm init}\|_F \le \min\Big(\frac{21}{121}, O\Big(\sqrt{\frac{\lambda_r^*}{\lambda_1^*}\frac{1}{\log(t/K)}}\Big)\Big)$, $m \ge \Omega\Big(\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^2 r^2\log(\frac{t}{\delta}) + r^2\log(\frac{K}{\delta}) + \log(\mu r)\Big)$, $t \ge \Omega(\mu^2 r^3 K\log(\frac{K}{\delta}))$ and $mt \ge \Omega\Big(\mu dr^2K\frac{\lambda_1^*}{\lambda_r^*}\Big(\log(\frac{t}{\delta}) + \Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^2\log^2(\frac{t}{\delta})\log(\frac{rK}{\delta})\Big)\Big)$. Then, for any $0 < \delta < 1$, after K iterations, AltMin-S (Algorithm 4) returns an orthonormal matrix $U \in \mathbb{R}^{d\times r}$, such that with a probability of at least $1-\delta$*

$$\frac{1}{\sqrt{r}}\|(\mathbf{I} - U^*(U^*)^\top)U\|_F \le O\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\mu drK\log(\frac{t}{\delta})\log(\frac{rK}{\delta})}{mt}}\Big) \tag{309}$$

and the algorithm uses an additional memory of size $O(d^2r^2)$.

A proof is in Section D.1.

AltMin-S: Initialized at U, the k-the step of alternating minimization-based AltMinGD-S (Algorithm 2) is:

$$\mathcal{T}_k = \Big\{ i \in [1+\frac{(k-1)t}{K}, \frac{tk}{K}] \mid \sigma_{\min}(U^\top S^{(i)}U) \ge 1/2 \text{ and } \sigma_{\max}(U^\top S^{(i)}U) \le 2 \Big\} \tag{310}$$

$$v^{(i)} \leftarrow (U^\top S^{(i)}U)^\dagger((U^\top S^{(i)}U^*)v^{*(i)} + U^\top z^{(i)}) , \qquad \text{for } i \in \mathcal{T}_k \tag{311}$$

$$\widetilde{U} \leftarrow U - \eta\Big(\sum_{i\in[t]} S_2^{(i)}(Uv^{(i)} - U^*v^{*(i)})(v^{(i)})^\top + z^{(i)}(v^{(i)})^\top\Big) , \tag{312}$$

$$U^+ \leftarrow \mathrm{QR}(\widetilde{U}) , \tag{313}$$

where U^+ is the next iterate, $S_1^{(i)}$, $S_2^{(i)}$, and $\mathcal{A}$ are defined in the same way as above for AltMin-S.

Theorem 11. *Let there be t linear regression tasks, each with m samples satisfying Assumptions 1 and 2, and $K = \Omega(\lceil\frac{\lambda_1^*}{\lambda_r^*}\log(\frac{mt}{(\lambda_1^*/\lambda_r^*)(\sigma/\sqrt{\lambda_r^*})\mu dr})\rceil)$, $\|(\mathbf{I} - U^*(U^*)^\top)U_{\rm init}\|_F \le \min\Big(\frac{21}{121}, O\Big(\frac{\lambda_r^*}{\lambda_1^*}\sqrt{\frac{1}{\log(t/K)}}\Big)\Big)$, $m \ge \Omega\Big(r^2\frac{\lambda_1^*}{\lambda_r^*}\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^2\log(\frac{t}{\delta}) + r^2\log(\frac{K}{\delta}) + \log(\mu r)\Big)$, $t \ge \Omega(\mu^2r^3K\log(\frac{K}{\delta}))$ and $mt \ge \Omega\Big(\mu dr^2K\log(\frac{t}{\delta})\Big(1 + \Big(\frac{\lambda_1^*}{\lambda_r^*}\Big)^2\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^2\log(\frac{t}{\delta})\log(\frac{rK}{\delta})\Big)\Big)$. Then, for any $0 < \delta < 1$, after K iterations and using the stepsize $\eta = (r/t)/2\lambda_1^*$, AltMinGD-S (Algorithm 2) returns an orthonormal matrix $U \in \mathbb{R}^{d\times r}$, such that with a probability of at least $1-\delta$*

$$\frac{1}{\sqrt{r}}\|(\mathbf{I} - U^*(U^*)^\top)U\|_F \le O\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\mu dr^2K\log(\frac{t}{\delta})\log(\frac{rK}{\delta})}{mt}}\Big) \tag{314}$$

A proof is in Section D.1.

D.1 Proofs of Theorem 10 and Theorem 11

Here we provide only the proof of Theorem 10 as the proof of Theorem 11 is very similar and straightforward, given the former.

First, in the following lemma, we prove that the task subset $\mathcal{T}_k$ has similar properties as the full task partition $[1 + t(k-1)/K, tk/K]$.

Lemma D.1 (Subset selection). *If $m \geq \Omega(r + \log(\mu r))$ and $t \geq \Omega(\mu^2\, r^2\, K \log(\frac{1}{\delta}))$, then with a probability of at least $1 - \delta/3$,*

$$|\mathcal{T}_k| = \Theta\Big(\frac{t}{K}\Big)\,, \quad \text{, and} \quad \|V^*\|_{\infty,2}^2 \leq O\Big(\frac{\mu\, r}{|\mathcal{T}_k|}\lambda_r(\sum_{i\in\mathcal{T}} v^{*(i)}(v^{*(i)})^\top)\Big) \tag{315}$$

$$\lambda_r(\sum_{i\in\mathcal{T}} v^{*(i)}(v^{*(i)})^\top) = \Theta(\lambda_r(\sum_{i\in\mathcal{P}_k} v^{*(i)}(v^{*(i)})^\top))\text{ , and} \tag{316}$$

$$\lambda_1(\sum_{i\in\mathcal{T}} v^{*(i)}(v^{*(i)})^\top) = \Theta(\lambda_1(\sum_{i\in\mathcal{P}_k} v^{*(i)}(v^{*(i)})^\top))\,, \tag{317}$$

where $\mathcal{P}_k = [1 + t(k-1)/K, tk/K]$ is the k-th K-way partition of $[t]$ after shuffling.

A proof is in Section D.2. Therefore, assuming that the above high-probability event holds, in the rest of the proof we can consider that $\mathcal{T}_k$ is equivalent to $\mathcal{P}_k$.

In the rest of the proof, when compared to the proof of Theorem 8, only the following Lemma (corresponding to Lemma B.2) analyzing the V update changes in its necessary condition.

Lemma D.2. *If $\|(\mathbf{I} - U^*(U^*)^\top)U\|_F \leq \min\Big(\frac{21}{121}, O\Big(\sqrt{\frac{\lambda_r^*}{\lambda_1^*}\frac{1}{\log(t/K)}}\Big)\Big)$ and $m \geq \Omega\Big(\Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^2 r^2 \log(\frac{t}{K\delta}) + r\log(\frac{1}{\delta})\Big)$, then with a probability of at least $1-\delta/3$,*

$$\|v^{(i)}\| \leq O\Big(\mu\,\lambda_r\Big)\,, \;\; \lambda_1 \leq 2\lambda_1^*\text{ , and } \lambda_r^*/2 \leq \lambda_r \leq 2\lambda_r^* \tag{318}$$

and

$$\sqrt{\frac{rK}{t}}\frac{\|H\|_F}{\sqrt{\lambda_r}} \leq O\Big(\sqrt{\frac{\log(\frac{t}{K\delta})}{\log(\frac{1}{\delta})}}\sqrt{\frac{\lambda_1^*}{\lambda_r^*}}\|(\mathbf{I} - U^*(U^*)^\top)U\|_F + \frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2\log(\frac{t}{K\delta})}{m}}\Big) \tag{319}$$

$$\sqrt{\frac{rK}{t}}\frac{\|H\|_{\infty,2}}{\sqrt{\lambda_r}} \leq O\Big(\sqrt{\frac{\log(\frac{t}{K\delta})}{\log(\frac{1}{\delta})}}\|(\mathbf{I} - U^*(U^*)^\top)U\|\sqrt{\frac{\mu r K}{t}} + \frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{r^2K\log(\frac{t}{K\delta})}{mt}}\Big) \tag{320}$$

A proof is in Section D.3.1. We omit the rest of the proof, as it is same as that of Theorem 8.

D.2 Analysis of task subset selection

Proof of Lemma D.1 (Subset selection). Let $\mathcal{P}_k = [1 + (k-1)t/K, tk/K]$ and

$$\mathcal{T}_k = \big\{\, i \in [1 + (k-1)t/K, tk/K] \mid \sigma_{\min}(U^\top S^{(i)} U) \geq 1/2 \text{ and } \sigma_{\max}(U^\top S^{(i)}U) \leq 2\big\}\,. \tag{321}$$

For all $i \in \mathcal{P}_k$, $X_i = \mathbb{I}(\sigma_{\min}(U^\top S^{(i)}U) \geq 1/2$ and $\sigma_{\max}(U^\top S^{(i)}U) \leq 2)$ be the indicator variable denoting whether index i was select into the subset $\widehat{\mathcal{T}}$.

By Lemma F.7 (by setting $a_j \leftarrow 1$, $x_j \leftarrow U^\top x_j^{(i)}$ for all $j \in [m]$, and $\delta \leftarrow 1/4\mu r$) X_i are i.i.d. Bernoulli random variables with mean $p \geq 1 - \frac{1}{4\mu r}$, if $c\max\Bigg(\sqrt{\frac{r\log(9)+\log(4\mu r)}{m}}, \frac{r\log(9)+\log(4\mu r)}{m}\Bigg) \leq 1/2$, which is satisfied by $m \geq \Omega(r + \log(\mu r))$, for all $i \in \mathcal{P}_k$.

By Hoeffding inequality for Bernoulli random variables, with a probability of at least $1-\delta/3$

$$||\mathcal{T}_k| - pt/K| = \big|\sum_{i\in\mathcal{P}_k} X_i - (1 - \frac{1}{4\,\mu\, r})\frac{t}{K}\big| \leq \frac{t}{K}\sqrt{\frac{K\log(\frac{3}{\delta})}{2t}} \leq \frac{t}{K}O\Big(\frac{1}{4\,\mu\,r}\Big) \tag{322}$$

where we used the fact that $t \geq \Omega(8K\mu^2\, r^2\, \log(\frac{3}{\delta}))$. Therefore

$$\frac{t}{K} - |\mathcal{T}_k| \leq \frac{t}{K}O\Big(\frac{1}{2\,\mu\,r}\Big)\text{ , and } \;\; |\mathcal{T}_k| \leq \Theta\Big(\frac{t}{K}\Big) \tag{323}$$

where we used the fact that $\mu \geq 1$ and $r \geq 1$.

$$\frac{r}{t}\big|z^\top(\sum_{i\in\mathcal{T}_k} v^{*(i)}(v^{*(i)})^\top)z - z^\top(\sum_{i\in\mathcal{P}_k} v^{*(i)}(v^{*(i)})^\top)z\big| \leq \frac{r}{t}(t-\widehat{t})\|V^*\|_{\infty,2}^2 \tag{324}$$

$$\leq \frac{r}{t}O\Big(\frac{t}{2\mu\, r}\Big)\cdot\|V^*\|_{\infty,2}^2 \leq \frac{\lambda_r}{2}\,, \tag{325}$$

for all $z \in \mathbb{R}^r$, where $\lambda_r = \lambda_r(\sum_{i\in\mathcal{P}_k} v^{*(i)}(v^{*(i)})^\top)$. Therefore

$$\lambda_r(\sum_{i\in\mathcal{T}} v^{*(i)}(v^{*(i)})^\top) = \Theta(\lambda_r(\sum_{i\in\mathcal{P}_k} v^{*(i)}(v^{*(i)})^\top))\text{ , and} \tag{326}$$

$$\lambda_1(\sum_{i\in\mathcal{T}} v^{*(i)}(v^{*(i)})^\top) = \Theta(\lambda_1(\sum_{i\in\mathcal{P}_k} v^{*(i)}(v^{*(i)})^\top)) \tag{327}$$

Using approximate incoherence of the partition $\mathcal{P}_k$ (Lemma B.1) we get

$$\|V^*\|_{\infty,2}^2 \leq O\Big(\frac{\mu r K}{t}\Big)\lambda_r(\sum_{i\in\mathcal{P}_k} v^{*(i)}(v^{*(i)})^\top) \tag{328}$$

$$= O\Big(\frac{\mu r K}{t}\Big)\min_{\|z\|=1} z^\top(\sum_{i\in\mathcal{P}_k} v^{*(i)}(v^{*(i)})^\top)z \tag{329}$$

$$\leq O\Big(\frac{\mu r K}{t}\Big)\min_{\|z\|=1} z^\top(\sum_{i\in\mathcal{T}_k} v^{*(i)}(v^{*(i)})^\top)z + O\Big(\frac{\mu r K}{t}\Big)(\frac{t}{K} - |\mathcal{T}_k|)\|V^*\|_{\infty,2}^2 \tag{330}$$

$$\leq O\Big(\frac{\mu r K}{t}\Big)\lambda_r(\sum_{i\in\mathcal{T}_k} v^{*(i)}(v^{*(i)})^\top) + \frac{1}{2}\|V^*\|_{\infty,2}^2 \tag{331}$$

$$\tag{332}$$

This implies that approximate incoherence holds for $\mathcal{T}_k$, $\|V^*\|_{\infty,2}^2 \leq O\Big(\frac{\mu r K}{t}\Big)\lambda_r(\sum_{i\in\mathcal{T}_k} v^{*(i)}(v^{*(i)})^\top) \leq O\big(\frac{\mu r}{|\mathcal{T}_k|}\lambda_r(\sum_{i\in\mathcal{T}_k} v^{(i)}(v^{(i)})^\top)\big)$.

□

D.3 Analysis of update on V

D.3.1 Proof of Lemma D.2

Proof of Lemma D.2. The proof is similar to that of Lemma B.2, but instead of using Lemma B.5 to bound some linear operators, we use the definition of selected task subset $\mathcal{T}_k$ and Lemma D.3 to get that $\|(U^\top S^{(i)}U)^\dagger\| \leq 2$ for all $i \in \mathcal{T}_k$ and with a probability of at least $1-\delta$,

$$\left.\begin{array}{c}\|U^\top S^{(i)}U_\perp U_\perp^\top U^* v^{*(i)}\| \leq \alpha\|U_\perp^\top U^* v^{*(i)}\|,\ \ \text{and}\\ \big\|U^\top z^{(i)}\big\| \leq \sigma\alpha\,,\end{array}\right\}\text{ for all } i\in\mathcal{T}_k \tag{333}$$

where $\alpha = c\sqrt{\frac{r\log(10t/\delta)}{m}}$. We omit the rest of the proof, as it is same as that of Lemma B.2. □

Here we bound the linear operators in the $v^{(i)}$ update.

Lemma D.3. *With a probability of at least $1-\delta$, the following are true for all $i \in [t]$*

$$\|U^\top S^{(i)}U_\perp(U_\perp)^\top U^* v^{*(i)}\| \leq \sqrt{\frac{2cr\log(10t/\delta)}{m}}\|U_\perp U^* v^{*(i)}\|\text{, and} \tag{334}$$

$$\big\|U^\top z^{(i)}\big\| \leq \sigma\sqrt{\frac{2cr\log(10t/\delta)}{m}} \tag{335}$$

Proof. Let $i \in [t]$. Let $b = (U_\perp)^\top U^* v^{*(i)} \in \mathbb{R}^r$

Let $\mathcal{S} = \{v \in \mathbb{R}^r \mid \|v\| = 1\}$ be the set of all real vectors of dimension r with unit Euclidean norm. For $\epsilon \leq 1$, there exists an ϵ-net, $N_\epsilon \subset \mathcal{S}$, of size $(1 + 2/\epsilon)^r$ with respect to the Euclidean norm [40, Lemma 5.2]. That is for any $v' \in \mathcal{S}$, there exists some $v \in N_\epsilon$ such that $\|v' - v\|_F \leq \epsilon$.

Consider a $v \in N_\epsilon$, such that $\|v\|_F = 1$. Now we will prove with high-probability that $\big\langle (U^\top S^{(i)} U_\perp) v, b \big\rangle$ is small. Consider the the following quadratic form

$$v^\top (U^\top S^{(i)} U_\perp) b = \frac{1}{m} \sum_{j \in [m]} v^\top (U^\top x_j^{(i)} (x_j^{(i)})^\top U_\perp) b \overset{d}{=} \|b\| \frac{1}{m} \sum_{j \in [m]} \widetilde{x}_j g_j \tag{336}$$

where $g_j \sim \mathcal{N}(0,1))$ are i.i.d. standard Gaussian random variables and $\widetilde{x}_j = v^\top U^\top x_j^{(i)} \in \mathbb{R}^d$. This follows from the fact that sets of columns of U and $U_\perp$ forms an orthonormal basis.

Note that g_j and $\widetilde{x}_j$ are independent, as U and $U_\perp$ are orthogonal and $U^\top S^{(i)} U$, does not depend on $U_\perp x_j^{(i)}$. We will use the properties of Gaussian random variables to prove that $\|\frac{1}{m} \sum_{j \in [m]} \widetilde{x}_j g_j\|$ concentrates around zero. Note that

$$\frac{1}{m} \sum_{j \in [m]} \widetilde{x}_j g_j \overset{d}{=} \frac{1}{m} \|\widetilde{x}\| g \ , \ \text{ where } g \sim \mathcal{N}(0,1) \tag{337}$$

Then with probability at least $1 - \delta/2t/|N_\epsilon|$, $|g|^2 \leq c \log(2t|N_\epsilon|/\delta)$. Additionally, by definition of $\mathcal{T}_k$ we have

$$\frac{1}{m} \|\widetilde{x}\|^2 = \frac{1}{m} \sum_{j \in [m]} \widetilde{x}_j^2 = v^\top U^\top \big(\frac{1}{m} \sum_{j \in [m]} x_j^{(i)} (x_j^{(i)})^\top \big) U v \leq \sigma_{\max}(U^\top S^{(i)} U) \leq 2 \tag{338}$$

Therefore

$$v^\top (U^\top S^{(i)} U_\perp) b \leq \frac{1}{\sqrt{m}} \|b\| \sqrt{2c} \sqrt{\log(2t|N_\epsilon|/\delta)} \tag{339}$$

For brevity, let $e = (U^\top S^{(i)} U_\perp) b$. Let $v' \in \mathcal{S} \subset \mathbb{R}^r$ be the unit vector parallel to e, such that $(v')^\top e = \|e\| = \max_{\|\widetilde{v}\|=1} \widetilde{v}^\top e$. Then there exists some $v \in N_\epsilon$ such that $\|v' - v\| \leq \epsilon$.

$$\|e\| = (v')^\top e = v^\top e + (v' - v)^\top e \leq v^\top e + \|v' - v\| \|e\| \leq v^\top e + \epsilon \|e\| \tag{340}$$

Re-arranging and setting $\epsilon = 1/2$, and $c \leftarrow 2c$, we get

$$\|(U^\top S^{(i)} U_\perp) b\| \leq \|b\| \sqrt{\frac{2cr \log(10t/\delta)}{m}}, \text{ with a probability of at least } 1 - \delta/2t \tag{341}$$

Using similar arguments we can also prove that with a probability of at least $1 - \delta$

$$\|U^\top z^{(i)}\| = \|\frac{1}{m} U^\top x_j^{(i)} \varepsilon_j^{(i)}\| \leq \sigma \sqrt{\frac{2cr \log(10t/\delta)}{m}}, \text{ with a probability of at least } 1 - \delta/2t \tag{342}$$

Finally taking the union bound over the two bounds over all the tasks in $\mathcal{T}$ gets us the desired result. □

E Corollaries of known results

Theorem 12 (Theorem 3, Tripuraneni et al. 2020)**.** *Let there be t linear regression tasks, each with m samples satisfying Assumptions 1 and 2, and*

$$mt \geq \widetilde{\Omega}\Big(\frac{\lambda_1^*}{\lambda_r^*} \mu d r + \Big(\frac{\sigma}{\sqrt{\lambda_r^*}} \Big)^4 d r^2 \Big) \tag{343}$$

then with a high probability of at least $1 - O((mt)^{-100})$, Method-of-Moments [39, Algorithm 1] outputs an orthonormal matrix $U \in \mathbb{R}^{d \times r}$ such that

$$\|(\mathbf{I} - U^*(U^*)^\top) U\|_2 \leq \tilde{O}\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*} \frac{\mu d r}{mt}} + \Big(\frac{\sigma}{\sqrt{\lambda_r^*}} \Big)^2 \sqrt{\frac{d r^2}{mt}} \Big) \tag{344}$$

and

$$\|(\mathbf{I}-U^*(U^*)^\top)U\|_F \le \tilde{O}\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{\mu d r^2}{mt}} + \Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^2\sqrt{\frac{dr^3}{mt}}\Big). \tag{345}$$

Proof. From the details of the proof of Theorem 3 in [39] we can derive that, with a high probability of at least $1-O((mt)^{-100})$,

$$\|(\mathbf{I}-U^*(U^*)^\top)U\|_2 \tag{346}$$

$$\le \tilde{O}\Big(\sqrt{\frac{dr^2\mathrm{tr}(W^*)\|V^*\|_{\infty,2}^2}{\lambda_r^{*2}\,mt^2}} + \frac{dr\|V^*\|_{\infty,2}^2}{\lambda_r^*\,mt} + \sigma\Big(\sqrt{\frac{dr^2\mathrm{tr}(W^*)}{\lambda_r^{*2}\,mt^2}} + \frac{dr\|V^*\|_{\infty,2}}{\lambda_r^*\,mt}\Big) + \sigma^2\Big(\sqrt{\frac{dr^2}{\lambda_r^{*2}\,mt}} + \frac{dr}{\lambda_r^*\,mt}\Big)\Big) \tag{347}$$

$$\le \tilde{O}\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{\mu dr}{mt}} + \frac{\mu dr}{mt} + \frac{\sigma}{\sqrt{\lambda_r^*}}\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{dr}{mt}} + \frac{\sqrt{\mu}\,dr}{mt}\Big) + \Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^2\Big(\sqrt{\frac{dr^2}{mt}} + \frac{dr}{mt}\Big) \tag{348}$$

$$\le \tilde{O}\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{\mu dr}{mt}} + \frac{\sigma}{\sqrt{\lambda_r^*}}\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{dr}{mt}} + \Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^2\sqrt{\frac{dr^2}{mt}}\Big) \tag{349}$$

$$\le \tilde{O}\Big(\sqrt{\frac{\lambda_1^*}{\lambda_r^*}\frac{\mu dr}{mt}} + \Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^2\sqrt{\frac{dr^2}{mt}}\Big) \tag{350}$$

where $\|V\|_{\infty,2} = \max_{i\in[t]}\|v^{(i)}\|$, and the second-last inequality uses the fact that $mt \ge \widetilde{\Omega}(\mu dr)$ and last inequality uses the fact that $\frac{\lambda_1^*}{\lambda_r^*} \le \mu r$. Additionally we require that

$$mt \ge \widetilde{\Omega}\Big(\frac{\lambda_1^*}{\lambda_r^*}\mu dr + \Big(\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)^4 dr^2\Big) \tag{351}$$

□

Theorem 13. *[39, Theorem 5] Let $r \le d/2$ and $mt \ge r(d-r)$, then for all V^*, w.p. $\ge 1/2$*

$$\inf_{\widehat{U}}\sup_{U\in \mathrm{Gr}_{r,d}} \frac{\|(\mathbf{I}-U^*(U^*)^\top)\widehat{U}\|_F}{\sqrt{r}} \;\ge\; \Omega\Big(\Big(\frac{\lambda_r^*}{\lambda_1^*}\frac{\sigma}{\sqrt{\lambda_r^*}}\Big)\sqrt{\frac{d\,r}{m\,t}}\Big)\,,$$

where $G_{r,d}$ is the Grassmannian manifold of r-dimensional subspaces in $\mathbb{R}^d$, the infimum for $\widehat{U}$ is taken over the set of all measurable functions that takes mt samples in total from the model in Section 2 satisfying Assumption 1 and 2.

Proof. The proof is very similar to that of Theorem 5 of [39]. The main difference is that instead of lower bounding error in spectral norm we have to bound the distance in the Frobenius norm. However, the rest of the details are almost the same, hence we omit a full proof. □

F Technical Lemmas

This section contains some technical lemmas used in this paper.

Lemma F.1. *For a real matrix $A \in \mathbb{R}^{m\times n}$ and a real symmetric positive semi-definite (PSD) matrix $B \in \mathbb{R}^{n\times n}$, the following holds true: $\sigma_{\min}^2(A)\lambda_{\min}(B) \le \lambda_{\min}(ABA^\top)$, where $\sigma_{\min}(\cdot)$ and $\lambda_{\min}(\cdot)$ represents the minimum singular value and minimum eigenvalue operators respectively.*

Proof. The proof directly follows from the definitons of $\sigma_{\min}$ and $\lambda_{\min}$. Since B is a PSD matrix, therefore $ABA^\top$ is also PSD, i.e. $\lambda_{\min}(ABA^\top) \ge 0$. This is because since B is PSD, it has a PSD matrix square root $B^{1/2}$ such that $B = (B^{1/2})^\top B^{1/2}$ and $B^{1/2}$ is PSD. Then

$$z^\top ABA^\top z = z^\top A(B^{1/2})^\top B^{1/2}A^\top z = \|B^{1/2}A^\top z\|^2 \ge 0 \tag{352}$$

First assume that $\sigma_{\min}(A) > 0$, then

$$\lambda_{\min}(ABA^\top) = \min_{\|z\|=1} z^\top ABA^\top z \tag{353}$$

$$= \sigma^2_{\min}(A) \min_{\|z\|=1} (\frac{A^\top z}{\sigma_{\min}(A)})^\top B(\frac{A^\top z}{\sigma_{\min}(A)}) \tag{354}$$

$$\geq \sigma^2_{\min}(A) \min_{1\leq\|z\|\leq\frac{\sigma_{\max}(A)}{\sigma_{\min}(A)}} z^\top Bz \tag{355}$$

$$\geq \sigma^2_{\min}(A) \min_{\|z\|=1} z^\top Bz \tag{356}$$

$$= \sigma^2_{\min}(A)\lambda_{\min}(B) \tag{357}$$

The second last inequality above follows from the fact that B is a PSD matrix,i.e. $\min_{\|z\|=1} z^\top Bz = \lambda_{\min}(B) \geq 0$. Secondly if $\sigma_{\min}(A) = 0$, then A is rank deficient and hence $ABA^\top$ is also rank deficient, i.e. $\lambda_{\min}(ABA^\top) = 0$. Therefore $\lambda_{\min}(ABA^\top) = 0 = \sigma^2_{\min}(A)\lambda_{\min}(B)$. □

Lemma F.2 (Weyl's inequality [1])**.** *For three real r-rank matrices, satisfying $A - B = C$, Weyl's inequality [1, Theorem 3.6], tells that*

$$\sigma_k(A) - \sigma_k(B) \leq \|C\|, \text{for all } k \in [r] \tag{358}$$

where $\sigma_k(\cdot)$ is the k-th largest singular value operator.

Lemma F.3 (a variant of Woodburry matrix identity [18])**.** *For linear operators A and B such that A and $A + B$ are invertible, then*

$$(A + B)^{-1} - A^{-1} = -A^{-1}B(A + B)^{-1} \tag{359}$$

Lemma F.4. *Let $U \in \mathbb{R}^{d\times r}$ and $U^* \in \mathbb{R}^{d\times r}$ be two orthonormal matrices. Let $\{\sin\theta_j(U, U^*)\}_{j=1}^r$ be the singular values of $(U^*)^\top U$. Then following are true.*

$$\|U - U^*(U^*)^\top U\|_F \geq \|\mathbf{I} - (U^*)^\top U\|_F\,, \tag{360}$$

$$\|U - U^*(U^*)^\top U\|_F \geq r - \|(U^*)^\top U\|_F^2 \geq \sum_{k\in[r]} \sin^2\theta_k(U, U^*), \tag{361}$$

$$\|(\mathbf{I} - U^*(U^*)^\top)U\| = \|(U^*{}_\perp)^\top U\| = \|U_\perp^\top U^*\| = \|(\mathbf{I} - U(U)^\top)U^*\|, \tag{362}$$

$$\|(\mathbf{I} - U^*(U^*)^\top)U\|_F = \|(U^*{}_\perp)^\top U\|_F = \|U_\perp^\top U^*\|_F = \|(\mathbf{I} - U(U)^\top)U^*\|_F, \text{ and} \tag{363}$$

$$\sigma_r((U^*)^\top U) \geq \sqrt{1 - \|(\mathbf{I} - U^*(U^*)^\top)U\|} \tag{364}$$

Proof.

$$\|U - U^*(U^*)^\top U\|_F^2 = \left\langle U - U^*(U^*)^\top U, U - U^*(U^*)^\top U\right\rangle \tag{365}$$

$$= \langle U, U\rangle - 2\left\langle U^*(U^*)^\top U, U\right\rangle + \left\langle U^*(U^*)^\top U, U^*(U^*)^\top U\right\rangle \tag{366}$$

$$= r - 2\mathrm{tr}(((U^*)^\top U)^\top((U^*)^\top U)) + \mathrm{tr}(((U^*)^\top U)^\top((U^*)^\top U)) \tag{367}$$

$$= r - \mathrm{tr}(((U^*)^\top U)^\top((U^*)^\top U)) \tag{368}$$

$$= r - \sum_{k\in[r]} \cos^2\theta_k(U, U^*) = \sum_{k\in[r]} \sin^2\theta_k(U, U^*) \geq \sin^2\theta_1(U, U^*) \tag{369}$$

$$\geq \sum_{k\in[r]} (1 - \cos^2\theta_k(U, U^*)) \tag{370}$$

$$\geq \sum_{k\in[r]} (1 - \cos\theta_k(U, U^*))^2 \tag{371}$$

$$= \|\mathbf{I} - (U^*)^\top U\|_F^2 \tag{372}$$

$$\|U_\perp^\top U^*\| = \sigma_{\max}(U_\perp^\top U^*) = \sqrt{\lambda_{\max}((U^*)^\top U_\perp U_\perp^\top U^*)} \tag{373}$$

$$= \sqrt{\lambda_{\max}((U^*)^\top U_\perp U_\perp^\top U_\perp U_\perp^\top U^*)} = \|U_\perp U_\perp^\top U^*\| = \|(\mathbf{I} - UU^\top)U^*\| \tag{374}$$

Note that for $\|z\|=1$

$$1 = z^\top U^\top U z = z^\top U^\top U^*(U^*)^\top U z + z^\top U^\top {U^*}_\perp ({U^*}_\perp)^\top U z \tag{375}$$

$$\implies 1 - z^\top U^\top U^*(U^*)^\top U z = z^\top U^\top {U^*}_\perp ({U^*}_\perp)^\top U z \tag{376}$$

$$\implies 1 - \min_{\|z\|=1} z^\top U^\top U^*(U^*)^\top U z = \max_{\|z\|=1} z^\top U^\top {U^*}_\perp ({U^*}_\perp)^\top U z \tag{377}$$

$$\implies 1 - \sigma^2_{\min}((U^*)^\top U) = \|({U^*}_\perp)^\top U\|^2 \tag{378}$$

Therefore

$$\sigma^2_{\min}(U^\top U^*) + \|U_\perp^\top U^*\|^2 = 1 = \sigma^2_{\min}((U^*)^\top U) + \|({U^*}_\perp)^\top U\|^2 \implies \|U_\perp^\top U^*\| = \|({U^*}_\perp)^\top U\| \tag{379}$$

Rest of the equality can be obtained in a similar fashion using the above two relations.

$$\|U_\perp^\top U^*\|_F^2 = \mathrm{tr}((U^*)^\top U_\perp U_\perp^\top U^*) = \mathrm{tr}((U^*)^\top (\mathbf{I} - UU^\top) U^*) \tag{380}$$

$$= \mathrm{tr}((U^*)^\top (\mathbf{I} - UU^\top)^2 U^*) \tag{381}$$

$$= \|(\mathbf{I} - UU^\top) U^*\|_F^2 \tag{382}$$

$$= \|(\mathbf{I} - U^*(U^*)^\top) U\|_F^2 = \|({U^*}_\perp)^\top U\|_F^2 \tag{383}$$

Let $E = (\mathbf{I} - U^*(U^*)^\top)U$ and $Q = (U^*)^\top U$. Then $U^\top E = \mathbf{I} - Q^\top Q$. Then by Weyl's inequality (Lemma F.2, by setting $A \leftarrow \mathbf{I}$, $B \leftarrow Q^\top Q$, and $C \leftarrow U^\top E$) we get that

$$1 - \sigma_r(Q)^2 = \sigma_r(\mathbf{I}) - \sigma_r(Q^\top Q) \le \|U^\top E\| \le \|U\|\|E\| \le \|(\mathbf{I} - U^*(U^*)^\top)U\| \tag{384}$$

This implies that $\sigma_r((U^*)^\top U) \ge \sqrt{1 - \|(\mathbf{I} - U^*(U^*)^\top)U\|}$ ☐

Lemma F.5 (Hanson-Wright inequality, Theorem 6.2.1 [41])**.** *Let $x_1, \dots, x_m \sim \mathcal{N}(0, \mathbf{I}_{d\times d})$ be m i.i.d. standard isotropic Gaussian random vectors of dimension d. Then, for some universal constant $c \ge 0$, the following holds true with a probability of at least $1-\delta$.*

$$\left|\frac{1}{m}\sum_{j=1}^m x_j^\top A_j x_j - \frac{1}{m}\sum_{j=1}^m \mathrm{tr} A_j\right| \le c \max\left(\sqrt{\sum_{j=1}^m \|A_j\|_F^2 \frac{\log(1/\delta)}{m^2}}, \max_{j=1,\dots,n} \|A_j\|_2 \frac{\log(1/\delta)}{m}\right) \tag{385}$$

Lemma F.6. *Let $x_1, \dots, x_m \sim \mathcal{N}(0, \mathbf{I}_{d\times d})$ be m i.i.d. standard isotropic Gaussian random vectors of dimension d. Then, for some universal constant $c \ge 0$, the following holds true with a probability of at least $1-\delta$.*

$$\left|\frac{1}{m}\sum_{j=1}^m a^\top (x_j x_j^\top) b - a^\top b\right| \le c\|a\|\|b\| \max\left(\sqrt{\frac{\log(1/\delta)}{m}}, \frac{\log(1/\delta)}{m}\right) \tag{386}$$

Proof. First notice that $a^\top(x_j x_j^\top)b = \mathrm{tr}(a^\top(x_j x_j^\top)b) = \mathrm{tr}(x_j^\top b a^\top x_j) = x_j^\top b a^\top x_j$ and $a^\top b = \mathrm{tr}(ba^\top)$. Then desired result follows from Lemma F.5, by setting $A_j = ba^\top$. . ☐

Lemma F.7. *Let $x_1, \dots, x_m \sim \mathcal{N}(0, \mathbf{I}_{d\times d})$ be m i.i.d. standard isotropic Gaussian random vectors of dimension d. Then, for some universal constant $c \ge 0$, the following holds true with a probability of at least $1-\delta$.*

$$\left\|\frac{1}{m}\sum_{j=1}^m a_j x_j x_j^\top - \frac{1}{m}\sum_{j=1}^m a_j \mathbf{I}\right\| \le c \max\left(\frac{\|a\|_2}{\sqrt{m}}\sqrt{\frac{d\log(9) + \log(1/\delta)}{m}}, \|a\|_\infty \frac{d\log(9) + \log(1/\delta)}{m}\right) \tag{387}$$

Proof. For $\epsilon \le 1$, consider a unit vector $u \in N_\epsilon$ from the ϵ-net of size $|N_\epsilon| = (1+2/\epsilon)^d$, of the sphere $\mathbb{S}^{d-1}$ [40, Lemma 5.2]. That is for any $u' \in \mathbb{S}^{d-1}$, there exists some $u \in N_\epsilon$ such that $\|u' - u\| \le \epsilon$.

Now we will prove a concentration for $\frac{1}{m}\sum_{j=1}^m a_j u^\top x_j x_j^\top u - \frac{1}{m}\sum_{j=1}^m a_j$. Notice that, $a_j u^\top (x_j x_j^\top) u = a\mathrm{tr}(u^\top (x_j x_j^\top) u) = a_j \mathrm{tr}(x_j^\top u u^\top x_j) = x_j^\top (a_j u u^\top) x_j$ and $\mathrm{tr}(a_j u u^\top) = a_j$. Then, by Hanson-Wright inequality (Lemma F.5), for some universal constant $c \ge 0$, the following holds true with a probability of at least $1 - \delta'$.

$$\left|\frac{1}{m}\sum_{j=1}^m a_j u^\top x_j x_j^\top u - \frac{1}{m}\sum_{j=1}^m a_j\right| \le c \max\left(\frac{\|a\|_2}{\sqrt{m}}\sqrt{\frac{\log(1/\delta')}{m}}, \|a\|_\infty \frac{\log(1/\delta')}{m}\right) \tag{388}$$

This implies that, through union bound, for the matrix $A' = \frac{1}{m}\sum_{j=1}^m a_j x_j x_j^\top - \frac{1}{m}\sum_{j=1}^m a_j \mathbf{I}$ the following holds true with probability at least $1 - \delta$

$$u^\top A' u \le c \max\left(\frac{\|a\|_2}{\sqrt{m}}\sqrt{\frac{\log(|N_\epsilon|/\delta)}{m}}, \|a\|_\infty \frac{\log(|N_\epsilon|/\delta)}{m}\right), \quad \text{any } u \in N_\epsilon \tag{389}$$

Let $u' \in \mathbb{S}^{d-1}$ be the top singular-value of A', then there exists some $u \in N_\epsilon$ such that $\|u' - u\| \le \epsilon$.

$$\sigma_{\max}(A') = (u')^\top A' u' = (u' - u)^\top A' u' + u^\top A' u + u^\top A' u \tag{390}$$

$$\le \|u' - u\| \sigma_{\max}(A') \|u'\| + \|u\| \sigma_{\max}(A') \|u' - u\| + u^\top A' u \tag{391}$$

$$\tag{392}$$

Re-arranging and setting $\epsilon = 1/4$ and setting $c \leftarrow 2c$, we get

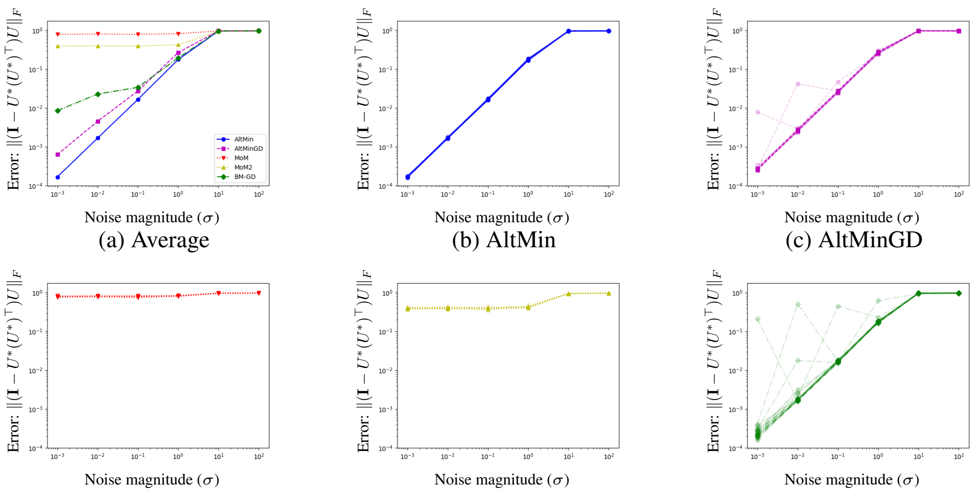
$$\sigma_{\max}(A') \le \frac{u^\top A' u}{1 - 2\epsilon} \le 2c \max\left(\frac{\|a\|_2}{\sqrt{m}}\sqrt{\frac{d\log(9) + \log(1/\delta)}{m}}, \|a\|_\infty \frac{d\log(9) + \log(1/\delta)}{m}\right) \tag{393}$$

□

G Experimental details

We empirically compare the performance of our methods AltMinGD (Algorithm 1) and exact minimization variant AltMin (Algorithm 3 in Appendix), two different versions of Method-of-Moments (MoM [39], MoM2 [26]), and simultaneous gradient descent on (U, V) using the Burer-Monteiro factorized loss (4) (BM-GD [39]). We generate data samples with dimension $d = 100$ and generate random subspace U^* of rank $r = 5$. We sample the task specific true regression parameter from the standard isotropic Gaussian distribution, i.e. $v^{*(i)} \sim \mathcal{N}(0, \mathbf{I})$. In all the figures, the magenta dashed line with square marker represents AltMinGD, the blue straight line with circular marker denotes the AltMin , the red dotted line with downwards pointing triangular marker denotes the MoM, the yellow dotted line with upwards pointing triangular marker represents the MoM2, and the green dashed and dotted line with diamond marker represents the BM-GD. In all the figures we plot the subspace estimation error of the output U of the algorithms. The error is calculated using the rescaled Frobenius norm $\|(\mathbf{I} - U^*(U^*)^\top)U\|_F/\sqrt{r}$, which takes a value in the interval $[0, 1]$.

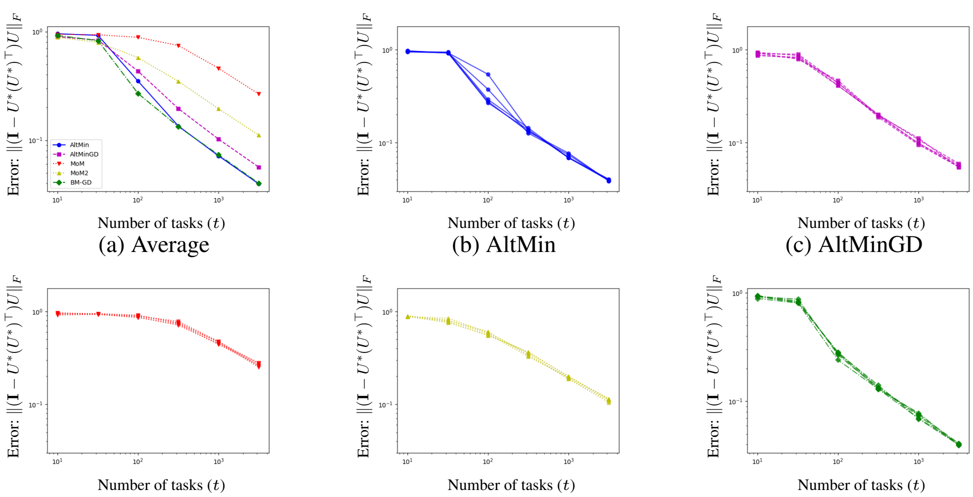
Figure 1a plots subspace distance against the standard deviation σ of the regression noise, $\varepsilon_j^{(i)} \sim \mathcal{N}(0, \sigma^2)$; see (2). We vary σ from 10^{-3} to 10^2, while fixing the number of tasks at $t = 200$ and the number of samples per task at $m = 25$. We initialize AltMinGD, AltMin, and BM-GD uniformly at random and run them for $K = 100$, $K = 20$ and $K = 500$ iterations, respectively. AltMinGD and BM-GD use stepsizes $\eta = 1.0$ and $\eta = 0.1$ respectively. In Figure 3a, we plot the average over 25 trials for each algorithm. the individual trials for each algorithm are plotted in Figure 3. Figure 1b plots the subspace error against the number of tasks t. We vary t from 10 to 3163, while the number of samples per task is fixed at $m = 25$ and $\sigma = 1$. We initialize AltMinGD, AltMin, and BM-GD uniformly at random and run them for $K = 200$, $K = 20$ and $K = 500$ iterations, respectively. AltMinGD and BM-GD use stepsizes $\eta = 1.0$ and $\eta = 0.1$ respectively. In Figure 4a, we plot the average over 5 trials for each algorithm. The individual trials for each algorithm are plotted in Figure 4. In Figure 1c, we plot the the error against the number samples per tasks m. We vary m from



(a) Average (b) AltMin (c) AltMinGD

(d) MoM (e) MoM2 (f) BM-GD

Figure 3: Individual trials of each algorithm when varying noise magnitude σ



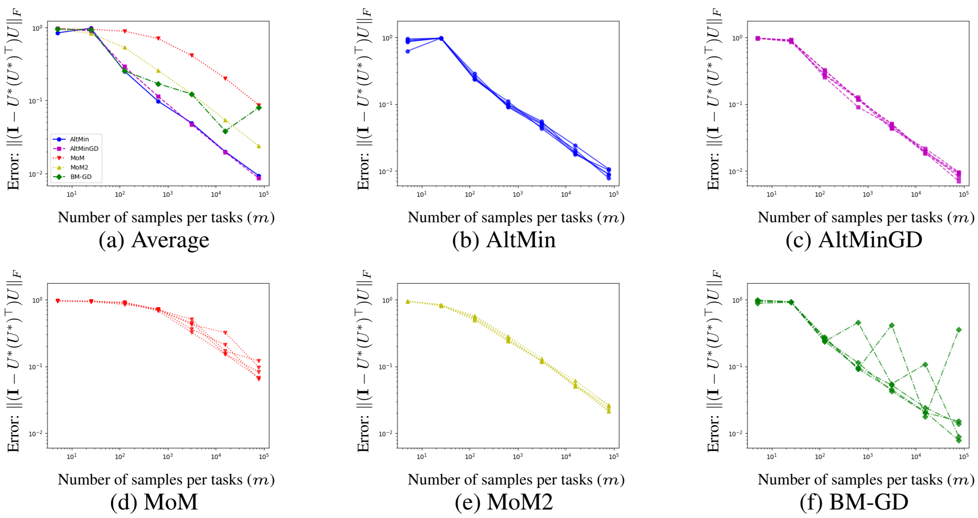
(a) Average (b) AltMin (c) AltMinGD

(d) MoM (e) MoM2 (f) BM-GD

Figure 4: Individual trials of each algorithm when varying number of tasks t

5 to 78125, while fixing the number of tasks at $t = 20$ and the standard deviation of the regression noise at $\sigma = 1$. We initialize AltMinGD, AltMin, and BM-GD uniformly at random from the set of all orthonormal rank-r matrices and run them for $K = 1000$, $K = 20$ and $K = 500$ iterations, respectively. AltMinGD and BM-GD use stepsizes $\eta = 0.2$ and $\eta = 0.1$ respectively. In Figure 5a, we plot the average over 5 trials for each algorithm. The individual trials for each algorithm are also plotted in Figure 5. We see that BM-GD is very ustable even at a lower or comparable stepsize than AltMinGD.

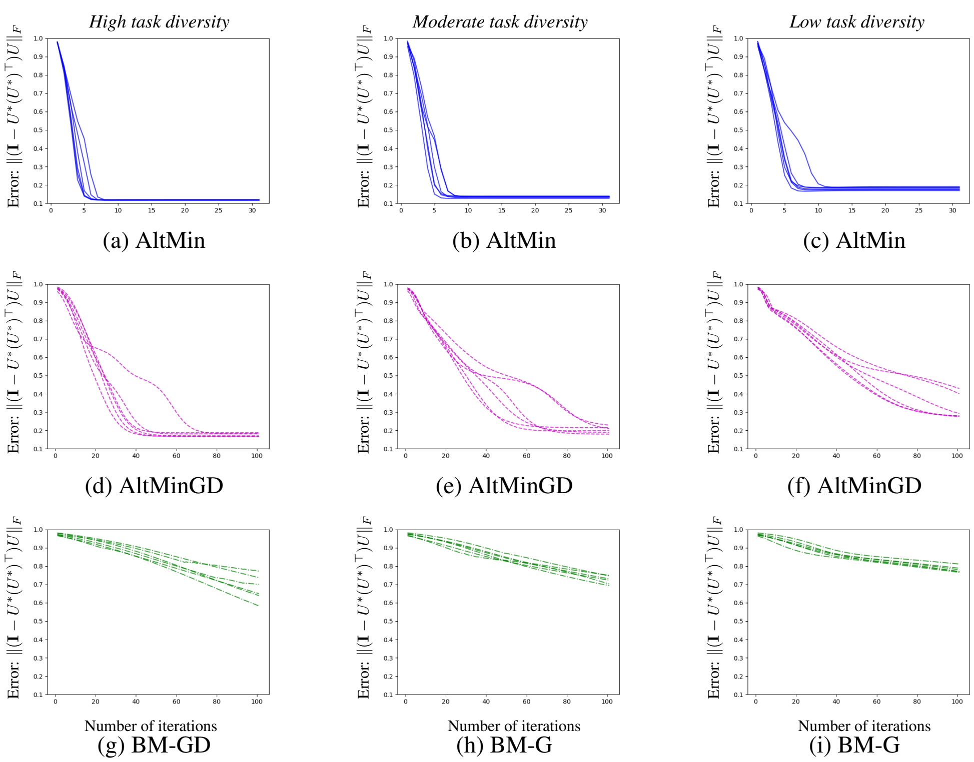
In Figure 2 we plot the subspace estimation error against the number of iterations of the iterative method AltMinGD, AltMin, and BM-GD for varying levels of task diversity/incoherence (Assumption 2). We vary task diversity while fixing the random noise magnitude at $\sigma = 1$, the number of tasks at $t = 200$, and the number of samples per task at $m = 25$. We vary the diversity by generating a fraction of the true task regression parameters from the standard isotropic Gaussian distribution, i.e. $v^{(i)} \sim \mathcal{N}(0, \mathbf{I})$, and setting the rest of them as $v^{(i)} = \sqrt{r}e_1 = [\sqrt{r}, 0, \ldots, 0]^\top$. For high task diversity all task parameters are generated randomly, for moderate task diversity 0.4 fraction of the task parameters are set as $\sqrt{r}e_1$, and for low task diversity 0.8 fraction of the task parameters are set



(a) Average (b) AltMin (c) AltMinGD

(d) MoM (e) MoM2 (f) BM-GD

Figure 5: Individual trials of each algorithm when varying number of samples per task m



(a) AltMin (b) AltMin (c) AltMin

(d) AltMinGD (e) AltMinGD (f) AltMinGD

(g) BM-GD (h) BM-G (i) BM-G

Figure 6: Individual trials of each iterative algorithm when plotting against the number of iterations for different task diversities.

as $\sqrt{r}e_1$. In Figure 2, we plot the average over 6 trials for each algorithm. The individual trials for each algorithm for each task diversity are plotted in Figure 6.