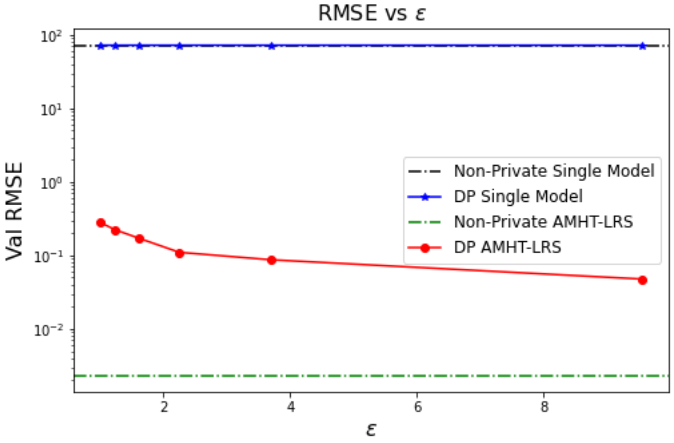
A ADDITIONAL EXPERIMENTS AND SETUP

A.1 SYNTHETIC EXPERIMENTS WITH PRIVATE ALGORITHMS

DP LRS Simulations: We conduct the experiment on a synthetic dataset: for each task $i \in [t]$, we generate $m = 100$ samples $\{(\mathbf{x}_j^{(i)}, y_j^{(i)})\}_{j\in[m]}$ where $\mathbf{x}_j^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d\times d})$, $y_j^{(i)} = \langle \mathbf{x}_j^{(i)}, \mathbf{u}^{\star} w^{\star(i)} + \mathbf{b}^{\star(i)} \rangle$ and $w^{\star(i)} = 1$ is fixed for simplicity. We select number of tasks $t = 5000$, $d = 10$, data dimension k, ζ, and both the column and row sparsity level of $\{\mathbf{b}^{\star(i)}\}_{i\in[t]}$ to be 2. We sample $\mathbf{u}^{\star}$ uniformly from the unit sphere and the non-zero elements of $\{\mathbf{b}^{\star(i)}\}_{i\in[t]}$ are sampled i.i.d. from $\mathcal{N}(0, 1)$ with the indices of zeros selected randomly.

We run the algorithm for 15 epochs and use RDP sequential composition to compute the privacy risk accumulated over the epochs. We set minimum possible clipping norm values for $\mathsf{A}_1, \mathsf{A}_2$ and A_3 s.t. a majority portion of samples don't get clipped. We fix $\delta = 10^{-5}$. For hyperparameter tuning, we perform a search pick the values which give the minimum RMSE on the validation set. Finally we plot the RMSE on the Test Set for different values of ϵ in .



(a) Overall comparison of RMSE.

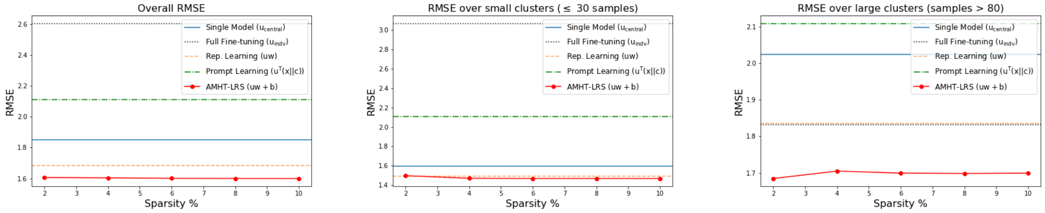
Figure 3: Comparison of Overall RMSE on the simulated data for both the private and non-private versions of AMHT-LRS and the Single Model Baseline.

We note that while DP AMHT-LRS performs quite well for each ϵ, both the private and non-private Single Model baselines fare badly even on higher values of ϵ. Further, DP AMHT-LRS is able to achieve RMSE comparable to its non-private version by $\epsilon \approx 2$ mark.

A.2 EXPERIMENTS WITH LINEAR MODELS

A.2.1 NETFLIX DATASET

We consider the Netflix Challenge dataset comprising of 17k users and 480k movies where ratings are provided as integers on a scale of $1 - 5$. We choose the top 200 users who have rated the most movies and top 200 movies that have been rated the most and consider the 200×200 rating matrix restricted to these sets of top users and movies - this rating matrix comprises approximately 35k ratings. We perform the train-validation split in the following way: for 70 users, we keep 10% of their ratings in the training set (small data/user); for 70 users, we kept 50% of their ratings in their training set (medium data/user) and for the rest of 60 users, we kept 90% of their ratings in their training set (large data/user). All the observed ratings restricted to the 200×200 ratings matrix that are absent in the training set is inserted into the validation set. By using standard low rank matrix completion techniques (Chen et al., 2020b), we complete the ratings matrix by minimizing the MSE w.r.t to the entries in the training set with a nuclear norm regularizer. Following this, we compute SVD $\mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{T}}$ and take the first 50 columns of $\mathbf{V}$; this results in a truncated 200×50 dimensional



(a) Overall RMSE (b) RMSE for data-starved tasks (c) RMSE for data surplus tasks

Figure 4: Decrease in RMSE on Netflix validation data for AMHT-LRS algorithm on increase in fine-tunable parameters. Note that AMHT-LRS outperforms other baselines for both data-starved and data-surplus tasks.

matrix $\widehat{\mathbf{V}}$ where each row corresponds to a 50-dimensional embedding of each movie. Note that we ensure there is no data leakage while creating the embeddings.

For each task representing each user, the samples consist of (movie embedding, rating) tuples; the response is the average rating of the movie given by users in that task. We use the training data to learn the different models (with some hyper-parameter tuning) mentioned earlier and use them to predict the ratings in the validation data.

Empirical Observations on Netflix validation data:

The overall average validation RMSE for AMHT-LRS and the different baselines that we consider is shown in Fig. 4a against percentage of fine-tunable parameters used by the model. As in the Movielens dataset, with respect to the single model as reference, in the linear rank-1 case, the representation learning and the prompt learning based baselines have 1 and 50 additional parameters per task respectively; they are unable to personalize well. In contrast, with only $4\%(=2)$ additional parameters per task, AMHT-LRS has smaller RMSE than fully fine-tuned model, which require 200x more parameters. Note that AMHT-LRS outperforms other baselines for both data-starved and data-surplus tasks.

A.2.2 JESTER DATASET

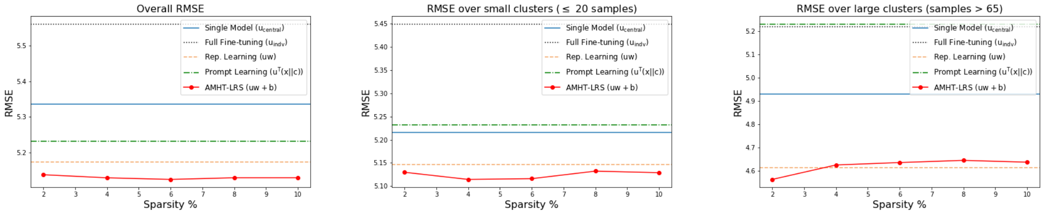
The Jester dataset comprises of 4.1m ratings from 73k users for 100 jokes with each rating being on a scale of -10.0 to $+10.0$. We choose 100 users who have rated all the 100 jokes and consider the 100×100 rating matrix restricted to these users and jokes - this rating matrix is entirely filled. Similar to the Netflix dataset, we perform the train-validation split in the following way: for 30 users, we keep 10% of their ratings in the training set (small data/user); for 40 users, we kept 50% of their ratings in their training set (medium data/user) and for the rest of 30 users, we kept 90% of their ratings in their training set (large data/user). We use the training data to learn the different models (with some hyper-parameter tuning) and use them to predict the ratings in the validation data.

Empirical Observations on Jester validation dataset:

The conclusions are mostly similar as in the case of Netflix dataset. The overall average validation RMSE for AMHT-LRS and the different baselines that we consider is shown in Fig. 5a against percentage of fine-tunable parameters used by the model. Again, with respect to the single model as reference, in the linear rank-1 case, the representation learning and the prompt learning based baselines have 1 and 50 additional parameters per task respectively; they are unable to personalize well. In contrast, with only $2\%(=1)$ additional parameter per task, AMHT-LRS has smaller RMSE than fully fine-tuned model, which require 100x more parameters and the other baselines. Note that our method outperforms other baselines for both data-starved and data-surplus tasks.

A.3 EXPERIMENTS WITH NEURAL NETWORKS

As described in Section 3, our techniques/approaches can be extended to complex classes of non-linear model. To demonstrate this, we fix the class of models to Neural Networks (denoted by $\mathsf{F} : \mathbb{R}^d \to \mathbb{R}$). Similar to Section 3, we consider the following baselines: 1) **Single Model ($\mathsf{F}(x; \mathsf{u}_{\mathbf{central}})$)**: learns a single Neural Network model for all tasks, 2) **Full Fine-tuning ($\mathsf{F}(x; \mathsf{u}_{\mathbf{indv}})$)** Learns a separate fully-trained Neural Network model for each task, 3) **Rep. Learning** ($\mathsf{F}(x; \mathsf{uw})$): Learns separate Neural Networks for each task such that the NN parameters of each task lie on a low dimensional manifold.



(a) Overall RMSE (b) RMSE for data-starved tasks (c) RMSE for data surplus tasks

Figure 5: Decrease in RMSE on Jester validation data for AMHT-LRS algorithm on increase in fine-tunable parameters. Note that AMHT-LRS outperforms other baselines.

4) **Prompt Learning ($\mathsf{F}(x \mid\mid c; \mathsf{u}_{\textbf{central}})$):** The covariate is concatenated with a trainable embedding of the corresponding task and a single Neural network model is trained with the modified covariates. AMHT-LRS ($\mathsf{F}(x; \mathsf{uw} + \mathsf{b}_{\text{sparse}})$) itself trains a separate Neural Network model for each task but assumes that the entire set of parameters of the Neural networks for each task can be represented as a Low Rank+Sparse matrix.

As in the case of linear models, we use the training data to learn the parameters of different models (with some hyper-parameter tuning) described above and use them to predict the ratings in the validation data. The overall average validation RMSE for AMHT-LRS and the 4 different baselines (modified for neural networks) that we consider against different amounts of sparsity in $\mathsf{b}_{\text{sparse}}$ is computed (shown in figures for each of three datasets that we experiment with). The memory footprint of the different methods (for each of the three datasets) has been provided in Table A.3.3.

More Details on Experimental Setup: To compute the single model and fully fine-tuned model metrics, we used batched gradient descent. To compute the low rank model metrics, we performed alternating optimization as per the algorithm described in (Thekumparampil et al., 2021). Finally, to compute AMHT-LRS metrics, we used rank-1 case of Algorithm 1 and used an L_2 regularization for each $\mathbf{b}^{(i,\ell)}$. For all the gradient based methods, we used Adam Optimizer with weight decay and learning rate scheduler. We experimented with Cosine Annealing and Decay on Plateau schedulers. We performed a search over learning rates, L_2 weight decay values and learning rate scheduler hyperparameters (decay factor for Decay on Plateau and window size for Cosine Annealing) and reported the model metrics which gave the best overall RMSE on the validation dataset.

A.3.1 NETFLIX DATASET

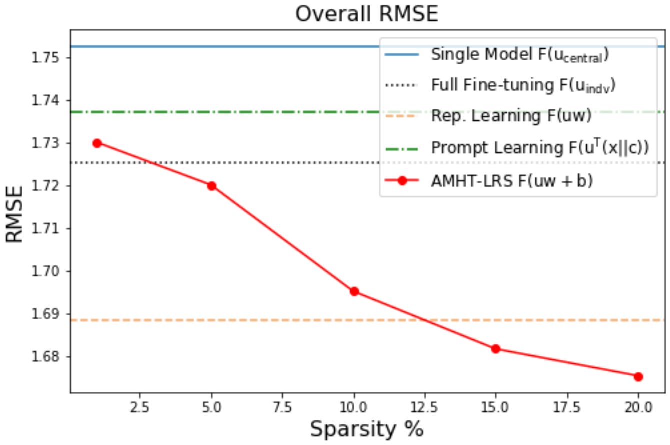
LRS with 2 layer Neural Net for Netflix: For the Neural Network (NN) experiments on Netflix dataset, we consider the function class F - a 2 layer Neural Net with a single hidden layer of 50 neurons and tanh activation. The training and the validation data on the Netflix dataset is same as created for the linear models (see Section A.2.1). The comparison of validation RMSE of AMHT-LRS and all the 4 baselines corresponding to the Netflix dataset is given in Figure 6a.

Observe that AMHT-LRS outperforms the rest of the baselines with a small memory overhead (see Table A.3.3). In particular, the improvement in performance is achieved along with a significant improvement in memory cost compared to Full fine-tuning - AMHT-LRS (with only 1% sparsity/tunable parameters for each user) outperforms the Full-Finetuning baseline with only 2% of the corresponding number of parameters.

A.3.2 JESTER DATASET

LRS with 2 layer Neural Net for Jester: For the Neural Network (NN) experiments on Jester dataset, we consider the function class F - a 2 layer Neural Net with a single hidden layer comprising 50 neurons and tanh activation. As before, the training and the validation data is the same that was created for the case of linear models (see Sec. A.2.2). The comparison of validation RMSE of AMHT-LRS and all the 4 baselines corresponding to the Jester dataset is given in Figure 7a.

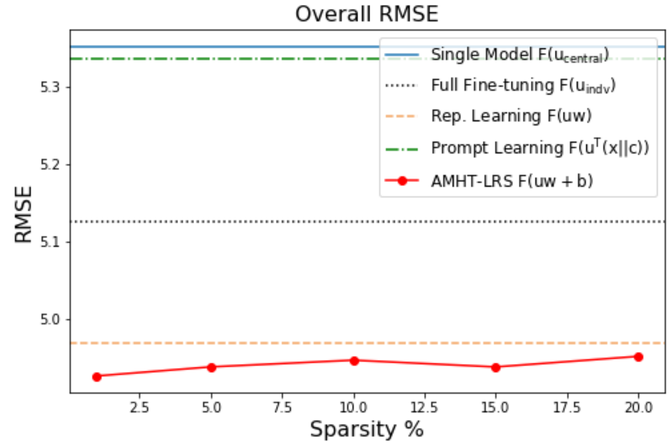
Again, we observe that AMHT-LRS outperforms the rest of the baselines with a small memory overhead (see Table A.3.3). As before, the improvement in performance is achieved along with a significant improvement in memory cost compared to Full fine-tuning - AMHT-LRS (with only 2%



(a) Overall comparison of RMSE.

Figure 6: Comparison of Overall RMSE on the Netflix validation data achieved by AMHT-LRS and the different baselines we consider where the training is modified for toy Neural Network models with a single hidden layer of 50 neurons and tanh activation. AMHT-LRS outperforms other baselines along with a significantly smaller memory footprint (see Table A.3.3 for exact numbers on model parameters).

sparsity/tuneable parameters for each user) outperforms the Full-Finetuning baseline with only 1.5% of the corresponding number of parameters.

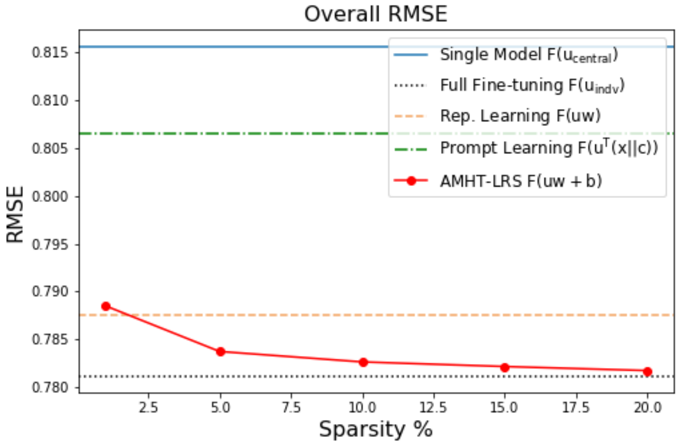


(a) Overall comparison of RMSE.

Figure 7: Comparison of Overall RMSE on the Jester validation data achieved by AMHT-LRS and the different baselines we consider where the training is modified for toy Neural Network models with a single hidden layer of 50 neurons and tanh activation. AMHT-LRS outperforms other baselines along with a significantly smaller memory footprint (see Table A.3.3 for exact numbers on model parameters).

A.3.3 MOVIELENS DATASET

LRS with 3 layer Neural Net for MovieLens: For the Neural Network (NN) experiments on MovieLens dataset, we consider the function class F - a 3 layer Neural Net with 2 hidden layers of 50 neurons each and tanh activation. The training and the validation data on the MovieLens dataset is created in a similar manner as discussed in Section 3. The comparison of validation RMSE of AMHT-LRS and all the 4 baselines corresponding to the MovieLens dataset is given in Figure 8a. Here, we can observe that AMHT-LRS has almost similar performance as the best performing baseline Full Fine-tuning $(\mathsf{F}(x;\mathsf{u}_{\text{indv}}))$ while outperforming the other baselines. However, note that the individual models $\mathsf{F}(x;\mathsf{u}_{\text{indv}})$ have a high memory overhead since every trained model per task has the same memory usage as a single Neural Network model. In particular AMHT-LRS (with only 20% sparsity/tunable parameters) matches the Full-finetuning baseline with approximately 20% of the corresponding number of parameters.



(a) Overall comparison of RMSE.

Figure 8: Comparison of RMSE on the MovieLens validation data achieved by AMHT-LRS and the different baselines we consider where the training is modified for toy Neural Network models with 2 hidden layers of 50 neurons each. AMHT-LRS has similar performance as individual models $\mathsf{F}(x; \mathsf{u}_{\text{indv}})$ trained for each task; however our models have a significantly smaller memory footprint (see Table A.3.3 for exact numbers on model parameters).

B WARM-UP: CENTRAL MODEL + FINE-TUNING

B.1 SPARSE FINE-TUNING OF CENTRAL MODEL

Inspired by parameter efficient transfer learning applications shown in Guo et al. (2020) where the authors propose learning a task-specific sparse vector, we consider the following simple variant of our problem in the noiseless rank-1 setting ($r = 1$) with $(\mathbf{W}^\star)^\mathsf{T}$ being a multiple of an all 1 t-dimensional vector. We will denote the representation vector by $\mathbf{u}^\star \in \mathbb{R}^d$ that is shared by all the tasks. Therefore, the ERM for this model is given by the LRS problem with $\mathbf{w}^{(i)} = 1$ for all $i \in [t]$. We can also pose this problem as the setting when the low rank representation of the datapoints corresponds to projection on a fixed unknown vector; there exists a central model (parameterized by the fixed unknown vector shared across tasks) and each task is a fine-tuned version of the central model. Our AM algorithm to solve the modified ERM problem is significantly simpler; in particular Steps 2-8 in Algorithm 1 is replaced by the following set of updates given estimates $\mathbf{u}^{(\ell-1)} \in \mathbb{R}^d$ (of $\mathbf{u}^\star$) and $\{\mathbf{b}^{(i,\ell-1)}\}_{i\in[t]}$ (of $\{\mathbf{b}^{\star(i)}\}_{i\in[t]}$) in the ℓ^{th} iteration with a suitable choice of $\Delta^{(\ell)}$:

$$\mathbf{c}^{(i,\ell)} \leftarrow \mathbf{b}^{(i,\ell-1)} - (m^{-1}\mathbf{X}^{(i)})^\mathsf{T}(\mathbf{X}^{(i)}(\mathbf{u}^{(\ell-1)} + \mathbf{b}^{(i,\ell-1)}) - \mathbf{y}^{(i)}) \tag{9}$$

$$\mathbf{b}^{(i,\ell)} \leftarrow \mathsf{HT}(\mathbf{c}^{(i,\ell)}, \Delta^{(\ell)}) \tag{10}$$

$$\mathbf{u}^{(\ell)} \leftarrow \Big(\sum_i (\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big)^{-1}\Big(\sum_i (\mathbf{X}^{(i)})^\mathsf{T}\Big(\mathbf{y}^{(i)} - \mathbf{X}^{(i)}\mathbf{b}^{(i,\ell)}\Big)\Big) \tag{11}$$

Notice that the updates in eq. 11 are only implemented once in each iteration (unlike Algorithm 2) which improves the run-time as well as the sample complexity of the algorithm by logarithmic factors. The detailed Algorithm is provided in Appendix B. We present the main theorem below:

Theorem 4. *Consider the LRS problem with t linear regression tasks and samples obtained by equation 1 where rank $r = 1$, $\sigma = 0$, $\mathbf{U}^\star \equiv \mathbf{u}^\star \in \mathbb{R}^d$ and $\mathbf{w}_i^\star \equiv w^\star \in \mathbb{R}$. Let model parameters $\{\mathbf{b}^{\star(i)}\}_{i\in[t]}$ satisfy assumption A1. Suppose Algorithm 1 with modified updates (eq. 11) is run for $\mathsf{L} = \log\Big(\epsilon_0^{-1}\Big(\max_{i\in[t]} \big|\big|\mathbf{b}^{\star(i)}\big|\big|_\infty + ||\mathbf{u}^\star||_\infty + \frac{||\mathbf{u}^\star||_2}{\sqrt{k}}\Big)\Big)$ iterations. Then, w.p. $\geq 1 - O(\delta_0)$, the outputs*

Dataset	Method	#Parameters
MovieLens	Single Model $\mathsf{F}(x; \mathsf{u}_{\text{central}})$	5,151
	Full Fine-tuning $\mathsf{F}(x; \mathsf{u}_{\text{indv}})$	1,241,391
	Rep. Learning $\mathsf{F}(x; \mathsf{uw})$	5,392
	Prompt Learning $\mathsf{F}(x \mid\mid c; \mathsf{u}_{\text{central}})$	19,701
	AMHT-LRS $\mathsf{F}(x; \mathsf{uw} + \mathsf{b}_{1\%\text{ sparse}})$	17,924
	AMHT-LRS $\mathsf{F}(x; \mathsf{uw} + \mathsf{b}_{5\%\text{ sparse}})$	67,570
	AMHT-LRS $\mathsf{F}(x; \mathsf{uw} + \mathsf{b}_{10\%\text{ sparse}})$	129,748
	AMHT-LRS $\mathsf{F}(x; \mathsf{uw} + \mathsf{b}_{15\%\text{ sparse}})$	191,685
	AMHT-LRS $\mathsf{F}(x; \mathsf{uw} + \mathsf{b}_{20\%\text{ sparse}})$	253,863
Jester	Single Model $\mathsf{F}(x; \mathsf{u}_{\text{central}})$	2,601
	Full Fine-tuning $\mathsf{F}(x; \mathsf{u}_{\text{indv}})$	260,100
	Rep. Learning $\mathsf{F}(x; \mathsf{uw})$	2,701
	Prompt Learning $\mathsf{F}(x \mid\mid c; \mathsf{u}_{\text{central}})$	10,101
	AMHT-LRS $\mathsf{F}(x; \mathsf{uw} + \mathsf{b}_{1\%\text{ sparse}})$	5,302
	AMHT-LRS $\mathsf{F}(x; \mathsf{uw} + \mathsf{b}_{5\%\text{ sparse}})$	15,706
	AMHT-LRS $\mathsf{F}(x; \mathsf{uw} + \mathsf{b}_{10\%\text{ sparse}})$	28,711
	AMHT-LRS $\mathsf{F}(x; \mathsf{uw} + \mathsf{b}_{15\%\text{ sparse}})$	41,716
	AMHT-LRS $\mathsf{F}(x; \mathsf{uw} + \mathsf{b}_{20\%\text{ sparse}})$	54,721
Netflix	Single Model ($\mathsf{F}(x; \mathsf{u}_{\text{central}})$)	2,601
	Full Fine-tuning $\mathsf{F}(x; \mathsf{u}_{\text{indv}})$	520,200
	Rep. Learning $\mathsf{F}(x; \mathsf{uw})$	2,801
	Prompt Learning $\mathsf{F}(x \mid\mid c; \mathsf{u}_{\text{central}})$	15,101
	AMHT-LRS $\mathsf{F}(x; \mathsf{uw} + \mathsf{b}_{1\%\text{ sparse}})$	8,003
	AMHT-LRS $\mathsf{F}(x; \mathsf{uw} + \mathsf{b}_{5\%\text{ sparse}})$	28,811
	AMHT-LRS $\mathsf{F}(x; \mathsf{uw} + \mathsf{b}_{10\%\text{ sparse}})$	54,821
	AMHT-LRS $\mathsf{F}(x; \mathsf{uw} + \mathsf{b}_{15\%\text{ sparse}})$	80,831
	AMHT-LRS $\mathsf{F}(x; \mathsf{uw} + \mathsf{b}_{20\%\text{ sparse}})$	106,841

Table 1: Comparison of number of model parameters for AMHT-LRS at different sparsity levels and the different baselines we consider. Note that our approach AMHT-LRS at 1% and 5% sparsity levels has substantially less number of parameters than Full fine-tuning $\mathsf{F}(x; \mathsf{u}_{\text{indv}})$ and comparable number of model parameters with the other baselines.

$\mathbf{u}^{(\mathsf{L})}, \{\mathbf{b}^{(i,\mathsf{L})}\}_{i\in[t]}$ *satisfy:* $\left|\left|\mathbf{u}^{(\mathsf{L})} - w^{\star}\mathbf{u}^{\star}\right|\right|_{\infty} \le O(\epsilon_0)$ *and* $\left|\left|\mathbf{b}^{(i,\mathsf{L})} - \mathbf{b}^{\star(i)}\right|\right|_{\infty} \le O(\epsilon_0)$ *for all* $i \in [t]$ *provided the total number of samples satisfy* $m = \widetilde{\Omega}(k),\ mt = \widetilde{\Omega}(d\sqrt{k})$ *and* $mt^2 = \widetilde{\Omega}(\zeta kd)$.

Remark 5. *Notice from Theorem 4 that our AM algorithm in the sparse fine-tuning setting enjoys global convergence guarantees and does not require any initialization conditions. Secondly, we do not need* $\mathbf{u}^{\star}$ *to satisfy any incoherence property for convergence guarantees of Theorem 4 (unlike Theorem 1). Therefore, Theorem 4 is interesting in itself and significantly improves on the guarantees of Theorem 1 directly applied to the special setting.*

B.2 DETAILED ANALYSIS AND PROOF OF THEOREM 4

In the fine-tuning model described in Section B.1, we consider a system comprising of t tasks, each of which (indexed by $i \in [t]$) is parameterized by an unknown task-specific sparse parameter vector $\mathbf{b}^{\star(i)} \in \mathbb{R}^d$ satisfying $\|\mathbf{b}^{\star(i)}\|_0 \le k$ along with a dense unknown parameter vector $\mathbf{u}^{\star} \in \mathbb{R}^d$ that is shared across all tasks. Now, for each task $i \in [t]$, we obtain samples $\{(\mathbf{x}_j^{(i)}, y_j^{(i)})\}_{j=1}^{m}$ according to the following model:

$$\mathbf{x}_j^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d) \text{ and } y_j^{(i)} \mid \mathbf{x}_j^{(i)} = \langle \mathbf{x}_j^{(i)}, \mathbf{u}^{\star} + \mathbf{b}^{\star(i)} \rangle \quad \text{for all } i \in [t], j \in [m] \tag{12}$$

Algorithm 3 AMHT-LRS (Central Model+Finetuning)

```
Require: Data $\{(\mathbf{x}_j^{(i)} \in \mathbb{R}^d, y_j^{(i)} \in \mathbb{R})\}_{j=1}^m$ for all $i \in [t]$, column sparsity $k$ of $\mathbf{B}$. Initial Error Bounds $\max_{i\in[t]} \|\mathbf{b}^{(0,\ell)} - \mathbf{b}^{\star(i)}\|_2 \le \alpha^{(0)}$, $\max_{i\in[t]} \|\mathbf{b}^{(i,0)} - \mathbf{b}^{\star(i)}\|_\infty \le \gamma^{(0)}$, $\|\mathbf{u}^{(0)} - \mathbf{u}^\star\|_\infty \le \beta^{(0)}$ and $\|\mathbf{u}^{(0)} - \mathbf{u}^\star\|_2 \le \tau^{(0)}$.
1:  Suitable constants $c_1, c_2, c_3 > 0$.
2:  for $\ell = 1, 2, \dots$ (Until Convergence) do
3:      $\Delta^{(\ell)} \leftarrow \beta^{(\ell-1)} + \frac{c_1}{\sqrt{k}}\Big(\tau^{(\ell-1)} + \alpha^{(\ell-1)}\Big)$
4:      $\mathbf{c}^{(i,\ell)} \leftarrow \mathbf{b}^{(i,\ell-1)} - \frac{1}{m} \cdot (\mathbf{X}^{(i)})^\mathsf{T}(\mathbf{X}^{(i)}(\mathbf{u}^{(\ell-1)} + \mathbf{b}^{(i,\ell-1)}) - \mathbf{y}^{(i)})$
5:      $\mathbf{b}^{(i,\ell)} \leftarrow \mathsf{HT}(\mathbf{c}^{(i,\ell)}, \Delta^{(\ell)})$
6:      $\mathbf{u}^{(\ell)} \leftarrow \Big(\frac{1}{mt}\sum_i\sum_j \mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}\Big)^{-1}\Big(\frac{1}{mt}\sum_i\sum_j \mathbf{x}_j^{(i)}(y_j^{(i)} - (\mathbf{x}_j^{(i)})^\mathsf{T}\mathbf{b}^{(i,\ell)})\Big)$
7:      Set, $\gamma^{(\ell)} \leftarrow 2\beta^{(\ell-1)} + \frac{2c_1}{\sqrt{k}}\tau^{(\ell-1)} + 2c_1\gamma^{(\ell-1)}$
8:      Set $\tau^{(\ell)} \leftarrow c_2\sqrt{k}\gamma^{(\ell)}$, $\beta^{(\ell)} \leftarrow c_3\gamma^{(\ell)}$ and $\alpha^{(\ell)} \leftarrow \sqrt{k}\gamma^{(\ell)}$
9:  end for
10: Return $\mathbf{w}^{(\ell)}$, $\mathbf{U}^{+(\ell)}$ and $\{\mathbf{b}^{(i,\ell)}\}_{i\in[t]}$.
```

We will assume that the model parameters $\{\mathbf{b}^{\star(i)}\}_{i\in[t]}$ satisfy Assumption A1. More importantly, we do not assume A2 and furthermore, we do not assume that $\mathbf{u}^\star$ is unit-norm. Since $\mathbf{u}^\star$ is not unit norm, we can write it as $\mathbf{u}^\star = \frac{\mathbf{u}^\star}{||\mathbf{u}^\star||_2}||\mathbf{u}^\star||_2$. In order to map it to the statement of Theorem 4 and the general problem statement in 1, we can immediately write $w^\star \leftarrow ||\mathbf{u}^\star||_2$ and $\mathbf{u}^\star \leftarrow \frac{\mathbf{u}^\star}{||\mathbf{u}^\star||_2}$ (since $\mathbf{u}^\star$ in the statement of Theorem 4 is unit-norm). Hence, we can simplify the notation significantly by assuming that $\mathbf{u}^\star$ is not unit norm and by subsuming the scalar $w^\star$ (which is same across all tasks for this special setting) with the norm of vector $\mathbf{u}^\star$.

Initialization and Notations: For $\ell = 0$, we will initialize $\mathbf{u}^{(0)} = \mathbf{0}$ and $\mathbf{b}^{(i,0)} = \mathbf{0}$ for all tasks indexed by $i \in [t]$. For any $\ell \ge 0$, at the beginning of the $(\ell+1)^{\text{th}}$ iteration, we will use $\alpha^{(\ell)}, \tau^{(\ell)}$ to denote known upper bounds on the ℓ_2-norm of the approximated parameters and $\gamma^{(\ell)}, \beta^{(\ell)}$ to denote known upper bounds on the ℓ_∞-norm of the approximated parameters that will hold with high probability as described below:

$$\max_{i\in[t]} \Big|\Big|\mathbf{b}^{(i,\ell)} - \mathbf{b}^{\star(i)}\Big|\Big|_2 \le \alpha^{(\ell)} \quad \text{and} \quad \max_{i\in[t]} \Big|\Big|\mathbf{b}^{(i,\ell)} - \mathbf{b}^{\star(i)}\Big|\Big|_\infty \le \gamma^{(\ell)},$$
$$\Big|\Big|\mathbf{u}^{(\ell)} - \mathbf{u}^\star\Big|\Big|_\infty \le \beta^{(\ell)} \quad \text{and} \quad \Big|\Big|\mathbf{u}^{(\ell)} - \mathbf{u}^\star\Big|\Big|_2 \le \tau^{(\ell)}.$$

Lemma 2. *For some constant $c > 0$ and for any iteration indexed by $\ell \in [\mathsf{L}]$, we can have the following updates*

$$\gamma^{(\ell)} = 2\beta^{(\ell-1)} + 2c\sqrt{\frac{\log(td/\delta_0)}{m}}\Big(\tau^{(\ell-1)} + \alpha^{(\ell-1)}\Big),$$
$$\alpha^{(\ell)} = 2\sqrt{k}\beta^{(\ell-1)} + 2c\sqrt{\frac{k\log(td/\delta_0)}{m}}\Big(\tau^{(\ell-1)} + \alpha^{(\ell-1)}\Big)$$
$$\mathsf{support}(\mathbf{b}^{(i,\ell)}) \subseteq \mathsf{support}(\mathbf{b}^{\star(i)}).$$

with probability $1 - O(\delta_0)$.

Proof. Fix any $i \in [t]$. It is easy to see that update step 4 of Algorithm 3 gives us

$$\mathbf{c}^{(i,\ell)} - \mathbf{b}^{\star(i)} = \Big(\mathbf{I} - \frac{1}{m}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big)\Big(\mathbf{b}^{(i,\ell-1)} - \mathbf{b}^{\star(i)}\Big) + \frac{1}{m}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}(\mathbf{u}^\star - \mathbf{u}^{(\ell-1)})$$
$$\implies \mathbf{c}^{(i,\ell)} - \mathbf{b}^{\star(i)} - \mathbf{u}^\star + \mathbf{u}^{(\ell-1)} = \Big(\mathbf{I} - \frac{1}{m}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big)\Big(\mathbf{b}^{(i,\ell-1)} - \mathbf{b}^{\star(i)}\Big)$$
$$+ \Big(\mathbf{I} - \frac{1}{m}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big)(\mathbf{u}^{(\ell-1)} - \mathbf{u}^\star). \tag{13}$$

Let $\mathbf{e}_s \in \mathbb{R}^d$ denote the $s^{\textsf{th}}$ basis vector for which the $s^{\textsf{th}}$ coordinate entry is 1 and all other coordinate entries are 0. Then, note that:

$$\begin{aligned}
&\left|\left(\mathbf{c}^{(i,\ell)} - \mathbf{b}^{\star(i)} - \mathbf{u}^{\star} + \mathbf{u}^{(\ell-1)}\right)_s\right| \\
&= \left|\mathbf{e}_s^{\mathsf{T}}\left(\mathbf{I} - \frac{1}{m}(\mathbf{X}^{(i)})^{\mathsf{T}}\mathbf{X}^{(i)}\right)\left(\mathbf{b}^{(i,\ell-1)} - \mathbf{b}^{\star(i)}\right) + \mathbf{e}_s^{\mathsf{T}}\left(\mathbf{I} - \frac{1}{m}(\mathbf{X}^{(i)})^{\mathsf{T}}\mathbf{X}^{(i)}\right)(\mathbf{u}^{(\ell-1)} - \mathbf{u}^{\star})\right| \\
&\le \left|\mathbf{e}_s^{\mathsf{T}}\left(\mathbf{I} - \frac{1}{m}(\mathbf{X}^{(i)})^{\mathsf{T}}\mathbf{X}^{(i)}\right)\left(\mathbf{b}^{(i,\ell-1)} - \mathbf{b}^{\star(i)}\right)\right| + \left|\mathbf{e}_s^{\mathsf{T}}\left(\mathbf{I} - \frac{1}{m}(\mathbf{X}^{(i)})^{\mathsf{T}}\mathbf{X}^{(i)}\right)(\mathbf{u}^{(\ell-1)} - \mathbf{u}^{\star})\right| \\
&\le \left|\frac{1}{m}\mathbf{e}_s^{\mathsf{T}}(\mathbf{X}^{(i)})^{\mathsf{T}}\mathbf{X}^{(i)}(\mathbf{b}^{(i,\ell-1)} - \mathbf{b}^{\star(i)}) - \mathbf{e}_s^{\mathsf{T}}(\mathbf{b}^{(i,\ell-1)} - \mathbf{b}^{\star(i)})\right| \\
&\qquad + \left|\frac{1}{m}\mathbf{e}_s^{\mathsf{T}}(\mathbf{X}^{(i)})^{\mathsf{T}}\mathbf{X}^{(i)}(\mathbf{u}^{(\ell-1)} - \mathbf{u}^{\star}) - \mathbf{e}_s^{\mathsf{T}}(\mathbf{u}^{(\ell-1)} - \mathbf{u}^{\star})\right| \\
&\le c\sqrt{\frac{\log(1/\delta_0)}{m}}\left(\tau^{(\ell-1)} + \alpha^{(\ell-1)}\right),
\end{aligned}$$

w.p. $\ge 1 - O(\delta_0)$, where we invoke Lemma 17 in the last step and plugging $\mathbf{a} = \mathbf{e}_s$ and $\mathbf{b} = \mathbf{b}^{(i,\ell-1)} - \mathbf{b}^{\star(i)}$ and $\mathbf{u}^{(\ell-1)} - \mathbf{u}^{\star}$ for the two terms respectively. Therefore, by taking a union bound over all entries $s \in [d]$, and a further union bound over all tasks (t of them), we can conclude that for all $i \in [t]$, we must have

$$\begin{aligned}
&\left|\left|\mathbf{c}^{(i,\ell)} - \mathbf{b}^{\star(i)} - \mathbf{u}^{\star} + \mathbf{u}^{(\ell-1)}\right|\right|_\infty \le c\sqrt{\frac{\log(td/\delta_0)}{m}}\left(\tau^{(\ell-1)} + \alpha^{(\ell-1)}\right) \\
\implies &\left|\left|\mathbf{c}^{(i,\ell)} - \mathbf{b}^{\star(i)}\right|\right|_\infty \le \beta^{(\ell-1)} + c\sqrt{\frac{\log(td/\delta_0)}{m}}\left(\tau^{(\ell-1)} + \alpha^{(\ell-1)}\right) \qquad (14)
\end{aligned}$$

w.p. $1 - O(\delta_0)$. Now, we have

$$\begin{aligned}
\mathbf{b}^{(i,l)} &= \mathsf{HT}(\mathbf{c}^{(i,\ell)}, \Delta^{(\ell)}) \\
\implies b_s^{(i,l)} &= \begin{cases} c_s^{(i,\ell)} & \text{if } |c_s^{(i,\ell)}| > \Delta^{(\ell)}, \\ 0 & \text{otherwise,} \end{cases} \qquad (15) \\
\implies |b_s^{(i,l)} - b_s^{\star(i)}| &= \begin{cases} |c_s^{(i,\ell)} - b_s^{\star(i)}| & \text{if } |c_s^{(i,\ell)}| > \Delta^{(\ell)}, \\ |b_s^{\star(i)}| & \text{otherwise.} \end{cases} \qquad (16)
\end{aligned}$$

Therefore if we set $\Delta^{(\ell)} = \beta^{(\ell-1)} + c\sqrt{\frac{\log(td/\delta_0)}{m}}\left(\tau^{(\ell-1)} + \alpha^{(\ell-1)}\right)$ (as described in Step 2 of the algorithm), then, by using equation 14 and equation 16, we have $\left|\left|\mathbf{b}^{(i,\ell)} - \mathbf{b}^{\star(i)}\right|\right|_\infty \le 2\Delta^{(\ell)}$ and therefore,

$$\begin{aligned}
\implies \left|\left|\mathbf{b}^{(i,\ell)} - \mathbf{b}^{\star(i)}\right|\right|_\infty &\le 2\beta^{(\ell-1)} + 2c\sqrt{\frac{\log(td/\delta_0)}{m}}\left(\tau^{(\ell-1)} + \alpha^{(\ell-1)}\right) = \gamma^{(\ell)} \qquad (17) \\
\text{and } \left|\left|\mathbf{b}^{(i,\ell)} - \mathbf{b}^{\star(i)}\right|\right|_2 &\le 2\sqrt{k}\beta^{(\ell-1)} + 2c\sqrt{\frac{k\log(td/\delta_0)}{m}}\left(\tau^{(\ell-1)} + \alpha^{(\ell-1)}\right) = \alpha^{(\ell)}, \qquad (18)
\end{aligned}$$

with probability $1 - O(\delta_0)$. Furthermore, from equation equation 14 we have for any coordinate s

$$\left|\left(\mathbf{c}^{(i,\ell)} - \mathbf{b}^{\star(i)}\right)_s\right| \le \Delta^{(\ell)}.$$

Thus, if $s \notin \mathsf{support}(\mathbf{b}^{\star(i)})$, then the above gives $|\mathbf{c}^{(i,\ell)}| \le \Delta^{(\ell)}$. Using this in equation 15 gives $b_s^{(i,l)} = 0$. Hence, for all $s \in [d]$, we must have $s \notin \mathsf{support}(\mathbf{b}^{\star(i)}) \implies s \notin \mathsf{support}(\mathbf{b}^{\star(i,\ell)})$ implying that $\mathsf{support}(\mathbf{b}^{(i,\ell})) \subseteq \mathsf{support}(\mathbf{b}^{\star(i)})$. Hence, the proof of the lemma is complete.

□

Lemma 3. *For some constant* $c > 0$ *and for any iteration indexed by* $\ell > 0$*, we have*

$$\tau^{(\ell)} = \frac{\sqrt{\frac{2\zeta k}{t}}\gamma^{(\ell)} + 4\alpha^{(\ell)}\sqrt{\frac{d\log(d/\delta_0)}{mt}}}{1 - c\sqrt{\frac{d\log(1/\delta_0)}{mt}}}$$

with probability $1 - O(\delta_0)$.

Proof. Update step 3 of the Algorithm for the ℓ^{th} iteration gives us

$$\mathbf{u}^{(\ell)} = \Big(\frac{1}{mt}\sum_i\sum_j \mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}\Big)^{-1}\Big(\frac{1}{mt}\sum_i\sum_j \mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}(\mathbf{u}^\star + \mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})\Big)$$

$$\implies \mathbf{u}^{(\ell)} - \mathbf{u}^\star = \underbrace{\Big(\frac{1}{mt}\sum_i\sum_j \mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}\Big)^{-1}}_{\mathbf{A}}\underbrace{\Big(\frac{1}{mt}\sum_i\sum_j \mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})\Big)}_{\mathbf{v}}$$

Let us denote the vector $\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}$ by $\mathbf{z}^{(i,\ell)}$ for simplicity. Notice that for any $h \in [d]$, we have

$$\begin{aligned} v_h &= \Big(\frac{1}{mt}\sum_i\sum_j \mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}\mathbf{z}^{(i,\ell)}\Big)_h \\ &= \frac{1}{mt}\sum_i\sum_j\Big(\Big(x_{j,h}^{(i)}\Big)^2 z_h^{(i,\ell)} + \sum_{u:u\neq h} x_{j,h}^{(i)}x_{j,u}^{(i)}z_u^{(i,\ell)}\Big). \end{aligned} \tag{19}$$

Now, note that the random variable $\Big(x_{j,h}^{(i)}\Big)^2 z_h^{(i,\ell)}$ is a $\Big(4(z_h^{(i,\ell)})^2, 4|z_h^{(i,\ell)}|\Big)$ sub-exponential random variable. Similarly, $x_{j,h}^{(i)}x_{j,u}^{(i)}z_u^{(i,\ell)}$ is a $\Big(2(z_u^{(i,\ell)})^2, \sqrt{2}|z_u^{(i,\ell)}|\Big)$ sub-exponential random variable. Therefore, we must have

$$\begin{aligned} &\Big(x_{j,h}^{(i)}\Big)^2 z_h^{(i,\ell)} + \sum_{u:u\neq h} x_{j,h}^{(i)}x_{j,u}^{(i)}z_u^{(i,\ell)} \\ &= \Big(4(z_h^{(i,\ell)})^2 + 2\sum_{u:u\neq h}(z_u^{(i,\ell)})^2, \max\Big(4|z_h^{(i,\ell)}|, \max_{u:u\neq h}(\sqrt{2}|z_u^{(i,\ell)}|)\Big)\Big) \\ &= \Big(4\|\mathbf{z}^{(i,\ell)}\|_2^2, 4\|\mathbf{z}^{(i,\ell)}\|_\infty\Big) \text{ sub-exponential random variable.} \end{aligned} \tag{20}$$

Furthermore,

$$\begin{aligned} \mathbb{E}\left[v_h\right] &= \frac{1}{mt}\sum_i\sum_j\Big(\mathbb{E}\Big[\Big(x_{j,h}^{(i)}\Big)^2 z_h^{(i,\ell)}\Big] + \mathbb{E}\Big[\sum_{u:u\neq h} x_{j,h}^{(i)}x_{j,u}^{(i)}z_u^{(i,\ell)}\Big]\Big) \\ &= \frac{1}{mt}\sum_i\sum_j\Big(z_h^{(i,\ell)} + 0\Big) \\ &= \frac{1}{t}\sum_i z_h^{(i,\ell)}. \end{aligned} \tag{21}$$

Using equation 20, equation 21 and Lemma 23 in equation 19 implies that

$$\begin{aligned} \Big|v_h - \frac{1}{t}\sum_i z_h^{(i,\ell)}\Big| &\le \max\Big(2\|\mathbf{z}^{(i,\ell)}\|_2\sqrt{\frac{2\log(1/\delta_0)}{mt}}, 2\|\mathbf{z}^{(i,\ell)}\|_\infty\frac{2\log(1/\delta_0)}{mt}\Big) \\ &\le \underbrace{\max\Big(2\alpha^{(\ell)}\sqrt{\frac{2\log(1/\delta_0)}{mt}}, 2\gamma^{(\ell)}\frac{2\log(1/\delta_0)}{mt}\Big)}_{\epsilon_h}. \end{aligned}$$

will be true with probability at least $1-\delta_0$. On taking a union bound over all $h \in [d]$, we will have that

$$\Big|v_h - \frac{1}{t}\sum_i z_h^{(i,\ell)}\Big| \le \underbrace{\max\Big(2\alpha^{(\ell)}\sqrt{\frac{2\log(d/\delta_0)}{mt}}, 2\gamma^{(\ell)}\frac{2\log(d/\delta_0)}{mt}\Big)}_{\epsilon_h}. \tag{22}$$

with probability $1-O(\delta_0)$. Note that $\|\mathbf{v}\|_2^2 = \sum_h v_h^2$. Hence, we have

$$\begin{aligned}\sum_h v_h^2 &\le \sum_h \Big(2\Big(\frac{1}{t}\sum_i z_h^{(i,\ell)}\Big)^2 + 2\epsilon_h^2\Big) \\ &\le 2\zeta\sum_h\sum_i(\frac{z_h^{(i,\ell)}}{t})^2 + 2\sum_h \epsilon_h^2 \\ &\le 2\zeta\sum_i\sum_h\Big(\frac{z_h^{(i,\ell)}}{t}\Big)^2 + 2\sum_h \epsilon_h^2 \\ &\le \frac{2\zeta}{t}(\alpha^{(\ell)})^2 + 8(\alpha^{(\ell)})^2\frac{2d\log(d/\delta_0)}{mt},\end{aligned}$$

where we use that $2\alpha^{(\ell)}\sqrt{\frac{2\log(d/\delta_0)}{mt}} > 2\gamma^{(\ell)}\frac{2\log(d/\delta_0)}{mt}$. Hence, with probability at least $1-O(\delta_0)$, we must have by using that $\alpha^{(\ell)} \le \sqrt{k}\gamma^{(\ell)}$

$$\|\mathbf{v}\|_2 \le \sqrt{\frac{2\zeta k}{t}}\gamma^{(\ell)} + 4\alpha^{(\ell)}\sqrt{\frac{d\log(d/\delta_0)}{mt}} \tag{23}$$

Furthermore, from Lemma 19, we have with probability $1-\delta_0$ for any iterations $\ell \in [\mathsf{L}]$

$$\|\frac{1}{mt}\sum_i\sum_j \mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^{\mathsf{T}} - \mathbf{I}\|_2 \le c\sqrt{\frac{d\log(1/\delta_0)}{mt}}. \tag{24}$$

implying that the minimum eigenvalue of the matrix $\frac{1}{mt}\sum_i\sum_j \mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^{\mathsf{T}} - \mathbf{I}$ is at least $1 - c\sqrt{\frac{d\log(1/\delta_0)}{mt}}$; hence the maximum eigenvalue of the matrix $(\frac{1}{mt}\sum_i\sum_j \mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^{\mathsf{T}} - \mathbf{I})^{-1}$ is at most $(1-c\sqrt{\frac{d\log(1/\delta_0)}{mt}})^{-1}$. Using equation 23 and equation 24, we get for any iterations $\ell \in [\mathsf{L}]$ with probability $1-O(\delta_0)$,

$$\|\mathbf{u}^{(\ell)} - \mathbf{u}^{\star}\|_2 \le \frac{\sqrt{\frac{2\zeta k}{t}}\gamma^{(\ell)} + 4\alpha^{(\ell)}\sqrt{\frac{d\log(d/\delta_0)}{mt}}}{1-c\sqrt{\frac{d\log(1/\delta_0)}{mt}}} \triangleq \tau^{(\ell)}. \tag{25}$$

□

Lemma 4. *For some constant* $c > 0$ *and for any iteration indexed by* $\ell > 0$*, we have*

$$\beta^{(\ell)} = \Big(\frac{\zeta}{t} + 2c\sqrt{\frac{\log(d/\delta_0)}{mt}}\sqrt{\frac{2\zeta k}{t}}\Big)\gamma^{(\ell)} + \Big(c\sqrt{\frac{\log(d/\delta_0)}{mt}} + 8c\sqrt{d}\frac{\log(d/\delta_0)}{mt}\Big)\alpha^{(\ell)}$$

with probability at least $1-O(\delta_0)$.

Proof. With probability at least $1-O(\delta_0)$, we have that $||\mathbf{E}||_2 \le \sqrt{\frac{d\log 9}{mt}}$. We fix $mt = \Omega(d)$ so that $||\mathbf{E}||_2 < 1$. Our goal is to bound the quantity $\|\mathbf{u}^{(\ell)} - \mathbf{u}^{\star}\|_\infty$ from above. Denoting $\mathbf{A} = \mathbf{I} + \mathbf{E}$ and using the fact that $(\mathbf{I}+\mathbf{E})^{-1} = \mathbf{I} - \mathbf{E} + \mathbf{E}^2 + \dots$ (since $||\mathbf{E}||_2 < 1$), by using Lemma 16 and taking a union bound over all entries $s \in [d]$, we have with probability at least $1 - \delta_0$,

$$\begin{aligned}
&\left|\left|\mathbf{v} - \frac{1}{t}\sum_i (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})\right|\right|_\infty \\
&= \max_s \left| \frac{1}{mt}\sum_i \sum_j \mathbf{e}_s^\mathsf{T} \mathbf{x}_j^{(i)} (\mathbf{x}_j^{(i)})^\mathsf{T} (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}) - \frac{1}{t}\sum_i \mathbf{e}_s^\mathsf{T} (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}) \right| \\
&= \max_s \left| \frac{1}{mt}\sum_i \sum_j (\mathbf{x}_j^{(i)})^\mathsf{T} (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}) \mathbf{e}_s^\mathsf{T} \mathbf{x}_j^{(i)} - \frac{1}{t}\sum_i \mathbf{e}_s^\mathsf{T} (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}) \right| \\
&\le c\sqrt{\sum_i \sum_j \|(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})\mathbf{e}_s^\mathsf{T}\|_\mathsf{F}^2 \frac{\log(d/\delta_0)}{m^2 t^2}} \\
&\le c\alpha^{(\ell)} \sqrt{\frac{\log(d/\delta_0)}{mt}}.
\end{aligned}$$

Hence with probability at least $1 - \delta_0$, we will have the following statement

$$||\mathbf{v}||_\infty \le c\alpha^{(\ell)} \sqrt{\frac{\log(d/\delta_0)}{mt}} + \frac{\zeta}{t}\gamma^{(\ell)}. \tag{26}$$

Since $\mathbf{u}^{(\ell)} - \mathbf{u}^\star = (\mathbf{I} + \mathbf{E})^{-1}\mathbf{v}$ with $\mathbf{E} = \frac{1}{mt}\sum_i \sum_j \mathbf{x}_j^{(i)} (\mathbf{x}_j^{(i)})^\mathsf{T} - \mathbf{I}$, we will have

$$\left|\left|\mathbf{u}^{(\ell)} - \mathbf{u}^\star\right|\right|_\infty \le \sum_{j=0}^{\infty} \left|\left|\mathbf{E}^j \mathbf{v}\right|\right|_\infty. \tag{27}$$

□

Let $\mathcal{V} \triangleq \{\mathbf{z} \in \mathbb{R}^d | \|\mathbf{z}\| = 1\}$. Then for $\epsilon \le 1$, there exists an ϵ-net, $N_\epsilon \subset \mathcal{Z}$, of size $(1 + 2/\epsilon)^d$ w.r.t the Euclidean norm, i.e. $\forall\, \mathbf{z} \in \mathcal{Z}$, $\exists\, \mathbf{z}' \in N_\epsilon$ s.t. $\|\mathbf{z} - \mathbf{z}'\|_2 \le \epsilon$. Now consider any $\mathbf{z} \in N_\epsilon$. Then, Lemma 17 with $\mathbf{a} = \mathbf{e}_s$ and $\mathbf{b} = \mathbf{z}$ and taking a union bound over all entries $s \in [d]$ gives

$$\begin{aligned}
&\left| \mathbf{e}_s^\mathsf{T} \Big(\frac{1}{mt}\sum_i \sum_j \mathbf{x}_j^{(i)} (\mathbf{x}_j^{(i)})^\mathsf{T} - \mathbf{I} \Big) \mathbf{z} \right| \le c\|\mathbf{z}\|_2^2 \max\Big(\sqrt{\frac{\log(1/\delta_0)}{mt}}, \frac{\log(1/\delta_0)}{mt} \Big) \\
\implies &\|\mathbf{E}\mathbf{z}\|_\infty \le c\max\Big(\sqrt{\frac{\log(d|N_\epsilon|/\delta_0)}{mt}}, \frac{\log(d|N_\epsilon|/\delta_0)}{mt} \Big) \\
&\le c\max\Big(\sqrt{\frac{\log(d(1+2/\epsilon)^d/\delta_0)}{mt}}, \frac{\log(d(1+2/\epsilon)^d/\delta_0)}{mt} \Big), \quad \forall \mathbf{v} \in N_\epsilon
\end{aligned} \tag{28}$$

Further, $\exists\, \mathbf{z} \in N_\epsilon$ s.t. $\|\mathbf{z}' - \mathbf{z}\|_2 \le \epsilon$. This implies that setting $\epsilon \leftarrow 1/4$ and $c \leftarrow 2c$ gives:

$$\begin{aligned}
\|\mathbf{E}\mathbf{z}'\|_\infty &\le \|\mathbf{E}(\mathbf{z} - \mathbf{z}')\|_\infty + \|\mathbf{E}\mathbf{z}\|_\infty \\
&\le \|\mathbf{E}(\mathbf{z} - \mathbf{z}')\|_2 + \|\mathbf{E}\mathbf{z}\|_\infty \\
&\le c\sqrt{\frac{d\log(d/\delta_0)}{mt}}.
\end{aligned} \tag{29}$$

with probability at least $1 - \delta_0$. Hence, with probability at least $1 - O(\delta_0)$, we have that $||\mathbf{E}||_2 \le \sqrt{\frac{d\log 9}{mt}}$ and $||\mathbf{E}\mathbf{z}||_\infty \le c\sqrt{\frac{d\log(d\delta_0^{-1})}{mt}}$ for all $\mathbf{z} \in \mathcal{V}$. Therefore, let us conditioned on these events in order to prove the next steps. We will show an upper bound on $\|\mathbf{A}^{-1}\mathbf{v}\|_\infty$.

$$\begin{aligned}
\|\mathbf{A}^{-1}\mathbf{v}\|_\infty &= \|(\mathbf{I} + \mathbf{E})^{-1}\mathbf{v}\|_\infty \\
&\le \sum_{j=0}^{\infty} \left|\left|\mathbf{E}^j \mathbf{v}\right|\right|_\infty.
\end{aligned} \tag{30}$$

We have with probability at least $1-\delta_0$

$$\begin{aligned}
\|\mathbf{E}^p\mathbf{v}\|_\infty &= \|\mathbf{E}\mathbf{E}^{p-1}\mathbf{v}\|_\infty \\
&= \|\Big(\mathbf{E}\|\mathbf{E}^{p-1}\mathbf{v}\|_2\Big)\Big(\frac{\mathbf{E}^{p-1}\mathbf{v}}{\|\mathbf{E}^{p-1}\mathbf{v}\|_2}\Big)\|_\infty \\
&= \|\mathbf{E}^{p-1}\mathbf{v}\|_2\|\Big(\mathbf{E}\Big)\Big(\frac{\mathbf{E}^{p-1}\mathbf{v}}{\|\mathbf{E}^{p-1}\mathbf{v}\|_2}\Big)\|_\infty \\
&\le \|\mathbf{E}^{p-1}\mathbf{v}\|_2 c\sqrt{\frac{d\log(d/\delta_0)}{mt}} \\
&\le \Big(c\sqrt{\frac{d\log(1/\delta_0)}{mt}}\Big)^{p-1} c\sqrt{\frac{d\log(d/\delta_0)}{mt}}\|\mathbf{v}\|_2
\end{aligned} \tag{31}$$

Therefore, if $mt=\Omega(d\log(d/\delta_0))$ by taking a union bound we must have with probability at least $1-\delta_0$,

$$\sum_{p=1}^{\infty}||\mathbf{E^p v}||_\infty = O\Big(\sqrt{\frac{d\log(d/\delta_0)}{mt}}\Big)||\mathbf{v}||_2. \tag{32}$$

Therefore we have w.p. $\ge 1-O(\delta_0)$

$$\Big|\Big|\mathbf{u}^{(\ell)}-\mathbf{u}^\star\Big|\Big|_\infty \le ||\mathbf{v}||_\infty + 2c\sqrt{\frac{d\log(d/\delta_0)}{mt}}||\mathbf{v}||_2. \tag{33}$$

Plugging the bounds of $||\mathbf{v}||_\infty$ and $||\mathbf{v}||_2$ from equation 26 and equation 23 in equation 33, we obtain that w.p. $\ge 1-\delta_0)$

$$\begin{aligned}
&\Big|\Big|\mathbf{u}^{(\ell)}-\mathbf{u}^\star\Big|\Big|_\infty \\
&\le c\alpha^{(\ell)}\sqrt{\frac{\log(d/\delta_0)}{mt}} + \frac{\zeta}{t}\gamma^{(\ell)} + 2c\sqrt{\frac{d\log(d/\delta_0)}{mt}}\Big(\sqrt{\frac{2\zeta k}{t}}\gamma^{(\ell)} + 4\alpha^{(\ell)}\sqrt{\frac{d\log(d/\delta_0)}{mt}}\Big) \\
&= \Big(\frac{\zeta}{t} + 2c\sqrt{\frac{d\log(d/\delta_0)}{mt}}\sqrt{\frac{2\zeta k}{t}}\Big)\gamma^{(\ell)} + \Big(c\sqrt{\frac{\log(d/\delta_0)}{mt}} + 8c\frac{d\log(d/\delta_0)}{mt}\Big)\alpha^{(\ell)} = \beta^{(\ell)}
\end{aligned} \tag{34}$$

Lemma 5. *After* L *iterations, for some constant* $c>0$. *we will have with probability* $1-O(\mathsf{L}\delta_0)$,

$$\begin{aligned}
\Big|\Big|\mathbf{u}^{(\mathsf{L})}-\mathbf{u}^\star\Big|\Big|_\infty &\le c_3 2^{\mathsf{L}-1}(c_3+c_1c_2+c_1)^{\mathsf{L}-1}\mathsf{Z}, \\
\Big|\Big|\mathbf{b}^{(i,\mathsf{L})}-\mathbf{b}^{\star(i)}\Big|\Big|_\infty &\le 2^{\mathsf{L}-1}(c_3+c_1c_2+c_1)^{\mathsf{L}-1}\mathsf{Z}, \\
\Big|\Big|\mathbf{b}^{(i,\mathsf{L})}-\mathbf{b}^{\star(i)}\Big|\Big|_2 &\le \sqrt{k}2^{\mathsf{L}-1}(c_3+c_1c_2+c_1)^{\mathsf{L}-1}\mathsf{Z}, \\
\Big|\Big|\mathbf{u}^{(\mathsf{L})}-\mathbf{u}^\star\Big|\Big|_2 &\le c_2\sqrt{k}2^{\mathsf{L}-1}(c_3+c_1c_2+c_1)^{\mathsf{L}-1}\mathsf{Z},
\end{aligned}$$

where

$$\begin{aligned}
\mathsf{Z} &= \Big(2\Big|\Big|\mathbf{u}^{(0)}-\mathbf{u}^\star\Big|\Big|_\infty + \frac{2c_1}{\sqrt{k}}\Big|\Big|\mathbf{u}^{(0)}-\mathbf{u}^\star\Big|\Big|_2 + 2c_1\max_{i\in[t]}\Big|\Big|\mathbf{b}^{(i,0)}-\mathbf{b}^{\star(i)}\Big|\Big|_\infty\Big) \\
c_1 &= c\sqrt{\frac{k\log(tdL/\delta_0)}{m}}, \\
c_2 &= \frac{\sqrt{\frac{2\zeta}{t}} + 4\sqrt{\frac{d\log(dL/\delta_0)}{mt}}}{1-c\sqrt{\frac{d\log(L/\delta_0)}{mt}}}, \\
c_3 &= \Big(\frac{\zeta}{t} + 2c\sqrt{\frac{d\log(d/\delta_0)}{mt}}\sqrt{\frac{2\zeta k}{t}}\Big) + \sqrt{k}\Big(c\sqrt{\frac{\log(d/\delta_0)}{mt}} + 8c\frac{d\log(d/\delta_0)}{mt}\Big).
\end{aligned}$$

Proof. Using Lemma 2 and the fact that $\alpha^{(\ell)} \leq \sqrt{k}\gamma^{(\ell)}$, we have for $\ell \geq 1$

$$\gamma^{(\ell)} \leq 2\beta^{(\ell-1)} + \frac{2c_1}{\sqrt{k}}\tau^{(\ell-1)} + 2c_1\gamma^{(\ell-1)}, \tag{35}$$

$$\tau^{(\ell)} \leq c_2\sqrt{k}\gamma^{(\ell)}, \tag{36}$$

$$\beta^{(\ell)} \leq c_3\gamma^{(\ell)}, \tag{37}$$

Using equation 36 and equation 37 in equation 35, we get

$$\begin{aligned}\gamma^{(\ell)} &\leq (2c_3 + 2c_1c_2 + 2c_1)\gamma^{(\ell-1)} \\ &= 2(c_3 + c_1c_2 + c_1)\gamma^{(\ell-1)} \\ &\leq \dots \\ &\leq 2^{\ell-1}(c_3 + c_1c_2 + c_1)^{\ell-1}\gamma^{(1)} \\ &\leq 2^{\ell-1}(c_3 + c_1c_2 + c_1)^{\ell-1}\Big(2\beta^{(0)} + \frac{2c_1}{\sqrt{k}}\tau^{(0)} + 2c_1\gamma^{(0)}\Big),\end{aligned} \tag{38}$$

where in the last step we plug in the value $\gamma^{(1)} \leq 2\beta^{(0)} + \frac{2c_1}{\sqrt{k}}\tau^{(0)} + 2c_1\gamma^{(0)}$ from equation 35 at $\ell = 1$.

Using equation 38 in equation 37 gives

$$\beta^{(\ell)} \leq c_3 2^{\ell-1}(c_3 + c_1c_2 + c_1)^{\ell-1}\Big(2\beta^{(0)} + \frac{2c_1}{\sqrt{k}}\tau^{(0)} + 2c_1\gamma^{(0)}\Big). \tag{39}$$

Using equation 38, equation 36 and $\alpha^{(\ell)} \leq \sqrt{k}\gamma^{(\ell)}$ further gives:

$$\alpha^{(\ell)} \leq \sqrt{k}2^{\ell-1}(c_3 + c_1c_2 + c_1)^{\ell-1}\Big(2\beta^{(0)} + \frac{2c_1}{\sqrt{k}}\tau^{(0)} + 2c_1\gamma^{(0)}\Big), \tag{40}$$

$$\tau^{(\ell)} \leq c_2\sqrt{k}2^{\ell-1}(c_3 + c_1c_2 + c_1)^{\ell-1}\Big(2\beta^{(0)} + \frac{2c_1}{\sqrt{k}}\tau^{(0)} + 2c_1\gamma^{(0)}\Big). \tag{41}$$

equation 38, equation 39,equation 40 and equation 41 give us the required result. ☐

Theorem (Restatement of Theorem 4). *Consider the LRS problem with t linear regression tasks and samples obtained by equation 1 where rank $r = 1$, $\sigma = 0$, $\mathbf{U}^\star \equiv \mathbf{u}^\star \in \mathbb{R}^d$ and $\mathbf{w}_i^\star \equiv w^\star \in \mathbb{R}$. Let model parameters $\{\mathbf{b}^{\star(i)}\}_{i\in[t]}$ satisfy assumption A1 with $\zeta = O(t)$. Suppose Algorithm 1 with modified updates (eqns. 9,10,11) is run for $\mathsf{L} = \log\Big(\epsilon_0^{-1}\Big(\max_{i\in[t]}\big|\big|\mathbf{b}^{\star(i)}\big|\big|_\infty + ||\mathbf{u}^\star||_\infty + \frac{||\mathbf{u}^\star||_2}{\sqrt{k}}\Big)\Big)$ iterations. Then, w.p. $\geq 1 - O(\delta_0)$, the outputs $\mathbf{u}^{(\mathsf{L})}, \{\mathbf{b}^{(i,\mathsf{L})}\}_{i\in[t]}$ satisfy:*

$$\Big|\Big|\mathbf{u}^{(\mathsf{L})} - w^\star\mathbf{u}^\star\Big|\Big|_\infty \leq O(\epsilon_0) \textit{ and } \Big|\Big|\mathbf{b}^{(i,\mathsf{L})} - \mathbf{b}^{\star(i)}\Big|\Big|_\infty \leq O(\epsilon_0) \textit{ for all } i \in [t]. \tag{42}$$

provided the total number of samples satisfy

$$m = \widetilde{\Omega}(k),\ mt = \widetilde{\Omega}(d\sqrt{k}) \textit{ and } mt^2 = \widetilde{\Omega}(\zeta kd).$$

Proof. In order to map 12 to the statement of Theorem 4 and the general problem statement in 1, recall that we can immediately write $w^\star \leftarrow ||\mathbf{u}^\star||_2$ and $\mathbf{u}^\star \leftarrow \frac{\mathbf{u}^\star}{||\mathbf{u}^\star||_2}$ (since $\mathbf{u}^\star$ in the statement of Theorem 4 is unit-norm). For the simplicity of notation, we had subsumed $w^\star$ within $||\mathbf{u}^\star||_2$. Therefore, we directly use Lemma 5 to prove our theorem where the result is stated after mapping back to the setting in 12 and the Theorem statement. ☐

C ALGORITHM AND PROOF OF THEOREM 1 (PARAMETER RECOVERY)

Assumption 3 (A3). *We assume that $||\mathbf{U}^\star||_{2,\infty} \leq \sqrt{\nu^\star/k}$ for some constant $\nu^\star > 0$.*

Algorithm 4 AMHT-LRS (Alternating Minimization for LRS in (2))

```
Require: Data $\{(\mathbf{x}_j^{(i)} \in \mathbb{R}^d, y_j^{(i)} \in \mathbb{R})\}_{j=1}^m$ for all $i \in [t]$, column sparsity $k$ of $\mathbf{B}$, $\left|\left|\Delta(\mathbf{U}^{+(0)}, \mathbf{U}^\star)\right|\right|_{\mathsf{F}} \le \mathsf{B}$, $\max_i \|\mathbf{b}^{(i,0)} - \mathbf{b}^{\star(i)}\|_\infty \le \gamma^{(0)}$, Parameters $\epsilon > 0$ and $\mathsf{A}$.
for $\ell = 1, 2, \dots$ do
    Set $T^{(\ell)} = \Omega\Big(\ell \log\Big(\frac{\gamma^{(\ell-1)}}{\epsilon}\Big)\Big)$
    for $i = 1, 2, \dots, t$ do
        $\mathbf{b}^{(i,\ell)} \leftarrow \mathsf{OptimizeSparseVector}((\mathbf{X}^{(i)}, \mathbf{y}^{(i)}), \mathbf{v} = \mathbf{U}^{+(\ell-1)}\mathbf{w}^{(i,\ell-1)}, \alpha = O\Big(c_4^{\ell-1}\frac{\mathsf{B}}{\sqrt{k}} + \mathsf{A}\Big), \beta = O(c_5^{\ell-1}\mathsf{B} + \mathsf{A}), \gamma = \gamma^{(\ell-1)} + \mathsf{A}, \mathsf{T} = T^{(\ell)})$
        {Use a fresh batch of data samples; $c_4, c_5$ are suitable constants}1
        $\mathbf{w}^{(i,\ell)} = \Big((\mathbf{X}^{(i)}\mathbf{U}^{+(\ell-1)})^\mathsf{T}(\mathbf{X}^{(i)}\mathbf{U}^{+(\ell-1)})\Big)^{-1}\Big((\mathbf{X}^{(i)}\mathbf{U}^{+(\ell-1)})^\mathsf{T}(\mathbf{y}^{(i)} - \mathbf{X}^{(i)}\mathbf{b}^{(i,\ell)})\Big)$ {Use a fresh batch of data samples}
    end for
    Set $\mathbf{A} := \sum_{i\in[t]}\Big(\mathbf{w}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T} \otimes \Big(\sum_{j=1}^m \mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}\Big)\Big)$ and $\mathbf{V} := \sum_{i\in[t]}(\mathbf{X}^{(i)})^\mathsf{T}\Big(\mathbf{y}^{(i)} - \mathbf{b}^{(i,\ell)}\Big)(\mathbf{w}^{(i,\ell)})^\mathsf{T}$ {Use a fresh batch of data samples}
    Compute $\mathbf{U}^{(\ell)} = \mathsf{vec}_{d\times r}^{-1}(\mathbf{A}^{-1}\mathsf{vec}(\mathbf{V}))$ and $\mathbf{U}^{+(\ell)} \leftarrow \mathsf{QR}(\mathbf{U}^{(\ell)})$ $\{\mathbf{U}^{(\ell)} = \mathbf{U}^{+(\ell)}\mathbf{R}\}$
    $\gamma^{(\ell)} \leftarrow (c_3)^{\ell-1}\epsilon\mathsf{B} + \mathsf{A}$ for a suitable constant $c_3 < 1$.
end for
Return $\mathbf{w}^{(\ell)}$, $\mathbf{U}^{+(\ell)}$ and $\{\mathbf{b}^{(i,\ell)}\}_{i\in[t]}$.
```

Algorithm 5 DP-AMHT-LRS (Private Alternating Minimization for LRS in (2))

```
Require: Data $\{(\mathbf{x}_j^{(i)} \in \mathbb{R}^d, y_j^{(i)} \in \mathbb{R})\}_{j=1}^m$ for all $i \in [t]$, column sparsity $k$ of $\mathbf{B}$, $\left|\left|\Delta(\mathbf{U}^{+(0)}, \mathbf{U}^\star)\right|\right|_{\mathsf{F}} \le \mathsf{B}$, $\max_i \|\mathbf{b}^{(i,0)} - \mathbf{b}^{\star(i)}\|_\infty \le \gamma^{(0)}$, Parameters $\epsilon > 0$ and $\mathsf{A}$.
for $\ell = 1, 2, \dots$ do
    Set $T^{(\ell)} = \Omega\Big(\ell \log\Big(\frac{\gamma^{(\ell-1)}}{\epsilon}\Big)\Big)$
    for $i = 1, 2, \dots, t$ do
        $\mathbf{b}^{(i,\ell)} \leftarrow \mathsf{OptimizeSparseVector}((\mathbf{X}^{(i)}, \mathbf{y}^{(i)}), \mathbf{v} = \mathbf{U}^{+(\ell-1)}\mathbf{w}^{(i,\ell-1)}, \alpha = O\Big(c_4^{\ell-1}\frac{\mathsf{B}}{\sqrt{k}} + \mathsf{A}\Big), \beta = O(c_5^{\ell-1}\mathsf{B} + \mathsf{A}), \gamma = \gamma^{(\ell-1)} + \mathsf{A}, \mathsf{T} = T^{(\ell)})$ for suitable constants $c_4, c_5$
        $\mathbf{w}^{(i,\ell)} = \Big((\mathbf{X}^{(i)}\mathbf{U}^{+(\ell-1)})^\mathsf{T}(\mathbf{X}^{(i)}\mathbf{U}^{+(\ell-1)})\Big)^{-1}\Big((\mathbf{X}^{(i)}\mathbf{U}^{+(\ell-1)})^\mathsf{T}(\mathbf{y}^{(i)} - \mathbf{X}^{(i)}\mathbf{b}^{(i,\ell)})\Big)$
    end for
    $\forall i, j : \widehat{\mathbf{x}_j^{(i)}} \leftarrow \mathsf{clip}_{A_1}\Big(\mathbf{x}_j^{(i)}\Big)$, $\widehat{y_j^{(i)}} \leftarrow \mathsf{clip}_{A_2}\Big(y_j^{(i)}\Big)$, $\widehat{(\mathbf{x}_j^{(i)})^\mathsf{T}\mathbf{b}^{(i,\ell)}} \leftarrow \mathsf{clip}_{A_3}\Big((\mathbf{x}_j^{(i)})^\mathsf{T}\mathbf{b}^{(i,\ell)}\Big)$ and $\widehat{\mathbf{w}^{(i,\ell)}} \leftarrow \mathsf{clip}_{A_w}\Big(\mathbf{w}^{(i,\ell)}\Big)$
    $\mathbf{A} := \frac{1}{mt}\Big(\sum_{i\in[t]}\Big(\widehat{\mathbf{w}^{(i,\ell)}}(\widehat{\mathbf{w}^{(i,\ell)}})^\mathsf{T} \otimes \Big(\sum_{j=1}^m \widehat{\mathbf{x}_j^{(i)}}(\widehat{\mathbf{x}_j^{(i)}})^\mathsf{T}\Big)\Big) + \mathbf{N}_1\Big)$
    $\mathbf{V} := \frac{1}{mt}\Big(\sum_{i\in[t]}\sum_{j\in[m]}\widehat{\mathbf{x}_j^{(i)}}\Big(\widehat{y_j^{(i)}} - \widehat{(\mathbf{x}_j^{(i)})^\mathsf{T}\mathbf{b}^{(i,\ell)}}\Big)(\widehat{\mathbf{w}^{(i,\ell)}})^\mathsf{T} + \mathbf{N}_2\Big)$
    $\mathbf{U}^{(\ell)} = \mathsf{vec}_{d\times r}^{-1}(\mathbf{A}^{-1}\mathsf{vec}(\mathbf{V}))$
    $\mathbf{U}^{+(\ell)} \leftarrow \mathsf{QR}(\mathbf{U}^{(\ell)})$ $\{\mathbf{U}^{(\ell)} = \mathbf{U}^{+(\ell)}\mathbf{R}\}$
    $\gamma^{(\ell)} \leftarrow (c_3)^{\ell-1}\epsilon\mathsf{B} + \mathsf{A}$ for a suitable constant $c_3 < 1$.
end for
Return $\mathbf{w}^{(\ell)}$, $\mathbf{U}^{+(\ell)}$ and $\{\mathbf{b}^{(i,\ell)}\}_{i\in[t]}$.
```

Note that Assumption A3 is weaker than Assumption A2 where $||\mathbf{U}^\star||_{2,\infty} \le \sqrt{\mu^\star/d}$ provided $k \le \frac{d\nu^\star}{\mu^\star}$. We will use Assumption A3 in place of A2 for simplicity of exposition and for sharper guarantees as well. Recall that in the general setting described in eq. 1, we obtain samples that are generated according to the following process:

$$\mathbf{x}_j^{(i)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d) \text{ and } y_j^{(i)} \mid \mathbf{x}_j^{(i)} = \langle \mathbf{x}_j^{(i)}, \mathbf{U}^\star \mathbf{w}^{\star(i)} + \mathbf{b}^{\star(i)} \rangle + z_j^{(i)} \quad \text{for all } i \in [t], j \in [m], \tag{43}$$

where each $z_j^{(i)} \sim \mathcal{N}(0, \sigma^2)$ denotes the independent measurement noise with known variance σ^2. For each task $i \in [t]$, we will denote the noise vector to be $\mathbf{z}^{(i)}$ such that its j^{th} co-ordinate is $\mathbf{z}_j^{(i)}$. Further, with some abuse of notation we will denote:

- $\lambda_j \equiv \lambda_j\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)$, $\lambda_j^\star = \lambda_j\Big(\frac{r}{t}(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^\star\Big) \equiv \lambda_j\Big(\frac{r}{t}(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^\star\Big) \; \forall\, j \in [r]$,
- $\mu \equiv \mu^{(\ell)}$ and $\mu^\star$ for the incoherence factors for $\mathbf{W}^{(\ell)}$ and $\mathbf{W}^\star$ respectively,
- $\nu \equiv \nu^{(\ell)}$ for the incoherence factor of $\mathbf{U}^{+(\ell)}$.

We will now prove Theorem 1 via an inductive argument. We will start with the base case.

C.1 BASE CASE

We initialize $\mathbf{W}^{(0)} = \mathbf{0}$ and recall $\Big|\Big|(\mathbf{I} - \mathbf{U}^\star(\mathbf{U}^\star)^\mathsf{T})\mathbf{U}^{+(0)}\Big|\Big|_\mathsf{F} = O\Big(\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big)$, $\Big|\Big|\mathbf{U}^{+(0)}\Big|\Big| \le \sqrt{\frac{\nu^{(0)}}{k}}$ where $\nu^{(0)}$ is an appropriate constant less than 1. We use Lemma 13 that is proved later in its full generality. We have by using Lemma 13 :

$$\Big|\Big|\mathbf{b}^{(i,1)} - \mathbf{b}^{\star(i)}\Big|\Big|_2 \le 2\varphi^{(i)} + \epsilon \text{ and } \Big|\Big|\mathbf{b}^{(i,1)} - \mathbf{b}^{\star(i)}\Big|\Big|_\infty \le \frac{1}{\sqrt{k}}\Big(2\varphi^{(i)} + \epsilon\Big)$$

with probability at least $1 - T^{(\ell)}\delta$, where $\varphi^{(i)}$ is an upper-bound on $\widehat{\varphi}^{(i)}$ s.t.

$$\begin{aligned}
\widehat{\varphi}^{(i)} &= 2\Big(\sqrt{k}\|\mathbf{U}^\star\mathbf{w}^{\star(i)}\|_\infty + c_1\|\mathbf{U}^\star\mathbf{w}^{\star(i)}\|_2 + \sigma\sqrt{\frac{k\log(d\delta^{-1})}{m}}\Big) \\
&\le 2\Big(\sqrt{k}\|\mathbf{U}^\star\|_{2,\infty}\|\mathbf{w}^{\star(i)}\|_2 + c_1\|\mathbf{w}^{\star(i)}\|_2 + \sigma\sqrt{\frac{k\log(d\delta^{-1})}{m}}\Big) \\
&\le 2\Big(\sqrt{\nu^\star} + c_1\Big)\|\mathbf{w}^{\star(i)}\|_2 + 2\sigma\sqrt{\frac{k\log(d\delta^{-1})}{m}}
\end{aligned}$$

Choosing $\epsilon = 4\Big(\sqrt{\nu^\star} + c_1\Big)$ gives us the required expression for $\ell = 0$. Hence, we have that for $c' = O\Big(\frac{1}{\mathsf{B}_{\mathbf{U}^{(0)}}}\frac{\lambda_1^\star}{\lambda_r^\star}\Big)$, we will have that

$$\Big|\Big|\mathbf{b}^{(i,0)} - \mathbf{b}^{\star(i)}\Big|\Big| \le c' \max(\epsilon, \Big|\Big|\mathbf{w}^{\star(i)}\Big|\Big|_2)\mathsf{B}_{\mathbf{U}^{(0)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} + 4\sigma\sqrt{\frac{k\log(d\delta^{-1})}{m}} \tag{44}$$

$$\Big|\Big|\mathbf{b}^{(i,0)} - \mathbf{b}^{\star(i)}\Big|\Big| \le c' \max(\epsilon, \Big|\Big|\mathbf{w}^{\star(i)}\Big|\Big|_2)\mathsf{B}_{\mathbf{U}^{(0)}}\sqrt{\frac{\lambda_r^\star}{k\lambda_1^\star}} + 4\sigma\sqrt{\frac{k\log(d\delta^{-1})}{m}} \tag{45}$$

C.2 INDUCTIVE STEP

We will begin with the inductive assumption. Note that these assumptions are true in the base case as well due to our initialization and optimizing the task-specific sparse vector. Let

$$\Lambda = \mathcal{O}\Big(\sqrt{\lambda_r^\star\mu^\star}\Big(\tfrac{\sigma_2 r}{mt\lambda_r^\star} + \tfrac{\sigma_1 r^{3/2}}{mt\lambda_r^\star}\sqrt{rd}\log rd + \sigma\sqrt{\tfrac{r^3 d\mu^\star \log^2(r\delta^{-1})}{mt\lambda_r^\star}}\Big) + \sigma\Big(\sqrt{\tfrac{r^3\log^2(r\delta^{-1})}{m\lambda_r^\star}}\Big) + \sqrt{\tfrac{k\log(d\delta^{-1})}{m}}\Big)\Big)$$

$$\Lambda' = \mathcal{O}\Big(\frac{\Lambda}{\sqrt{\mu^\star\lambda_r^\star}}\Big).$$

Assumption 4 (Inductive Assumption)**.** *At the beginning of the* ℓ^{th} *iteration, we will use* $q^{(\ell-1)}, \mathsf{B}_{\mathbf{u}^{+(\ell-1)}}$ *to describe the following upper bounds on the quantities of interest:*

$$\text{1) } 1/2 < \lambda_{\min}(\mathbf{Q}^{(\ell-1)}) \leq \lambda_{\max}(\mathbf{Q}^{(\ell-1)}) < 1, \text{ where } \mathbf{Q}^{(\ell-1)} := \langle (\mathbf{U}^\star)^\mathsf{T}\mathbf{U}^{+(\ell-1)} \rangle \tag{46}$$

$$\text{2) } \|\Delta(\mathbf{U}^{+(\ell-1)}, \mathbf{U}^\star)\|_\mathsf{F} = \|(\mathbf{I} - \mathbf{U}^\star(\mathbf{U}^\star)^\mathsf{T})\mathbf{U}^{+(\ell-1)}\|_\mathsf{F} = \|\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_\mathsf{F}$$

$$\leq \mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} + \Lambda', \tag{47}$$

$$\text{3) } \|\mathbf{b}^{(i,\ell)} - \mathbf{b}^{\star(i)}\|_2 \leq c'\|(\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\|_2$$

$$\leq c'\max\{\epsilon, \|\mathbf{w}^{\star(i)}\|_2\}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} + \Lambda, \tag{48}$$

$$\text{4) } \|\mathbf{b}^{(i,\ell)} - \mathbf{b}^{\star(i)}\|_\infty \leq c'\|(\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\|_2/\sqrt{k}$$

$$\leq c'\max\{\epsilon, \|\mathbf{w}^{\star(i)}\|_2\}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star k}} + \frac{\Lambda}{\sqrt{k}}, \tag{49}$$

$$\text{5) } \|\mathbf{U}^{+(\ell-1)}\|_{2,\infty} \leq \sqrt{\nu^{(\ell-1)}/k}, \tag{50}$$

where $\nu^{(\ell-1)} < \frac{1}{181c}$*,* $c' > 0$ *and* $\Lambda' < 1/1000$*. Note that* Λ, Λ' *are fixed and do not change with iterations.*

Note that the base case satisfies the inductive assumption for our problem. Let us denote $\mathbf{h}^{(i,\ell)} \triangleq \mathbf{w}^{(i,\ell)} - (\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}$ and and $(\mathbf{H}^{(\ell)})^\mathsf{T} = \begin{bmatrix}\mathbf{h}^{(1,\ell)} & \mathbf{h}^{(2,\ell)} & \dots & \mathbf{h}^{(t,\ell)}\end{bmatrix}_{\mathbb{R}^{r\times t}}$.

Lemma 6. *For some constant* $c > 0$ *and for any iteration indexed by* $\ell > 0$*, we have*

$$\|\mathbf{h}^{(i,\ell)}\|_2 \leq \frac{1}{1 - c\sqrt{\frac{r\log(1/\delta_0)}{m}}}\Big\{\|\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_\mathsf{F}\|(\mathbf{Q}^{(\ell-1)})^{-1}\|\|\mathbf{w}^{\star(i)}\|_2\cdot$$

$$\Big(\|\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_\mathsf{F} + c\sqrt{\frac{\log(r/\delta_0)}{m}}\|\mathbf{U}^{+(\ell-1)}\|_\mathsf{F}\Big)$$

$$+ \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\Big(\sqrt{k}\|\mathbf{U}^{+(\ell-1)}\|_{2,\infty} + c\sqrt{\frac{\log(r/\delta_0)}{m}}\|\mathbf{U}^{+(\ell-1)}\|_\mathsf{F}\Big) + \sigma\sqrt{\frac{r\log^2(r\delta^{-1})}{m}}\Big\},$$

$$\|\mathbf{H}^{(\ell)}\|_\mathsf{F} \leq \frac{1}{1 - c\sqrt{\frac{r\log(1/\delta_0)}{m}}}\cdot\Big\{\|\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_\mathsf{F}\|(\mathbf{Q}^{(\ell-1)})^{-1}\|\sqrt{\frac{t}{r}\lambda_1^\star}\cdot$$

$$\Big(\|\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_\mathsf{F} + c\sqrt{\frac{\log(r/\delta_0)}{m}}\|\mathbf{U}^{(\ell-1)}\|_\mathsf{F}\Big)$$

$$+ \sqrt{k\zeta}\|\mathbf{U}^\star\|_\mathsf{F}\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_\infty + \sqrt{\sum_{i\in[t]}\Big(c\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\sqrt{\frac{\log(r/\delta_0)}{m}}\|\mathbf{U}^{+(\ell-1)}\|_\mathsf{F}\Big)^2}$$

$$+ \sigma\sqrt{\frac{rt\log^2(r\delta^{-1})}{m}}\Big\}$$

with probability at least $1 - \delta_0$*, where* $\mathbf{h}^{(i,\ell)} = \mathbf{w}^{(i,\ell)} - (\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}$.

Proof. According to the update step equation 5, we have

$$\mathbf{w}^{(i,\ell)} - (\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)} = \Big(\frac{(\mathbf{X}^{(i)}\mathbf{U}^{+(\ell-1)})^\mathsf{T}(\mathbf{X}^{(i)}\mathbf{U}^{+(\ell-1)})}{m}\Big)^{-1}\cdot$$

$$\Big(\frac{(\mathbf{X}^{(i)}\mathbf{U}^{+(\ell-1)})^\mathsf{T}(\mathbf{y}^{(i)} - \mathbf{X}^{(i)}\mathbf{b}^{(i,\ell)})}{m}\Big) - (\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}$$

$$\begin{aligned}&\iff \mathbf{h}^{(i,\ell)}\\ &:= \underbrace{\Big(\frac{(\mathbf{X}^{(i)}\mathbf{U}^{+(\ell-1)})^\mathsf{T}(\mathbf{X}^{(i)}\mathbf{U}^{+(\ell-1)})}{m}\Big)^{-1}}_{\mathbf{A}} \cdot \\ &\quad \underbrace{\Big(\frac{(\mathbf{X}^{(i)}\mathbf{U}^{+(\ell-1)})^\mathsf{T}(\mathbf{y}^{(i)}-\mathbf{X}^{(i)}\mathbf{b}^{(i,\ell)})}{m} - \frac{(\mathbf{X}^{(i)}\mathbf{U}^{+(\ell-1)})^\mathsf{T}(\mathbf{X}^{(i)}\mathbf{U}^{+(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}}{m}\Big)}_{\mathbf{z}}.\end{aligned} \tag{51}$$

Therefore,

$$\|\mathbf{h}^{(i,\ell)}\|_\infty \le \|\mathbf{A}\|\|\mathbf{z}\|_\infty \qquad \text{and} \qquad \|\mathbf{h}^{(i,\ell)}\|_2 \le \|\mathbf{A}\|\|\mathbf{z}\|_2. \tag{52}$$

We will analyse the terms $\mathbf{A}$ and $\mathbf{z}$ separately.

Analysis of A:

Note that:

$$\begin{aligned}\mathbf{A}^{-1} &= \frac{1}{m}(\mathbf{X}^{(i)}\mathbf{U}^{+(\ell-1)})^\mathsf{T}(\mathbf{X}^{(i)}\mathbf{U}^{+(\ell-1)})\\ &= \frac{1}{m}(\mathbf{U}^{+(\ell-1)})^\mathsf{T}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\mathbf{U}^{+(\ell-1)}\\ &= \frac{1}{m}(\mathbf{U}^{+(\ell-1)})^\mathsf{T}\Big(\sum_j \mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}\Big)\mathbf{U}^{+(\ell-1)}\\ &= \frac{1}{m}\sum_j(\mathbf{U}^{+(\ell-1)})^\mathsf{T}\mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}\mathbf{U}^{+(\ell-1)}.\end{aligned} \tag{53}$$

Now, let $\mathcal{V} \triangleq \{\mathbf{v} \in \mathbb{R}^r | \|\mathbf{v}\| = 1\}$. Then for $\epsilon \le 1$, there exists an ϵ-net, $N_\epsilon \subset \mathcal{V}$, of size $(1 + 2/\epsilon)^r$ w.r.t the Euclidean norm, i.e. $\forall\, \mathbf{v} \in \mathcal{V}$, $\exists\, \mathbf{v}' \in N_\epsilon$ s.t. $\|\mathbf{v} - \mathbf{v}'\|_2 \le \epsilon$. Then for any $\mathbf{v} \in N_\epsilon$,

$$\begin{aligned}&\mathbf{v}^\mathsf{T}\Big(\frac{1}{m}\sum_j(\mathbf{U}^{+(\ell-1)})^\mathsf{T}\mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}\mathbf{U}^{+(\ell-1)}\Big)\mathbf{v}\\ &\quad= \frac{1}{m}\sum_j\Big(\mathbf{v}^\mathsf{T}(\mathbf{U}^{+(\ell-1)})^\mathsf{T}\Big)\mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}\Big(\mathbf{U}^{+(\ell-1)}\mathbf{v}\Big)\\ &\quad= \frac{1}{m}\sum_j\Big(\mathbf{U}^{+(\ell-1)}\mathbf{v}\Big)^\mathsf{T}\mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}\Big(\mathbf{U}^{+(\ell-1)}\mathbf{v}\Big).\end{aligned} \tag{54}$$

Further, note that

$$\begin{aligned}\Big(\mathbf{U}^{+(\ell-1)}\mathbf{v}\Big)^\mathsf{T}\Big(\mathbf{U}^{+(\ell-1)}\mathbf{v}\Big) &= \mathsf{Tr}\Big(\Big(\mathbf{U}^{+(\ell-1)}\mathbf{v}\Big)^\mathsf{T}\Big(\mathbf{U}^{+(\ell-1)}\mathbf{v}\Big)\Big)\\ &= \mathsf{Tr}\Big(\mathbf{v}^\mathsf{T}\Big((\mathbf{U}^{+(\ell-1)})^\mathsf{T}\mathbf{U}^{+(\ell-1)}\Big)\mathbf{v}\Big).\end{aligned} \tag{55}$$

Using equation 54 and equation 55 in Lemma 17 with $\mathbf{a} = \mathbf{b} = \mathbf{U}^{+(\ell-1)}\mathbf{v}$ gives

$$\begin{aligned}&\left|\mathbf{v}^\mathsf{T}\Big(\frac{1}{m}\sum_j(\mathbf{U}^{+(\ell-1)})^\mathsf{T}\mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}\mathbf{U}^{+(\ell-1)} - (\mathbf{U}^{+(\ell-1)})^\mathsf{T}\mathbf{U}^{+(\ell-1)}\Big)\mathbf{v}\right|\\ &\qquad \le c\|\mathbf{U}^{+(\ell-1)}\mathbf{v}\|_2^2 \max\Big(\sqrt{\frac{\log(1/\delta_0)}{m}}, \frac{\log(1/\delta_0)}{m}\Big)\\ &\implies \|\mathbf{v}^\mathsf{T}\mathbf{E}\mathbf{v}\| \le c\|\mathbf{U}^{+(\ell-1)}\mathbf{v}\|_2^2 \max\Big(\sqrt{\frac{\log(|N_\epsilon|/\delta_0)}{m}}, \frac{\log(|N_\epsilon|/\delta_0)}{m}\Big)\\ &\qquad \le c\|\mathbf{U}^{+(\ell-1)}\mathbf{v}\|_2^2 \max\Big(\sqrt{\frac{\log((1+2/\epsilon)^r/\delta_0)}{m}}, \frac{\log((1+2/\epsilon)^r/\delta_0)}{m}\Big) \quad \forall \mathbf{v} \in N_\epsilon\end{aligned} \tag{56}$$

where $\mathbf{E} \triangleq \frac{1}{m}\sum_j (\mathbf{U}^{+(\ell-1)})^\top \mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\top \mathbf{U}^{+(\ell-1)} - (\mathbf{U}^{+(\ell-1)})^\top \mathbf{U}^{+(\ell-1)}$. Since $\mathbf{E}$ is symmetric, therefore $\|\mathbf{E}\| = (\mathbf{v}')^\top \mathbf{E}\mathbf{v}'$ where $\mathbf{v}' \in \mathsf{V}$ is the largest eigenvector of $\mathbf{E}$. Further, $\exists\ \mathbf{v} \in N_\epsilon$ s.t. $\|\mathbf{v}' - \mathbf{v}\| \le \epsilon$. This implies

$$\begin{aligned}
\|\mathbf{E}\| = (\mathbf{v}')^\top \mathbf{E}\mathbf{v}' &= (\mathbf{v}'-\mathbf{v})^\top \mathbf{E}\mathbf{v} + (\mathbf{v}')^\top \mathbf{E}(\mathbf{v}'-\mathbf{v}) + \mathbf{v}^\top \mathbf{E}\mathbf{v} \\
&\le \|\mathbf{v}'-\mathbf{v}\|\|\mathbf{E}\|\|\mathbf{v}\| + \|\mathbf{v}'\|\|\mathbf{E}\|\|\mathbf{v}'-\mathbf{v}\| + \mathbf{v}^\top \mathbf{E}\mathbf{v} \\
&\le 2\epsilon\|\mathbf{E}\| + \mathbf{v}^\top \mathbf{E}\mathbf{v} \\
\implies \|\mathbf{E}\| &\le \frac{\mathbf{v}^\top \mathbf{E}\mathbf{v}}{1-2\epsilon}. \qquad (57)
\end{aligned}$$

Using equation 56 and equation 57 and setting $\epsilon \leftarrow 1/4$ and $c \leftarrow 2c\sqrt{\log(9)}$ then gives:

$$\|\mathbf{E}\| \le c\|\mathbf{U}^{+(\ell-1)}\mathbf{v}\|_2^2 \max\sqrt{\frac{r\log(1/\delta_0)}{m}} \qquad (58)$$

Using equation 58 in equation 53 then gives

$$\begin{aligned}
\|\mathbf{A}^{-1}\| &\ge \|\mathbf{U}^{+(\ell-1)}\mathbf{v}\|_2^2\Big(1 - c\sqrt{\frac{r\log(1/\delta_0)}{m}}\Big) \\
&\ge \lambda_{\min}\Big((\mathbf{U}^{+(\ell-1)})^\top \mathbf{U}^{+(\ell-1)}\Big)\Big(1 - c\sqrt{\frac{r\log(1/\delta_0)}{m}}\Big) \\
\implies \|\mathbf{A}\| &\le \frac{1}{\lambda_{\min}\Big((\mathbf{U}^{+(\ell-1)})^\top \mathbf{U}^{+(\ell-1)}\Big)\Big(1 - c\sqrt{\frac{r\log(1/\delta_0)}{m}}\Big)} \\
&= \frac{1}{1 - c\sqrt{\frac{r\log(1/\delta_0)}{m}}} \qquad (59)
\end{aligned}$$

since $(\mathbf{U}^{+(\ell-1)})^\top \mathbf{U}^{+(\ell-1)} = \mathbf{I}$.

Analysis of z:

Similarly, we have

$$\begin{aligned}
\mathbf{z} &= \frac{1}{m}(\mathbf{X}^{(i)}\mathbf{U}^{+(\ell-1)})^\top(\mathbf{y}^{(i)} - \mathbf{X}^{(i)}\mathbf{b}^{(i,\ell)}) - \frac{1}{m}(\mathbf{X}^{(i)}\mathbf{U}^{+(\ell-1)})^\top(\mathbf{X}^{(i)}\mathbf{U}^{+(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)} \\
&= \underbrace{\frac{1}{m}(\mathbf{U}^{+(\ell-1)})^\top(\mathbf{X}^{(i)})^\top\mathbf{X}^{(i)}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)} - \mathbf{U}^{+(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}}_{:=\mathbf{d}_1^{(i,\ell)}} \\
&\quad + \underbrace{\frac{1}{m}(\mathbf{U}^{+(\ell-1)})^\top(\mathbf{X}^{(i)})^\top\mathbf{X}^{(i)}(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})}_{:=\mathbf{d}_2^{(i,\ell)}} + \underbrace{\frac{1}{m}(\mathbf{U}^{+(\ell-1)})^\top(\mathbf{X}^{(i)})^\top\mathbf{z}^{(i)}}_{\mathbf{d}_3^{(i,\ell)}}.
\end{aligned}$$

Analysis of $\mathbf{d}_3^{(i,\ell)}$:

Let us condition on the vector $\mathbf{z}^{(i)}$. In that case $(\mathbf{X}^{(i)})^\top\mathbf{z}^{(i)}$ is a $d \times 1$ vector, each of whose entry is generated independently according to $\mathcal{N}(0, \left|\left|\mathbf{z}^{(i)}\right|\right|_2^2)$. Therefore, if we consider any vector $\mathbf{v}$ satisfying $\left|\left|\mathbf{v}\right|\right|_2 = 1$, we have

$$\mathbf{v}^\top(\mathbf{X}^{(i)})^\top\mathbf{z}^{(i)} \sim \mathcal{N}(0, \left|\left|\mathbf{z}^{(i)}\right|\right|_2^2)$$

and therefore, with probability $1-\delta$, we must have

$$\left|\left|\frac{1}{m}(\mathbf{U}^{+(\ell-1)})^\top(\mathbf{X}^{(i)})^\top\mathbf{z}^{(i)}\right|\right|_2 \le \sigma\sqrt{\frac{r\log^2(r\delta^{-1})}{m}}.$$

Analysis of $\mathbf{d}_1^{(i,\ell)}$:

$$\begin{aligned}\mathbf{d}_1^{(i,\ell)} &= \frac{1}{m}(\mathbf{U}^{+(\ell-1)})^\mathsf{T}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)} - \mathbf{U}^{+(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\\ &= \underbrace{(\mathbf{U}^{+(\ell-1)})^\mathsf{T}\Big(\frac{1}{m}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)} - \mathbf{I}\Big)(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)} - \mathbf{U}^{+(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}}_{\mathbf{d}_{1,1}^{(i,\ell)}}\\ &\quad + \underbrace{(\mathbf{U}^{+(\ell-1)})^\mathsf{T}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}) - \mathbf{U}^{+(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}}_{\mathbf{d}_{1,2}^{(i,\ell)}}. \end{aligned} \tag{60}$$

Note that

$$\begin{aligned}\mathbb{E}\left[\mathbf{d}_{1,1}^{(i,\ell)}\right] &= \mathbb{E}\left[(\mathbf{U}^{+(\ell-1)})^\mathsf{T}\Big(\frac{1}{m}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)} - \mathbf{I}\Big)(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)} - \mathbf{U}^{+(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\right]\\ &= \mathbf{0}.\end{aligned}$$

Further,

$$\begin{aligned}(z_{1,1}^{(i,\ell)})_k = \frac{1}{m}\sum_j &(\mathbf{u}^{(k,\ell-1)})^\mathsf{T}\mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)} - \mathbf{U}^{+(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\\ &- (\mathbf{u}^{(k,\ell-1)})^\mathsf{T}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)} - \mathbf{U}^{+(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}.\end{aligned}$$

Using Lemma 17 in the above with $\mathbf{a} = \mathbf{u}^{(k,\ell-1)}$ and $\mathbf{b} = (\mathbf{U}^\star\mathbf{Q} - \mathbf{U}^{+(\ell-1)})\mathbf{Q}^{-1}\mathbf{w}^{\star(i)}$, we get

$$(z_{1,1}^{(i,\ell)})_k \le c\sqrt{\frac{\log(1/\delta_0)}{m}}\|\mathbf{u}^{(k,\ell-1)}\|_2\|(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)} - \mathbf{U}^{+(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\|_2. \tag{61}$$

Taking the Union Bound overall entries $k \in [r]$, we have

$$\begin{aligned}\|\mathbf{d}_{1,1}^{(i,\ell)}\|_2 &= \sqrt{\sum_{k\in[r]}|(z_{1,1}^{(i,\ell)})_k|^2}\\ &\le c\sqrt{\frac{\log(r/\delta_0)}{m}}\sqrt{\sum_{k\in[r]}\|\mathbf{u}^{(k,\ell-1)}\|_2^2}\cdot\|(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)} - \mathbf{U}^{+(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\|_2\\ &= c\sqrt{\frac{\log(r/\delta_0)}{m}}\|\mathbf{U}^{+(\ell-1)}\|_\mathsf{F}\|(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)} - \mathbf{U}^{+(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\|_2.\end{aligned} \tag{62}$$

Further,

$$\begin{aligned}\|\mathbf{d}_{1,2}^{(i,\ell)}\|_2 &= \|(\mathbf{U}^{+(\ell-1)})^\mathsf{T}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)} - \mathbf{U}^{+(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\|_2\\ &= \|(\mathbf{U}^{+(\ell-1)})^\mathsf{T}(\mathbf{I} - \mathbf{U}^\star(\mathbf{U}^\star)^\mathsf{T})\mathbf{U}^{+(\ell-1)}(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\|_2\\ &= \|(\mathbf{U}^{+(\ell-1)})^\mathsf{T}(\mathbf{I} - \mathbf{U}^\star(\mathbf{U}^\star)^\mathsf{T})^2\mathbf{U}^{+(\ell-1)}(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\|_2\\ &= \Big\|\Big((\mathbf{U}^{+(\ell-1)})^\mathsf{T} - (\mathbf{U}^{+(\ell-1)})^\mathsf{T}\mathbf{U}^\star(\mathbf{U}^\star)^\mathsf{T}\Big)\cdot\\ &\qquad\Big(\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star(\mathbf{U}^\star)^\mathsf{T}\mathbf{U}^{+(\ell-1)}\Big)(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\Big\|_2\\ &\le \|\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star(\mathbf{U}^\star)^\mathsf{T}\mathbf{U}^{+(\ell-1)}\|_\mathsf{F}^2\|(\mathbf{U}^{(\ell-1)} - \mathbf{U}^\star(\mathbf{U}^\star)^\mathsf{T}\mathbf{U}^{+(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\|_2\\ &= \|\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_\mathsf{F}\|(\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\|_2\\ &\le \|\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_\mathsf{F}^2\|(\mathbf{Q}^{(\ell-1)})^{-1}\|\|\mathbf{w}^{\star(i)}\|_2.\end{aligned} \tag{63}$$

We will also use the sharper bound below later for finding the Frobenius norm of $\mathbf{H}^{(\ell)}$

$$\|\mathbf{d}_{1,2}^{(i,\ell)}\|_2 = \|(\mathbf{U}^{+(\ell-1)})^\mathsf{T} - (\mathbf{U}^{+(\ell-1)})^\mathsf{T}\mathbf{U}^\star(\mathbf{U}^\star)^\mathsf{T}\|_\mathsf{F}\|(\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\|_2. \tag{64}$$

Using equation 62 and equation 63 in equation 60, we have

$$\begin{aligned}\|\mathbf{d}_1^{(i,\ell)}\|_2 \le{}& c\sqrt{\frac{\log(r/\delta_0)}{m}}\|\mathbf{U}^{+(\ell-1)}\|_{\mathsf{F}}\|(\mathbf{U}^{\star}\mathbf{Q}^{(\ell-1)} - \mathbf{U}^{+(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\|_2 \\ &+ \|\mathbf{U}^{+(\ell-1)} - \mathbf{U}^{\star}\mathbf{Q}^{(\ell-1)}\|_{\mathsf{F}}^2\|(\mathbf{Q}^{(\ell-1)})^{-1}\|\|\mathbf{w}^{\star(i)}\|_2 \\ \le{}& \|\mathbf{U}^{+(\ell-1)} - \mathbf{U}^{\star}\mathbf{Q}^{(\ell-1)}\|_{\mathsf{F}}\|(\mathbf{Q}^{(\ell-1)})^{-1}\|\|\mathbf{w}^{\star(i)}\|_2 \cdot \\ & \Big(\|\mathbf{U}^{+(\ell-1)} - \mathbf{U}^{\star}\mathbf{Q}^{(\ell-1)}\|_{\mathsf{F}} + c\sqrt{\frac{\log(r/\delta_0)}{m}}\|\mathbf{U}^{+(\ell-1)}\|_{\mathsf{F}}\Big) \end{aligned} \tag{65}$$

As before, using equation 62 and equation 64, we have

$$\begin{aligned}\|\mathbf{d}_1^{(i,\ell)}\|_2 \le{}& \|(\mathbf{U}^{+(\ell-1)} - \mathbf{U}^{\star}\mathbf{Q}^{(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\|_2 \cdot \\ & \Big(\|\mathbf{U}^{+(\ell-1)} - \mathbf{U}^{\star}\mathbf{Q}^{(\ell-1)}\|_{\mathsf{F}} + c\sqrt{\frac{\log(r/\delta_0)}{m}}\|\mathbf{U}^{(\ell-1)}\|_{\mathsf{F}}\Big). \end{aligned} \tag{66}$$

Analysis of $\mathbf{d}_2^{(i,\ell)}$:
Note that

$$\begin{aligned}\mathbf{d}_2^{(i,\ell)} &= \frac{1}{m}(\mathbf{U}^{+(\ell-1)})^{\mathsf{T}}(\mathbf{X}^{(i)})^{\mathsf{T}}\mathbf{X}^{(i)}(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}) \\ &= \underbrace{(\mathbf{U}^{+(\ell-1)})^{\mathsf{T}}\Big(\frac{1}{m}(\mathbf{X}^{(i)})^{\mathsf{T}}\mathbf{X}^{(i)} - \mathbf{I}\Big)(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})}_{\mathbf{d}_{2,1}^{(i,\ell)}} + \underbrace{(\mathbf{U}^{+(\ell-1)})^{\mathsf{T}}(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})}_{\mathbf{d}_{2,2}^{(i,\ell)}} \end{aligned}$$

and $\mathbb{E}\left[\mathbf{d}_{2,1}^{(i,\ell)}\right] = \mathbb{E}\left[(\mathbf{U}^{+(\ell-1)})^{\mathsf{T}}\Big(\frac{1}{m}(\mathbf{X}^{(i)})^{\mathsf{T}}\mathbf{X}^{(i)} - \mathbf{I}\Big)(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})\right] = \mathbf{0}$.

Further,

$$(z_{2,1}^{(i,\ell)})_k = \frac{1}{m}\sum_j (\mathbf{u}^{(k,\ell-1)})^{\mathsf{T}}\mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^{\mathsf{T}}(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}) - (\mathbf{u}^{(k,\ell-1)})^{\mathsf{T}}(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}).$$

Using Lemma 17 in the above with $\mathbf{a} = \mathbf{u}^{(k,\ell-1)}$ and $\mathbf{b} = (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})$, we get

$$(z_{2,1}^{(i,\ell)})_k \le c\sqrt{\frac{\log(1/\delta_0)}{m}}\|\mathbf{u}^{(k,\ell-1)}\|_2\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2. \tag{67}$$

Taking the Union Bound overall entries $k \in [r]$, using the above we have

$$\begin{aligned}\|\mathbf{d}_{2,1}^{(i,\ell)}\|_2 &= \sqrt{\sum_{k\in[r]} |(z_{2,1}^{(i,\ell)})_k|^2} \\ &\le c\sqrt{\frac{\log(r/\delta_0)}{m}}\sqrt{\sum_{k\in[r]}\|\mathbf{u}^{(k,\ell-1)}\|_2^2} \cdot \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2 \\ &= c\sqrt{\frac{\log(r/\delta_0)}{m}}\|\mathbf{U}^{+(\ell-1)}\|_{\mathsf{F}}\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2. \end{aligned} \tag{68}$$

Further,

$$\begin{aligned}\|\mathbf{d}_{2,2}^{(i,\ell)}\|_2 &= \|(\mathbf{U}^{+(\ell-1)})^{\mathsf{T}}(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})\|_2 \\ &= \|\Big((\mathbf{U}^{+(\ell-1)})^{\mathsf{T}}(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})\Big)_{\mathsf{supp}(\mathbf{b}^{\star(i)})}\|_2 \\ &\le \|\mathbf{U}^{+(\ell-1)}_{\mathsf{supp}(\mathbf{b}^{\star(i)})}\|_2\|(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})_{\mathsf{supp}(\mathbf{b}^{\star(i)})}\|_2 \\ &\le \sqrt{k}\|\mathbf{U}^{+(\ell-1)}\|_{2,\infty}\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2. \end{aligned} \tag{69}$$

Using equation 68 and equation 69 we have

$$\|\mathbf{d}_2^{(i,\ell)}\|_2 \le \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\Big(\sqrt{k}\|\mathbf{U}^{+(\ell-1)}\|_{2,\infty} + c\sqrt{\frac{\log(r/\delta_0)}{m}}\|\mathbf{U}^{+(\ell-1)}\|_{\mathsf{F}}\Big). \tag{70}$$

Using equation 59, equation 65 and equation 70 we have

$$\begin{aligned}\|\mathbf{h}^{(i,\ell)}\|_2 \leq \frac{1}{1-c\sqrt{\frac{r\log(1/\delta_0)}{m}}} \cdot \Big\{&\|\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_{\mathsf{F}}\|(\mathbf{Q}^{(\ell-1)})^{-1}\|\|\mathbf{w}^{\star(i)}\|_2 \cdot \\ &\Big(\|\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_{\mathsf{F}} + c\sqrt{\frac{\log(r/\delta_0)}{m}}\|\mathbf{U}^{(\ell-1)}\|_{\mathsf{F}}\Big) \\ &+ \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\Big(\sqrt{k}\|\mathbf{U}^{+(\ell-1)}\|_{2,\infty} + c\sqrt{\frac{\log(r/\delta_0)}{m}}\|\mathbf{U}^{+(\ell-1)}\|_{\mathsf{F}}\Big) + \sigma\sqrt{\frac{r\log^2(r\delta^{-1})}{m}}\Big\}.\end{aligned} \tag{71}$$

Using equation 59, equation 66 and equation 70, we also have the sharper bound

$$\begin{aligned}\|\mathbf{h}^{(i,\ell)}\|_2 \leq \frac{1}{1-c\sqrt{\frac{r\log(1/\delta_0)}{m}}} \Big\{&\|(\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\|_2 \cdot \\ &\Big(\|\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_{\mathsf{F}} + c\sqrt{\frac{\log(r/\delta_0)}{m}}\|\mathbf{U}^{(\ell-1)}\|_{\mathsf{F}}\Big) \\ &+ \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\Big(\sqrt{k}\|\mathbf{U}^{+(\ell-1)}\|_{2,\infty} + c\sqrt{\frac{\log(r/\delta_0)}{m}}\|\mathbf{U}^{+(\ell-1)}\|_{\mathsf{F}}\Big) + \sigma\sqrt{\frac{r\log^2(r\delta^{-1})}{m}}\Big\}.\end{aligned} \tag{72}$$

Further note that $\sum_{i\in[t]}\|(\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\|_2^2$:

$$\begin{aligned}&= \sum_{i\in[t]} \mathsf{Tr}\Big(\Big((\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\Big)^{\mathsf{T}} \cdot \\ &\qquad \Big((\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\Big)\Big) \\ &= \sum_{i\in[t]} \mathsf{Tr}\Big((\mathbf{w}^{\star(i)})^{\mathsf{T}}\Big((\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\Big)^{\mathsf{T}} \cdot \\ &\qquad \Big((\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\Big)\mathbf{w}^{\star(i)}\Big) \\ &= \mathsf{Tr}\Big(\Big(\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\Big)^{\mathsf{T}} \cdot \\ &\qquad \Big((\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\Big)\sum_{i\in[t]}\mathbf{w}^{\star(i)}(\mathbf{w}^{\star(i)})^{\mathsf{T}}\Big) \\ &\leq \mathsf{Tr}\Big(\Big(\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\Big)^{\mathsf{T}} \cdot \\ &\qquad \Big((\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\Big)\Big)\lambda_{\max}\Big(\sum_{i\in[t]}\mathbf{w}^{\star(i)}(\mathbf{w}^{\star(i)})^{\mathsf{T}}\Big) \\ &= \|\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_{\mathsf{F}}^2\|(\mathbf{Q}^{(\ell-1)})^{-1}\|^2 \cdot \frac{t}{r}\lambda_1^\star,\end{aligned} \tag{73}$$

and $\sum_{i\in[t]} \|(\mathbf{U}^{+(\ell-1)})^{\mathsf{T}}(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})\|_2^2$

$$
\begin{aligned}
&= \sum_{i\in[t]}\sum_{j\in[r]} \Big((\mathbf{U}^{+(\ell-1,j)})^{\mathsf{T}}(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})\Big)^2 \\
&= \sum_{i\in[t]}\sum_{j\in[r]} \Big(\sum_{p\in\mathsf{supp}(\mathbf{b}^{\star(i)})} \mathbf{U}_p^{+(\ell-1,j)}(\mathbf{b}_p^{\star(i)} - \mathbf{b}_p^{(i,\ell)})\Big)^2 \\
&\le k\sum_{i\in[t]}\sum_{j\in[r]}\sum_{p\in\mathsf{supp}(\mathbf{b}^{\star(i)})} (\mathbf{U}_p^{+(\ell-1,j)})^2(\mathbf{b}_p^{\star(i)} - \mathbf{b}_p^{(i,\ell)})^2 \\
&\le k\zeta\sum_{j\in[r]}\sum_{p\in\mathsf{supp}(\mathbf{b}^{\star(i)})} (\mathbf{U}_p^{+(\ell-1,j)})^2\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_\infty^2 \\
&= k\zeta\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_\infty^2\sum_{j\in[r]}\sum_{p\in\mathsf{supp}(\mathbf{b}^{\star(i)})} (\mathbf{U}_p^{+(\ell-1,j)})^2 \\
&\le k\zeta\|\mathbf{U}^\star\|_{\mathsf{F}}^2\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_\infty^2.
\end{aligned}
$$

The above two equations with equation 72 imply

$$
\begin{aligned}
\|\mathbf{H}^{(\ell)}\|_{\mathsf{F}} \le{}& \frac{1}{1-c\sqrt{\frac{r\log(1/\delta_0)}{m}}}\Big\{\|\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_{\mathsf{F}}\|(\mathbf{Q}^{(\ell-1)})^{-1}\|\sqrt{\frac{t}{r}\lambda_1^\star}\cdot \\
&\Big(\|\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_{\mathsf{F}} + c\sqrt{\frac{\log(r/\delta_0)}{m}}\|\mathbf{U}^{(\ell-1)}\|_{\mathsf{F}}\Big) \\
&+ \sqrt{k\zeta}\|\mathbf{U}^\star\|_{\mathsf{F}}\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_\infty + \sqrt{\sum_{i\in[t]}\Big(c\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\sqrt{\frac{\log(r/\delta_0)}{m}}\|\mathbf{U}^{+(\ell-1)}\|_{\mathsf{F}}\Big)^2} \\
&+ \sigma\sqrt{\frac{rt\log^2(r\delta^{-1})}{m}}\Big\}. \qquad (74)
\end{aligned}
$$

□

Corollary 1. *If* $\mathsf{B}_{\mathbf{U}^{(\ell-1)}} = \mathcal{O}\Big(\frac{1}{1\sqrt{r\mu^\star}}\Big)$, $\sqrt{\frac{r\log(1/\delta_0)}{m}} = \mathcal{O}(1)$, $\sqrt{\nu^{(\ell-1)}} = \mathcal{O}\Big(\frac{1}{\sqrt{r\mu^\star}}\Big)$, $\sqrt{\frac{r^2\log(r/\delta_0)}{m}} = \mathcal{O}\Big(\frac{1}{\sqrt{\mu^\star}}\Big)$, $\epsilon < \sqrt{\mu^\star\lambda_r^\star}$, $\sqrt{\frac{r^2\zeta}{t}} = \mathcal{O}\Big(\frac{1}{\sqrt{\mu^\star}}\Big)$, $\Lambda' = \mathcal{O}\Big(\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big)$ *and Assumption 4 holds for iteration* $\ell - 1$, *then w.p.* $1 - \mathcal{O}(\delta_0)$

$$
\begin{aligned}
\|\mathbf{h}^{(i,\ell)}\|_2 &= \mathcal{O}\Big(\frac{\max\{\epsilon,\|\mathbf{w}^{\star(i)}\|_2\}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\frac{\Lambda'\|\mathbf{w}^{\star(i)}\|_2}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\frac{\Lambda}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\sigma\sqrt{\frac{r\log^2(r\delta^{-1})}{m}}\Big), \\
\|\mathbf{H}^{(\ell)}\|_{\mathsf{F}} &\le \mathcal{O}\Big(\frac{\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\frac{\lambda_r^\star}{\lambda_1^\star}\sqrt{\frac{t}{r}\lambda_1^\star}}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\Lambda'\sqrt{\frac{t}{r}\lambda_r^\star}\frac{1}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\frac{1}{\sqrt{r\mu^\star}}\sqrt{\frac{t}{r}}\Lambda\Big) + \mathcal{O}\Big(\sigma\sqrt{\frac{rt\log^2(r\delta^{-1})}{m}}\Big).
\end{aligned}
$$

Proof. The proof follows from plugging the various constant bounds of the lemma statement and Inductive Assumption 4 in the expressions of Lemma 6:

$$
\begin{aligned}
\|\mathbf{h}^{(i,\ell)}\|_2 \le{}& \frac{1}{1-c\sqrt{\frac{r\log(1/\delta_0)}{m}}}\Big\{\|\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_{\mathsf{F}}\|(\mathbf{Q}^{(\ell-1)})^{-1}\|\|\mathbf{w}^{\star(i)}\|_2\cdot \\
&\Big(\|\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_{\mathsf{F}} + c\sqrt{\frac{\log(r/\delta_0)}{m}}\|\mathbf{U}^{+(\ell-1)}\|_{\mathsf{F}}\Big) \\
&+ \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\Big(\sqrt{k}\|\mathbf{U}^{+(\ell-1)}\|_{2,\infty} + c\sqrt{\frac{\log(r/\delta_0)}{m}}\|\mathbf{U}^{+(\ell-1)}\|_{\mathsf{F}}\Big) + \sigma\sqrt{\frac{r\log^2(r\delta^{-1})}{m}}\Big\}.
\end{aligned}
$$

$$
\begin{aligned}
&\leq \frac{1}{1 - c\sqrt{\frac{r\log(1/\delta_0)}{m}}} \cdot \\
&\Big\{ \Big(\mathsf{B}_{\mathbf{U}^{(\ell-1)}} \sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} + \Lambda' \Big) \cdot 2\|\mathbf{w}^{\star(i)}\|_2 \Big(\Big(\mathsf{B}_{\mathbf{U}^{(\ell-1)}} \sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} + \Lambda' \Big) + c\sqrt{\frac{\log(r/\delta_0)}{m}} \cdot \sqrt{r} \Big) \\
&+ \Big(c' \max\{\epsilon, \|\mathbf{w}^{\star(i)}\|_2\} \mathsf{B}_{\mathbf{U}^{(\ell-1)}} \sqrt{\tfrac{\lambda_r^\star}{\lambda_1^\star}} + \Lambda \Big) \Big(\sqrt{\nu^{(\ell-1)}} + c\sqrt{\tfrac{\log(r/\delta_0)}{m}} \sqrt{r} \Big) + \sigma \sqrt{\tfrac{r\log^2(r\delta^{-1})}{m}} \Big\}.
\end{aligned} \tag{75}
$$

Using $\Lambda' = \mathcal{O}\Big(\frac{1}{\sqrt{r\mu^\star}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big)$, the above becomes

$$
\begin{aligned}
&\leq \frac{\max\{\epsilon, \|\mathbf{w}^{\star(i)}\|_2\} \mathsf{B}_{\mathbf{U}^{(\ell-1)}} \sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}}{1 - c\sqrt{\frac{r\log(1/\delta_0)}{m}}} \cdot \\
&\quad \Big\{ 2\Big(\mathsf{B}_{\mathbf{U}^{(\ell-1)}} + \mathcal{O}\Big(\frac{1}{\sqrt{r\mu^\star}} \sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} \Big) + c\sqrt{\frac{r\log(r/\delta_0)}{m}} \Big) + c'\Big(\sqrt{\nu^{(\ell-1)}} + c\sqrt{\frac{r\log(r/\delta_0)}{m}} \Big) \Big\} \\
&+ \frac{1}{1 - c\sqrt{\frac{r\log(1/\delta_0)}{m}}} \Big\{ 2\Lambda' \|\mathbf{w}^{\star(i)}\|_2 \Big(\mathsf{B}_{\mathbf{U}^{(\ell-1)}} + \mathcal{O}\Big(\frac{1}{\sqrt{r\mu^\star}} \sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} \Big) + c\sqrt{\frac{r\log(r/\delta_0)}{m}} \Big) \\
&\quad + \Lambda\Big(\sqrt{\nu^{(\ell-1)}} + c\sqrt{\frac{r\log(r/\delta_0)}{m}} \Big) + \sigma\sqrt{\frac{r\log^2(r\delta^{-1})}{m}} \Big\}.
\end{aligned} \tag{76}
$$

Further, using $c\sqrt{\frac{r\log(1/\delta_0)}{m}} = \mathcal{O}(1)$, $\mathsf{B}_{\mathbf{U}^{(\ell-1)}} = \mathcal{O}\Big(\frac{1}{\sqrt{r\mu^\star}}\Big)$, $\sqrt{\nu^{(\ell-1)}} = \mathcal{O}\Big(\frac{1}{\sqrt{r\mu^\star}}\Big)$, $\sqrt{\frac{r^2\log(r/\delta_0)}{m}} = \mathcal{O}\Big(\frac{1}{\sqrt{\mu^\star}}\Big)$, $\lambda_r^\star \leq \lambda_1^\star$, $r \geq 1$ in the above, we get

$$
= \mathcal{O}\Big(\frac{\max\{\epsilon, \|\mathbf{w}^{\star(i)}\|_2\} \mathsf{B}_{\mathbf{U}^{(\ell-1)}} \sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}}{\sqrt{r\mu^\star}} \Big) + \mathcal{O}\Big(\frac{\Lambda' \|\mathbf{w}^{\star(i)}\|_2}{\sqrt{r\mu^\star}} \Big) + \mathcal{O}\Big(\frac{\Lambda}{\sqrt{r\mu^\star}} \Big) + \mathcal{O}\Big(\sigma\sqrt{\frac{r\log^2(r\delta^{-1})}{m}} \Big). \tag{77}
$$

Similarly using the Inductive Assumption expressions from 4, we also have

$$
\begin{aligned}
\|\mathbf{H}^{(\ell)}\|_{\mathsf{F}} &\leq \frac{1}{1 - c\sqrt{\frac{r\log(1/\delta_0)}{m}}} \Big\{ \|\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_{\mathsf{F}} \|(\mathbf{Q}^{(\ell-1)})^{-1}\| \sqrt{\frac{t}{r}\lambda_1^\star} \cdot \\
&\quad \Big(\|\mathbf{U}^{+(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_{\mathsf{F}} + c\sqrt{\frac{\log(r/\delta_0)}{m}} \|\mathbf{U}^{(\ell-1)}\|_{\mathsf{F}} \Big) \\
&\quad + \sqrt{k\zeta} \|\mathbf{U}^\star\|_{\mathsf{F}} \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_\infty + \sqrt{\sum_{i\in[t]} \Big(c\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2 \sqrt{\frac{\log(r/\delta_0)}{m}} \|\mathbf{U}^{+(\ell-1)}\|_{\mathsf{F}} \Big)^2} \\
&\quad + \sigma\sqrt{\frac{rt\log^2(r\delta^{-1})}{m}} \Big\}
\end{aligned} \tag{78}
$$

$$
\begin{aligned}
&\leq \frac{1}{1 - c\sqrt{\frac{r\log(1/\delta_0)}{m}}} \Big\{ \Big(\mathsf{B}_{\mathbf{U}^{(\ell-1)}} \sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} + \Lambda' \Big) \cdot 2\sqrt{\frac{t}{r}\lambda_1^\star} \Big(\mathsf{B}_{\mathbf{U}^{(\ell-1)}} \sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} + \Lambda' + c\sqrt{\frac{\log(r/\delta_0)}{m}} \sqrt{r} \Big) \\
&\quad + \sqrt{k\zeta}\sqrt{r} \Big(c' \max\{\epsilon, \|\mathbf{w}^{\star(i)}\|_2\} \mathsf{B}_{\mathbf{U}^{(\ell-1)}} \sqrt{\frac{\lambda_r^\star}{\lambda_1^\star k}} + \frac{\Lambda}{\sqrt{k}} \Big) \\
&\quad + c\sqrt{\frac{\log(r/\delta_0)}{m}} \sqrt{r} \cdot \Big(c' \max\{\epsilon, \|\mathbf{w}^{\star(i)}\|_2\} \mathsf{B}_{\mathbf{U}^{(\ell-1)}} \sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} + \Lambda \Big) \sqrt{t} + \sigma\sqrt{\frac{rt\log^2(r\delta^{-1})}{m}} \Big\}.
\end{aligned} \tag{79}
$$

As before, using the bound on Λ', the above becomes

$$\begin{aligned}
&\leq \frac{\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}}{1-c\sqrt{\frac{r\log(1/\delta_0)}{m}}}\Big\{2\sqrt{\frac{t}{r}\lambda_1^\star}\Big(\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}+\mathcal{O}\Big(\frac{1}{\sqrt{r\mu^\star}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big)+c\sqrt{\frac{r\log(r/\delta_0)}{m}}\Big) \\
&\quad+\sqrt{k\zeta}\sqrt{r}c'\max\{\epsilon,\|\mathbf{w}^{\star(i)}\|_2\}\frac{1}{\sqrt{k}}+c\sqrt{\frac{r\log(r/\delta_0)}{m}}\cdot c'\sqrt{t}\max\{\epsilon,\sqrt{\mu^\star\lambda_r^\star}\}\Big\} \\
&\quad\frac{1}{1-c\sqrt{\frac{r\log(1/\delta_0)}{m}}}\cdot\Big\{2\Lambda'\sqrt{\frac{t}{r}\lambda_1^\star}\Big(\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}+\mathcal{O}\Big(\frac{1}{\sqrt{r\mu^\star}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big)+c\sqrt{\frac{\log(r/\delta_0)}{m}}\sqrt{r}\Big) \\
&\quad+\sqrt{r\zeta}\Lambda+c\sqrt{\frac{\log(r/\delta_0)}{m}}\sqrt{r}\cdot\Lambda\sqrt{t}+\sigma\sqrt{\frac{rt\log^2(r\delta^{-1})}{m}}\Big\}
\end{aligned} \tag{80}$$

Further, using $c\sqrt{\frac{r\log(1/\delta_0)}{m}}=\mathcal{O}(1)$, $\mathsf{B}_{\mathbf{U}^{(\ell-1)}}=\mathcal{O}\Big(\frac{1}{\sqrt{r\mu^\star}}\Big)$, $\sqrt{\nu^{(\ell-1)}}=\mathcal{O}\Big(\frac{1}{\sqrt{r\mu^\star}}\Big)$, $\sqrt{\frac{r^2\log(r/\delta_0)}{m}}=\mathcal{O}\Big(\frac{1}{\sqrt{\mu^\star}}\Big)$, $\lambda_r^\star\leq\lambda_1^\star$, $r\geq 1$ and $\sqrt{\frac{r^2\zeta}{t}}=\mathcal{O}\Big(\frac{1}{\sqrt{\mu^\star}}\Big)$ in the above, we get

$$\begin{aligned}
&\leq \frac{\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\frac{\lambda_r^\star}{\lambda_1^\star}\sqrt{\frac{t}{r}\lambda_1^\star}}{1-c\sqrt{\frac{r\log(1/\delta_0)}{m}}}\Big\{2\Big(\mathsf{B}_{\mathbf{U}^{(\ell-1)}}+\frac{1}{181\sqrt{r\mu^\star}}+c\sqrt{\frac{r\log(r/\delta_0)}{m}}\sqrt{\frac{\lambda_1^\star}{\lambda_r^\star}}\Big) \\
&\quad+\frac{1}{\sqrt{\lambda_r^\star}}\max\{\epsilon,\sqrt{\mu^\star\lambda_r^\star}\}c'\Big(\sqrt{\frac{r^2\zeta}{t}}+c\sqrt{\frac{r^2\log(r/\delta_0)}{m}}\cdot\Big)\Big\} \\
&\quad+\mathcal{O}(1)\cdot\Big\{2\Lambda'\sqrt{\frac{t}{r}\lambda_r^\star}\cdot\mathcal{O}\Big(\frac{1}{\sqrt{r\mu^\star}}\Big)+\sqrt{r\zeta}\Lambda+\mathcal{O}\Big(\frac{1}{\sqrt{r\mu^\star}}\Big)\cdot\Lambda\sqrt{t}+\sigma\sqrt{\frac{rt\log^2(r\delta^{-1})}{m}}\Big\}
\end{aligned} \tag{81}$$

$$=\mathcal{O}\Big(\frac{\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\frac{\lambda_r^\star}{\lambda_1^\star}\sqrt{\frac{t}{r}\lambda_1^\star}}{\sqrt{r\mu^\star}}\Big)+\mathcal{O}\Big(\Lambda'\sqrt{\frac{t}{r}\lambda_r^\star}\frac{1}{\sqrt{r\mu^\star}}\Big)+\mathcal{O}\Big(\frac{1}{\sqrt{r\mu^\star}}\sqrt{\frac{t}{r}}\Lambda\Big)+\mathcal{O}\Big(\sigma\sqrt{\frac{rt\log^2(r\delta^{-1})}{m}}\Big). \tag{82}$$

□

Corollary 2. *If Assumption 4 and Corollary 1 hold, and* $\Lambda'=\mathcal{O}\Big(\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big)$, $\Lambda=\mathcal{O}(\sqrt{\lambda_r^\star})$ *and* $\sigma\sqrt{\frac{r^2\log^2(r\delta^{-1})}{m}}=\mathcal{O}(\sqrt{\lambda_r^\star})$ *then,* $\mathbf{W}^{(\ell)}$ *is incoherent w.p. probability* $1-\mathcal{O}(\delta_0)$ *for* $\exists c''>0$*, s.t.*

$$\begin{gathered}
\|\mathbf{w}^{(i,\ell)}\|_2\leq(2+c'')\sqrt{\mu^\star\lambda_r\Big(\frac{r}{t}(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^\star\Big)}, \\
(1-c'')\sqrt{\lambda_r^\star\Big(\frac{r}{t}(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^\star\Big)}\leq\sqrt{\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)}\leq(1+c'')\sqrt{\lambda_r^\star\Big(\frac{r}{t}(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^\star\Big)}, \\
\|\mathbf{W}^{(\ell)}\|\leq(1+c'')\sqrt{\frac{t}{r}}\sqrt{\lambda_1\Big(\frac{r}{t}(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^\star\Big)}, \\
\sqrt{\mu^{(\ell)}}:=\frac{2+c''}{1-c''}\sqrt{\mu^\star}\leq(1+2c'')\sqrt{\mu^\star}.
\end{gathered}$$

Proof. Using Triangle Inequality, Assumption 4 and Corollary 1, we have

$$\begin{aligned}
\|\mathbf{w}^{(i,\ell)}\|_2 &\leq \|(\mathbf{Q}^{(\ell)})^{-1}\mathbf{w}^{\star(i)}\|_2 + \|\mathbf{h}^{(i,\ell)}\|_2 \\
&\leq 2\|\mathbf{w}^{(\star)}\|_2 + \mathcal{O}\Big(\frac{\max\{\epsilon, \|\mathbf{w}^{\star(i)}\|_2\}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\frac{\Lambda'\|\mathbf{w}^{\star(i)}\|_2}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\frac{\Lambda}{\sqrt{r\mu^\star}}\Big) \\
&\qquad + \mathcal{O}\Big(\sigma\sqrt{\frac{r\log^2(r\delta^{-1})}{m}}\Big). \\
&= \Big(2 + \mathcal{O}\Big(\frac{\mathsf{B}_{\mathbf{U}^{(\ell-1)}}}{\sqrt{r\mu^\star}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big) + \Big(\frac{\Lambda'}{\sqrt{r\mu^\star}}\Big) + \Big(\frac{1}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}(1)\Big)\sqrt{\mu^\star\lambda_r^\star} \\
&\leq (2 + c'')\sqrt{\mu^\star\lambda_r^\star},
\end{aligned} \tag{83}$$

for some $c'' > 0$, where in the last two lines, we use the fact that $\epsilon < \sqrt{\mu^\star\lambda_r^\star}$, $\Lambda' = \mathcal{O}\Big(\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big)$, $\Lambda = \Big(\sqrt{\mu^\star\lambda_r^\star}\Big)$ and $\sigma\sqrt{\frac{r\log^2(r\delta^{-1})}{m}} = \mathcal{O}(\sqrt{\mu^\star\lambda_r^\star})$ and $r, \mu^\star > 1$. Using Lemma 21 with $\mathbf{A} = (\mathbf{Q}^{(\ell-1)})^{-1}$ and $\mathbf{B} = \frac{r}{t}(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^\star$ we have

$$\sigma_{\min}^2\Big((\mathbf{Q}^{(\ell-1)})^{-1}\Big)\lambda_r\Big(\frac{r}{t}(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^\star\Big) \leq \frac{r}{t}\lambda_r\Big((\mathbf{Q}^{(\ell-1)})^{-1}(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^\star(\mathbf{Q}^{(\ell-1)})^{-\mathsf{T}}\Big) \tag{84}$$

Further, since $\sigma_{\min}\Big((\mathbf{Q}^{(\ell-1)})^{-1}\Big) = \sigma_{\min}\Big(((\mathbf{U}^\star)^\mathsf{T}\mathbf{U}^{+(\ell-1)})^{-1}\Big) > 1$, we have $\forall\, j \in [r]$

$$\lambda_j\Big(\frac{r}{t}(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^\star\Big) \leq \sigma_{\min}^2\Big((\mathbf{Q}^{(\ell-1)})^{-1}\Big)\lambda_j\Big(\frac{r}{t}(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^\star\Big) \tag{85}$$

Using equation 85 in equation 84 we get

$$\begin{aligned}
\lambda_j\Big(\frac{r}{t}(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^\star\Big) &\leq \frac{r}{t}\lambda_j\Big((\mathbf{Q}^{(\ell-1)})^{-1}(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^\star(\mathbf{Q}^{(\ell-1)})^{-\mathsf{T}}\Big) \\
&\leq \frac{r}{t}\sigma_j^2\Big((\mathbf{Q}^{(\ell-1)})^{-1}(\mathbf{W}^\star)^\mathsf{T}\Big)
\end{aligned} \tag{86}$$

Now, since $\mathbf{W}^\star(\mathbf{Q}^{(\ell-1)})^{-\mathsf{T}} = \mathbf{W}^{(\ell)} + \mathbf{W}^\star(\mathbf{Q}^{(\ell-1)})^{-\mathsf{T}} - \mathbf{W}^{(\ell)}$, Using Lemma 18 with $\mathbf{A} = \mathbf{W}^\star(\mathbf{Q}^{(\ell-1)})^{-\mathsf{T}}$, $\mathbf{B} = \mathbf{W}^{(\ell)}$ and $\mathbf{C} = \mathbf{W}^\star(\mathbf{Q}^{(\ell-1)})^{-\mathsf{T}} - \mathbf{W}^{(\ell)}$, we have

$$\Big|\sigma_j\Big(\mathbf{W}^\star(\mathbf{Q}^{(\ell-1)})^{-\mathsf{T}}\Big) - \sigma_j\Big(\mathbf{W}^{(\ell)}\Big)\Big| \leq \|\mathbf{W}^\star(\mathbf{Q}^{(\ell-1)})^{-\mathsf{T}} - \mathbf{W}^{(\ell)}\|. \tag{87}$$

Using equation 86 and Corollary 1 in equation 87, we have

$$\begin{aligned}
&\Big|\sqrt{\lambda_j\Big(\frac{r}{t}(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^\star\Big)} - \sqrt{\lambda_j\Big(\frac{r}{t}(\mathbf{W}^{(i,\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)}\Big| \leq \sqrt{\frac{r}{t}}\|\mathbf{W}^\star(\mathbf{Q}^{(\ell-1)})^{-\mathsf{T}} - \mathbf{W}^{(\ell)}\| \\
&\leq \sqrt{\frac{r}{t}}\|\mathbf{H}^{(\ell)}\| \\
&\leq \sqrt{\frac{r}{t}}\|\mathbf{H}^{(\ell)}\|_\mathsf{F} \\
&\leq \sqrt{\tfrac{r}{t}}\Big\{\mathcal{O}\Big(\frac{\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\frac{\lambda_r^\star}{\lambda_1^\star}\sqrt{\frac{t}{r}\lambda_1^\star}}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\Lambda'\sqrt{\tfrac{t}{r}\lambda_r^\star}\tfrac{1}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\tfrac{1}{\sqrt{r\mu^\star}}\sqrt{\tfrac{t}{r}}\Lambda\Big) + \mathcal{O}\Big(\sigma\sqrt{\tfrac{rt\log^2(r\delta^{-1})}{m}}\Big)\Big\} \\
&= \mathcal{O}\Big(\frac{\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\frac{\lambda_r^\star}{\lambda_1^\star}\sqrt{\lambda_1^\star}}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\frac{\Lambda'\sqrt{\lambda_r^\star}}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\frac{\Lambda}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\sigma\sqrt{\frac{r^2\log^2(r\delta^{-1})}{m}}\Big) \\
&= \Big\{\mathcal{O}\Big(\frac{\mathsf{B}_{\mathbf{U}^{(\ell-1)}}}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\frac{\Lambda'}{\sqrt{r\mu^\star}}\Big) + \Big(\frac{\Lambda}{\sqrt{r\mu^\star\lambda_r^\star}}\Big) + \Big(\sigma\sqrt{\frac{r^2\log^2(r\delta^{-1})}{m\lambda_r^\star}}\Big\}\sqrt{\lambda_r^\star} \\
&= \mathcal{O}(\sqrt{\lambda_r^\star})
\end{aligned} \tag{88}$$

where in the last two steps we use $\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} < 1, \Lambda' = \mathcal{O}\Big(\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big), \Lambda = \mathcal{O}(\sqrt{\lambda_r^\star})$ and $\sigma\sqrt{\frac{r^2\log^2(r\delta^{-1})}{m}} = \mathcal{O}(\sqrt{\lambda_r^\star})$. Note that for $j = r$, the above implies for some $c'' > 0$

$$\begin{aligned}
&\Big|\sqrt{\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)} - \sqrt{\lambda_r^\star\Big(\frac{r}{t}(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^\star\Big)}\Big| \le c''\sqrt{\lambda_r^\star\Big(\frac{r}{t}(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^\star\Big)} \\
&\iff (1-c'')\sqrt{\lambda_r^\star\Big(\frac{r}{t}(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^\star\Big)} \le \sqrt{\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)} \le (1+c'')\sqrt{\lambda_r^\star\Big(\frac{r}{t}(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^\star\Big)}.
\end{aligned} \tag{89}$$

and for $j = 1$,

$$\begin{aligned}
\sqrt{\lambda_1\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)} &\le \sqrt{\lambda_1\Big(\frac{r}{t}(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^\star\Big)} + c''\sqrt{\lambda_r^\star\Big(\frac{r}{t}(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^\star\Big)} \\
\implies \|\mathbf{W}^{(\ell)}\| &\le (1+c'')\sqrt{\frac{t}{r}}\sqrt{\lambda_1\Big(\frac{r}{t}(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^\star\Big)} \qquad (90) \\
&= \mathcal{O}\Big(\sqrt{\frac{t}{r}}\sqrt{\lambda_1\Big(\frac{r}{t}(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^\star\Big)}\Big) \qquad (91)
\end{aligned}$$

☐

Lemma 7. *If* $\|\mathbf{U}^{(\ell)} - \mathbf{U}^\star(\mathbf{U}^\star)^\mathsf{T}\mathbf{U}^{(\ell)}\| \le \|\mathbf{U}^{(\ell)} - \mathbf{U}^\star(\mathbf{U}^\star)^\mathsf{T}\mathbf{U}^{(\ell)}\|_\mathsf{F} \le \frac{3}{4}$ *then* $\frac{1}{2} \le \|\mathbf{Q}^{(\ell)}\| \le 1$

Proof. **Upper Bound:**

$$\|\mathbf{Q}^{(\ell)}\| = \|(\mathbf{U}^\star)^\mathsf{T}\mathbf{U}^{(\ell)}\| \le \|\mathbf{U}^\star\|\|\mathbf{U}^{(\ell)}\| \le 1,$$

since both $\mathbf{U}^\star, \mathbf{U}^{(\ell)} \in \mathbb{R}^{d\times r}$ are orthonormal.

Lower Bound:

Now, let $\mathbf{E} := \mathbf{U}^{(\ell)} - \mathbf{U}^\star(\mathbf{U}^\star)^\mathsf{T}\mathbf{U}^{(\ell)}$ and $\mathbf{Q} = (\mathbf{U}^\star)^\mathsf{T}\mathbf{U}^{(\ell)}$. Then $(\mathbf{U}^{(\ell)})^\mathsf{T}\mathbf{E} = \mathbf{I} - ((\mathbf{U}^\star)^\mathsf{T}\mathbf{U}^{(\ell)})^\mathsf{T}(\mathbf{U}^\star)^\mathsf{T}\mathbf{U}^{(\ell)} = \mathbf{I} - (\mathbf{Q}^{(\ell)})^\mathsf{T}\mathbf{Q}^{(\ell)}$. Then using Lemma 18 with $\mathbf{A} = \mathbf{I}, \mathbf{B} = (\mathbf{Q}^{(\ell)})^\mathsf{T}\mathbf{Q}^{(\ell)}$ and $\mathbf{C} = (\mathbf{U}^{(\ell)})^\mathsf{T}\mathbf{E}$, we get that

$$\begin{aligned}
\sigma_k(\mathbf{I}) - \sigma_k((\mathbf{Q}^{(\ell)})^\mathsf{T}\mathbf{Q}^{(\ell)}) &\le \|(\mathbf{U}^{(\ell)})^\mathsf{T}\mathbf{E}\| \le \|\mathbf{U}^{(\ell)}\|\|\mathbf{E}\| \\
\implies 1 - \sigma_k^2(\mathbf{Q}^{(\ell)}) &\le \|\mathbf{U}^{(\ell)} - \mathbf{U}^\star(\mathbf{U}^\star)^\mathsf{T}\mathbf{U}^{(\ell)}\| \\
\implies \sigma_k(\mathbf{Q}^{(\ell)}) &\ge \sqrt{1 - \|\mathbf{U}^{(\ell)} - \mathbf{U}^\star(\mathbf{U}^\star)^\mathsf{T}\mathbf{U}^{(\ell)}\|}.
\end{aligned}$$

Therefore, $\|\mathbf{U}^{(\ell)} - \mathbf{U}^\star(\mathbf{U}^\star)^\mathsf{T}\mathbf{U}^{(\ell)}\|_\mathsf{F} \le \frac{3}{4} \implies \sigma_k(\mathbf{Q}^{(\ell)}) \ge \frac{1}{2}\ \forall\ k \in [r]$. ☐

Lemma 8. *Let* $\mathbf{V} = \frac{1}{mt}\sum_i\sum_j \mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}\mathbf{R}^{(i)}$ *where both* $\mathbf{V}, \mathbf{R}^{(i)} \in \mathbb{R}^{d\times r}$. *Then,*

$$\|\mathbf{V}\|_\mathsf{F}^2 \le 2\sum_p\sum_q\Big(\frac{1}{t}\sum_i \mathbf{R}_{p,q}^{(i,\ell)}\Big)^2 + 16\|\mathbf{R}^{(i,\ell)}\|_\mathsf{F}^2\frac{d\log(rd/\delta_0)}{mt}.$$

Proof. Notice that then for any $(p, q) \in [d] \times [r]$, we have

$$\begin{aligned}
\mathbf{V}_{p,q} &= \Big(\frac{1}{mt}\sum_i\sum_j \mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}\mathbf{R}^{(i,\ell)}\Big)_{p,q} \\
&= \frac{1}{mt}\sum_i\sum_j \mathbf{x}_{j,p}^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}\mathbf{R}^{(i,\ell,q)} \\
&= \frac{1}{mt}\sum_i\sum_j \mathbf{x}_{j,p}^{(i)}\Big(\sum_{u=1}^{d}\mathbf{x}_{j,u}^{(i)}\mathbf{R}_{u,q}^{(i,\ell)}\Big) \\
&= \frac{1}{mt}\sum_i\sum_j\Big(\Big(\mathbf{x}_{j,p}^{(i)}\Big)^2\mathbf{R}_{p,q}^{(i,\ell)} + \sum_{u:u\neq p}\mathbf{x}_{j,p}^{(i)}\mathbf{x}_{j,u}^{(i)}\mathbf{R}_{u,q}^{(i,\ell)}\Big).
\end{aligned} \tag{92}$$

Now, note that the random variable $\left(\mathbf{x}_{j,p}^{(i)}\right)^2\mathbf{R}_{p,q}^{(i,\ell)}$ is a $\left(4(\mathbf{R}_{p,q}^{(i,\ell)})^2, 4|\mathbf{R}_{p,q}^{(i,\ell)}|\right)$ sub-exponential random variable. Similarly, $\mathbf{x}_{j,p}^{(i)}\mathbf{x}_{j,u}^{(i)}\mathbf{R}_{u,q}^{(i,\ell)}$ is a $\left(2(\mathbf{R}_{u,q}^{(i,\ell)})^2, \sqrt{2}|\mathbf{R}_{u,q}^{(i,\ell)}|\right)$ sub-exponential random variable. Therefore,

$$\begin{aligned}
&\left(\mathbf{x}_{j,p}^{(i)}\right)^2\mathbf{R}_{p,q}^{(i,\ell)} + \sum_{u:u\neq h}\mathbf{x}_{j,p}^{(i)}\mathbf{x}_{j,u}^{(i)}\mathbf{R}_{u,q}^{(i,\ell)} \\
&= \left(4(\mathbf{R}_{p,q}^{(i,\ell)})^2 + 2\sum_{u:u\neq p}(\mathbf{R}_{u,q}^{(i,\ell)})^2, \max\left(4|\mathbf{R}_{p,q}^{(i,\ell)}|, \max_{u:u\neq p}(\sqrt{2}|\mathbf{R}_{u,q}^{(i,\ell)}|)\right)\right) \\
&= \left(4\|\mathbf{R}^{(i,\ell,q)}\|_2^2, 4\|\mathbf{R}^{(i,\ell,q)}\|_\infty\right) \text{ sub-exponential random variable.} \qquad (93)
\end{aligned}$$

Furthermore,

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{V}_{p,q}\right] \\
&= \frac{1}{mt}\sum_i\sum_j\left(\mathbb{E}\left[\left(\mathbf{x}_{j,p}^{(i)}\right)^2\mathbf{R}_{p,q}^{(i,\ell)}\right] + \mathbb{E}\left[\sum_{u:u\neq p}\mathbf{x}_{j,p}^{(i)}\mathbf{x}_{j,u}^{(i)}\mathbf{R}_{u,q}^{(i,\ell)}\right]\right) \\
&= \frac{1}{mt}\sum_i\sum_j\left(\mathbf{R}_{p,q}^{(i,\ell)} + \mathbf{0}\right) \\
&= \frac{1}{t}\sum_i\mathbf{R}_{p,q}^{(i,\ell)}. \qquad (94)
\end{aligned}$$

Using equation 93, equation 94 and Lemma 23 in equation 92 gives

$$\left|\mathbf{V}_{p,q} - \frac{1}{t}\sum_i\mathbf{R}_{p,q}^{(i,\ell)}\right| \leq \underbrace{\max\left(2\|\mathbf{R}^{(i,\ell,q)}\|_2\sqrt{\frac{2\log(1/\delta_0)}{mt}}, 2\|\mathbf{R}^{(i,\ell,q)}\|_\infty\frac{2\log(1/\delta_0)}{mt}\right)}_{\epsilon_{p,q}}. \qquad (95)$$

Note that $\|\mathbf{V}\|_\mathsf{F}^2 = \sum_p\sum_q\mathbf{V}_{p,q}^2$. Hence taking a union bound over all entries $(p,q)\in[d]\times[r]$, we have

$$\begin{aligned}
\sum_p\sum_q\mathbf{V}_{p,q}^2 &\leq \sum_p\sum_q 2\left(\left(\frac{1}{t}\sum_i\mathbf{R}_{p,q}^{(i,\ell)}\right)^2 + \epsilon_{p,q}^2\right) \\
&\leq 2\sum_p\sum_q\left(\frac{1}{t}\sum_i\mathbf{R}_{p,q}^{(i,\ell)}\right)^2 + 2\sum_p\sum_q\left(2\|\mathbf{R}^{(i,\ell,q)}\|_2\sqrt{\frac{2\log(rd/\delta_0)}{mt}}\right)^2 \\
&\leq 2\sum_p\sum_q\left(\frac{1}{t}\sum_i\mathbf{R}_{p,q}^{(i,\ell)}\right)^2 + 8\sum_p\sum_q\|\mathbf{R}^{(i,\ell,q)}\|_2^2\frac{2\log(rd/\delta_0)}{mt} \\
&\leq 2\sum_p\sum_q\left(\frac{1}{t}\sum_i\mathbf{R}_{p,q}^{(i,\ell)}\right)^2 + 8\|\mathbf{R}^{(i,\ell)}\|_\mathsf{F}^2\frac{2d\log(rd/\delta_0)}{mt}
\end{aligned}$$

where we use that $2\|\mathbf{R}^{(i,\ell,q)}\|_2\sqrt{\frac{2\log(rd/\delta_0)}{mt}} > 2\|\mathbf{R}^{(i,\ell,q)}\|_\infty\frac{2\log(rd/\delta_0)}{mt}$. Hence, with probability at least $1-\delta_0$, we have

$$\|\mathbf{V}\|_\mathsf{F}^2 \leq 2\sum_p\sum_q\left(\frac{1}{t}\sum_i\mathbf{R}_{p,q}^{(i,\ell)}\right)^2 + 16\|\mathbf{R}^{(i,\ell)}\|_\mathsf{F}^2\frac{d\log(rd/\delta_0)}{mt}. \qquad (96)$$

□

Lemma 9. *Let*

$$\begin{aligned}
\mathbf{A}_{rd\times rd} &= \frac{1}{mt}\sum_{i\in[t]}\left(\mathbf{w}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}\otimes(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\right) \\
&= \frac{1}{mt}\sum_{i\in[t]}\left(\mathbf{w}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}\otimes\left(\sum_{j=1}^m\mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}\right)\right)
\end{aligned}$$

s.t. $\mathbf{A}^{-1} = \mathbb{E}\left[\mathbf{A}\right] + \mathbf{E}$ *where* $\mathbf{E}$ *is the error matrix due to perturbation. Then for vectors* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{rd}$*, we have*

$$\left|\mathbf{a}^{\mathsf{T}}\mathbf{E}\mathbf{b}\right| \leq c\|\mathbf{a}\|_2\|\mathbf{b}\|_2\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4\frac{\log(1/\delta_0)}{m^2t^2}}.$$

Proof. We can rewrite the vectors $\mathbf{a}$ and $\mathbf{b}$ s.t. $\mathbf{a}^{\mathsf{T}} = \left[\mathbf{a}_1^{\mathsf{T}}, \mathbf{a}_2^{\mathsf{T}}, \dots, \mathbf{a}_r^{\mathsf{T}}\right]$ and $\mathbf{b}^{\mathsf{T}} = \left[\mathbf{b}_1^{\mathsf{T}}, \mathbf{b}_2^{\mathsf{T}}, \dots, \mathbf{b}_r^{\mathsf{T}}\right]$ respectively where each $\mathbf{a}_i, \mathbf{b}_i \in \mathbb{R}^d$. Note that

$$\mathbb{E}\left[\mathbf{A}\right] = \frac{1}{t}\sum_{i\in[t]}\Big(\mathbf{w}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^{\mathsf{T}} \otimes \mathbf{I}\Big), \tag{97}$$

$$\begin{aligned}
\text{and } \mathbf{a}^{\mathsf{T}}\mathbf{A}\mathbf{b} &= \mathbf{a}^{\mathsf{T}}\Big(\frac{1}{mt}\sum_{i\in[t]}\Big(\mathbf{w}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^{\mathsf{T}} \otimes \Big(\sum_{j=1}^{m}\mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^{\mathsf{T}}\Big)\Big)\Big)\mathbf{b} \\
&= \frac{1}{mt}\sum_{i\in[t]}\sum_{j\in[m]}\sum_{p\in[r]}\sum_{q\in[r]}\mathbf{a}_p^{\mathsf{T}}\Big(w_p^{(i,\ell)}w_q^{(i,\ell)}\mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^{\mathsf{T}}\Big)\mathbf{b}_q \\
&= \frac{1}{mt}\sum_{i\in[t]}\sum_{j\in[m]}(\mathbf{x}_j^{(i)})^{\mathsf{T}}\Big(\sum_{p\in[r]}\sum_{q\in[r]}w_p^{(i,\ell)}w_q^{(i,\ell)}\mathbf{a}_p(\mathbf{b}_q)^{\mathsf{T}}\Big)\mathbf{x}_j^{(i)}. \qquad (98)
\end{aligned}$$

Furthermore,

$$\begin{aligned}
\mathsf{Tr}\Big(\sum_{p\in[r]}\sum_{q\in[r]}w_p^{(i,\ell)}w_q^{(i,\ell)}\mathbf{a}_p(\mathbf{b}_q)^{\mathsf{T}}\Big) &= \mathsf{Tr}\Big(\mathbf{a}^{\mathsf{T}}\Big(\mathbf{w}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^{\mathsf{T}} \otimes \mathbf{I}\Big)\mathbf{b}\Big) \\
&= \mathbf{a}^{\mathsf{T}}\Big(\mathbf{w}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^{\mathsf{T}} \otimes \mathbf{I}\Big)\mathbf{b} \qquad (99)
\end{aligned}$$

and

$$\begin{aligned}
&\|\sum_{p\in[r]}\sum_{q\in[r]}w_p^{(i,\ell)}w_q^{(i,\ell)}\mathbf{a}_p(\mathbf{b}_q)^{\mathsf{T}}\|_{\mathsf{F}} \\
&= \|\Big(\sum_{p\in[r]}w_p^{(i,\ell)}\mathbf{a}_p\Big)\Big(\sum_{q\in[r]}w_q^{(i,\ell)}\mathbf{b}_q\Big)^{\mathsf{T}}\|_{\mathsf{F}} \\
&= \|\sum_{p\in[r]}w_p^{(i,\ell)}\mathbf{a}_p\|_2\|\sum_{q\in[r]}w_q^{(i,\ell)}\mathbf{b}_q\|_2 \\
&\leq \Big(\sum_{p\in[r]}|w_p^{(i,\ell)}|\|\mathbf{v}_p\|_2\Big)\Big(\sum_{q\in[r]}|w_q^{(i,\ell)}|\|\mathbf{b}_q\|_2\Big) \\
&\leq \sqrt{\Big(\sum_{p\in[r]}(w_p^{(i,\ell)})^2\Big)\Big(\sum_{p\in[r]}\|\mathbf{a}_p\|_2^2\Big)}\sqrt{\Big(\sum_{q\in[r]}(w_q^{(i,\ell)})^2\Big)\Big(\sum_{q\in[r]}\|\mathbf{b}_q\|_2^2\Big)} \\
&= \|\mathbf{w}^{(i,\ell)}\|_2^2\|\mathbf{a}\|_2\|\mathbf{b}\|_2. \qquad (100)
\end{aligned}$$

Therefore, using equation 98, equation 99 and equation 100 in Lemma 16 we get

$$\begin{aligned}
\Big|\mathbf{a}^{\mathsf{T}}\Big(\mathbf{A} - \frac{1}{mt}\sum_{i\in[t]}\sum_{j\in[m]}\Big(\mathbf{w}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^{\mathsf{T}} \otimes \mathbf{I}\Big)\Big)\mathbf{b}\Big| &\leq c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4\|\mathbf{a}\|_2^2\|\mathbf{b}\|_2^2\frac{\log(1/\delta_0)}{m^2t^2}} \\
\iff \Big|\mathbf{a}^{\mathsf{T}}\mathbf{E}\mathbf{b}\Big| &\leq c\|\mathbf{a}\|_2\|\mathbf{b}\|_2\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4\frac{\log(1/\delta_0)}{m^2t^2}}.
\end{aligned} \tag{101}$$

□

Lemma 10. *For some constant* $c > 0$ *and for any iteration indexed by* $\ell > 0$, *we have*

$$
\begin{aligned}
&\|\mathbf{U}^{(\ell)} - \mathbf{U}^{\star}\mathbf{Q}^{(\ell-1)}\|_{\mathsf{F}} \\
&\le \frac{1}{\frac{1}{r}\lambda_r\left(\frac{r}{t}(\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\right) - c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(1/\delta_0)}{m^2t^2}} - \frac{\sigma_1}{mt}\left(2\sqrt{rd}+4\sqrt{\log rd}\right)} \cdot \\
&\quad \Big\{\frac{2}{t}\|\mathbf{U}^{\star}\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\|_{\mathsf{F}} + \sqrt{\frac{4\zeta}{t}}(\max_i\|\mathbf{w}^{(i,\ell)}\|_2)\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2 \\
&\quad + 4\Big(\|\mathbf{U}^{\star}\mathbf{Q}^{(\ell-1)}\|\|\mathbf{h}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2 + \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2\Big)\sqrt{\frac{d\log(rd/\delta_0)}{mt}} \\
&\quad + \frac{\sigma_2}{mt}6\sqrt{rd\log(rd)} + \frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)\sqrt{r} + \frac{2\sigma\sqrt{d\mu^{\star}\lambda_r^{\star}}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big\},
\end{aligned}
$$

$$
\begin{aligned}
&\|\mathbf{U}^{(\ell)}\|_{\mathsf{F}} \\
&\le \frac{1}{\frac{1}{r}\lambda_r\left(\frac{r}{t}(\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\right) - c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(1/\delta_0)}{m^2t^2}} - \frac{\sigma_1}{mt}\left(2\sqrt{rd}+4\sqrt{\log rd}\right)} \cdot \\
&\quad \Big\{\frac{2}{t}\|\mathbf{U}^{\star}(\mathbf{W}^{\star})^{\mathsf{T}}\mathbf{W}^{(\ell)}\|_{\mathsf{F}} + \sqrt{\frac{4\zeta}{t}}(\max_i\|\mathbf{w}^{(i,\ell)}\|_2)\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2 \\
&\quad + 4\Big(\|\mathbf{U}^{\star}\|\|\mathbf{w}^{\star(i)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2 + \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2\Big)\sqrt{\frac{d\log(rd/\delta_0)}{mt}} \\
&\quad + \frac{\sigma_2}{mt}6\sqrt{rd\log(rd)} + \frac{2\sigma\sqrt{d\mu^{\star}\lambda_r^{\star}}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big\},
\end{aligned}
$$

$$
\begin{aligned}
&\Big\|\Delta(\mathbf{U}^{+(\ell)}, \mathbf{U}^{\star})\Big\|_{\mathsf{F}} \\
&\le \Big\{\frac{c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(1/\delta_0)}{m^2t^2}} + \frac{\sigma_1}{mt}\left(2\sqrt{rd}+4\sqrt{\log rd}\right)}{\frac{1}{r}\lambda_r\left(\frac{r}{t}(\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\right) - c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(1/\delta_0)}{m^2t^2}} - \frac{\sigma_1}{mt}\left(2\sqrt{rd}+4\sqrt{\log rd}\right)} \cdot \\
&\quad \Big(\frac{2}{t}\|\mathbf{U}^{\star}\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\|_{\mathsf{F}}\Big) \\
&+ \frac{1}{\frac{1}{r}\lambda_r\left(\frac{r}{t}(\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\right) - c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(1/\delta_0)}{m^2t^2}} - \frac{\sigma_1}{mt}\left(2\sqrt{rd}+4\sqrt{\log rd}\right)} \cdot \\
&\quad \Big\{\Big(\sqrt{\frac{4\zeta}{t}}(\max_i\|\mathbf{w}^{(i,\ell)}\|_2)\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\Big) \\
&+ 4\Big(\|\mathbf{U}^{\star}\mathbf{Q}^{(\ell-1)}\|\|\mathbf{h}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2 + \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2\Big)\sqrt{\frac{d\log(rd/\delta_0)}{mt}} \\
&\quad \frac{\sigma_2}{mt}6\sqrt{rd\log(rd)} + \frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)\sqrt{r} + \frac{2\sigma\sqrt{d\mu^{\star}\lambda_r^{\star}}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big\}\Big\}\|\mathbf{R}^{-1}\|
\end{aligned}
$$

with probability at least $1 - \mathcal{O}(\delta_0)$.

Proof. **Analysis of $\|\mathbf{U}^{(\ell)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_\mathsf{F}$:**
Update step for $\mathbf{U}$ of the Algorithm without DP Noise for the ℓ^th iteration gives us

$$\begin{aligned}
&\sum_{i\in[t]}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\mathbf{U}^{(\ell)}\mathbf{w}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}\\
&= \sum_{i\in[t]}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big(\mathbf{U}^\star\mathbf{w}^{\star(i)} + (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})\Big)(\mathbf{w}^{(i,\ell)})^\mathsf{T} + \sum_{i\in[t]}(\mathbf{X}^{(i)})^\mathsf{T}\xi^{(i)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}\\
\implies &\sum_{i\in[t]}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}(\mathbf{U}^{(\ell)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\mathbf{w}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}\\
&= \sum_{i\in[t]}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big(\mathbf{U}^\star(\mathbf{w}^{\star(i)} - \mathbf{Q}^{(\ell-1)}\mathbf{w}^{(i,\ell)}) + (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})\Big)(\mathbf{w}^{(i,\ell)})^\mathsf{T}\\
&\quad + \sum_{i\in[t]}(\mathbf{X}^{(i)})^\mathsf{T}\xi^{(i)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}.
\end{aligned}$$

Using Lemma 22, the above can be written as:

$$\begin{aligned}
\mathbf{A}\mathsf{vec}(\mathbf{U}^{(\ell)}) &= \mathsf{vec}(\mathbf{V}' + \Xi),\\
\text{and } \mathbf{A}\mathsf{vec}(\mathbf{U}^{(\ell)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}) &= \mathsf{vec}(\mathbf{V} + \Xi),
\end{aligned}\tag{102}$$

where

$$\begin{aligned}
\mathbf{A}_{rd\times rd} &= \frac{1}{mt}\sum_{i\in[t]}\Big(\mathbf{w}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T} \otimes (\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big)\\
&= \frac{1}{mt}\sum_{i\in[t]}\Big(\mathbf{w}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T} \otimes \Big(\sum_{j=1}^{m}\mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}\Big)\Big),\\
\mathbf{V}_{d\times r} &= \frac{1}{mt}\sum_{i\in[t]}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big(\mathbf{U}^\star(\mathbf{w}^{\star(i)} - \mathbf{Q}^{(\ell-1)}\mathbf{w}^{(i,\ell)}) + (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})\Big)(\mathbf{w}^{(i,\ell)})^\mathsf{T}\\
&= \frac{1}{mt}\sum_{i\in[t]}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big(- \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\mathbf{h}^{(i,\ell)} + (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})\Big)(\mathbf{w}^{(i,\ell)})^\mathsf{T},\\
\mathbf{V}'_{d\times r} &= \frac{1}{mt}\sum_{i\in[t]}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big(\mathbf{U}^\star\mathbf{w}^{\star(i)} + (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})\Big)(\mathbf{w}^{(i,\ell)})^\mathsf{T},
\end{aligned}$$

$\mathbf{h}^{(i,\ell)} = \mathbf{w}^{(i,\ell)} - (\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}$, $\xi_j^{(i)} \sim \mathcal{N}(0,\sigma^2)$ and $\Xi_{d\times r} = \frac{1}{mt}\sum_{i\in[t]}(\mathbf{X}^{(i)})^\mathsf{T}\xi^{(i)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}$.
Now introducing DP noise we get:

$$\begin{aligned}
&\mathsf{vec}(\mathbf{U}^{(\ell)}) = \Big(\mathbf{A} + \frac{\mathbf{N}_1}{mt}\Big)^{-1}\Big(\mathsf{vec}\Big(\mathbf{V}' + \frac{\mathbf{N}_2}{mt} + \Xi\Big)\Big)\\
&\implies \mathsf{vec}(\mathbf{U}^{(\ell)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\\
&= \Big(\mathbf{A} + \frac{\mathbf{N}_1}{mt}\Big)^{-1}\Big(\mathsf{vec}\Big(\mathbf{V}' + \Xi + \frac{\mathbf{N}_2}{mt}\Big) - \Big(\mathbf{A} + \frac{\mathbf{N}_1}{mt}\Big)\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\Big)\\
&= \Big(\mathbf{A} + \frac{\mathbf{N}_1}{mt}\Big)^{-1}\cdot\\
&\qquad\Big(\Big(\mathsf{vec}(\mathbf{V}') - \mathbf{A}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\Big) + \mathsf{vec}\Big(\frac{\mathbf{N}_2}{mt}\Big) - \frac{\mathbf{N}_1}{mt}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}) + \mathsf{vec}(\Xi)\Big)\\
&= \Big(\mathbf{A} + \frac{\mathbf{N}_1}{mt}\Big)^{-1}\Big(\mathsf{vec}(\mathbf{V}) + \mathsf{vec}\Big(\frac{\mathbf{N}_2}{mt}\Big) - \frac{\mathbf{N}_1}{mt}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\Big) + \mathsf{vec}(\Xi)\Big),
\end{aligned}\tag{103}$$

where

$$\begin{aligned}
\mathbf{N}_1 &\sim \sigma_1\mathcal{MN}_{rd\times rd}(\mathbf{0}, \mathbf{I}_{rd\times rd}, \mathbf{I}_{rd\times rd}),\\
\mathbf{N}_2 &\sim \sigma_2\mathcal{MN}_{d\times r}(\mathbf{0}, \mathbf{I}_{d\times d}, \mathbf{I}_{r\times r})
\end{aligned}$$

where $\mathcal{MN}$ denotes the Matrix Normal Distribution. Note that equation 103 gives:

$$\begin{aligned}
&\|\mathsf{vec}(\mathbf{U}^{(\ell)} - \mathbf{U}^{\star}\mathbf{Q}^{(\ell-1)})\|_2 \\
&\quad \leq \left\|\Big(\mathbf{A} + \frac{\mathbf{N}_1}{mt}\Big)^{-1}\right\| \left\|\mathsf{vec}(\mathbf{V}) + \mathsf{vec}\Big(\frac{\mathbf{N}_2}{mt}\Big) - \frac{\mathbf{N}_1}{mt}\mathsf{vec}(\mathbf{U}^{\star}\mathbf{Q}^{(\ell-1)}) + \mathsf{vec}(\boldsymbol{\Xi})\right\|_2 \\
\iff &\|\mathbf{U}^{(\ell)} - \mathbf{U}^{\star}\mathbf{Q}^{(\ell-1)}\|_{\mathsf{F}} \\
&\quad \leq \left\|\Big(\mathbf{A} + \frac{\mathbf{N}_1}{mt}\Big)^{-1}\right\| \Big(\|\mathbf{V}\|_{\mathsf{F}} + \left\|\frac{\mathbf{N}_2}{mt}\right\|_{\mathsf{F}} + \left\|\frac{\mathbf{N}_1}{mt}\right\|_2 \|\mathbf{U}^{\star}\mathbf{Q}^{(\ell-1)}\|_{\mathsf{F}} + \|\boldsymbol{\Xi}\|_{\mathsf{F}}\Big). \qquad (104)
\end{aligned}$$

First, with high probability, we will bound the following quantity

$$\|\boldsymbol{\Xi}\|_{2,\infty} = \left\|\frac{1}{mt}\sum_{i\in[t]}(\mathbf{X}^{(i)})^{\mathsf{T}}\xi^{(i)}(\mathbf{w}^{(i,\ell)})^{\mathsf{T}}\right\|_{2,\infty}$$

Condition on the vector $\xi^{(i)}$. Notice that the random vector $(\mathbf{X}^{(i)})^{\mathsf{T}}\xi^{(i)}$ is a d dimensional vectors where each entry is independently generated according to $\mathcal{N}(0, \left|\left|\xi^{(i)}\right|\right|_2^2)$. Therefore for a fixed row indexed by s, we have ℓ_2 norm of the s^{th} row of $\frac{1}{mt}\sum_{i\in[t]}(\mathbf{X}^{(i)})^{\mathsf{T}}\xi^{(i)}(\mathbf{w}^{(i,\ell)})^{\mathsf{T}}$ is going to be $\frac{1}{mt}\left|\left|\mathbf{W}^{(\ell)}\right|\right|_{\mathsf{F}}\left|\left|\xi^{(i)}\right|\right|_2\sqrt{\log(dr\delta^{-1})}$. Hence, we have that with probability $1-\delta_0$ (provided that $m = \widetilde{\Omega}(\sigma^2)$),

$$\left\|\frac{1}{mt}\sum_{i\in[t]}(\mathbf{X}^{(i)})^{\mathsf{T}}\xi^{(i)}(\mathbf{w}^{(i,\ell)})^{\mathsf{T}}\right\|_{2,\infty} \leq \frac{2\sigma\sqrt{\mu^{\star}\lambda_r^{\star}}\log(2rdmt/\delta_0)}{\sqrt{mt}} \qquad (105)$$

and

$$\left\|\frac{1}{mt}\sum_{i\in[t]}(\mathbf{X}^{(i)})^{\mathsf{T}}\xi^{(i)}(\mathbf{w}^{(i,\ell)})^{\mathsf{T}}\right\|_{\mathsf{F}} \leq \frac{2\sigma\sqrt{d\mu^{\star}\lambda_r^{\star}}\log(2rdmt/\delta_0)}{\sqrt{mt}}. \qquad (106)$$

Now, We will analyse the two multiplicands separately.

$$\begin{aligned}
\mathbb{E}\left[\mathbf{V}\right] &= \frac{1}{t}\sum_{i\in[t]}\Big(-\mathbf{U}^{\star}\mathbf{Q}^{(\ell-1)}\mathbf{h}^{(i,\ell)} + (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})\Big)(\mathbf{w}^{(i,\ell)})^{\mathsf{T}} \\
&= \frac{1}{t}\sum_{i\in[t]}\Big(-\mathbf{U}^{\star}\mathbf{Q}^{(\ell-1)}\mathbf{h}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^{\mathsf{T}} + (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^{\mathsf{T}}\Big) \\
&= \frac{-1}{t}\mathbf{U}^{\star}\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)} + \frac{1}{t}\sum_{i\in[t]}(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^{\mathsf{T}}.
\end{aligned}$$

Using Lemma 8 with $\mathbf{R}^{(i,\ell)} = \Big(- \mathbf{U}^\star \mathbf{Q}^{(\ell-1)}\mathbf{h}^{(i,\ell)} + (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})\Big)(\mathbf{w}^{(i,\ell)})^\mathsf{T}$, we have with probability at least $1-\delta_0$

$$
\begin{aligned}
&\sum_p \sum_q \mathbf{V}_{p,q}^2 \\
&\leq 2 \sum_p \sum_q \Big(\frac{1}{t}\sum_i \mathbf{R}_{p,q}^{(i,\ell)}\Big)^2 + 16 \|\mathbf{R}^{(i,\ell)}\|_\mathsf{F}^2 \frac{d\log(rd/\delta_0)}{mt} \\
&\leq 2 \sum_p \sum_q \Big(\frac{1}{t}\sum_i \Big(- \mathbf{U}^\star \mathbf{Q}^{(\ell-1)}\mathbf{h}^{(i,\ell)} + (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})\Big)(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big)_{p,q}^2 \\
&\qquad + 16 \|\mathbf{R}^{(i,\ell)}\|_\mathsf{F}^2 \frac{d\log(rd/\delta_0)}{mt} \\
&\leq 2 \sum_p \sum_q \frac{1}{t^2}\Big(- \mathbf{U}^\star \mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)} + \sum_i (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big)_{p,q}^2 \\
&\qquad + 16 \|\mathbf{R}^{(i,\ell)}\|_\mathsf{F}^2 \frac{d\log(rd/\delta_0)}{mt} \\
&\leq \frac{4}{t^2}\sum_p \sum_q \Big(\mathbf{U}^\star \mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)_{p,q}^2 + \frac{4}{t^2}\sum_p \sum_q \Big(\sum_i (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big)_{p,q}^2 \\
&\qquad + 16 \|\mathbf{R}^{(i,\ell)}\|_\mathsf{F}^2 \frac{d\log(rd/\delta_0)}{mt} \\
&\leq \frac{4}{t^2}\|\mathbf{U}^\star \mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\|_\mathsf{F}^2 + \frac{4}{t}\sum_p\sum_q\sum_i (\mathbf{b}_p^{\star(i)} - \mathbf{b}_p^{(i,\ell)})^2 (\mathbf{w}_s^{(i,\ell)})^2 \\
&\qquad + 16 \|\mathbf{R}^{(i,\ell)}\|_\mathsf{F}^2 \frac{d\log(rd/\delta_0)}{mt} \\
&\leq \frac{4}{t^2}\|\mathbf{U}^\star \mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\|_\mathsf{F}^2 + \frac{4}{t}\zeta(\max_i \|\mathbf{w}^{(i,\ell)}\|_2^2)\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2^2 + 16 \|\mathbf{R}^{(i,\ell)}\|_\mathsf{F}^2 \frac{d\log(rd/\delta_0)}{mt}.
\end{aligned}
$$

$$
\begin{aligned}
\implies \|\mathbf{V}\|_\mathsf{F} \leq \frac{2}{t}\|\mathbf{U}^\star \mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\|_\mathsf{F} + \sqrt{\frac{4\zeta}{t}}(\max_i \|\mathbf{w}^{(i,\ell)}\|_2)\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2 \\
+ 4\|\mathbf{R}^{(i,\ell)}\|_\mathsf{F}\sqrt{\frac{d\log(rd/\delta_0)}{mt}}. \qquad (107)
\end{aligned}
$$

Since $\mathbf{R}^{(i,\ell)} = -\mathbf{U}^\star \mathbf{Q}^{(\ell-1)}\mathbf{h}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T} + (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}$, we have

$$
\begin{aligned}
\|\mathbf{R}^{(i,\ell)}\|_\mathsf{F} &= \| -\mathbf{U}^\star \mathbf{Q}^{(\ell-1)}\mathbf{h}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T} + (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\|_\mathsf{F} \\
&\leq \|\mathbf{U}^\star \mathbf{Q}^{(\ell-1)}\mathbf{h}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}\|_\mathsf{F} + \|(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\|_\mathsf{F} \\
&\leq \|\mathbf{U}^\star \mathbf{Q}^{(\ell-1)}\|\|\mathbf{h}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2 + \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2 \qquad (108)
\end{aligned}
$$

Using equation 108 in equation 107 gives

$$
\begin{aligned}
\|\mathbf{V}\|_\mathsf{F} \leq \frac{2}{t}\|\mathbf{U}^\star \mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\|_\mathsf{F} + \sqrt{\frac{4\zeta}{t}}(\max_i \|\mathbf{w}^{(i,\ell)}\|_2)\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2 \\
+ 4\Big(\|\mathbf{U}^\star \mathbf{Q}^{(\ell-1)}\|\|\mathbf{h}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2 + \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2\Big)\sqrt{\frac{d\log(rd/\delta_0)}{mt}} \\
(109)
\end{aligned}
$$

Now, let $\mathcal{V} \triangleq \{\mathbf{v} \in \mathbb{R}^{rd} | \|\mathbf{v}\|_2 = 1\}$. Then for $\epsilon \leq 1$, there exists an ϵ-net, $N_\epsilon \subset \mathcal{V}$, of size $(1+2/\epsilon)^{rd}$ w.r.t the Euclidean norm, i.e. $\forall\, \mathbf{v} \in \mathcal{V}$, $\exists\, \mathbf{v}' \in N_\epsilon$ s.t. $\|\mathbf{v} - \mathbf{v}'\|_2 \leq \epsilon$. Now consider any $\mathbf{v}^\mathsf{T} = \big[\mathbf{v}_1^\mathsf{T}, \mathbf{v}_2^\mathsf{T}, \ldots, \mathbf{v}_r^\mathsf{T}\big] \in N_\epsilon$ where each $\mathbf{v}_i \in \mathbb{R}^d$. Then using Lemma 9 with $\mathbf{a} = \mathbf{b} = \mathbf{v}$, we get:

$$
\Big|\mathbf{v}^\mathsf{T}\Big(\mathbf{A} - \frac{1}{mt}\sum_{i\in[t]}\sum_{j\in[m]}\Big(\mathbf{w}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T} \otimes \mathbf{I}\Big)\Big)\mathbf{v}\Big| \leq c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4 \frac{\log(1/\delta_0)}{m^2t^2}}
$$

$$\begin{aligned}\implies \|\mathbf{v}^\mathsf{T}\mathbf{E}\mathbf{v}\| &\leq c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{\log(|N_\epsilon|/\delta_0)}{m^2t^2}}\\ &\leq c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{\log((1+2/\epsilon)^{rd}/\delta_0)}{m^2t^2}}, \quad \forall \mathbf{v}\in N_\epsilon \end{aligned} \tag{110}$$

w.p. $1-\delta_)$ where $\mathbf{E} \triangleq \mathbf{A} - \frac{1}{mt}\sum_{i\in[t]}\sum_{j\in[m]}\Big(\mathbf{w}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}\otimes\mathbf{I}\Big)$. Since $\mathbf{E}$ is symmetric, therefore $\|\mathbf{E}\| = (\mathbf{v}')^\mathsf{T}\mathbf{E}\mathbf{v}'$ where $\mathbf{v}'\in\mathsf{V}$ is the largest eigenvector of $\mathbf{E}$. Further, $\exists\ \mathbf{v}\in N_\epsilon$ s.t. $\|\mathbf{v}'-\mathbf{v}\|\leq\epsilon$. This implies:

$$\begin{aligned}\|\mathbf{E}\| = (\mathbf{v}')^\mathsf{T}\mathbf{E}\mathbf{v}' &= (\mathbf{v}'-\mathbf{v})^\mathsf{T}\mathbf{E}\mathbf{v} + (\mathbf{v}')^\mathsf{T}\mathbf{E}(\mathbf{v}'-\mathbf{v}) + \mathbf{v}^\mathsf{T}\mathbf{E}\mathbf{v}\\ &\leq \|\mathbf{v}'-\mathbf{v}\|\|\mathbf{E}\|\|\mathbf{v}\| + \|\mathbf{v}'\|\|\mathbf{E}\|\|\mathbf{v}'-\mathbf{v}\| + \mathbf{v}^\mathsf{T}\mathbf{E}\mathbf{v}\\ &\leq 2\epsilon\|\mathbf{E}\| + \mathbf{v}^\mathsf{T}\mathbf{E}\mathbf{v}\\ \implies \|\mathbf{E}\| &\leq \frac{\mathbf{v}^\mathsf{T}\mathbf{E}\mathbf{v}}{1-2\epsilon}. \end{aligned} \tag{111}$$

Using equation 110 and equation 111 and setting $\epsilon\leftarrow 1/4$ and $c\leftarrow 2c\sqrt{\log(9)}$ then gives:

$$\|\mathbf{E}\| \leq c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(1/\delta_0)}{m^2t^2}} \tag{112}$$

Using equation 112 then gives

$$\begin{aligned}\|\mathbf{A}\| &\geq \min_{\mathbf{v}}\Big(\frac{1}{mt}\sum_{i\in[t]}\sum_{j\in[m]}\mathbf{v}^\mathsf{T}\Big(\mathbf{w}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}\otimes\mathbf{I}\Big)\mathbf{v}\Big) - c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(1/\delta_0)}{m^2t^2}}\\ &\geq \lambda_r\Big(\frac{1}{mt}\sum_{i\in[t]}\sum_{j\in[m]}\mathbf{w}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}\otimes\mathbf{I}\Big) - c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(1/\delta_0)}{m^2t^2}}\\ &\geq \frac{1}{r}\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big) - c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(1/\delta_0)}{m^2t^2}}\end{aligned}$$

Next, note that the matrix N/mt can be written as $\alpha\mathcal{N}(0,\mathbf{I}_{rd\times rd})$ where $\alpha=\frac{\sigma_1}{mt}$. Therefore, with probability at least $1-(rd)^{-8}$, the minimum eigenvalue of the matrix is at least $-\frac{4\sigma_1\sqrt{\log rd}}{mt}$. Further we have using standard gaussian concentration inequalities,

$$\Big|\Big|\frac{\mathbf{N}_1}{mt}\Big|\Big| \leq \frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big), \tag{113}$$

$$\Big|\Big|\frac{\mathbf{N}_2}{mt}\Big|\Big| \leq \frac{\sigma_2}{mt}\Big(\sqrt{d}+\sqrt{r}+4\sqrt{\log rd}\Big), \tag{114}$$

$$\begin{aligned}\Big|\Big|\frac{\mathbf{N}_2}{mt}\Big|\Big|_\mathsf{F} &= \frac{\sigma_2}{mt}\|\mathbf{N}_2\|_\mathsf{F}\\ &\leq \frac{\sigma_2}{mt}\sqrt{\sum_{i\in[d]}\sum_{j\in[r]}2\log((rd)^{2\cdot 9})}\\ &= \frac{\sigma_2}{mt}6\sqrt{rd\log(rd)}.\end{aligned} \tag{115}$$

Hence, the minimum eigenvalue of the matrix $\mathbf{A}+\frac{\mathbf{N}}{mt}$ is bounded from below by

$$\lambda_{\min}\Big(\mathbf{A}+\frac{\mathbf{N}}{mt}\Big) \geq \frac{1}{r}\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big) \tag{116}$$

$$- c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(1/\delta_0)}{m^2t^2}} - \frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big). \tag{117}$$

Therefore, the maximum eigenvalue of $(\mathbf{A} + \frac{\mathbf{N}_1}{mt})^{-1}$ is bounded from above by, $\left|\left|\left(\mathbf{A} + \frac{\mathbf{N}_1}{mt}\right)^{-1}\right|\right|$

$$\leq \frac{1}{\frac{1}{r}\lambda_r\left(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\right) - c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(1/\delta_0)}{m^2t^2}} - \frac{\sigma_1}{mt}\left(2\sqrt{rd} + 4\sqrt{\log rd}\right)}. \tag{118}$$

Using equation 109, equation 118, equation 113, equation 114, equation 115 and equation 106 in equation 104 gives $\|\mathbf{U}^{(\ell)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_\mathsf{F}$

$$\begin{aligned}
&\leq \frac{1}{\frac{1}{r}\lambda_r\left(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\right) - c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(1/\delta_0)}{m^2t^2}} - \frac{\sigma_1}{mt}\left(2\sqrt{rd} + 4\sqrt{\log rd}\right)}. \\
&\quad \Big\{\frac{2}{t}\|\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\|_\mathsf{F} + \sqrt{\frac{4\zeta}{t}}(\max_i \|\mathbf{w}^{(i,\ell)}\|_2)\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2 \\
&\quad + 4\Big(\|\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|\|\mathbf{h}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2 + \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2\Big)\sqrt{\frac{d\log(rd/\delta_0)}{mt}} \\
&\quad + \frac{\sigma_2}{mt}6\sqrt{rd\log(rd)} + \frac{\sigma_1}{mt}\Big(2\sqrt{rd} + 4\sqrt{\log rd}\Big)\sqrt{r} + \frac{2\sigma\sqrt{d\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big\}.
\end{aligned} \tag{119}$$

Analysis of $\left|\left|\mathbf{U}^{(\ell)}\right|\right|_\mathsf{F}$:

The analysis will follow along similar lines as in the previous section except that we will now have:

$$\mathsf{vec}(\mathbf{U}^{(\ell)}) = \Big(\mathbf{A} + \frac{\mathbf{N}_1}{mt}\Big)^{-1}\mathsf{vec}\Big(\mathbf{V}' + \frac{\mathbf{N}_2}{mt} + \mathbf{\Xi}\Big) \tag{120}$$

where

$$\begin{aligned}
\mathbf{A}_{rd\times rd} &= \frac{1}{mt}\sum_{i\in[t]}\Big(\mathbf{w}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T} \otimes (\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big) \\
&= \frac{1}{mt}\sum_{i\in[t]}\Big(\mathbf{w}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T} \otimes \Big(\sum_{j=1}^{m}\mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}\Big)\Big), \\
\mathbf{V}'_{d\times r} &= \frac{1}{mt}\sum_{i\in[t]}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big(\mathbf{U}^\star\mathbf{w}^{\star(i)} + (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})\Big)(\mathbf{w}^{(i,\ell)})^\mathsf{T}.
\end{aligned}$$

i.e. we have the term $-\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\mathbf{h}^{(i,\ell)}$ replaced by $\mathbf{U}^\star\mathbf{w}^{\star(i)}$. The above gives:

$$\begin{aligned}
&\|\mathsf{vec}(\mathbf{U}^{(\ell)}\|_2 \leq \left|\left|\Big(\mathbf{A} + \frac{\mathbf{N}_1}{mt}\Big)^{-1}\right|\right|\left|\left|\mathsf{vec}\Big(\mathbf{V}' + \frac{\mathbf{N}_2}{mt} + \mathbf{\Xi}\Big)\right|\right|_2 \\
\iff &\|\mathbf{U}^{(\ell)}\|_\mathsf{F} \leq \left|\left|\Big(\mathbf{A} + \frac{\mathbf{N}_1}{mt}\Big)^{-1}\right|\right|\Big(\|\mathbf{V}'\|_\mathsf{F} + \left|\left|\frac{\mathbf{N}_2}{mt}\right|\right|_\mathsf{F} + \|\mathbf{\Xi}\|_\mathsf{F}\Big).
\end{aligned} \tag{121}$$

We can compute the above following similar lines as before.

$$\begin{aligned}
\mathbb{E}\left[\mathbf{V}'\right] &= \frac{1}{t}\sum_{i\in[t]}\Big(\mathbf{U}^\star\mathbf{w}^{\star(i)} + (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})\Big)(\mathbf{w}^{(i,\ell)})^\mathsf{T} \\
&= \frac{1}{t}\sum_{i\in[t]}\Big(\mathbf{U}^\star\mathbf{w}^{\star(i)}(\mathbf{w}^{(i,\ell)})^\mathsf{T} + (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big) \\
&= \frac{1}{t}\mathbf{U}^\star(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^{(\ell)} + \frac{1}{t}\sum_{i\in[t]}(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}.
\end{aligned}$$

Using Lemma 8 with $\mathbf{R}^{(i,\ell)} = \Big(\mathbf{U}^\star \mathbf{w}^{\star(i)} + (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})\Big)(\mathbf{w}^{(i,\ell)})^\mathsf{T}$, we have with probability at least $1 - \delta_0$

$$
\begin{aligned}
&\sum_p \sum_q (\mathbf{V}'_{p,q})^2 \\
&\leq 2 \sum_p \sum_q \Big(\frac{1}{t} \sum_i \mathbf{R}^{(i,\ell)}_{p,q}\Big)^2 + 16 \|\mathbf{R}^{(i,\ell)}\|_\mathsf{F}^2 \frac{d \log(rd/\delta_0)}{mt} \\
&\leq 2 \sum_p \sum_q \Big(\frac{1}{t} \sum_i \Big(\mathbf{U}^\star \mathbf{w}^{\star(i)} + (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})\Big)(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big)^2_{p,q} + 16 \|\mathbf{R}^{(i,\ell)}\|_\mathsf{F}^2 \frac{d \log(rd/\delta_0)}{mt} \\
&\leq 2 \sum_p \sum_q \frac{1}{t^2} \Big(\mathbf{U}^\star (\mathbf{W}^\star)^\mathsf{T} \mathbf{W}^{(\ell)} + \sum_i (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big)^2_{p,q} + +16 \|\mathbf{R}^{(i,\ell)}\|_\mathsf{F}^2 \frac{d \log(rd/\delta_0)}{mt} \\
&\leq \frac{4}{t^2} \sum_p \sum_q \Big(\mathbf{U}^\star (\mathbf{W}^\star)^\mathsf{T} \mathbf{W}^{(\ell)}\Big)^2_{p,q} + \frac{4}{t^2} \sum_p \sum_q \Big(\sum_i (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big)^2_{p,q} \\
&\qquad + +16 \|\mathbf{R}^{(i,\ell)}\|_\mathsf{F}^2 \frac{d \log(rd/\delta_0)}{mt} \\
&\leq \frac{4}{t^2} \|\mathbf{U}^\star (\mathbf{W}^\star)^\mathsf{T} \mathbf{W}^{(\ell)}\|_\mathsf{F}^2 + \frac{4}{t} \sum_p \sum_q \sum_i (\mathbf{b}^{\star(i)}_p - \mathbf{b}^{(i,\ell)}_p)^2 (\mathbf{w}^{(i,\ell)}_s)^2 + 16 \|\mathbf{R}^{(i,\ell)}\|_\mathsf{F}^2 \frac{d \log(rd/\delta_0)}{mt} \\
&\leq \frac{4}{t^2} \|\mathbf{U}^\star (\mathbf{W}^\star)^\mathsf{T} \mathbf{W}^{(\ell)}\|_\mathsf{F}^2 + \frac{4}{t} \zeta (\max_i \|\mathbf{w}^{(i,\ell)}\|_2^2) \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2^2 + 16 \|\mathbf{R}^{(i,\ell)}\|_\mathsf{F}^2 \frac{d \log(rd/\delta_0)}{mt}.
\end{aligned}
$$

$$
\begin{aligned}
\implies \|\mathbf{V}'\|_\mathsf{F} \leq \frac{2}{t} \|\mathbf{U}^\star (\mathbf{W}^\star)^\mathsf{T} \mathbf{W}^{(\ell)}\|_\mathsf{F} &+ \sqrt{\frac{4\zeta}{t}} (\max_i \|\mathbf{w}^{(i,\ell)}\|_2) \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2 \\
&+ 4 \|\mathbf{R}^{(i,\ell)}\|_\mathsf{F} \sqrt{\frac{d \log(rd/\delta_0)}{mt}}.
\end{aligned} \tag{122}
$$

Since $\mathbf{R}^{(i,\ell)} = \mathbf{U}^\star \mathbf{w}^{\star(i)} (\mathbf{w}^{(i,\ell)})^\mathsf{T} + (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}$, we have

$$
\begin{aligned}
\|\mathbf{R}^{(i,\ell)}\|_\mathsf{F} &= \|\mathbf{U}^\star \mathbf{w}^{\star(i)} (\mathbf{w}^{(i,\ell)})^\mathsf{T} + (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\|_\mathsf{F} \\
&\leq \|\mathbf{U}^\star \mathbf{w}^{\star(i)} (\mathbf{w}^{(i,\ell)})^\mathsf{T}\|_\mathsf{F} + \|(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\|_\mathsf{F} \\
&\leq \|\mathbf{U}^\star\| \|\mathbf{w}^{\star(i)}\|_2 \|\mathbf{w}^{(i,\ell)}\|_2 + \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2 \|\mathbf{w}^{(i,\ell)}\|_2
\end{aligned} \tag{123}
$$

Using equation 123 in equation 122 gives

$$
\begin{aligned}
\|\mathbf{V}'\|_\mathsf{F} \leq \frac{2}{t} \|\mathbf{U}^\star (\mathbf{W}^\star)^\mathsf{T} \mathbf{W}^{(\ell)}\|_\mathsf{F} &+ \sqrt{\frac{4\zeta}{t}} (\max_i \|\mathbf{w}^{(i,\ell)}\|_2) \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2 \\
&+ 4 \Big(\|\mathbf{U}^\star\| \|\mathbf{w}^{\star(i)}\|_2 \|\mathbf{w}^{(i,\ell)}\|_2 + \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2 \|\mathbf{w}^{(i,\ell)}\|_2\Big) \sqrt{\frac{d \log(rd/\delta_0)}{mt}}
\end{aligned} \tag{124}
$$

Using equation 124. equation 115 and equation 118 in equation 121 gives $\|\mathbf{U}^{(\ell)}\|_\mathsf{F}$

$$
\begin{aligned}
&\leq \frac{1}{\frac{1}{r} \lambda_r \Big(\frac{r}{t} (\mathbf{W}^{(\ell)})^\mathsf{T} \mathbf{W}^{(\ell)}\Big) - c \sqrt{\sum_{i \in [t]} \sum_{j \in [m]} \|\mathbf{w}^{(i,\ell)}\|_2^4} \sqrt{\frac{rd \log(1/\delta_0)}{m^2 t^2}} - \frac{\sigma_1}{mt} \Big(2\sqrt{rd} + 4\sqrt{\log rd}\Big)} \cdot \\
&\quad \Big\{ \frac{2}{t} \|\mathbf{U}^\star (\mathbf{W}^\star)^\mathsf{T} \mathbf{W}^{(\ell)}\|_\mathsf{F} + \sqrt{\frac{4\zeta}{t}} (\max_i \|\mathbf{w}^{(i,\ell)}\|_2) \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2 \\
&\quad + 4 \Big(\|\mathbf{U}^\star\| \|\mathbf{w}^{\star(i)}\|_2 \|\mathbf{w}^{(i,\ell)}\|_2 + \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2 \|\mathbf{w}^{(i,\ell)}\|_2\Big) \sqrt{\frac{d \log(rd/\delta_0)}{mt}} \\
&\quad + \frac{\sigma_2}{mt} 6 \sqrt{rd \log(rd)} + \frac{2\sigma \sqrt{d\mu^\star \lambda_r^\star} \log(2rdmt/\delta_0)}{\sqrt{mt}} \Big\}.
\end{aligned} \tag{125}
$$

Analysis of $\|\Delta(\mathbf{U}^{+(\ell)}, \mathbf{U}^\star)\|_\mathsf{F}$:

$$
\begin{aligned}
\|\Delta(\mathbf{U}^{+(\ell)},\mathbf{U}^\star)\|_{\mathsf{F}} &= \|(\mathbf{I}-\mathbf{U}^\star(\mathbf{U}^\star)^\mathsf{T})\mathbf{U}^{+(\ell)}\|_{\mathsf{F}} \\
&= \min_{\mathbf{Q}^+}\|\mathbf{U}^{+(\ell)}-\mathbf{U}^\star\mathbf{Q}^+\|_{\mathsf{F}} \\
&= \|\mathbf{U}^{(\ell)}-\mathbf{U}^\star(\mathbf{U}^\star)^\mathsf{T}\mathbf{U}^{(\ell)}\|_{\mathsf{F}}\|\mathbf{R}^{-1}\| \qquad (126)
\end{aligned}
$$

From equation 103, we have:

$$
\begin{aligned}
&\mathsf{vec}(\mathbf{U}^{(\ell)}-\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}) = \Big(\mathbf{A}+\frac{\mathbf{N}_1}{mt}\Big)^{-1}\Big(\mathsf{vec}(\mathbf{V})+\Big(\mathsf{vec}\Big(\frac{\mathbf{N}_2}{mt}\Big)-\frac{\mathbf{N}_1}{mt}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\Big)\Big) \\
&= \Big(\mathbf{A}+\frac{\mathbf{N}_1}{mt}\Big)^{-1}\cdot \\
&\quad \Big(\mathsf{vec}\Big(\frac{1}{mt}\sum_{i\in[t]}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big(-\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\mathbf{h}^{(i,\ell)}+(\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)})\Big)(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big) \\
&+\Big(\mathsf{vec}\Big(\frac{\mathbf{N}_2}{mt}\Big)-\frac{\mathbf{N}_1}{mt}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\Big)\Big)
\end{aligned}
$$

Note that $\mathsf{vec}^{-1}\Big(\mathbb{E}\Big[\Big(\mathbf{A}+\frac{\mathbf{N}_1}{mt}\Big)^{-1}\mathsf{vec}\Big(\frac{1}{mt}\sum_{i\in[t]}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big(-\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\mathbf{h}^{(i,\ell)}\Big)(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big)\Big]\Big) = \mathsf{vec}^{-1}\Big(\mathbb{E}\left[\mathbf{A}\right]^{-1}\mathsf{vec}\Big(\frac{-1}{t}\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)\Big)$ lies in the subspace parallel to $\mathbf{U}^\star$ and therefore does not contribute to the distance $\|\Delta(\mathbf{U}^{(\ell)},\mathbf{U}^\star)\|_{\mathsf{F}}$. Subtracting this in equation 126, we get $\|\Delta(\mathbf{U}^{+(\ell)},\mathbf{U}^\star)\|_{\mathsf{F}}$

$$
\begin{aligned}
&\le \|\mathbf{U}^{(\ell)}-\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}+\mathsf{vec}^{-1}\Big(\mathbb{E}\left[\mathbf{A}\right]^{-1}\mathsf{vec}\Big(\frac{1}{t}\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)\Big)\|\mathbf{R}^{-1}\| && (127)\\
&= \|\mathsf{vec}\Big(\mathbf{U}^{(\ell)}-\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\Big)+\mathbb{E}\left[\mathbf{A}\right]^{-1}\mathsf{vec}\Big(\frac{1}{t}\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)\|_2\|\mathbf{R}^{-1}\| && (128)\\
&= \|\Big(\mathbf{A}+\frac{\mathbf{N}_1}{mt}\Big)^{-1}\Big(\mathsf{vec}(\mathbf{V})+\Big(\mathsf{vec}\Big(\frac{\mathbf{N}_2}{mt}\Big)-\frac{\mathbf{N}_1}{mt}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\Big)\Big) && (129)\\
&+\mathbb{E}\left[\mathbf{A}\right]^{-1}\mathsf{vec}\Big(\frac{1}{t}\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)\|_2\|\mathbf{R}^{-1}\| && (130)\\
&= \|\mathbf{R}^{-1}\|\Big\{\Big\|\Big(\mathbf{A}+\frac{\mathbf{N}_1}{mt}\Big)^{-1}\Big(\mathsf{vec}(\mathbf{V})+\mathsf{vec}\Big(\frac{\mathbf{N}_2}{mt}\Big)-\frac{\mathbf{N}_1}{mt}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\Big) && (131)\\
&\qquad +\Big(\mathbf{A}+\frac{\mathbf{N}_1}{mt}\Big)\mathbb{E}\left[\mathbf{A}\right]^{-1}\mathsf{vec}\Big(\frac{1}{t}\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)\Big\|_2\Big\} && (132)
\end{aligned}
$$

$$
\begin{aligned}
&\le \Bigg\{\frac{c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(1/\delta_0)}{m^2t^2}}+\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)}{\frac{1}{r}\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)-c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(1/\delta_0)}{m^2t^2}}-\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)}\cdot \\
&\quad \Big(\frac{2}{t}\|\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\|_{\mathsf{F}}\Big) \\
&+\frac{1}{\frac{1}{r}\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)-c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(1/\delta_0)}{m^2t^2}}-\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)}\cdot \\
&\quad \Big\{\Big(\sqrt{\frac{4\zeta}{t}}(\max_i\|\mathbf{w}^{(i,\ell)}\|_2)\|\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)}\|_2\Big) \\
&\quad +4\Big(\|\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|\|\mathbf{h}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2+\|\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2\Big)\sqrt{\frac{d\log(rd/\delta_0)}{mt}} \\
&\quad \frac{\sigma_2}{mt}6\sqrt{rd\log(rd)}+\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)\sqrt{r}+\frac{2\sigma\sqrt{d\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big\}\Bigg\}\|\mathbf{R}^{-1}\|. \qquad (133)
\end{aligned}
$$

□

Corollary 3. *If* $\mathsf{B}_{\mathbf{U}^{(\ell-1)}} = \mathcal{O}\Big(\frac{1}{\sqrt{r\mu^\star}}\Big)$, $\sqrt{\frac{r\log(1/\delta_0)}{m}} = \mathcal{O}(1)$, $\sqrt{\nu^{(\ell-1)}} = \mathcal{O}\Big(\frac{1}{\sqrt{r\mu^\star}}\Big)$, $\sqrt{\frac{r^2\log(r/\delta_0)}{m}} = \mathcal{O}\Big(\frac{1}{\sqrt{\mu^\star}}\Big)$, $\epsilon < \sqrt{\mu^\star\lambda_r^\star}$, $\sqrt{\frac{r^2\zeta}{t}} = \mathcal{O}\Big(\frac{1}{\sqrt{\mu^\star}}\Big)$, $\Lambda' = \mathcal{O}\Big(\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big)$, $\Lambda = \mathcal{O}(\sqrt{\lambda_r^\star})$, $\sigma\sqrt{\frac{r^2\log^2(r\delta^{-1})}{m}} = \mathcal{O}(\sqrt{\lambda_r^\star})$, $\sqrt{\frac{r^3 d\log(1/\delta_0)}{mt}} = \min\{\mathcal{O}\Big(\frac{1}{\mu^\star}\Big), \mathcal{O}\Big(\frac{1}{\mu^\star\lambda_r^\star}\Big)\}$, $\frac{\sigma_1}{mt}\Big(2\sqrt{rd} + 4\sqrt{\log rd}\Big) = \min\{\mathcal{O}\Big(\frac{\lambda_r^\star}{r}\Big), \mathcal{O}\Big(\frac{1}{r}\Big)\}$, $mt = \widetilde{\Omega}(dr^2\mu^\star(1 + \frac{1}{\lambda_r^\star}))$, $\zeta = \widetilde{O}(t(\mu^\star\lambda_r^\star)^{-1})$, $m = \widetilde{\Omega}(\sigma^2 r^3/\lambda_r^\star)$ *and Assumption 4 holds for iteration* $\ell - 1$*, then, with probability* $1 - \mathcal{O}(\delta_0)$*,*

$$
\begin{aligned}
&\|\mathbf{U}^{(\ell)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_{\mathsf{F}} \\
&= \mathcal{O}\Big(\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big) + \mathcal{O}\Big(\Lambda' + \frac{\Lambda}{\sqrt{\mu^\star\lambda_r^\star}} + \frac{\sigma_2 r}{mt\lambda_r^\star}\sqrt{rd\log(rd)} + \frac{\sigma_1 r^{3/2}}{mt\lambda_r^\star}\sqrt{rd\log rd} \\
&\quad + \sigma\Big(\sqrt{\frac{r^3 d\mu^\star\log^2(r\delta^{-1})}{mt\lambda_r^\star}} + \sqrt{\frac{r^3\log^2(r\delta^{-1})}{m\lambda_r^\star}}\Big)\Big),
\end{aligned}
$$

and $\Big\|\Delta(\mathbf{U}^{+(\ell)}, \mathbf{U}^\star)\Big\|_{\mathsf{F}}$

$$
\begin{aligned}
&\le \mathcal{O}\Big(\|\mathbf{R}^{-1}\|\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big) \\
&+ \mathcal{O}\Big(\|\mathbf{R}^{-1}\|\Big\{\tfrac{\Lambda'\sqrt{\lambda_r^\star}}{r} + \tfrac{\Lambda}{r} + \tfrac{\sigma_2 r}{mt\lambda_r^\star}\sqrt{rd\log(rd)} + \tfrac{\sigma_1 r\sqrt{r}}{mt\lambda_r^\star}\sqrt{rd\log rd} + \sigma\Big(\sqrt{\tfrac{r^3 d\mu^\star\log^2(rdmt/\delta_0)}{mt\lambda_r^\star}}\Big)\Big\}\Big).
\end{aligned}
$$

Proof. The proof follows from plugging the various constant bounds from the corollary statement and Inductive Assumption 4 in the expressions of Lemma 10. Note that, $\|\mathbf{U}^{(\ell)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_{\mathsf{F}}$

$$
\begin{aligned}
&\le \frac{1}{\frac{1}{r}\lambda_r - c\mu\lambda_r\sqrt{\frac{rd\log(1/\delta_0)}{mt}} - \frac{\sigma_1}{mt}\Big(2\sqrt{rd} + 4\sqrt{\log rd}\Big)} \cdot \\
&\quad \Big\{\frac{2}{t}\|\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|\|(\mathbf{H}^{(\ell)})^\top\mathbf{W}^{(\ell)}\|_{\mathsf{F}} + \sqrt{\frac{4\zeta}{t}}(\max_i\|\mathbf{w}^{(i,\ell)}\|_2\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2) \\
&\quad + 4\Big(\|\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|\|\mathbf{h}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2 + \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2\Big)\sqrt{\frac{d\log(rd/\delta_0)}{mt}} \\
&\quad + \frac{\sigma_2}{mt}6\sqrt{rd\log(rd)} + \frac{\sigma_1}{mt}\Big(2\sqrt{rd} + 4\sqrt{\log rd}\Big)\sqrt{r} + \frac{2\sigma\sqrt{d\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big\}.
\end{aligned}
\tag{134}
$$

Using Assumption 4 for $\mathbf{b}^{(i,\ell)}$ and $\mathbf{Q}^{(\ell-1)}$ terms, the fact that $\mathbf{U}^\star$ is orthonormal and eigenvalue ratios and incoherence bounds for $\mathbf{H}^{(\ell)}$ and $\mathbf{W}^{(\ell)}$ from Corollaries 1 and 2, the above becomes, $\|\mathbf{U}^{(\ell)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_{\mathsf{F}}$

$$
\begin{aligned}
&\le \frac{1}{\frac{1}{r}\lambda_r - c\mu\lambda_r\sqrt{\frac{rd\log(1/\delta_0)}{mt}} - \frac{\sigma_1}{mt}\Big(2\sqrt{rd} + 4\sqrt{\log rd}\Big)} \cdot \\
&\Big\{\tfrac{2}{t}\Big(\mathcal{O}\Big(\tfrac{\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\frac{\lambda_r^\star}{\lambda_1^\star}\sqrt{\frac{t}{r}\lambda_1^\star}}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\Lambda'\sqrt{\tfrac{t}{r}\lambda_r^\star}\tfrac{1}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\tfrac{1}{\sqrt{r\mu^\star}}\sqrt{\tfrac{t}{r}}\Lambda\Big) + \mathcal{O}\Big(\sigma\sqrt{\tfrac{rt\log^2(r\delta^{-1})}{m}}\Big)\Big) \cdot \\
&\quad \mathcal{O}(\sqrt{t\mu^\star\lambda_r^\star}) + \sqrt{\frac{4\zeta}{t}} \cdot \mathcal{O}(\sqrt{\mu^\star\lambda_r^\star}) \cdot \Big(c'\sqrt{\mu^\star\lambda_r^\star}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} + \Lambda\Big)
\end{aligned}
$$

$$\begin{aligned}
&+4\sqrt{\frac{d\log(rd/\delta_0)}{mt}}\mathcal{O}(\sqrt{\mu^\star\lambda_r^\star})\Big(\mathcal{O}\Big(\frac{\max\{\epsilon,\|\mathbf{w}^{\star(i)}\|_2\}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}}{\sqrt{r\mu^\star}}\Big)+\mathcal{O}\Big(\frac{\Lambda'\|\mathbf{w}^{\star(i)}\|_2}{\sqrt{r\mu^\star}}\Big) \\
&\quad+\mathcal{O}\Big(\frac{\Lambda}{\sqrt{r\mu^\star}}\Big)+\mathcal{O}\Big(\sigma\sqrt{\frac{r\log^2(r\delta^{-1})}{m}}\Big)+c'\max\{\epsilon,\|\mathbf{w}^{\star(i)}\|_2\}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}+\Lambda\Big) \\
&+\frac{\sigma_2}{mt}6\sqrt{rd\log(rd)}+\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)\sqrt{r}+\frac{2\sigma\sqrt{d\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big\} \\
&=J_1+J_2
\end{aligned} \tag{135}$$

where J_1 denotes the terms which arise from analysing the problem in the noiseless setting and J_2 denotes the contribution of noise terms $(\sigma_1,\sigma_2,\sigma,\Lambda,\Lambda')$. We will analyse both separately. Note that J_1

$$\begin{aligned}
&=\frac{1}{\frac{1}{r}\lambda_r-\frac{\mu}{\mu^\star}\cdot\frac{c\mu^\star\lambda_r}{r}\cdot\sqrt{\frac{r^3d\log(1/\delta_0)}{mt}}-\frac{\lambda_r}{r}\cdot\frac{\lambda_r^\star}{\lambda_r}\cdot\frac{\sigma_1}{mtr\lambda_r^\star}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)}\cdot \\
&\Big\{\frac{2}{t}\cdot\mathcal{O}\Big(\frac{\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\frac{\lambda_r^\star}{\lambda_1^\star}\sqrt{\frac{t}{r}\lambda_1^\star}}{\sqrt{r\mu^\star}}\cdot\sqrt{t\mu^\star\lambda_r^\star}\Big)+\mathcal{O}\Big(\sqrt{\frac{\zeta}{t}}\cdot\sqrt{\mu^\star\lambda_r^\star}\cdot\sqrt{\mu^\star\lambda_r^\star}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big) \\
&+\mathcal{O}\Big(\sqrt{\tfrac{d\log(rd/\delta_0)}{mt}}\cdot\sqrt{\mu^\star\lambda_r^\star}\cdot\Big(\tfrac{\max\{\epsilon,\|\mathbf{w}^{\star(i)}\|_2\}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}}{\sqrt{r\mu^\star}}+\max\{\epsilon,\|\mathbf{w}^{\star(i)}\|_2\}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\tfrac{\lambda_r^\star}{\lambda_1^\star}}\Big)\Big)\Big\}.
\end{aligned}$$

Using $\mathsf{B}_{\mathbf{U}^{(\ell-1)}}=\mathcal{O}\Big(\frac{1}{\sqrt{r\mu^\star}}\Big)$, $\lambda_r^\star\le\lambda_1^\star$, $r\ge 1$,$\mu^\star\ge 1$, the bracketed expression in the above simplifies to

$$\begin{aligned}
\le\ &\frac{\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}}{\frac{1}{r}\lambda_r-\frac{\mu}{\mu^\star}\frac{c\mu^\star\lambda_r}{r}\sqrt{\frac{r^3d\log(1/\delta_0)}{mt}}-\frac{\lambda_r}{r}\cdot\frac{\lambda_r^\star}{\lambda_r}\cdot\frac{\sigma_1}{mtr\lambda_r^\star}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)}\cdot \\
&\mathcal{O}\Big(\frac{\lambda_r^\star}{r}+\sqrt{\frac{\zeta}{t}}\mu^\star\lambda_r^\star+\sqrt{\frac{d\log(rd/\delta_0)}{mt}}\mu^\star\lambda_r^\star\Big(\frac{1}{\sqrt{r\mu^\star}}+1\Big)\Big).
\end{aligned}$$

Further, using $\sqrt{\frac{r^3d\log(1/\delta_0)}{mt}}=\mathcal{O}(\frac{1}{\mu^\star})$, $\frac{\sigma_1}{mtr}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)=\mathcal{O}(\lambda_r^\star)$, $\sqrt{\frac{r^2\zeta}{t}}=\mathcal{O}\Big(\frac{1}{\mu^\star}\Big)$, and eigenvalue and incoherence ratios from Corollary 2 in the above, we have

$$\begin{aligned}
J_1&=\frac{\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}}{1-\mathcal{O}(1)}\cdot\mathcal{O}\Big(\frac{1}{\sqrt{r\mu^\star}}+1\Big) \\
&=\mathcal{O}\Big(\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big).
\end{aligned} \tag{136}$$

Similarly, we have J_2

$$\begin{aligned}
&=\frac{1}{\frac{1}{r}\lambda_r-\frac{\mu}{\mu^\star}\frac{c\mu^\star\lambda_r}{r}\sqrt{\frac{r^3d\log(1/\delta_0)}{mt}}-\frac{\lambda_r}{r}\cdot\frac{\lambda_r^\star}{\lambda_r}\cdot\frac{\sigma_1}{mtr\lambda_r^\star}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)}\cdot \\
&\quad\Big\{\frac{2}{t}\mathcal{O}\Big(\sqrt{\frac{t}{r}\lambda_r^\star}\frac{\Lambda'}{\sqrt{r\mu^\star}}+\frac{1}{\sqrt{r\mu^\star}}\sqrt{\frac{t}{r}}\Lambda+\sigma\sqrt{\frac{rt\log^2(r\delta^{-1})}{m}}\Big)\cdot\mathcal{O}(\sqrt{t\mu^\star\lambda_r^\star}) \\
&\quad+\mathcal{O}\Big(\sqrt{\tfrac{\zeta}{t}}\cdot\sqrt{\mu^\star\lambda_r^\star}\Lambda\Big)+\mathcal{O}\Big(\sqrt{\tfrac{d\log(rd/\delta_0)}{mt}}\sqrt{\mu^\star\lambda_r^\star}\Big(\tfrac{\Lambda'\|\mathbf{w}^{\star(i)}\|_2}{\sqrt{r\mu^\star}}+\Lambda+\tfrac{1}{\sqrt{r\mu^\star}}\sigma\sqrt{\tfrac{r\log^2(r\delta^{-1})}{m}}\Big)\Big) \\
&\quad+\frac{\sigma_2}{mt}6\sqrt{rd\log(rd)}+\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)\sqrt{r}+\frac{2\sigma\sqrt{d\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big\}.
\end{aligned}$$

Now, using $\sqrt{\frac{r^3d\log(1/\delta_0)}{mt}}=\mathcal{O}(\frac{1}{\mu^\star})$, $\frac{\sigma_1}{mtr}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)=\mathcal{O}(\lambda_r^\star)$, eigenvalue and incoherence ratios from Corollary 2 for the denominator term and $\sqrt{\frac{r^2\zeta}{t}}=\mathcal{O}\Big(\frac{1}{\mu^\star}\Big)$, $\lambda_r^\star\le\lambda_1^\star$, $r\ge 1$,$\mu^\star\ge 1$

for the bracketed terms in the above, we get J_2

$$= \frac{1}{1-\mathcal{O}(1)}\Big\{\mathcal{O}\Big(\Lambda' + \frac{\Lambda}{\sqrt{\lambda_r^\star}} + \sigma\sqrt{\frac{r^3\mu^\star\log^2(r\delta^{-1})}{m\lambda_r^\star}}\Big) + \mathcal{O}\Big(\sqrt{\frac{\zeta}{t}}\cdot\sqrt{\mu^\star\lambda_r^\star}\Lambda\Big)$$
$$+ \mathcal{O}\Big(\sqrt{\frac{r^2 d\log(rd/\delta_0)}{mt}}\sqrt{\mu^\star}\Big(\frac{\Lambda'}{\sqrt{r}} + \frac{\Lambda}{\sqrt{\lambda_r^\star}} + \frac{\sigma}{\sqrt{r\mu^\star\lambda_r^\star}}\sqrt{\frac{r\log^2(r\delta^{-1})}{m}}\Big)\Big)$$
$$+ \frac{\sigma_2 r}{mt\lambda_r^\star}6\sqrt{rd\log(rd)} + \frac{\sigma_1 r}{mt\lambda_r^\star}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)\sqrt{r} + \frac{2\sigma r\sqrt{d\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}\lambda_r^\star}\Big\}. \tag{137}$$

Rearranging the terms in above gives

$$= \mathcal{O}\Big(\Big\{\Lambda'\Big(1+\sqrt{\frac{r^2 d\log(rd/\delta_0)}{mt}}\frac{\sqrt{\mu^\star}}{\sqrt{r}}\Big) + \Lambda\Big(\frac{1}{\sqrt{\lambda_r^\star}} + \sqrt{\frac{\zeta}{t}}\cdot\sqrt{\mu^\star\lambda_r^\star} + \sqrt{\frac{r^2 d\log(rd/\delta_0)}{mt}}\frac{\sqrt{\mu^\star}}{\sqrt{\lambda_r^\star}}\Big)$$
$$+ \frac{\sigma_2 r}{mt\lambda_r^\star}\sqrt{rd\log(rd)} + \frac{\sigma_1 r}{mt\lambda_r^\star}\Big(\sqrt{rd}+\sqrt{\log rd}\Big)\sqrt{r}$$
$$+ \sigma\Big(\sqrt{\tfrac{r^3\log^2(r\delta^{-1})}{m\lambda_r^\star}} + \sqrt{\tfrac{r^2 d\log(rd/\delta_0)}{mt}}\cdot\tfrac{1}{\sqrt{r\lambda_r^\star}}\sqrt{\tfrac{r\log^2(r\delta^{-1})}{m}} + \tfrac{r\sqrt{d\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}\lambda_r^\star}\Big)\Big\}. \tag{138}$$

Substituting $mt = \widetilde{\Omega}(dr^2\mu^\star(1+\frac{1}{\lambda_r^\star}))$, $\zeta = \widetilde{O}(t(\mu^\star\lambda_r^\star)^{-1})$, we have that

$$J_2 = \mathcal{O}\Big(\Lambda' + \frac{\Lambda}{\sqrt{\mu^\star\lambda_r^\star}} + \frac{\sigma_2 r}{mt\lambda_r^\star}\sqrt{rd\log(rd)} + \frac{\sigma_1 r^{3/2}}{mt\lambda_r^\star}\sqrt{rd\log rd}$$
$$+ \sigma\Big(\sqrt{\frac{r^3 d\mu^\star\log^2(r\delta^{-1})}{mt\lambda_r^\star}} + \sqrt{\frac{r^3\log^2(r\delta^{-1})}{m\lambda_r^\star}}\Big)\Big). \tag{139}$$

Using equation 136, equation 139 in equation 135 gives us the required norm bound for $\|\mathbf{U}^{(\ell)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_{\mathsf{F}}$.

Similarly, we can simplify the following $\big|\big|\Delta(\mathbf{U}^{+(\ell)},\mathbf{U}^\star)\big|\big|_{\mathsf{F}}$ bound expression from the Lemma statement.

$$\leq \Big\{\frac{c\mu\lambda_r\sqrt{\frac{rd\log(1/\delta_0)}{mt}} + \frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)}{\frac{1}{r}\lambda_r - c\mu\lambda_r\sqrt{\frac{rd\log(1/\delta_0)}{mt}} - \frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)}\cdot\Big(\frac{2}{t}\|\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\|_{\mathsf{F}}\Big)$$
$$+ \frac{1}{\frac{1}{r}\lambda_r - c\mu\lambda_r\sqrt{\frac{rd\log(1/\delta_0)}{mt}} - \frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)}\cdot$$
$$\Big\{\Big(\sqrt{\frac{4\zeta}{t}}(\max_i\|\mathbf{w}^{(i,\ell)}\|_2)\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\Big)+$$
$$4\Big(\|\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|\|\mathbf{h}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2 + \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2\Big)\sqrt{\frac{d\log(rd/\delta_0)}{mt}}$$
$$\frac{\sigma_2}{mt}6\sqrt{rd\log(rd)} + \frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)\sqrt{r} + \frac{2\sigma\sqrt{d\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big\}\Big\}\|\mathbf{R}^{-1}\| \tag{140}$$

Using the stated bounds on $\sqrt{\frac{r^3 d\log(1/\delta_0)}{mt}}$, $\frac{\sigma_1}{mtr}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)$, Corollary 2 as well as $\mathbf{H}^{(\ell)}$ and $\mathbf{W}^{(\ell)}$ bounds from Corollaries 1 and 2 in the above, we have

$$
\begin{aligned}
&\le \|\mathbf{R}^{-1}\|\Big\{\frac{\mathcal{O}(1)}{1-\mathcal{O}(1)}\frac{1}{\lambda_r^\star}\cdot\frac{2}{t}\mathcal{O}\Big(\Big(\frac{\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\frac{\lambda_r^\star}{\lambda_1^\star}\sqrt{\frac{t}{r}\lambda_1^\star}}{\sqrt{r\mu^\star}}+\sqrt{\frac{t}{r}\lambda_r^\star}\frac{\Lambda'}{\sqrt{r\mu^\star}} \\
&\quad+\frac{1}{\sqrt{r\mu^\star}}\sqrt{\frac{t}{r}}\Lambda+\sigma\sqrt{\frac{r\log^2(r\delta^{-1})}{m}}\Big)\sqrt{t\mu^\star\lambda_r^\star}\Big)+\frac{1}{1-\mathcal{O}(1)}\frac{r}{\lambda_r}\cdot \\
&\quad\Big\{\sqrt{\frac{4\zeta}{t}}\cdot\mathcal{O}(\sqrt{\mu^\star\lambda_r^\star})\Big(c'\max\{\epsilon,\|\mathbf{w}^{\star(i)}\|_2\}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}+\Lambda\Big) \\
&\quad+4\sqrt{\frac{d\log(rd/\delta_0)}{mt}}\cdot\mathcal{O}(\sqrt{\mu^\star\lambda_r^\star})\cdot\mathcal{O}\Big(\frac{\max\{\epsilon,\|\mathbf{w}^{\star(i)}\|_2\}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}}{\sqrt{r\mu^\star}}+\frac{\Lambda'\|\mathbf{w}^{\star(i)}\|_2}{\sqrt{r\mu^\star}} \\
&\quad+\frac{\Lambda}{\sqrt{r\mu^\star}}+\sigma\sqrt{\frac{r\log^2(r\delta^{-1})}{m}}+\max\{\epsilon,\|\mathbf{w}^{\star(i)}\|_2\}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}+\Lambda\Big) \\
&\quad+\frac{\sigma_2}{mt}6\sqrt{rd\log(rd)}+\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)\sqrt{r}+\frac{2\sigma\sqrt{d\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big\}\Big\} \qquad (141) \\
&= J_1'+J_2' \qquad (142)
\end{aligned}
$$

where as before, J_1' denotes the terms which arise from analysing the problem in the noiseless setting and J_2' denotes the contribution of noise terms $(\sigma_1,\sigma_2,\sigma,\Lambda,\Lambda')$. Analysing both the terms separately, we have J_1'

$$
\begin{aligned}
&= \|\mathbf{R}^{-1}\|\Big\{\mathcal{O}(1)\frac{1}{\lambda_r^\star}\cdot\frac{2}{t}\mathcal{O}\Big(\frac{\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\frac{\lambda_r^\star}{\lambda_1^\star}\sqrt{\frac{t}{r}\lambda_1^\star}}{\sqrt{r\mu^\star}}\Big)\sqrt{t}\mathcal{O}(\sqrt{\mu^\star\lambda_r^\star}) \\
&\quad+\mathcal{O}\Big(\frac{r}{\lambda_r}\Big)\Big\{\sqrt{\frac{4\zeta}{t}}\cdot\mathcal{O}(\sqrt{\mu^\star\lambda_r^\star})\cdot\max\{\epsilon,\|\mathbf{w}^{\star(i)}\|_2\}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} \\
&\quad+4\sqrt{\frac{d\log(rd/\delta_0)}{mt}}\cdot\mathcal{O}(\sqrt{\mu^\star\lambda_r^\star})\cdot\mathcal{O}\Big(\frac{\max\{\epsilon,\|\mathbf{w}^{\star(i)}\|_2\}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}}{\sqrt{r\mu^\star}}\Big)\Big\}\Big\}. \qquad (143) \\
&= \mathcal{O}\Big(\|\mathbf{R}^{-1}\|\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big\{\frac{1}{r}+\Big\{\sqrt{\frac{r^2\zeta}{t}}\mu^\star+\sqrt{\frac{r^2 d\log(rd/\delta_0)}{mt}}\cdot\frac{\mu^\star}{\sqrt{r\mu^\star}}\Big\}\Big\}\Big) \\
&= \mathcal{O}\Big(\|\mathbf{R}^{-1}\|\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big), \qquad (144)
\end{aligned}
$$

where in the last two steps we use the fact that $\lambda_r^\star \le \lambda_1^\star$, $r \ge 1$,$\mu^\star \ge 1$, $\sqrt{\frac{r^2\zeta}{t}} = \mathcal{O}\Big(\frac{1}{\mu^\star}\Big)$ and $\sqrt{\frac{r^3 d\log(1/\delta_0)}{mt}} = \mathcal{O}(\frac{1}{\mu^\star})$. Similarly we have J_2'

$$
\begin{aligned}
&= \|\mathbf{R}^{-1}\|\Big\{\mathcal{O}(1)\cdot\frac{2}{t}\mathcal{O}\Big(\sqrt{\frac{t}{r}\lambda_r^\star}\frac{\Lambda'}{\sqrt{r\mu^\star}}+\frac{1}{\sqrt{r\mu^\star}}\sqrt{\frac{t}{r}}\Lambda+\sigma\sqrt{\frac{r\log^2(r\delta^{-1})}{m}}\Big)\cdot\sqrt{t}\mathcal{O}(\sqrt{\mu^\star\lambda_r^\star}) \\
&\quad+\mathcal{O}\Big(\frac{r}{\lambda_r}\Big)\cdot\Big\{\sqrt{\frac{4\zeta}{t}}\cdot\mathcal{O}(\sqrt{\mu^\star\lambda_r^\star})\Lambda \\
&\quad+4\sqrt{\frac{d\log(rd/\delta_0)}{mt}}\cdot\mathcal{O}(\sqrt{\mu^\star\lambda_r^\star})\mathcal{O}\Big(\frac{\Lambda'\|\mathbf{w}^{\star(i)}\|_2}{\sqrt{r\mu^\star}}+\frac{\Lambda}{\sqrt{r\mu^\star}}+\sigma\sqrt{\frac{r\log^2(r\delta^{-1})}{m}}+\Lambda\Big)
\end{aligned}
$$

$$+\frac{\sigma_2}{mt}6\sqrt{rd\log(rd)}+\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)\sqrt{r}+\frac{2\sigma\sqrt{d\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big\}\Big\}. \tag{145}$$

Rearranging the terms in above gives

$$\begin{aligned}
&=\mathcal{O}\Big(\|\mathbf{R}^{-1}\|\Big\{\Lambda'\Big(\frac{\sqrt{\lambda_r^\star}}{r}+\sqrt{\frac{r^2d\log(rd/\delta_0)}{mt}}\cdot\frac{\mu^\star}{\sqrt{r\mu^\star}}\Big)\\
&\quad+\Lambda\Big(\frac{1}{r}+\sqrt{\frac{r^2\zeta}{t}}\cdot\frac{\sqrt{\mu^\star\lambda_r^\star}}{\lambda_r^\star}+\sqrt{\frac{d\log(rd/\delta_0)}{mt}}\frac{\sqrt{\mu^\star\lambda_r^\star}}{\lambda_r^\star}\cdot\Big(1+\frac{1}{\sqrt{r\mu^\star}}\Big)\Big)\\
&\quad+\frac{\sigma_2 r}{mt\lambda_r^\star}\sqrt{rd\log(rd)}+\frac{\sigma_1 r}{mt\lambda_r^\star}\Big(\sqrt{rd}+\sqrt{\log rd}\Big)\sqrt{r}\\
&\quad+\sigma\frac{r\sqrt{\mu^\star\lambda_r^\star}}{\lambda_r^\star}\Big(\sqrt{\frac{r\log^2(r\delta^{-1})}{mt}}+\sqrt{\frac{d\log(rd/\delta_0)}{mt}}\sqrt{\frac{r\log^2(r\delta^{-1})}{m}}+\frac{\sqrt{d}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big)\Big\}\Big).
\end{aligned} \tag{146}$$

Assuming $mt=\widetilde{\Omega}(dr^2\mu^\star(1+\frac{1}{\lambda_r^\star}))$, $mt=\widetilde{\Omega}\Big(\frac{\sqrt{dr^3}}{\lambda_r^\star}\max(\sigma_1\sqrt{r},\sigma_2)\Big)$, $m=\widetilde{\Omega}(\sigma^2r^3/\lambda_r^\star)$, we have the above as

$${}'_2=\mathcal{O}\Big(\|\mathbf{R}^{-1}\|\Big\{\tfrac{\Lambda'\sqrt{\lambda_r^\star}}{r}+\tfrac{\Lambda}{r}+\tfrac{\sigma_2 r}{mt\lambda_r^\star}\sqrt{rd\log(rd)}+\tfrac{\sigma_1 r\sqrt{r}}{mt\lambda_r^\star}\sqrt{rd\log rd}+\sigma\Big(\sqrt{\tfrac{r^3d\mu^\star\log^2(rdmt/\delta_0)}{mt\lambda_r^\star}}\Big)\Big\}\Big) \tag{147}$$

Using equation 144 and equation 147 in equation 142 gives us the required bound for $\big|\big|\Delta(\mathbf{U}^{+(\ell)},\mathbf{U}^\star)\big|\big|_{\mathsf{F}}$. ☐

Corollary 4. *If* $\frac{1}{2}\le\sigma_{\min}(\mathbf{Q}^{(\ell-1)})\le\sigma_{\max}(\mathbf{Q}^{(\ell-1)})<1$, $\frac{2r}{\lambda_r^\star}\Big(\frac{\sigma_2}{mt}6\sqrt{rd\log(rd)}+\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)\sqrt{r}+\frac{2\sigma\sqrt{d\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big)=\mathcal{O}(1)$ *then* $\mathbf{R}$ *is invertible and* $\|\mathbf{R}^{-1}\|\le 2+c''$ *and* $||\mathbf{R}||\le 1+c''$ *for some* $c''>0$ *w.p.* $1-\mathcal{O}(\delta_0)$.

Proof. $\forall\,\mathbf{z}\in\mathbb{R}^r$, we have:

$$\begin{aligned}
\|\mathbf{U}^{+(\ell)}\mathbf{R}\mathbf{z}\|_2&=\|\mathbf{U}^{(\ell)}\mathbf{z}\|\\
&=\|\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\mathbf{z}-\mathsf{vec}^{-1}\Big(\mathbb{E}\left[\mathbf{A}\right]^{-1}\mathsf{vec}\Big(\frac{1}{t}\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)\Big)\\
&\quad+\Big(\mathbf{U}^{(\ell)}-\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}+\mathsf{vec}^{-1}\Big(\mathbb{E}\left[\mathbf{A}\right]^{-1}\mathsf{vec}\Big(\frac{1}{t}\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)\Big)\Big)\mathbf{z}\Big\|_2
\end{aligned} \tag{148}$$

Using equation 148, we have:

$$\begin{aligned}
&\min_{\|\mathbf{z}\|_2=1}\|\mathbf{U}^{+(\ell)}\mathbf{R}\mathbf{z}\|_2\\
&\ge\min\nolimits_{\|\mathbf{z}\|_2=1}\sqrt{\mathbf{z}^\mathsf{T}(\mathbf{Q}^{(\ell-1)})^\mathsf{T}(\mathbf{U}^\star)^\mathsf{T}\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\mathbf{z}}-\|\mathsf{vec}^{-1}\Big(\mathbb{E}\left[\mathbf{A}\right]^{-1}\mathsf{vec}\Big(\tfrac{1}{t}\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)\Big)\|\\
&\quad-\|\mathbf{U}^{(\ell)}-\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}+\mathsf{vec}^{-1}\Big(\mathbb{E}\left[\mathbf{A}\right]^{-1}\mathsf{vec}\Big(\frac{1}{t}\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)\Big)\| \qquad (149)\\
&\ge\min_{\|\mathbf{z}\|_2=1}\sqrt{\mathbf{z}^\mathsf{T}(\mathbf{Q}^{(\ell-1)})^\mathsf{T}\mathbf{Q}^{(\ell-1)}\mathbf{z}}-\|\mathsf{vec}^{-1}\Big(\mathbb{E}\left[\mathbf{A}\right]^{-1}\mathsf{vec}\Big(\frac{1}{t}\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)\Big)\|_\mathsf{F}\\
&\quad-\|\mathbf{U}^{(\ell)}-\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}-\mathsf{vec}^{-1}\Big(\mathbb{E}\left[\mathbf{A}\right]^{-1}\mathsf{vec}\Big(\frac{1}{t}\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)\Big)\|_\mathsf{F}\\
&\ge\sigma_{\min}(\mathbf{Q}^{(\ell-1)})-\|\mathbb{E}\left[\mathbf{A}\right]^{-1}\|\|\frac{1}{t}\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\|_\mathsf{F}\\
&\quad-\|\mathbf{U}^{(\ell)}-\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}+\mathsf{vec}^{-1}\Big(\mathbb{E}\left[\mathbf{A}\right]^{-1}\mathsf{vec}\Big(\frac{1}{t}\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)\Big)\|_\mathsf{F}
\end{aligned}$$

Using the bounds from Corollary 1 and 3 and Inductive Assumption 4 in the above we get,

$$\begin{aligned}
&\geq \frac{1}{2} - \frac{r}{t\lambda_r}\sqrt{t\mu\lambda_r} \cdot \mathcal{O}\Big(\frac{\mathsf{B}_{\mathbf{U}^{(\ell-1)}} \frac{\lambda_r^\star}{\lambda_1^\star}\sqrt{\frac{t}{r}\lambda_1^\star}}{\sqrt{r\mu^\star}} + \sqrt{\frac{t}{r}\lambda_r^\star}\frac{\Lambda'}{\sqrt{r\mu^\star}} + \frac{\Lambda}{\sqrt{r\mu^\star}}\sqrt{\frac{t}{r}} + \sigma\sqrt{\frac{r\log^2(r\delta^{-1})}{m}}\Big) \\
&\quad - \mathcal{O}\Big(\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} + \Lambda' + \frac{\Lambda}{\sqrt{\mu^\star\lambda_r^\star}} + \frac{\sigma_2 r}{mt\lambda_r^\star}\sqrt{rd\log(rd)} \\
&\quad + \frac{\sigma_1 r\sqrt{r}}{mt\lambda_r^\star}\sqrt{rd\log rd} + \sigma\sqrt{\frac{r^3 d\mu^\star\log^2(rdmt/\delta_0)}{mt\lambda_r^\star}}\Big). \qquad (150)
\end{aligned}$$

Rearranging the terms in the above gives

$$\begin{aligned}
&\geq \frac{1}{2} - \mathcal{O}\Big(\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\Big(\sqrt{\frac{\mu}{\mu^\star}}\sqrt{\frac{\lambda_r}{\lambda_r^\star}} + 1\Big) + \Lambda'\Big(\sqrt{\frac{\mu}{\mu^\star}}\sqrt{\frac{\lambda_r}{\lambda_r^\star}} + 1\Big) + \frac{\Lambda}{\mu^\star\lambda_r^\star}\Big(\sqrt{\frac{\mu}{\mu^\star}}\sqrt{\frac{\lambda_r}{\lambda_r^\star}} + 1\Big) \\
&\quad + \frac{\sigma_2 r}{mt\lambda_r^\star}\sqrt{rd\log(rd)} + \frac{\sigma_1 r\sqrt{r}}{mt\lambda_r^\star}\sqrt{rd\log rd} \\
&\quad + \sigma\Big(\sqrt{\frac{\mu}{\mu^\star}}\sqrt{\frac{\lambda_r}{\lambda_r^\star}}\sqrt{\frac{\mu^\star r^2\log^2(r\delta^{-1})}{mt\lambda_r^\star}} + \sqrt{\frac{r^3 d\mu^\star\log^2(rdmt/\delta_0)}{mt\lambda_r^\star}} + \sqrt{\frac{r^3\log^2(r\delta^{-1})}{m\lambda_r^\star}}\Big)\Big) \\
&\geq \frac{1}{2} - c''. \qquad (151)
\end{aligned}$$

where $c'' > 0$ Here, we need to use that $\Lambda' < 10^{-3}, \Lambda < 10^{-3}\mu^\star\lambda_r^\star$, $mt = \Omega(\max(\sigma_2, \sigma_1\sqrt{r})\sqrt{rd\log d}/\lambda_r^\star)$ and $mt = \widetilde{\Omega}(\sigma^2 dr^3\mu^\star/\lambda_r^\star)$, $m = \widetilde{\Omega}(\sigma^2 r^3/\lambda_r^\star)$.

Similarly, using equation 148, we also have:

$$\begin{aligned}
&\max_{\|\mathbf{z}\|_2=1} \|\mathbf{U}^{+(\ell)}\mathbf{R}\mathbf{z}\|_2 \\
&\leq \max_{\|\mathbf{z}\|_2=1} \sqrt{\mathbf{z}^\mathsf{T}(\mathbf{Q}^{(\ell-1)})^\mathsf{T}(\mathbf{U}^\star)^\mathsf{T}\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\mathbf{z}} + \|\mathsf{vec}^{-1}\Big(\mathbb{E}\left[\mathbf{A}\right]^{-1}\mathsf{vec}\Big(\frac{1}{t}\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)\Big)\| \\
&\quad + \|\mathbf{U}^{(\ell)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)} + \mathsf{vec}^{-1}\Big(\mathbb{E}\left[\mathbf{A}\right]^{-1}\mathsf{vec}\Big(\frac{1}{t}\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)\Big)\| \qquad (152) \\
&\leq \max_{\|\mathbf{z}\|_2=1} \sqrt{\mathbf{z}^\mathsf{T}(\mathbf{Q}^{(\ell-1)})^\mathsf{T}\mathbf{Q}^{(\ell-1)}\mathbf{z}} + \|\mathsf{vec}^{-1}\Big(\mathbb{E}\left[\mathbf{A}\right]^{-1}\mathsf{vec}\Big(\frac{1}{t}\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)\Big)\|_\mathsf{F} \\
&\quad + \|\mathbf{U}^{(\ell)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)} - \mathsf{vec}^{-1}\Big(\mathbb{E}\left[\mathbf{A}\right]^{-1}\mathsf{vec}\Big(\frac{1}{t}\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)\Big)\|_\mathsf{F} \\
&\leq \sigma_{\max}(\mathbf{Q}^{(\ell-1)}) + \|\mathbb{E}\left[\mathbf{A}\right]^{-1}\|\|\frac{1}{t}\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\|_\mathsf{F} \\
&\quad + \|\mathbf{U}^{(\ell)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)} + \mathsf{vec}^{-1}\Big(\mathbb{E}\left[\mathbf{A}\right]^{-1}\mathsf{vec}\Big(\frac{1}{t}\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)\Big)\|_\mathsf{F}.
\end{aligned}$$

Directly using the bounds of $\|\mathbb{E}\left[\mathbf{A}\right]^{-1}\|\|\frac{1}{t}\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\|_\mathsf{F}$ and $\|\mathbf{U}^{(\ell)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)} + \mathsf{vec}^{-1}\Big(\mathbb{E}\left[\mathbf{A}\right]^{-1}\mathsf{vec}\Big(\frac{1}{t}\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)\Big)\|_\mathsf{F}$ from the previous calculations as well inductive Assumption 4, we get

$$\begin{aligned}
&\max_{\|\mathbf{z}\|_2=1} \|\mathbf{U}^{+(\ell)}\mathbf{R}\mathbf{z}\|_2 \leq 1 + c'' \\
&\qquad \implies \|\mathbf{R}\| \leq \max_{\|\mathbf{z}\|_2=1} \|\mathbf{R}\mathbf{z}\|_2 = \max_{\|\mathbf{z}\|_2=1} \|\mathbf{U}^{+(\ell)}\mathbf{R}\mathbf{z}\|_2 \leq 1 + c''. \qquad (153)
\end{aligned}$$

Bounds equation 151 and equation 153 complete the proof. □

Lemma 11.

$$\begin{aligned}
&\|\mathbf{U}^{(\ell)}\|_{2,\infty} \\
&\leq \|\mathbf{U}^\star\|_{2,\infty} \|\frac{1}{t}\sum_{i\in[t]} \Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1} \mathbf{w}^{\star(i)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}\|_2 \\
&\quad + c\|\mathbf{U}^\star\|_\mathsf{F} \|\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1}\| \|\mathbf{w}^{\star(i)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}\|_\mathsf{F} \sqrt{\frac{\log(1/\delta_0)}{mt}} \\
&\quad + \zeta(\max_i \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_\infty \|\mathbf{w}^{(i,\ell)}\|_2) \|\Big((\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1}\|_2 \\
&\quad + c\|\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1}\|_\mathsf{F} \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2 \|\mathbf{w}^{(i,\ell)}\|_2 \sqrt{\frac{\log(1/\delta_0)}{mt}} \\
&\quad + \tfrac{2\sigma_2}{mt}\sqrt{\log(rd/\delta_0)} \cdot \frac{r\sqrt{r}}{\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)} + \frac{2\sigma\sqrt{\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}} \frac{r}{\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)} \\
&\quad + \frac{\sqrt{r}\Big(c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(rd/\delta_0)}{m^2t^2}} + \frac{\sigma_1}{mt}\Big(2\sqrt{rd} + 2\sqrt{2rd\log(2rd/\delta_0)}\Big)\Big)}{\frac{1}{r}\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)} \cdot \\
&\quad \Big\{\frac{2}{t}\|\mathbf{U}^\star(\mathbf{W}^\star)^\mathsf{T}\mathbf{W}^{(\ell)}\|_\mathsf{F} + \sqrt{\frac{4\zeta}{t}}(\max_i \|\mathbf{w}^{(i,\ell)}\|_2)\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2 \\
&\quad + 4\Big(\|\mathbf{U}^\star\|\|\mathbf{w}^{\star(i)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2 + \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2\Big)\sqrt{\frac{d\log(rd/\delta_0)}{mt}} \\
&\quad + \frac{\sigma_2}{mt}6\sqrt{rd\log(rd)} + \frac{2\sigma\sqrt{d\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big\}
\end{aligned}$$

and $\|\mathbf{U}^{(\ell)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_{2,\infty}$

$$\begin{aligned}
&\leq \|\mathbf{U}^\star\|_{2,\infty} \|\frac{1}{t}\sum_{i\in[t]} \Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1} \mathbf{h}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}\|_2 \\
&\quad + c\|\mathbf{U}^\star\|_\mathsf{F} \|\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1}\| \|\mathbf{h}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}\|_\mathsf{F} \sqrt{\frac{\log(1/\delta_0)}{mt}} \\
&\quad + \zeta(\max_i \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_\infty \|\mathbf{w}^{(i,\ell)}\|_2) \|\Big((\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1}\|_2 \\
&\quad + c\|\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1}\|_\mathsf{F} \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2 \|\mathbf{w}^{(i,\ell)}\|_2 \sqrt{\frac{\log(1/\delta_0)}{mt}} \\
&\quad + \frac{2\sigma_2}{mt}\sqrt{\log(rd/\delta_0)} \cdot \frac{r\sqrt{r}}{\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)} + \frac{r^2\sigma_1\sqrt{rd\cdot 2\log(2r^2d^2/\delta_0)}}{mt\lambda_r\Big(\frac{r}{t}\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)} \\
&\quad + \frac{2\sigma\sqrt{\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}} \frac{r}{\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)} \\
&\quad + \frac{\sqrt{r}\Big(c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(rd/\delta_0)}{m^2t^2}} + \frac{\sigma_1}{mt}\Big(2\sqrt{rd} + 2\sqrt{2rd\log(2rd/\delta_0)}\Big)\Big)}{\frac{1}{r}\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)} \cdot \\
&\quad \Big\{\frac{2}{t}\|\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\|_\mathsf{F} + \sqrt{\frac{4\zeta}{t}}(\max_i \|\mathbf{w}^{(i,\ell)}\|_2)\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2 \\
&\quad + 4\Big(\|\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|\|\mathbf{h}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2 + \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2\Big)\sqrt{\frac{d\log(rd/\delta_0)}{mt}} \\
&\quad + \frac{\sigma_2}{mt}6\sqrt{rd\log(rd)} + \frac{\sigma_1}{mt}\Big(2\sqrt{rd} + 4\sqrt{\log rd}\Big)\sqrt{r} + \frac{2\sigma\sqrt{d\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big\}
\end{aligned}$$

w.p. $1 - \mathcal{O}(\delta_0)$

Proof. Recall that

$$\mathsf{vec}(\mathbf{U}^{(\ell)}) = \Big(\underbrace{\mathbf{A}' + \frac{\mathbf{N}_1}{mt}}_{\mathbf{A}}\Big)^{-1}\mathsf{vec}\Big(\mathbf{V} + \frac{\mathbf{N}_2}{mt} + \mathbf{\Xi}\Big) \tag{154}$$

where

$$\begin{aligned}
\mathbf{A}'_{rd\times rd} &= \frac{1}{mt}\sum_{i\in[t]}\Big(\mathbf{w}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T} \otimes (\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big)\\
&= \frac{1}{mt}\sum_{i\in[t]}\Big(\mathbf{w}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T} \otimes \Big(\sum_{j=1}^{m}\mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}\Big)\Big),\\
\mathbf{V}'_{d\times r} &= \frac{1}{mt}\sum_{i\in[t]}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big(\mathbf{U}^\star\mathbf{w}^{\star(i)} + (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})\Big)(\mathbf{w}^{(i,\ell)})^\mathsf{T}.
\end{aligned}$$

Note that

$$\begin{aligned}
\mathbb{E}\left[\mathbf{A}\right] &= \frac{1}{t}\sum_{i\in[t]}\Big(\mathbf{w}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T} \otimes \mathbf{I}\Big) + \mathbf{0}\\
&= \frac{1}{t}\cdot(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)} \otimes \mathbf{I}.
\end{aligned} \tag{155}$$

Let $\mathbf{C} := \mathbb{E}\left[\mathbf{A}\right]^{-1}$ and $\mathbf{A}' = \mathbf{I} + \mathbf{E}$. Then, we have:

$$\begin{aligned}
\|\mathbf{C}\|_2 &= \|\mathbb{E}\left[\mathbf{A}\right]^{-1}\|_2\\
&\le \frac{1}{\lambda_{\min}\Big(\frac{1}{t}\cdot(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)} \otimes \mathbf{I}\Big)}\\
&\le \frac{1}{\frac{1}{r}\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)}.
\end{aligned} \tag{156}$$

and

$$\begin{aligned}
\mathbf{A}^{-1} &= \Big(\mathbb{E}\left[\mathbf{A}\right]\mathbb{E}\left[\mathbf{A}\right]^{-1}\mathbf{A}\Big)^{-1}\\
&= \Big(\mathbb{E}\left[\mathbf{A}\right]\Big(\mathbf{I} + \mathbb{E}\left[\mathbf{A}\right]^{-1}\mathbf{E} + \mathbb{E}\left[\mathbf{A}\right]^{-1}\frac{\mathbf{N}_1}{mt}\Big)\Big)^{-1}\\
&= \Big(\mathbf{I} + \mathbb{E}\left[\mathbf{A}\right]^{-1}\mathbf{E} + \mathbb{E}\left[\mathbf{A}\right]^{-1}\frac{\mathbf{N}_1}{mt}\Big)^{-1}\mathbb{E}\left[\mathbf{A}\right]^{-1}\\
&= \Big(\mathbf{I} + \mathbf{C}\mathbf{E} + \frac{\mathbf{C}\mathbf{N}_1}{mt}\Big)^{-1}\mathbf{C}\\
&= \sum_{p=0}^{\infty}(-1)^p\Big(\mathbf{C}\mathbf{E} + \frac{\mathbf{C}\mathbf{N}_1}{mt}\Big)^p\mathbf{C}\\
&= \mathbf{C} + \sum_{p=1}^{\infty}(-1)^p\Big(\mathbf{C}\mathbf{E} + \frac{\mathbf{C}\mathbf{N}_1}{mt}\Big)^p\mathbf{C}
\end{aligned}$$

since $\|\mathbf{C}\mathbf{E}\| < 1$. Now, let $\mathcal{Z} \triangleq \{\mathbf{z} \in \mathbb{R}^{rd} | \|\mathbf{z}\|_2 = 1\}$. Then for $\epsilon \le 1$, there exists an ϵ-net, $N_\epsilon \subset \mathcal{Z}$, of size $(1 + 2/\epsilon)^{rd}$ w.r.t the Euclidean norm, i.e. $\forall\ \mathbf{z} \in \mathcal{Z}$, $\exists\ \mathbf{z}' \in N_\epsilon$ s.t. $\|\mathbf{z} - \mathbf{z}'\|_2 \le \epsilon$. We will now bound $\Big\|\Big(\mathbf{C}\mathbf{E} + \frac{\mathbf{C}\mathbf{N}_1}{mt}\Big)\mathbf{z}\Big\|_\infty$ $\forall\ \mathbf{z} \in \mathbf{Z}$. Now consider any $\mathbf{z}^\mathsf{T} = \left[\mathbf{z}_1^\mathsf{T}, \mathbf{z}_2^\mathsf{T}, \dots, \mathbf{z}_r^\mathsf{T}\right] \in N_\epsilon$ where each $\mathbf{z}_i \in \mathbb{R}^d$. Then for s-th standard basis vector $\mathbf{e}_s \in \mathbb{R}^{rd}$, we have using Lemma 9

$$\begin{aligned}
|\mathbf{e}_s^{\mathsf{T}}\mathbf{C}\mathbf{E}\mathbf{z}| &\le c\|\mathbf{C}^{\mathsf{T}}\mathbf{e}_s\|_2\|\mathbf{z}\|_2\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4\frac{\log(1/\delta_0)}{m^2t^2}} \\
&\le c\|\mathbf{C}\|_2\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4\frac{\log(1/\delta_0)}{m^2t^2}} \qquad (157) \\
\text{and } \Big|\mathbf{e}_s^{\mathsf{T}}\frac{\mathbf{C}\mathbf{N}_1}{mt}\mathbf{z}\Big| &\le \|\mathbf{C}^{\mathsf{T}}\mathbf{e}_s\|_2\|\frac{\mathbf{N}_1}{mt}\|_2\|\mathbf{z}\|_2 \\
&\le \|\mathbf{C}\|_2\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+2\sqrt{2\log(1/\delta_0)}\Big).
\end{aligned}$$

$$\begin{aligned}
&\implies \Big\|\mathbf{C}\Big(\mathbf{E}+\frac{\mathbf{N}_1}{mt}\Big)\mathbf{z}\Big\|_\infty \\
&\le \|\mathbf{C}\|_2\Big(c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{\log(|N_\epsilon|/\delta_0)}{m^2t^2}}+\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+\sqrt{2\log(2|N_\epsilon|/\delta_0)}\Big)\Big) \\
&\le \|\mathbf{C}\|_2\Big(c\sqrt{\textstyle\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\tfrac{\log((1+2/\epsilon)^{rd}/\delta_0)}{m^2t^2}}+\tfrac{\sigma_1}{mt}\Big(2\sqrt{rd}+\sqrt{2\log(2(1+2/\epsilon)^{rd}/\delta_0)}\Big)\Big).
\end{aligned}$$

Further, $\exists\ \mathbf{z}\in N_\epsilon$ s.t. $\|\mathbf{z}'-\mathbf{z}\|_2\le\epsilon$. This implies that setting $\epsilon\leftarrow 1/4$ and $c\leftarrow 2c\sqrt{\log(9)}$ gives $\Big\|\mathbf{C}\Big(\mathbf{E}+\frac{\mathbf{N}_1}{mt}\Big)\mathbf{z}'\Big\|_\infty$

$$\begin{aligned}
&\le \Big\|\mathbf{C}\Big(\mathbf{E}+\frac{\mathbf{N}_1}{mt}\Big)(\mathbf{z}-\mathbf{z}')\Big\|_\infty+\Big\|\mathbf{C}\Big(\mathbf{E}+\frac{\mathbf{N}_1}{mt}\Big)\mathbf{z}\Big\|_\infty \\
&\le \Big\|\mathbf{C}\Big(\mathbf{E}+\frac{\mathbf{N}_1}{mt}\Big)(\mathbf{z}-\mathbf{z}')\Big\|_2+\Big\|\mathbf{C}\Big(\mathbf{E}+\frac{\mathbf{N}_1}{mt}\Big)\mathbf{z}\Big\|_\infty \\
&\le \Big\|\mathbf{C}\Big(\mathbf{E}+\frac{\mathbf{N}_1}{mt}\Big)\Big\|_2\epsilon+\Big\|\mathbf{C}\Big(\mathbf{E}+\frac{\mathbf{N}_1}{mt}\Big)\mathbf{z}\Big\|_\infty \\
&\le \|\mathbf{C}\|\cdot\frac{1}{4}\cdot\Big(c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(rd/\delta_0)}{m^2t^2}}+\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{rd\log(rd/\delta_0)}\Big)\Big) \\
&\quad+\|\mathbf{C}\|_2\Big(c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(rd/\delta_0)}{m^2t^2}}+\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{rd\log(rd/\delta_0)}\Big)\Big) \\
&\le \|\mathbf{C}\|_2\Big(c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(rd/\delta_0)}{m^2t^2}}+\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{rd\log(rd/\delta_0)}\Big)\Big). \qquad (158)
\end{aligned}$$

with probability at least $1-\delta_0$ where we use equation 112, equation 113 and the fact that $\|\mathbf{M}\mathbf{N}\|_2\le\|\mathbf{M}\|_2\|\mathbf{N}\|_2$. Hence, with probability at least $1-O(\delta_0)$, we have $\|\mathbf{C}\mathbf{E}\|_2$ and $\|\mathbf{C}\mathbf{E}\mathbf{z}\|_\infty$ for all $\mathbf{z}\in\mathcal{Z}$. Therefore, let us condition on these events in order to prove the next steps. We will now show an upper bound on $\Big\|\mathbf{A}^{-1}\mathsf{vec}\Big(\mathbf{V}'+\frac{\mathbf{N}_2}{mt}\Big)\Big\|_\infty$. Note that $\Big\|\mathbf{A}^{-1}\mathsf{vec}\Big(\mathbf{V}'+\frac{\mathbf{N}_2}{mt}+\mathbf{\Xi}\Big)\Big\|_\infty$

$$\begin{aligned}
&= \Big\|\mathbf{C}\mathsf{vec}\Big(\mathbf{V}'+\frac{\mathbf{N}_2}{mt}+\mathbf{\Xi}\Big)+\sum_{p=1}^{\infty}(-1)^p\Big(\mathbf{C}\mathbf{E}+\frac{\mathbf{C}\mathbf{N}_1}{mt}\Big)^p\mathbf{C}\mathsf{vec}\Big(\mathbf{V}'+\frac{\mathbf{N}_2}{mt}+\mathbf{\Xi}\Big)\Big\|_\infty \\
&\le \Big\|\mathbf{C}\mathsf{vec}\Big(\mathbf{V}'+\frac{\mathbf{N}_2}{mt}+\mathbf{\Xi}\Big)\Big\|_\infty+\sum_{p=1}^{\infty}\Big\|\Big(\mathbf{C}\mathbf{E}+\frac{\mathbf{C}\mathbf{N}_1}{mt}\Big)^p\mathbf{C}\mathsf{vec}\Big(\mathbf{V}'+\frac{\mathbf{N}_2}{mt}+\mathbf{\Xi}\Big)\Big\|_\infty. \qquad (159)
\end{aligned}$$

We have with probability at least $1-\delta_0$, $\sum_{p=1}^{\infty}\Big\|\Big(\mathbf{CE}+\frac{\mathbf{CN}_1}{mt}\Big)^p\mathsf{vec}\Big(\mathbf{V}'+\frac{\mathbf{N}_2}{mt}+\mathbf{\Xi}\Big)\Big\|_\infty$

$$
\begin{aligned}
&=\textstyle\sum_{p=1}^{\infty}\Bigg\|\Big(\mathbf{CE}+\frac{\mathbf{CN}_1}{mt}\Big)\cdot\Big\|\Big(\mathbf{CE}+\frac{\mathbf{CN}_1}{mt}\Big)^{p-1}\mathsf{vec}\Big(\mathbf{V}'+\frac{\mathbf{N}_2}{mt}+\mathbf{\Xi}\Big)\Big\|_2\cdot\frac{\Big(\mathbf{CE}+\frac{\mathbf{CN}_1}{mt}\Big)^{p-1}\mathsf{vec}\Big(\mathbf{V}'+\frac{\mathbf{N}_2}{mt}+\mathbf{\Xi}\Big)}{\Big\|\Big(\mathbf{CE}+\frac{\mathbf{CN}_1}{mt}\Big)^{p-1}\mathsf{vec}\Big(\mathbf{V}'+\frac{\mathbf{N}_2}{mt}+\mathbf{\Xi}\Big)\Big\|_2}\Bigg\|_\infty\\
&=\textstyle\sum_{p=1}^{\infty}\Big\|\Big(\mathbf{CE}+\frac{\mathbf{CN}_1}{mt}\Big)^{p-1}\mathsf{vec}\Big(\mathbf{V}'+\frac{\mathbf{N}_2}{mt}+\mathbf{\Xi}\Big)\Big\|_2\Bigg\|\Big(\mathbf{CE}+\frac{\mathbf{CN}_1}{mt}\Big)\cdot\frac{\Big(\mathbf{CE}+\frac{\mathbf{CN}_1}{mt}\Big)^{p-1}\mathsf{vec}\Big(\mathbf{V}'+\frac{\mathbf{N}_2}{mt}+\mathbf{\Xi}\Big)}{\Big\|\Big(\mathbf{CE}+\frac{\mathbf{CN}_1}{mt}\Big)^{p-1}\mathsf{vec}\Big(\mathbf{V}'+\frac{\mathbf{N}_2}{mt}+\mathbf{\Xi}\Big)\Big\|_2}\Bigg\|_\infty\\
&\le\sum_{p=1}^{\infty}\Big\|\mathbf{CE}+\frac{\mathbf{CN}_1}{mt}\Big\|^{p-1}\Big\|\mathsf{vec}\Big(\mathbf{V}'+\frac{\mathbf{N}_2}{mt}+\mathbf{\Xi}\Big)\Big\|_2\cdot\\
&\qquad\|\mathbf{C}\|_2\Big(c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(rd/\delta_0)}{m^2t^2}}+\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{rd\log(rd/\delta_0)}\Big)\Big) \qquad (160)\\
&\le\sum_{p=1}^{\infty}\Bigg(\frac{c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(rd/\delta_0)}{m^2t^2}}+\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)}{\frac{1}{r}\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\top\mathbf{W}^{(\ell)}\Big)}\Bigg)^{p-1}\\
&\qquad\|\mathbf{C}\|_2\Big(c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(1/\delta_0)}{m^2t^2}}\\
&\qquad+\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+2\sqrt{2rd\log(2/\delta_0)}\Big)\Big)\Big\|\mathsf{vec}\Big(\mathbf{V}'+\frac{\mathbf{N}_2}{mt}+\mathbf{\Xi}\Big)\Big\|_2\\
&\le\sum_{p=1}^{\infty}\Big(c'\sqrt{\frac{r^3d\log(rd/\delta_0)}{mt}}+c''\frac{\sigma_1}{mt\lambda_r}\sqrt{r^3d\log(rd/\delta_0)}\Big)^{p-1}\cdot\\
&\qquad\|\mathbf{C}\|_2\Big(c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(1/\delta_0)}{m^2t^2}}\\
&\qquad+\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+2\sqrt{2rd\log(2/\delta_0)}\Big)\Big)\Big\|\mathsf{vec}\Big(\mathbf{V}'+\frac{\mathbf{N}_2}{mt}+\mathbf{\Xi}\Big)\Big\|_2, \qquad (161)
\end{aligned}
$$

where in equation 160 we use equation 158 and in equation 161 the fact that $mt=\Omega(\max\{r^3d\log(rd/\delta_0),\frac{\sigma_1}{\lambda_r}\sqrt{r^3d\log(rd/\delta_0)}\})$. By taking a union bound we must have with probability at least $1-\delta_0$, $\sum_{p=1}^{\infty}\Big\|\Big(\mathbf{CE}+\frac{\mathbf{CN}_1}{mt}\Big)^p\mathsf{vec}\Big(\mathbf{V}'+\frac{\mathbf{N}_2}{mt}+\mathbf{\Xi}\Big)\Big\|_\infty$

$$
\begin{aligned}
&=\mathcal{O}\Big(\|\mathbf{C}\|_2\Big(c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(1/\delta_0)}{m^2t^2}}+\frac{\sigma_1}{mt}\Big(2\sqrt{rd}\\
&\qquad+2\sqrt{2rd\log(2/\delta_0)}\Big)\Big)\Big\|\mathsf{vec}\Big(\mathbf{V}'+\frac{\mathbf{N}_2}{mt}+\mathbf{\Xi}\Big)\Big\|_2\Big). \qquad (162)
\end{aligned}
$$

Using the above and equation 156 in equation 159, we have $\Big\|\mathbf{A}^{-1}\mathsf{vec}\Big(\mathbf{V}'+\frac{\mathbf{N}_2}{mt}\Big)\Big\|_\infty$

$$
\begin{aligned}
&\le\|\mathbf{C}\mathsf{vec}\Big(\mathbf{V}'+\frac{\mathbf{N}_2}{mt}\mathbf{\Xi}\Big)\|_\infty+\|\mathbf{C}\|_2\Big(c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(1/\delta_0)}{m^2t^2}}\\
&\qquad+\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+2\sqrt{2rd\log(2/\delta_0)}\Big)\Big)\Big\|\mathsf{vec}\Big(\mathbf{V}'+\frac{\mathbf{N}_2}{mt}+\mathbf{\Xi}\Big)\Big\|_2
\end{aligned}
$$

$$\begin{aligned}
&\le \left\|\mathbf{C}\mathsf{vec}\Big(\mathbf{V}' + \frac{\mathbf{N}_2}{mt} + \mathbf{\Xi}\Big)\right\|_\infty \\
&\quad + \frac{\left(c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(rd/\delta_0)}{m^2t^2}} + \frac{\sigma_1}{mt}\left(2\sqrt{rd}+2\sqrt{2rd\log(2rd/\delta_0)}\right)\right)}{\frac{1}{r}\lambda_r\left(\frac{r}{t}(\mathbf{W}^{(\ell)})^\top\mathbf{W}^{(\ell)}\right)} \cdot \\
&\quad \left\|\mathbf{V}' + \frac{\mathbf{N}_2}{mt} + \mathbf{\Xi}\right\|_{\mathsf{F}}.
\end{aligned} \tag{163}$$

Similarly, we also have the following bound $\|\mathbf{U}^{(\ell)}\|_{2,\infty} = \left\|\mathsf{vec}^{-1}\Big(\mathbf{A}^{-1}\mathsf{vec}\Big(\mathbf{V}' + \frac{\mathbf{N}_2}{mt} + \mathbf{\Xi}\Big)\Big)\right\|_{2,\infty}$

$$\begin{aligned}
&\le \left\|\mathsf{vec}^{-1}\Big(\mathbf{C}\mathsf{vec}\Big(\mathbf{V}' + \frac{\mathbf{N}_2}{mt} + \mathbf{\Xi}\Big)\Big)\right\|_{2,\infty} \\
&\quad + \frac{\sqrt{r}\left(c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(rd/\delta_0)}{m^2t^2}} + \frac{\sigma_1}{mt}\left(2\sqrt{rd}+2\sqrt{2rd\log(2rd/\delta_0)}\right)\right)}{\frac{1}{r}\lambda_r\left(\frac{r}{t}(\mathbf{W}^{(\ell)})^\top\mathbf{W}^{(\ell)}\right)} \cdot \\
&\quad \left\|\mathbf{V}' + \frac{\mathbf{N}_2}{mt} + \mathbf{\Xi}\right\|_{\mathsf{F}}.
\end{aligned} \tag{164}$$

Now $\mathbf{C}\mathsf{vec}(\mathbf{V}')$

$$\begin{aligned}
&= \Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\top\mathbf{W}^{(\ell)} \otimes \mathbf{I}\Big)^{-1}\mathsf{vec}\Big(\frac{1}{mt}\sum_{i\in[t]}(\mathbf{X}^{(i)})^\top\mathbf{X}^{(i)}\Big(\mathbf{U}^\star\mathbf{w}^{\star(i)} + (\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})\Big)(\mathbf{w}^{(i,\ell)})^\top\Big) \\
&= \Big(\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\top\mathbf{W}^{(\ell)}\Big)^{-1} \otimes \mathbf{I}\Big)\cdot \\
&\quad \frac{1}{mt}\sum_{i\in[t]}\Big(\mathsf{vec}\Big((\mathbf{X}^{(i)})^\top\mathbf{X}^{(i)}\mathbf{U}^\star\mathbf{w}^{\star(i)}(\mathbf{w}^{(i,\ell)})^\top\Big) + \mathsf{vec}\Big((\mathbf{X}^{(i)})^\top\mathbf{X}^{(i)}(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\top\Big)\Big) \\
&= \Big(\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\top\mathbf{W}^{(\ell)}\Big)^{-1} \otimes \mathbf{I}\Big)\cdot \\
&\quad \frac{1}{mt}\sum_{i\in[t]}\Big(\Big(\mathbf{w}^{\star(i)}(\mathbf{w}^{(i,\ell)})^\top \otimes (\mathbf{X}^{(i)})^\top\mathbf{X}^{(i)}\Big)\mathsf{vec}(\mathbf{U}^\star) \\
&\qquad\quad + \mathsf{vec}\Big((\mathbf{X}^{(i)})^\top\mathbf{X}^{(i)}(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\top\Big)\Big) \\
&= \frac{1}{mt}\sum_{i\in[t]}\Big(\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\top\mathbf{W}^{(\ell)}\Big)^{-1} \otimes \mathbf{I}\Big)\cdot\Big(\mathbf{w}^{\star(i)}(\mathbf{w}^{(i,\ell)})^\top \otimes (\mathbf{X}^{(i)})^\top\mathbf{X}^{(i)}\Big)\mathsf{vec}(\mathbf{U}^\star) \\
&\quad + \frac{1}{mt}\sum_{i\in[t]}\Big(\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\top\mathbf{W}^{(\ell)}\Big)^{-1} \otimes \mathbf{I}\Big)\mathsf{vec}\Big((\mathbf{X}^{(i)})^\top\mathbf{X}^{(i)}(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\top\Big) \\
&= \underbrace{\frac{1}{mt}\sum_{i\in[t]}\Big(\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\top\mathbf{W}^{(\ell)}\Big)^{-1}\mathbf{w}^{\star(i)}(\mathbf{w}^{(i,\ell)})^\top \otimes (\mathbf{X}^{(i)})^\top\mathbf{X}^{(i)}\Big)\mathsf{vec}(\mathbf{U}^\star)}_{\mathsf{vec}(\mathbf{V}_1')} \\
&\quad + \underbrace{\frac{1}{mt}\sum_{i\in[t]}\Big(\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\top\mathbf{W}^{(\ell)}\Big)^{-1} \otimes \mathbf{I}\Big)\mathsf{vec}\Big((\mathbf{X}^{(i)})^\top\mathbf{X}^{(i)}(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\top\Big)}_{\mathsf{vec}(\mathbf{V}_2')}.
\end{aligned} \tag{165}$$

Analysis for $\mathsf{vec}(\mathbf{V}_1')$:

Let $\mathbf{J}^{(i,\ell)} := \left(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\right)^{-1}\mathbf{w}^{\star(i)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}$. Then we can rewrite $\mathbf{V}_1'$ as,

$$
\begin{aligned}
\mathsf{vec}(\mathbf{V}_1') &= \frac{1}{mt}\sum_{i\in[t]}\left(\mathbf{J}^{(i,\ell)} \otimes (\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\right)\mathsf{vec}(\mathbf{U}^\star) && (166)\\
\iff \mathbf{V}_1' &= \frac{1}{mt}\sum_{i\in[t]}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\mathbf{U}^\star(\mathbf{J}^{(i,\ell)})^\mathsf{T} && (167)\\
\implies \mathbb{E}\left[\mathsf{vec}(\mathbf{V}_1')\right] &= \Big(\frac{1}{t}\sum_{i\in[t]}\mathbf{J}^{(i,\ell)} \otimes \mathbf{I}\Big)\mathsf{vec}(\mathbf{U}^\star), && (168)\\
\iff \mathbb{E}\left[\mathbf{V}_1'\right] &= \mathbf{U}^\star\Big(\frac{1}{t}\sum_{i\in[t]}\mathbf{J}^{(i,\ell)}\Big)^\mathsf{T} && (169)\\
\implies \|\mathbb{E}\left[\mathbf{V}_1'\right]\|_{2,\infty} &= \|\mathbf{U}^\star\Big(\frac{1}{t}\sum_{i\in[t]}\mathbf{J}^{(i,\ell)}\Big)^\mathsf{T}\|_{2,\infty}\\
&= \max_{p\in[d]}\|\mathbf{U}_p^\star\|_2\|\frac{1}{t}\sum_{i\in[t]}\mathbf{J}^{(i,\ell)}\|_2\\
&\le \|\mathbf{U}^\star\|_{2,\infty}\|\frac{1}{t}\sum_{i\in[t]}\mathbf{J}^{(i,\ell)}\|_2 && (170)
\end{aligned}
$$

Now consider the s-th standard basis vector $\mathbf{e}_s \in \mathbb{R}^{rd}$ s.t. s falls under the $q^\star$-th fragment ($q \in [r]$), i.e. $\forall\ \mathbf{a} \in \mathbb{R}^{rd}$, if we write denote $\mathbf{a}^\mathsf{T} = \left[\mathbf{a}_1^\mathsf{T}, \mathbf{a}_2^\mathsf{T}, \dots, \mathbf{a}_r^\mathsf{T}\right]$ where each $\mathbf{a}_i \in \mathbb{R}^d$, then $\mathbf{e}_s^\mathsf{T}\mathbf{a} = \left[\mathbf{0}_1^\mathsf{T}\dots, \underbrace{[0\dots, 1_s, \dots 0]}_{q^\star}, \dots \mathbf{0}_r^\mathsf{T}\right]\begin{bmatrix}\mathbf{a}_1\\ \dots\\ \mathbf{a}_{q^\star}\\ \dots\\ \mathbf{a}_r\end{bmatrix} = [0\dots, 1_s, \dots 0]\,\mathbf{a}_{q^\star} = \mathsf{vec}(\mathbf{a})_s$. Then following along similar lines of Lemma 9, we have

$$
\begin{aligned}
\mathbf{e}_s^\mathsf{T}\mathsf{vec}(\mathbf{V}_1') &= \mathbf{e}_s^\mathsf{T}\frac{1}{mt}\sum_{i\in[t]}\left(\mathbf{J}^{(i,\ell)} \otimes (\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\right)\mathsf{vec}(\mathbf{U}^\star)\\
&= \frac{1}{mt}\sum_{i\in[t]}\sum_{j\in[m]}\sum_{p\in[r]}\sum_{q\in[r]}\mathbf{e}_p^\mathsf{T}\left(J_{p,q}^{(i,\ell)}\mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}\right)\mathsf{vec}(\mathbf{U}^\star)_q\\
&= \frac{1}{mt}\sum_{i\in[t]}\sum_{j\in[m]}(\mathbf{x}_j^{(i)})^\mathsf{T}\Big(\sum_{p\in[r]}\sum_{q\in[r]}J_{p,q}^{(i,\ell)}\mathsf{vec}(\mathbf{U}^\star)_p\mathbf{e}_q^\mathsf{T}\Big)\mathbf{x}_j^{(i)} && (171)
\end{aligned}
$$

and

$$
\begin{aligned}
\|\sum_{p\in[r]}\sum_{q\in[r]}J_{p,q}^{(i,\ell)}\mathsf{vec}(\mathbf{U}^\star)_p\mathbf{e}_q^\mathsf{T}\|_\mathsf{F} &= \|\sum_{p\in[r]}J_{p,q^\star}^{(i,\ell)}\mathsf{vec}(\mathbf{U}^\star)_p\mathbf{e}_{q^\star}\|_\mathsf{F}\\
&= \|\sum_{p\in[r]}J_{p,q^\star}^{(i,\ell)}\mathsf{vec}(\mathbf{U}^\star)_p\|_2\|\mathbf{e}_{q^\star}^\mathsf{T}\|_2\\
&\le \|\sum_{p\in[r]}|J_{p,q^\star}^{(i,\ell)}|\mathsf{vec}(\mathbf{U}^\star)_p\|_2\\
&\le \sqrt{\Big(\sum_{p\in[r]}(J_{p,q^\star}^{(i,\ell)})^2\Big)\Big(\sum_{p\in[r]}\|\mathsf{vec}(\mathbf{U}^\star)_p\|_2^2\Big)}\\
&= \|\mathbf{J}^{(i,\ell,q^\star)}\|_2\|\mathsf{vec}(\mathbf{U}^\star)\|_2\\
&= \|\mathbf{J}^{(i,\ell,q^\star)}\|_2\|\mathbf{U}^\star\|_\mathsf{F}. && (172)
\end{aligned}
$$

Thus, using equation 171 and equation 172 in Lemma 16 we have

$$\left|\mathbf{e}_s^{\mathsf{T}}\Big(\mathsf{vec}(\mathbf{V}_1') - \mathbb{E}\left[\mathsf{vec}(\mathbf{V}_1')\right]\Big)\right| \leq c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{J}^{(i,\ell,q^\star)}\|_2^2\|\mathbf{U}^\star\|_{\mathsf{F}}^2\frac{\log(1/\delta_0)}{m^2t^2}}. \tag{173}$$

Now, note that

$$\begin{aligned}
\|\mathbf{V}_1' - \mathbb{E}\left[\mathbf{V}_1'\right]\|_{2,\infty} &= \max_{p\in[d]}\|(\mathbf{V}_1')_p - \mathbb{E}\left[\mathbf{V}_1'\right]_p\|_2 \\
&= \max_{p\in[d]}\sqrt{\sum_{q\in[r]}|(V_1)_{p,q} - \mathbb{E}\left[V_1\right]_{p,q}|_2^2} \\
&= \max_{p\in[d]}\sqrt{\sum_{s=\{(q-1)d+p:q\in[r]\}}|\mathbf{e}_s^{\mathsf{T}}(\mathsf{vec}(\mathbf{V}_1') - \mathbb{E}\left[\mathsf{vec}(\mathbf{V}_\mathbf{1}')\right]|^2} \\
&\leq \max_{p\in[d]}\sqrt{\sum_{s=\{(q-1)d+p:q\in[r]\}}\Big(c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{J}^{(i,\ell,q^\star)}\|_2^2\|\mathbf{U}^\star\|_{\mathsf{F}}^2\frac{\log(1/\delta_0)}{m^2t^2}}\Big)^2},
\end{aligned}$$

where in the last step we use equation 173. Now note that as per the notation discussed above, $s = (q-1)d + p$ lies in the q-th (= $q^\star$) segment. Since $q \in [r]$, therefore summation over s is equivalent to summation over $q^\star \in [r]$. Using this fact, the above becomes:

$$\begin{aligned}
\|\mathbf{V}_1' - \mathbb{E}\left[\mathbf{V}_1'\right]\|_{2,\infty} &\leq \max_{p\in[d]}\sqrt{\sum_{q^\star\in[r]}\Big(c^2\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{J}^{(i,\ell,q^\star)}\|_2^2\|\mathbf{U}^\star\|_{\mathsf{F}}^2\frac{\log(1/\delta_0)}{m^2t^2}\Big)} \\
&\leq \max_{p\in[d]} c\|\mathbf{U}^\star\|_{\mathsf{F}}\sqrt{\sum_{q^\star\in[r]}\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{J}^{(i,\ell,q^\star)}\|_2^2}\sqrt{\frac{\log(1/\delta_0)}{m^2t^2}} \\
&\leq c\|\mathbf{U}^\star\|_{\mathsf{F}}\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{J}^{(i,\ell)}\|_{\mathsf{F}}^2}\sqrt{\frac{\log(1/\delta_0)}{m^2t^2}}. \qquad (174)
\end{aligned}$$

Therefore, using equation 170 in the above, we have

$$\begin{aligned}
\|\mathbf{V}_1'\|_{2,\infty} &\leq \|\mathbb{E}\left[\mathbf{V}_1'\right]\|_{2,\infty} + c\|\mathbf{U}^\star\|_{\mathsf{F}}\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{J}^{(i,\ell)}\|_{\mathsf{F}}^2}\sqrt{\frac{\log(1/\delta_0)}{m^2t^2}} \\
&\leq \|\mathbf{U}^\star\|_{2,\infty}\|\frac{1}{t}\sum_{i\in[t]}\mathbf{J}^{(i,\ell)}\|_2 + c\|\mathbf{U}^\star\|_{\mathsf{F}}\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{J}^{(i,\ell)}\|_{\mathsf{F}}^2}\sqrt{\frac{\log(1/\delta_0)}{m^2t^2}} \\
&\leq \|\mathbf{U}^\star\|_{2,\infty}\|\frac{1}{t}\sum_{i\in[t]}\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\Big)^{-1}\mathbf{w}^{\star(i)}(\mathbf{w}^{(i,\ell)})^{\mathsf{T}}\|_2 \\
&\quad + c\|\mathbf{U}^\star\|_{\mathsf{F}}\|\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\Big)^{-1}\|\|\mathbf{w}^{\star(i)}(\mathbf{w}^{(i,\ell)})^{\mathsf{T}}\|_{\mathsf{F}}\sqrt{\frac{\log(1/\delta_0)}{mt}}. \qquad (175)
\end{aligned}$$

Analysis for $\mathsf{vec}(\mathbf{V}_2')$:

Let $\mathbf{L}^{(i,\ell)} := \left(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\right)^{-1}$. Then,

$$
\begin{aligned}
\mathsf{vec}(\mathbf{V}_2') &= \frac{1}{mt}\sum_{i\in[t]}\Big(\mathbf{L}^{(i,\ell)}\otimes\mathbf{I}\Big)\mathsf{vec}\Big((\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}(\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big) && (176)\\
&= \frac{1}{mt}\sum_{i\in[t]}\Big(\mathbf{L}^{(i,\ell)}\otimes\mathbf{I}_{d\times d}\Big)\Big(\mathbf{I}_{r\times r}\otimes(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big)\mathsf{vec}\Big((\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big)\\
&= \frac{1}{mt}\sum_{i\in[t]}\Big(\mathbf{L}^{(i,\ell)}\otimes(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big)\mathsf{vec}\Big((\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big) && (177)\\
\iff \mathbf{V}_2' &= \frac{1}{mt}\sum_{i\in[t]}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}(\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big(\mathbf{L}^{(i,\ell)}\Big)^\mathsf{T}, && (178)\\
\implies \mathbb{E}\left[\mathsf{vec}(\mathbf{V}_2')\right] &= \frac{1}{t}\sum_{i\in[t]}\Big(\mathbf{L}^{(i,\ell)}\otimes\mathbf{I}\Big)\mathsf{vec}\Big((\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big)\\
\iff \mathbb{E}\left[\mathbf{V}_2'\right] &= \frac{1}{t}\sum_{i\in[t]}\Big((\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big)\Big(\mathbf{L}^{(i,\ell)}\Big)^\mathsf{T} && (179)\\
\implies \|\mathbb{E}\left[\mathbf{V}_2'\right]\|_{2,\infty} &= \|\frac{1}{t}\sum_{i\in[t]}\Big((\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big)\Big(\mathbf{L}^{(i,\ell)}\Big)^\mathsf{T}\|_{2,\infty}\\
&\le \frac{1}{t}\sum_{i\in[t]}\max_{p\in[d]}\|\Big((\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big)_p\|_2\|\mathbf{L}^{(i,\ell)}\|_2\\
&\le \frac{1}{t}\sum_{i\in[t]}\max_{p\in[d]}|\mathbf{b}_p^{\star(i)}-\mathbf{b}_p^{(i,\ell)}|\|\mathbf{w}^{(i,\ell)}\|_2\|\mathbf{L}^{(i,\ell)}\|_2\\
&\le \zeta(\max_i\|\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)}\|_\infty\|\mathbf{w}^{(i,\ell)}\|_2)\|\Big((\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1}\|_2 && (180)
\end{aligned}
$$

Further, we have

$$
\begin{aligned}
\mathbf{e}_s^\mathsf{T}\mathsf{vec}(\mathbf{V}_2') &= \mathbf{e}_s^\mathsf{T}\frac{1}{mt}\sum_{i\in[t]}\Big(\mathbf{L}^{(i,\ell)}\otimes(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big)\mathsf{vec}\Big((\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big)\\
&= \frac{1}{mt}\sum_{i\in[t]}\sum_{j\in[m]}\sum_{p\in[r]}\sum_{q\in[r]}\mathbf{e}_p^\mathsf{T}\Big(L_{p,q}^{(i,\ell)}\mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}\Big)\mathsf{vec}\Big((\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big)_q\\
&= \frac{1}{mt}\sum_{i\in[t]}\sum_{j\in[m]}(\mathbf{x}_j^{(i)})^\mathsf{T}\Big(\sum_{p\in[r]}\sum_{q\in[r]}L_{p,q}^{(i,\ell)}\mathsf{vec}\Big((\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big)_p\mathbf{e}_q^\mathsf{T}\Big)\mathbf{x}_j^{(i)} && (181)
\end{aligned}
$$

and $\|\sum_{p\in[r]}\sum_{q\in[r]}L_{p,q}^{(i,\ell)}\mathsf{vec}\Big((\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big)_p\mathbf{e}_q^\mathsf{T}\|_\mathsf{F}$

$$
\begin{aligned}
&= \|\sum_{p\in[r]}L_{p,q^\star}^{(i,\ell)}\mathsf{vec}\Big((\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big)_p\mathbf{e}_{q^\star}\|_\mathsf{F}\\
&= \|\sum_{p\in[r]}L_{p,q^\star}^{(i,\ell)}\mathsf{vec}\Big((\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big)_p\|_2\|\mathbf{e}_{q^\star}^\mathsf{T}\|_2\\
&\le \|\sum_{p\in[r]}|L_{p,q^\star}^{(i,\ell)}|\mathsf{vec}\Big((\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big)_p\|_2\\
&\le \sqrt{\Big(\sum_{p\in[r]}(L_{p,q^\star}^{(i,\ell)})^2\Big)\Big(\sum_{p\in[r]}\|\mathsf{vec}\Big((\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big)_p\|_2^2\Big)}\\
&= \|\mathbf{L}^{(i,\ell,q^\star)}\|_2\|\mathsf{vec}\Big((\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big)\|_2\\
&= \|\mathbf{L}^{(i,\ell,q^\star)}\|_2\|(\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\|_\mathsf{F} && (182)
\end{aligned}
$$

Thus, using equation 181 and equation 182 in Lemma 16 we have

$$\left|\mathbf{e}_s^\mathsf{T}\left(\mathsf{vec}(\mathbf{V}_2') - \mathbb{E}\left[\mathsf{vec}(\mathbf{V}_2')\right]\right)\right| \le c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]} \|\mathbf{L}^{(i,\ell,q^\star)}\|_2^2 \|(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\|_\mathsf{F}^2 \frac{\log(1/\delta_0)}{m^2t^2}} \tag{183}$$

Now, note that $\|\mathbf{V}_2' - \mathbb{E}\left[\mathbf{V}_2'\right]\|_{2,\infty}$

$$\begin{aligned}
&= \max_{p\in[d]} \|(\mathbf{V}_1')_p - \mathbb{E}\left[\mathbf{V}_1'\right]_p\|_2 \\
&= \max_{p\in[d]} \sqrt{\sum_{q\in[r]} |(V_1)_{p,q} - \mathbb{E}\left[V_1\right]_{p,q}|_2^2} \\
&= \max_{p\in[d]} \sqrt{\sum_{s=\{(q-1)d+p:q\in[r]\}} |\mathbf{e}_s^\mathsf{T}(\mathsf{vec}(\mathbf{V}_1') - \mathbb{E}\left[\mathsf{vec}(\mathbf{V_1'}\right]|^2} \\
&\le \max_{p\in[d]} \sqrt{\sum_{s=\{(q-1)d+p:q\in[r]\}} \left(c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]} \|\mathbf{L}^{(i,\ell,q^\star)}\|_2^2 \|(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\|_\mathsf{F}^2 \frac{\log(1/\delta_0)}{m^2t^2}}\right)^2}
\end{aligned}$$

Now note that as per the notation discussed above, $s = (q-1)d + p$ lies in the q-th (= $q^\star$) segment. Since $q \in [r]$, therefore summation over s is equivalent to summation over $q^\star \in [r]$. Using this fact the above becomes, $\|\mathbf{V}_2' - \mathbb{E}\left[\mathbf{V}_2'\right]\|_{2,\infty}$

$$\begin{aligned}
&\le \max_{p\in[d]} \sqrt{\sum_{q^\star\in[r]} \left(c^2 \sum_{i\in[t]}\sum_{j\in[m]} \|\mathbf{L}^{(i,\ell,q^\star)}\|_2^2 \|(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\|_\mathsf{F}^2 \frac{\log(1/\delta_0)}{m^2t^2}\right)} \\
&\le \max_{p\in[d]} c \sqrt{\sum_{q^\star\in[r]}\sum_{i\in[t]}\sum_{j\in[m]} \|\mathbf{L}^{(i,\ell,q^\star)}\|_2^2 \|(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\|_\mathsf{F}^2} \sqrt{\frac{\log(1/\delta_0)}{m^2t^2}} \\
&\le \max_{p\in[d]} c \sqrt{\sum_{i\in[t]}\sum_{j\in[m]} \|\mathbf{L}^{(i,\ell)}\|_\mathsf{F}^2 \|(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\|_\mathsf{F}^2} \sqrt{\frac{\log(1/\delta_0)}{m^2t^2}} \\
&= \max_{p\in[d]} c \sqrt{\sum_{i\in[t]}\sum_{j\in[m]} \|\left(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\right)^{-1}\|_\mathsf{F}^2 \|(\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\|_\mathsf{F}^2} \sqrt{\frac{\log(1/\delta_0)}{m^2t^2}} \\
&\le c\|\left(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\right)^{-1}\|_\mathsf{F} \sqrt{\sum_{i\in[t]}\sum_{j\in[m]} \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2^2 \|\mathbf{w}^{(i,\ell)}\|_2^2} \sqrt{\frac{\log(1/\delta_0)}{m^2t^2}} \\
&\le c\|\left(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\right)^{-1}\|_\mathsf{F} \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2 \|\mathbf{w}^{(i,\ell)}\|_2 \sqrt{\frac{\log(1/\delta_0)}{mt}}.
\end{aligned} \tag{184}$$

Therefore, using equation 180 in the above, we have $\|\mathbf{V}_2'\|_{2,\infty}$

$$\begin{aligned}
&\le \|\mathbb{E}\left[\mathbf{V}_2'\right]\|_{2,\infty} + c\|\left(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\right)^{-1}\|_\mathsf{F} (\max_i \sqrt{\zeta k}\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_\infty \|\mathbf{w}^{(i,\ell)}\|_2) \sqrt{\frac{\log(1/\delta_0)}{mt^2}} \\
&\le \zeta(\max_i \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_\infty \|\mathbf{w}^{(i,\ell)}\|_2) \|\left((\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\right)^{-1}\|_2 \\
&\qquad c\|\left(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\right)^{-1}\|_\mathsf{F} \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2 \|\mathbf{w}^{(i,\ell)}\|_2 \sqrt{\frac{\log(1/\delta_0)}{mt}}.
\end{aligned} \tag{185}$$

Analysis for $\mathbf{C}\mathsf{vec}\left(\frac{\mathbf{N}_2}{mt}\right)$:

Note that:

$$\begin{aligned}
\mathbf{C}\mathsf{vec}\Big(\frac{\mathbf{N}_2}{mt}\Big) &= \Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)} \otimes \mathbf{I}\Big)^{-1}\mathsf{vec}\Big(\frac{\mathbf{N}_2}{mt}\Big) \\
&= \Big(\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1} \otimes \mathbf{I}\Big)\mathsf{vec}\Big(\frac{\mathbf{N}_2}{mt}\Big) := \mathsf{vec}(\mathbf{V}'_3) \\
\implies \mathbf{V}'_3 &= \mathbf{I}\cdot\frac{\mathbf{N}_2}{mt}\cdot\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-\mathsf{T}} \\
\implies \|\mathbf{V}'_3\|_{2,\infty} &= \|\frac{\mathbf{N}_2}{mt}\cdot\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1}\|_{2,\infty} \\
&\le \frac{1}{mt}\|\mathbf{N}_2\|_{2,\infty}\Big\|\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1}\Big\|_2 \\
&\le \frac{1}{mt}\|\mathbf{N}_2\|_{2,\infty}\frac{r}{\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)} \\
&\le \frac{2\sigma_2}{mt}\sqrt{\log(rd/\delta_0)}\cdot\frac{r\sqrt{r}}{\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)}.
\end{aligned} \tag{186}$$

Analysis for $\mathbf{C}\mathsf{vec}(\boldsymbol{\Xi})$:
Note that:

$$\begin{aligned}
\mathbf{C}\mathsf{vec}(\boldsymbol{\Xi}) &= \Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)} \otimes \mathbf{I}\Big)^{-1}\mathsf{vec}(\boldsymbol{\Xi}) \\
&= \Big(\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1} \otimes \mathbf{I}\Big)\mathsf{vec}(\boldsymbol{\Xi}) := \mathsf{vec}(\mathbf{V}'_\xi) \\
\implies \mathbf{V}'_\xi &= \mathbf{I}\cdot\boldsymbol{\Xi}\cdot\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-\mathsf{T}} \\
\implies \|\mathbf{V}'_\xi\|_{2,\infty} &= \|\boldsymbol{\Xi}\cdot\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1}\|_{2,\infty} \\
&\le \|\boldsymbol{\Xi}\|_{2,\infty}\Big\|\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1}\Big\|_2 \\
&\le \frac{2\sigma\sqrt{\mu^\star\lambda^\star_r\log(2rdmt/\delta_0)}}{\sqrt{mt}}\frac{r}{\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)}
\end{aligned} \tag{187}$$

Combining $\mathbf{V}'_1$, $\mathbf{V}'_2$, $\mathbf{V}'_3$ and $\mathbf{V}'_\xi$ from equation 175, equation 185, equation 186 and equation 187 respectively in equation 165, we have:

$$\begin{aligned}
\mathbf{C}\mathsf{vec}(\mathbf{V}') &= \mathsf{vec}(\mathbf{V}'_1) + \mathsf{vec}(\mathbf{V}'_2) \\
\iff \mathbf{C}\mathsf{vec}\Big(\mathbf{V}' + \frac{\mathbf{N}_2}{mt} + \boldsymbol{\Xi}\Big) &= \mathsf{vec}(\mathbf{V}'_1) + \mathsf{vec}(\mathbf{V}'_2) + \mathsf{vec}(\mathbf{V}'_3) + \mathsf{vec}(\mathbf{V}'_\xi)
\end{aligned}$$

$$\begin{aligned}
\implies &\Big\|\mathsf{vec}^{-1}\Big(\mathbf{C}\mathsf{vec}\Big(\mathbf{V}' + \frac{\mathbf{N}_2}{mt} + \boldsymbol{\Xi}\Big)\Big)\Big\|_{2,\infty} \\
&= \|\mathbf{V}'_1\|_{2,\infty} + \|\mathbf{V}'_2\|_{2,\infty} + \|\mathbf{V}'_3\|_{2,\infty} + \|\mathbf{V}'_\xi\|_{2,\infty}
\end{aligned}$$

$$
\begin{aligned}
&\leq \|\mathbf{U}^{\star}\|_{2,\infty}\|\frac{1}{t}\sum_{i\in[t]}\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\Big)^{-1}\mathbf{w}^{\star(i)}(\mathbf{w}^{(i,\ell)})^{\mathsf{T}}\|_2 \\
&\qquad + c\|\mathbf{U}^{\star}\|_{\mathsf{F}}\|\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\Big)^{-1}\|\|\mathbf{w}^{\star(i)}(\mathbf{w}^{(i,\ell)})^{\mathsf{T}}\|_{\mathsf{F}}\sqrt{\frac{\log(1/\delta_0)}{mt}} \\
&+ \zeta(\max_i\|\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)}\|_{\infty}\|\mathbf{w}^{(i,\ell)}\|_2)\|\Big((\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\Big)^{-1}\|_2 \\
&\qquad + c\|\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\Big)^{-1}\|_{\mathsf{F}}\|\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2\sqrt{\frac{\log(1/\delta_0)}{mt}} \\
&+ \frac{2\sigma_2}{mt}\sqrt{\log(rd/\delta_0)}\cdot\frac{r\sqrt{r}}{\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\Big)} + \frac{2\sigma\sqrt{\mu^{\star}\lambda_r^{\star}}\log(2rdmt/\delta_0)}{\sqrt{mt}}\frac{r}{\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\Big)}
\end{aligned}
\tag{188}
$$

Therefore, using equation 188, equation 109 and equation 115 in equation 164, we have $\|\mathbf{U}^{(\ell)}\|_{2,\infty}$

$$
\begin{aligned}
&\leq \|\mathbf{U}^{\star}\|_{2,\infty}\|\frac{1}{t}\sum_{i\in[t]}\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\Big)^{-1}\mathbf{w}^{\star(i)}(\mathbf{w}^{(i,\ell)})^{\mathsf{T}}\|_2 \\
&+ c\|\mathbf{U}^{\star}\|_{\mathsf{F}}\|\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\Big)^{-1}\|\|\mathbf{w}^{\star(i)}(\mathbf{w}^{(i,\ell)})^{\mathsf{T}}\|_{\mathsf{F}}\sqrt{\frac{\log(1/\delta_0)}{mt}} \\
&+ \zeta(\max_i\|\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)}\|_{\infty}\|\mathbf{w}^{(i,\ell)}\|_2)\|\Big((\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\Big)^{-1}\|_2 \\
&+ c\|\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\Big)^{-1}\|_{\mathsf{F}}\|\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2\sqrt{\frac{\log(1/\delta_0)}{mt}} \\
&+ \frac{2\sigma_2}{mt}\sqrt{\log(rd/\delta_0)}\cdot\frac{r\sqrt{r}}{\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\Big)} + \frac{2\sigma\sqrt{\mu^{\star}\lambda_r^{\star}}\log(2rdmt/\delta_0)}{\sqrt{mt}}\frac{r}{\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\Big)}
\end{aligned}
$$

$$
\begin{aligned}
&\qquad + \frac{\sqrt{r}\Big(c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(rd/\delta_0)}{m^2t^2}} + \frac{\sigma_1}{mt}\Big(2\sqrt{rd}+2\sqrt{2rd\log(2rd/\delta_0)}\Big)\Big)}{\frac{1}{r}\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\Big)}\cdot \\
&\Big\{\frac{2}{t}\|\mathbf{U}^{\star}(\mathbf{W}^{\star})^{\mathsf{T}}\mathbf{W}^{(\ell)}\|_{\mathsf{F}} + \sqrt{\frac{4\zeta}{t}}(\max_i\|\mathbf{w}^{(i,\ell)}\|_2)\|\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)}\|_2 \\
&\quad + 4\Big(\|\mathbf{U}^{\star}\|\|\mathbf{w}^{\star(i)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2 + \|\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2\Big)\sqrt{\frac{d\log(rd/\delta_0)}{mt}} \\
&\quad + \frac{\sigma_2}{mt}6\sqrt{rd\log(rd)} + \frac{2\sigma\sqrt{d\mu^{\star}\lambda_r^{\star}}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big\}.
\end{aligned}
\tag{189}
$$

Calculation for $\|\mathbf{U}^{(\ell)}-\mathbf{U}^{\star}\mathbf{Q}^{(\ell-1)}\|_{2,\infty}$:

The analysis will follow along similar lines as in the previous section except that we will now have:

$$
\mathsf{vec}(\mathbf{U}^{(\ell)}-\mathbf{U}^{\star}\mathbf{Q}^{(\ell-1)}) = \Big(\underbrace{\mathbf{A}'+\frac{\mathbf{N}_1}{mt}}_{\mathbf{A}}\Big)^{-1}\Big(\mathsf{vec}(\mathbf{V}) + \Big(\mathsf{vec}\Big(\frac{\mathbf{N}_2}{mt}\Big) - \frac{\mathbf{N}_1}{mt}\mathsf{vec}(\mathbf{U}^{\star}\mathbf{Q}^{(\ell-1)})\Big)\Big)
\tag{190}
$$

where

$$
\begin{aligned}
\mathbf{A'}_{rd\times rd} &= \frac{1}{mt}\sum_{i\in[t]}\Big(\mathbf{w}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}\otimes(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big)\\
&= \frac{1}{mt}\sum_{i\in[t]}\Big(\mathbf{w}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}\otimes\Big(\sum_{j=1}^{m}\mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}\Big)\Big),\\
\mathbf{V}_{d\times r} &= \frac{1}{mt}\sum_{i\in[t]}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big(\mathbf{U}^\star(\mathbf{w}^{\star(i)}-\mathbf{Q}^{(\ell-1)}\mathbf{w}^{(i,\ell)})+(\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)})\Big)(\mathbf{w}^{(i,\ell)})^\mathsf{T}\\
&= \frac{1}{mt}\sum_{i\in[t]}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big(-\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\mathbf{h}^{(i,\ell)}+(\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)})\Big)(\mathbf{w}^{(i,\ell)})^\mathsf{T},
\end{aligned}
$$

i.e. we have the term $\mathbf{U}^\star\mathbf{w}^{\star(i)}$ replaced by $-\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\mathbf{h}^{(i,\ell)}$ where $\mathbf{h}^{(i,\ell)} = \mathbf{w}^{(i,\ell)} - (\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}$.

Therefore, following along similar lines as for the analysis of $\|\mathbf{U}^{(\ell)}\|_{2,\infty}$, we have $\|\mathbf{U}^{(\ell)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_{2,\infty}$

$$
\begin{aligned}
&\le \left\|\mathsf{vec}^{-1}\Big(\mathbf{C}\Big(\mathsf{vec}(\mathbf{V})+\Big(\mathsf{vec}\Big(\frac{\mathbf{N}_2}{mt}\Big)-\frac{\mathbf{N}_1}{mt}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\Big)\Big)\Big)\right\|_{2,\infty}\\
&\quad+\frac{\sqrt{r}\Big(c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(rd/\delta_0)}{m^2t^2}}+\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+2\sqrt{2rd\log(2rd/\delta_0)}\Big)\Big)}{\frac{1}{r}\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)}\cdot\\
&\quad\left\|\Big(\mathsf{vec}(\mathbf{V})+\Big(\mathsf{vec}\Big(\frac{\mathbf{N}_2}{mt}\Big)-\frac{\mathbf{N}_1}{mt}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\Big)\Big)\right\|_2\\
&\le \left\|\mathsf{vec}^{-1}\Big(\mathbf{C}\Big(\mathsf{vec}(\mathbf{V})+\Big(\mathsf{vec}\Big(\frac{\mathbf{N}_2}{mt}\Big)-\frac{\mathbf{N}_1}{mt}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\Big)\Big)\Big)\right\|_{2,\infty}\\
&\quad+\frac{\sqrt{r}\Big(c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(rd/\delta_0)}{m^2t^2}}+\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+2\sqrt{2rd\log(2rd/\delta_0)}\Big)\Big)}{\frac{1}{r}\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)}\cdot\\
&\quad\Big(\|\mathbf{V}\|_\mathsf{F}+\left\|\frac{\mathbf{N}_2}{mt}\right\|_\mathsf{F}+\left\|\frac{\mathbf{N}_1}{mt}\right\|_2\|\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_\mathsf{F}\Big).
\end{aligned}
\tag{191}
$$

Now, note that:

$$
\begin{aligned}
\mathbf{C}\mathsf{vec}(\mathbf{V}) = &\underbrace{\frac{1}{mt}\sum_{i\in[t]}\Big(\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1}\mathbf{h}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}\otimes(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big)\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)})}_{\mathsf{vec}(\mathbf{V}_1)}\\
&+\underbrace{\frac{1}{mt}\sum_{i\in[t]}\Big(\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1}\otimes\mathbf{I}\Big)\mathsf{vec}\Big((\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}(\mathbf{b}^{\star(i)}-\mathbf{b}^{(i,\ell)})(\mathbf{w}^{(i,\ell)})^\mathsf{T}\Big)}_{\mathsf{vec}(\mathbf{V}_2)}
\end{aligned}
\tag{192}
$$

Let $\mathbf{J}^{(i,\ell)} := \left(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\right)^{-1}\mathbf{h}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}$. Then equation 175 in this case becomes

$$\begin{aligned}
\|\mathbf{V}_1\|_{2,\infty} &\leq \|\mathbf{U}^\star\|_{2,\infty}\|\frac{1}{t}\sum_{i\in[t]}\mathbf{J}^{(i,\ell)}\|_2 + c\|\mathbf{U}^\star\|_\mathsf{F}\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{J}^{(i,\ell)}\|_\mathsf{F}^2}\sqrt{\frac{\log(1/\delta_0)}{m^2t^2}} \\
&\leq \|\mathbf{U}^\star\|_{2,\infty}\|\frac{1}{t}\sum_{i\in[t]}\left(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\right)^{-1}\mathbf{h}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}\|_2 \\
&\quad + c\|\mathbf{U}^\star\|_\mathsf{F}\|\left(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\right)^{-1}\|\|\mathbf{h}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}\|_\mathsf{F}\sqrt{\frac{\log(1/\delta_0)}{mt}} \qquad (193)
\end{aligned}$$

while equation 185, equation 186 and equation 187 remain the same

$$\begin{aligned}
\|\mathbf{V}_2\|_{2,\infty} &\leq \zeta(\max_i\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_\infty\|\mathbf{w}^{(i,\ell)}\|_2)\|\left((\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\right)^{-1}\|_2 \\
&\quad c\|\left(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\right)^{-1}\|_\mathsf{F}\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2\sqrt{\frac{\log(1/\delta_0)}{mt}}. \qquad (194) \\
\|\mathbf{V}_3\|_{2,\infty} &\leq \frac{2\sigma_2}{mt}\sqrt{\log(rd/\delta_0)}\cdot\frac{r\sqrt{r}}{\lambda_r\left(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\right)}, \qquad (195) \\
\|\mathbf{V}_\xi\|_{2,\infty} &\leq \frac{2\sigma\sqrt{\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\frac{r}{\lambda_r\left(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\right)} \qquad (196)
\end{aligned}$$

We also have the additional term $\mathbf{V}_4$ s.t.

$$\begin{aligned}
\mathbf{C}\cdot\frac{\mathbf{N}_1}{mt}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}) &= \left(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\otimes\mathbf{I}\right)^{-1}\frac{\mathbf{N}_1}{mt}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}) := \mathsf{vec}(\mathbf{V}_4) \qquad (197) \\
\iff \mathbf{V}_4 &= \mathbf{I}\cdot\mathsf{vec}^{-1}\left(\frac{\mathbf{N}_1}{mt}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\right)\cdot\left(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\right)^{-\mathsf{T}} \\
\implies \|\mathbf{V}_4\|_{2,\infty} &\leq \left\|\mathsf{vec}^{-1}\left(\frac{\mathbf{N}_1}{mt}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\right)\cdot\left(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\right)^{-\mathsf{T}}\right\|_{2,\infty} \\
&\leq \left\|\mathsf{vec}^{-1}\left(\frac{\mathbf{N}_1}{mt}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\right)\right\|_{2,\infty}\left\|\left(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\right)^{-\mathsf{T}}\right\|_2 \\
&\leq \left\|\mathsf{vec}^{-1}\left(\frac{\mathbf{N}_1}{mt}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\right)\right\|_{2,\infty}\frac{r}{\lambda_r\left(\frac{r}{t}\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\right)}. \qquad (198)
\end{aligned}$$

Now, $\mathbf{N}_1 \sim \sigma_1\mathcal{MN}(\mathbf{0}, \mathbf{I}_{rd\times rd}, \mathbf{I}_{rd\times rd}) \implies \frac{\mathbf{N}_1}{mt}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}) = \frac{1}{mt}\begin{bmatrix}\mathbf{N}_{1,1}^\mathsf{T}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}) \\ \vdots \\ \mathbf{N}_{1,j}^\mathsf{T}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}) \\ \vdots \\ \mathbf{N}_{1,rd}^\mathsf{T}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\end{bmatrix}$ where each $\mathbf{N}_{1,j} \sim \sigma_1\mathcal{N}(\mathbf{0}, \mathbf{I}_{rd\times rd})$. Therefore,

$$\left\|\mathsf{vec}^{-1}\left(\frac{\mathbf{N}_1}{mt}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\right)\right\|_{2,\infty}^2 = \frac{1}{m^2t^2}\max_{p\in[d]}\sum_{q\in[r]}\left(\mathbf{N}_{1,p+q}^\mathsf{T}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\right)^2, \qquad (199)$$

where each $(\mathbf{N}_{1,p+q})_{e,f}(\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}))_{e,f} \in \mathcal{N}(0,\sigma_1^2)\ \forall\ e,f \in d\times r$. Therefore, using standard gaussian concentration results and taking union bound over e,f, we have w.p. atleast $1-\delta_0$,

$$\begin{aligned}\Big(\mathbf{N}_{1,p+q}^\mathsf{T}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\Big)^2 &= \Big(\sum_e\sum_f(\mathbf{N}_{1,p+q})_{e,f}(\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}))_{e,f}\Big)^2\\ &\le rd\sum_e\sum_f|(\mathbf{N}_{1,p+q})_{e,f}|^2|(\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}))_{e,f}|^2\\ &\le rd\sum_e\sum_f\sigma_1^2\cdot 2\log(2rd/\delta_0)|(\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}))_{e,f}|^2\\ &= rd\sigma_1^2\cdot 2\log(2rd/\delta_0)\|\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\|_2^2. \qquad (200)\end{aligned}$$

Using equation 200 in equation 199 and taking a Union Bound $\forall\ p,q$, we have w.p. $\ge 1-\delta_0$

$$\begin{aligned}&\Big\|\mathsf{vec}^{-1}\Big(\frac{\mathbf{N}_1}{mt}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\Big)\Big\|_{2,\infty}^2 \le \frac{1}{m^2t^2}\max_{p\in[d]}\sum_{q\in[r]}r^2d\sigma_1^2\cdot 2\log(2r^2d^2/\delta_0)\|\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\|_2^2\\ \implies &\Big\|\mathsf{vec}^{-1}\Big(\frac{\mathbf{N}_1}{mt}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)})\Big)\Big\|_{2,\infty} \le \frac{1}{mt}r\sigma_1\sqrt{d\cdot 2\log(2r^2d^2/\delta_0)}\|\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_\mathsf{F}. \qquad (201)\end{aligned}$$

Using equation 201 in equation 198 we get

$$\|\mathbf{V}_4\|_{2,\infty} \le \frac{r^2\sigma_1\sqrt{rd\cdot 2\log(2r^2d^2/\delta_0)}}{mt\lambda_r\Big(\frac{r}{t}\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)}. \qquad (202)$$

Combining $\mathbf{V}_1$, $\mathbf{V}_2$, $\mathbf{V}_3$, $\mathbf{V}_4$ and $\mathbf{V}_\xi$ from equation 193, equation 194, equation 195, equation 202 and equation 196 respectively in equation 192, we have $\Big\|\mathsf{vec}^{-1}\Big(\mathbf{C}\Big(\mathsf{vec}(\mathbf{V}) + \mathsf{vec}\Big(\frac{\mathbf{N}_2}{mt}\Big) - \frac{\mathbf{N}_1}{mt}\mathsf{vec}(\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}) + \mathsf{vec}(\mathbf{\Xi})\Big)\Big\|_{2,\infty}$

$$\begin{aligned}&= \|\mathbf{V}_1\|_{2,\infty} + \|\mathbf{V}_2\|_{2,\infty} + \|\mathbf{V}_3\|_{2,\infty} + \|\mathbf{V}_4\|_{2,\infty} + \|\mathbf{V}_\xi\|_{2,\infty}\\ &\le \|\mathbf{U}^\star\|_{2,\infty}\|\frac{1}{t}\sum_{i\in[t]}\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1}\mathbf{h}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}\|_2\\ &\quad + c\|\mathbf{U}^\star\|_\mathsf{F}\|\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1}\|\|\|\mathbf{h}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}\|_\mathsf{F}\sqrt{\frac{\log(1/\delta_0)}{mt}}\\ &\quad + \zeta(\max_i\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_\infty\|\mathbf{w}^{(i,\ell)}\|_2)\|\Big((\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1}\|_2\\ &\quad + c\|\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1}\|_\mathsf{F}\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2\sqrt{\frac{\log(1/\delta_0)}{mt}}\\ &\quad + \frac{2\sigma_2}{mt}\sqrt{\log(rd/\delta_0)}\cdot\frac{r\sqrt{r}}{\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)} + \frac{r^2\sigma_1\sqrt{rd\cdot 2\log(2r^2d^2/\delta_0)}}{mt\lambda_r\Big(\frac{r}{t}\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)}\\ &\quad + \frac{2\sigma\sqrt{\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\frac{r}{\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)}. \qquad (203)\end{aligned}$$

Therefore, using equation 203, equation 109, equation 115 and equation 106 in equation 191, we have $\|\mathbf{U}^{(\ell)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_{2,\infty}$

$$
\begin{aligned}
&\le \|\mathbf{U}^\star\|_{2,\infty}\|\frac{1}{t}\sum_{i\in[t]}\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1}\mathbf{h}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}\|_2 \\
&\quad + c\|\mathbf{U}^\star\|_\mathsf{F}\|\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1}\|\|\mathbf{h}^{(i,\ell)}(\mathbf{w}^{(i,\ell)})^\mathsf{T}\|_\mathsf{F}\sqrt{\frac{\log(1/\delta_0)}{mt}} \\
&\quad + \zeta(\max_i\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_\infty\|\mathbf{w}^{(i,\ell)}\|_2)\|\Big((\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1}\|_2 \\
&\quad + c\|\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)^{-1}\|_\mathsf{F}\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2\sqrt{\frac{\log(1/\delta_0)}{mt}} \\
&\quad + \frac{2\sigma_2}{mt}\sqrt{\log(rd/\delta_0)}\cdot\frac{r\sqrt{r}}{\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)} + \frac{r^2\sigma_1\sqrt{rd\cdot 2\log(2r^2d^2/\delta_0)}}{mt\lambda_r\Big(\frac{r}{t}\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)} \\
&\quad + \frac{2\sigma\sqrt{\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\frac{r}{\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)} \\
&\quad + \frac{\sqrt{r}\Big(c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(rd/\delta_0)}{m^2t^2}} + \frac{\sigma_1}{mt}\Big(2\sqrt{rd} + 2\sqrt{2rd\log(2rd/\delta_0)}\Big)\Big)}{\frac{1}{r}\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\Big)}\cdot \\
&\quad \Big\{\frac{2}{t}\|\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}(\mathbf{H}^{(\ell)})^\mathsf{T}\mathbf{W}^{(\ell)}\|_\mathsf{F} + \sqrt{\frac{4\zeta}{t}}(\max_i\|\mathbf{w}^{(i,\ell)}\|_2)\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2 \\
&\quad + 4\Big(\|\mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|\|\mathbf{h}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2 + \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2\Big)\sqrt{\frac{d\log(rd/\delta_0)}{mt}} \\
&\quad + \frac{\sigma_2}{mt}6\sqrt{rd\log(rd)} + \frac{\sigma_1}{mt}\Big(2\sqrt{rd} + 4\sqrt{\log rd}\Big)\sqrt{r} + \frac{2\sigma\sqrt{d\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big\}.
\end{aligned}
\tag{204}
$$

equation 189 and equation 204 give us the required result. □

Corollary 5. *If* $\mathsf{B}_{\mathbf{U}^{(\ell-1)}} = \mathcal{O}\big(\frac{1}{\sqrt{r\mu^\star}}\}\big)$, $\sqrt{\frac{r\log(1/\delta_0)}{m}} = \mathcal{O}(1)$, $\sqrt{\nu^{(\ell-1)}} = \mathcal{O}\big(\frac{1}{\sqrt{r\mu^\star}}$, $\sqrt{\frac{r^2\log(r/\delta_0)}{m}} = \mathcal{O}\big(\frac{1}{\sqrt{\mu^\star}}\big)$, $\epsilon < \sqrt{\mu^\star\lambda_r^\star}$, $\sqrt{\frac{r^2\zeta}{t}} = \min\{\mathcal{O}\big(\frac{1}{\sqrt{\mu^\star}}\big), \mathcal{O}\big(\frac{1}{\mu^\star}\big)\}$, $\Lambda' = \mathcal{O}\Big(\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big)$, $\Lambda = \mathcal{O}(\sqrt{\lambda_r^\star})$, $\sigma\sqrt{\frac{r^2\log^2(r\delta^{-1})}{m}} = \mathcal{O}(\sqrt{\lambda_r^\star})$, $\sqrt{\frac{r^3d\log(1/\delta_0)}{mt}} = \min\{\mathcal{O}\big(\frac{1}{\mu^\star}\big), \mathcal{O}\big(\frac{1}{\mu^\star\lambda_r^\star}\big), \mathcal{O}\big(\frac{1}{\sqrt{r}(\mu^\star)^2\sqrt{\mu^\star}\lambda_r^\star\sqrt{k}}\big)\}$, $\frac{\sigma_1}{mt}\big(2\sqrt{rd} + 4\sqrt{\log rd}\big) = \min\{\mathcal{O}(\lambda_r^\star), \mathcal{O}\big(\frac{1}{r}\big), \mathcal{O}\big(\frac{1}{r\sqrt{r}\mu^\star\sqrt{\mu^\star}\sqrt{k}}\big)\}$, $\sqrt{\nu^\star} = \mathcal{O}\big(\frac{1}{\sqrt{\mu^\star}}\frac{\lambda_r^\star}{\lambda_1^\star}\big)$, $\Lambda = \mathcal{O}(\sqrt{\lambda_r^\star})$, $mt = \Omega(\sigma_2 r(\mu^\star)^{1/2}\sqrt{rd\log d}/\lambda_r^\star)$ *and* $mt = \widetilde{\Omega}(\sigma^2dr^3\mu^\star(1 + 1/\lambda_r^\star))$, $t = \widetilde{\Omega}(\zeta(\mu^\star\lambda_r^\star)\max(1, \lambda_r^\star/r))$, $m = \widetilde{\Omega}(\sigma^2r^3/\lambda_r^\star)$. $\sqrt{\frac{k}{d}} = \mathcal{O}(1)$ *and Assumption 4 holds for iteration* $\ell - 1$*, then, w.p.* $1 - \mathcal{O}(\delta_0)$*,*

$$
\mathbf{U}_{2,\infty}^{(\ell)} = \mathcal{O}\Big(\frac{1}{\sqrt{k\mu^\star}}\Big) + \frac{1}{\sqrt{k}}\Big\{\mathcal{O}\Big(\frac{\Lambda}{\sqrt{\mu^\star\lambda_r^\star}}\Big) + \mathcal{O}\Big(\frac{\sigma_2 r}{mt\lambda_r^\star}\sqrt{rd\log(rd)}\Big) + \mathcal{O}\Big(\sigma\sqrt{\frac{r^3d\log^2(2rdmt/\delta_0)}{mt\lambda_r^\star}}\Big\}
$$

and $\|\mathbf{U}^{(\ell)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_{2,\infty}$

$$
\begin{aligned}
&= \mathcal{O}\Big(\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star k}}\Big) + \frac{1}{\sqrt{k}}\Big\{\mathcal{O}(\Lambda' + \mathcal{O}\Big(\frac{\Lambda}{\sqrt{\mu^\star\lambda_r^\star}}\Big) + \mathcal{O}\Big(\frac{\sigma_2 r}{mt\lambda_r^\star}\sqrt{rd\log(rd)}\Big) \\
&\quad + \mathcal{O}\Big(\frac{\sigma_1 r^{3/2}}{mt\lambda_r^\star}\sqrt{rd\log rd}\Big) + \mathcal{O}\Big(\sigma\sqrt{\frac{r^3d\mu^\star\log^2(r\delta^{-1})}{mt\lambda_r^\star}}\Big)\Big\}
\end{aligned}
$$

Proof. From the Lemma statement we have, $\big|\big|\mathbf{U}^{(\ell)}\big|\big|_{2,\infty}$

$$
\begin{aligned}
&\leq \|\mathbf{U}^{\star}\|_{2,\infty} \| \frac{1}{t} \sum_{i\in[t]} \Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\Big)^{-1} \mathbf{w}^{\star(i)}(\mathbf{w}^{(i,\ell)})^{\mathsf{T}}\|_2 \\
&+ c\|\mathbf{U}^{\star}\|_{\mathsf{F}} \|\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\Big)^{-1}\| \|\mathbf{w}^{\star(i)}(\mathbf{w}^{(i,\ell)})^{\mathsf{T}}\|_{\mathsf{F}} \sqrt{\frac{\log(1/\delta_0)}{mt}} \\
&+ \zeta(\max_i \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_\infty \|\mathbf{w}^{(i,\ell)}\|_2) \|\Big((\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\Big)^{-1}\|_2 \\
&+ c\|\Big(\frac{1}{t}(\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\Big)^{-1}\|_{\mathsf{F}} \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2 \|\mathbf{w}^{(i,\ell)}\|_2 \sqrt{\frac{\log(1/\delta_0)}{mt}} \\
&+ \frac{2\sigma_2}{mt}\sqrt{\log(rd/\delta_0)} \cdot \frac{r\sqrt{r}}{\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\Big)} + \frac{2\sigma\sqrt{\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}} \frac{r}{\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\Big)} \\
&+ \frac{\sqrt{r}\Big(c\sqrt{\sum_{i\in[t]}\sum_{j\in[m]}\|\mathbf{w}^{(i,\ell)}\|_2^4}\sqrt{\frac{rd\log(rd/\delta_0)}{m^2t^2}} + \frac{\sigma_1}{mt}\Big(2\sqrt{rd} + 2\sqrt{2rd\log(2rd/\delta_0)}\Big)\Big)}{\frac{1}{r}\lambda_r\Big(\frac{r}{t}(\mathbf{W}^{(\ell)})^{\mathsf{T}}\mathbf{W}^{(\ell)}\Big)} \cdot \\
&\Big\{\frac{2}{t}\|\mathbf{U}^{\star}(\mathbf{W}^{\star})^{\mathsf{T}}\mathbf{W}^{(\ell)}\|_{\mathsf{F}} + \sqrt{\frac{4\zeta}{t}}(\max_i \|\mathbf{w}^{(i,\ell)}\|_2)\|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2 \\
&+ 4\Big(\|\mathbf{U}^{\star}\|\|\mathbf{w}^{\star(i)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2 + \|\mathbf{b}^{\star(i)} - \mathbf{b}^{(i,\ell)}\|_2\|\mathbf{w}^{(i,\ell)}\|_2\Big)\sqrt{\frac{d\log(rd/\delta_0)}{mt}} \\
&+ \frac{\sigma_2}{mt}6\sqrt{rd\log(rd)} + \frac{2\sigma\sqrt{d\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big\}
\end{aligned} \tag{205}
$$

As done in the analysis for Corollary 3, using Assumption 4 to plug in values for $\mathbf{b}^{(i,\ell)}$ and $\mathbf{Q}^{(\ell-1)}$, the fact that $\mathbf{U}^\star$ is orthonormal and norm and incoherence bounds for $\mathbf{H}^{(\ell)}$ and $\mathbf{W}^{(\ell)}$ from Corollaries 1 and 2, the above becomes $\mathbf{U}_{2,\infty}^{(\ell)}$

$$
\begin{aligned}
&\leq \sqrt{\frac{\nu^\star}{k}}\frac{\sqrt{\lambda_1\lambda_1^\star}}{\lambda_r} + c\frac{r\sqrt{r}}{\lambda_r}\sqrt{\mu\lambda_r}\sqrt{\mu^\star\lambda_r^\star}\sqrt{\frac{\log(1/\delta_0)}{mt}} + \zeta\Big(c'\sqrt{\mu^\star\lambda_r^\star}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star k}} + \frac{\Lambda}{\sqrt{k}}\Big)\frac{r\sqrt{\mu\lambda_r}}{t\lambda_r} \\
&+ c\frac{r\sqrt{r}}{\lambda_r}\Big(c'\sqrt{\mu^\star\lambda_r^\star}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} + \Lambda\Big)\sqrt{\mu\lambda_r}\sqrt{\frac{\log(1/\delta_0)}{mt}} \\
&+ \frac{2\sigma_2}{mt}\sqrt{\log(rd/\delta_0)} \cdot \frac{r\sqrt{r}}{\lambda_r} + \frac{2\sigma\sqrt{\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\frac{r}{\lambda_r} \\
&+ \sqrt{r}\Big(c\mu\sqrt{\frac{r^3d\log(rd/\delta_0)}{mt}} + \frac{\sigma_1 r}{mt\lambda_r}\Big(2\sqrt{rd} + 2\sqrt{2rd\log(2rd/\delta_0)}\Big)\Big) \cdot \\
&\Big\{2\sqrt{\mu^\star\lambda_r^\star}\sqrt{\mu\lambda_r} + \sqrt{\frac{4\zeta}{t}}\sqrt{\mu\lambda_r}\Big(c'\sqrt{\mu^\star\lambda_r^\star}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} + \Lambda\Big) \\
&+ 4\sqrt{\mu\lambda_r}\Big(\sqrt{\mu^\star\lambda_r^\star} + \Big(c'\sqrt{\mu^\star\lambda_r^\star}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} + \Lambda\Big)\Big)\sqrt{\frac{d\log(rd/\delta_0)}{mt}} \\
&+ \frac{\sigma_2}{mt}6\sqrt{rd\log(rd)} + \frac{2\sigma\sqrt{d\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big\}
\end{aligned} \tag{206}
$$

$$
= J_1 + J_2 \tag{207}
$$

where J_1 denotes the terms which arise from analysing the problem in the noiseless setting and J_2 denotes the contribution of noise terms $(\sigma_1, \sigma_2, \sigma, \Lambda, \Lambda')$. Now, J_1

$$\begin{aligned}
&= \sqrt{\frac{\nu^\star}{k}}\frac{\sqrt{\lambda_1\lambda_1^\star}}{\lambda_r} + c\sqrt{r}\frac{r}{\lambda_r}\sqrt{\mu\lambda_r}\sqrt{\mu^\star\lambda_r^\star}\sqrt{\frac{\log(1/\delta_0)}{mt}} + \zeta\cdot c'\sqrt{\mu^\star\lambda_r^\star}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star k}}\cdot\sqrt{\mu\lambda_r}\frac{r}{t\lambda_r} \\
&\quad + c\frac{r\sqrt{r}}{\lambda_r}\cdot c'\sqrt{\mu^\star\lambda_r^\star}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\cdot\sqrt{\mu\lambda_r}\sqrt{\frac{\log(1/\delta_0)}{mt}} \\
&\quad + \sqrt{r}\Big(c\mu\sqrt{\frac{r^3 d\log(rd/\delta_0)}{mt}} + \frac{\sigma_1 r}{mt\lambda_r}\Big(2\sqrt{rd} + 2\sqrt{2rd\log(2rd/\delta_0)}\Big)\Big)\cdot \\
&\quad \Big\{2\sqrt{\mu^\star\lambda_r^\star}\sqrt{\mu\lambda_r} + \sqrt{\frac{4\zeta}{t}}\sqrt{\mu\lambda_r}\cdot c'\sqrt{\mu^\star\lambda_r^\star}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} \\
&\quad + 4\sqrt{\mu\lambda_r}\Big(\sqrt{\mu^\star\lambda_r^\star} + c'\sqrt{\mu^\star\lambda_r^\star}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big)\sqrt{\frac{d\log(rd/\delta_0)}{mt}}\}. \qquad (208)
\end{aligned}$$

Using $\sqrt{\nu^\star} = \mathcal{O}\Big(\frac{\lambda_r^\star}{\lambda_1^\star\sqrt{\mu^\star}}\Big)$ and rearranging the terms in the above we get J_1

$$\begin{aligned}
&= \frac{1}{\sqrt{k\mu^\star}}\Big\{\mathcal{O}\Big(\frac{\lambda_r^\star}{\lambda_r}\sqrt{\frac{\lambda_1^\star}{\lambda_1}}\Big) + \sqrt{\frac{k}{d}}\sqrt{\frac{\lambda_r^\star}{\lambda_r}}\sqrt{\frac{\mu}{\mu^\star}}\sqrt{\frac{r^3 d\log(1/\delta_0)}{mt}} + \mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\frac{r\zeta}{t}c'\sqrt{\mu^\star\mu}\sqrt{\frac{\lambda_r^\star}{\lambda_r}} \\
&\quad + \mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}cc'\sqrt{\frac{\lambda_r^\star}{\lambda_r}}\sqrt{\mu^\star\mu}\sqrt{\mu^\star}\sqrt{\frac{r^3 d\log(1/\delta_0)}{mt}}\sqrt{\frac{k}{d}} \\
&\quad + \sqrt{r}\Big(c\mu\sqrt{\frac{r^3 d\log(rd/\delta_0)}{mt}} + \frac{\sigma_1 r}{mt\lambda_r}\Big(2\sqrt{rd} + 2\sqrt{2rd\log(2rd/\delta_0)}\Big)\Big) \\
&\quad \sqrt{\mu^\star}\sqrt{\mu^\star\lambda_r^\star}\sqrt{\mu\lambda_r}\Big\{2 + \sqrt{\frac{4\zeta}{t}}c'\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} + 4\Big(1 + c'\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big)\sqrt{\frac{d\log(rd/\delta_0)}{mt}}\Big\}\Big\}.
\end{aligned}$$

Using $\sqrt{\frac{r^3 d\log(1/\delta_0)}{mt}} = \mathcal{O}\Big(\frac{1}{\mu^\star}\Big)$, $\sqrt{\frac{k}{d}} = \mathcal{O}(1)$, $\mathsf{B}_{\mathbf{U}^{(\ell-1)}} = \mathcal{O}\Big(\frac{1}{\sqrt{r\mu^\star}}\Big)$, $\sqrt{\frac{r^2\log(r/\delta_0)}{m}} = \mathcal{O}\Big(\frac{1}{\sqrt{\mu^\star}}\Big)$, $\sqrt{\frac{r^3 d\log(1/\delta_0)}{mt}} = \mathcal{O}\Big(\frac{1}{\sqrt{r}(\mu^\star)^2\sqrt{\mu^\star}\lambda_r^\star\sqrt{k}}\Big)$, $\frac{\sigma_1}{mt}\Big(2\sqrt{rd} + 4\sqrt{\log rd}\Big) = \mathcal{O}\Big(\frac{1}{r\sqrt{r}\mu^\star\sqrt{\mu^\star}\sqrt{k}}\Big)$, $\sqrt{\frac{r^2\zeta}{t}} = \mathcal{O}\Big(\frac{1}{\sqrt{\mu^\star}}\Big)$, eigenvalue ratios and incoherence bounds for $\mathbf{H}^{(\ell)}$ and $\mathbf{W}^{(\ell)}$ from Corollaries 1 and 2 as well as $\mathsf{B}_{\mathbf{U}^{(\ell-1)}} = \mathcal{O}\Big(\frac{1}{\sqrt{r\mu^\star}}\Big)$, $\lambda_r^\star \leq \lambda_1^\star$, $r \geq 1$,$\mu^\star \geq 1$ for the bracketed terms, we get

$$\begin{aligned}
&= \frac{1}{\sqrt{k\mu^\star}}\Big\{\mathcal{O}(1)\Big\} \\
&= \mathcal{O}\Big(\frac{1}{\sqrt{k\mu^\star}}\Big). \qquad (209)
\end{aligned}$$

Similarly, we have J_2

$$\begin{aligned}
&\leq \zeta\frac{\Lambda}{\sqrt{k}}\sqrt{\mu\lambda_r}\frac{r}{t\lambda_r} + c\frac{r\sqrt{r}}{\lambda_r}\Lambda\sqrt{\mu\lambda_r}\sqrt{\frac{\log(1/\delta_0)}{mt}} + \frac{2\sigma_2}{mt}\sqrt{\log(rd/\delta_0)}\cdot\frac{r\sqrt{r}}{\lambda_r} \\
&+ \tfrac{2\sigma\sqrt{\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)r}{\sqrt{mt}\lambda_r} + \sqrt{r}\Big(c\mu\sqrt{\tfrac{r^3 d\log(rd/\delta_0)}{mt}} + \tfrac{\sigma_1 r}{mt\lambda_r}\Big(2\sqrt{rd} + 2\sqrt{2rd\log(2rd/\delta_0)}\Big)\Big)\cdot \\
&\Big\{\sqrt{\frac{4\zeta}{t}}\sqrt{\mu\lambda_r}\Lambda + 4\sqrt{\mu\lambda_r}\Lambda\sqrt{\frac{d\log(rd/\delta_0)}{mt}} + \frac{\sigma_2}{mt}6\sqrt{rd\log(rd)} + \frac{2\sigma\sqrt{d\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big\}. \qquad (210)
\end{aligned}$$

Using $\sqrt{\frac{r^3 d\log(1/\delta_0)}{mt}} = \mathcal{O}\Big(\frac{1}{\sqrt{r}(\mu^\star)^2\sqrt{\mu^\star}\lambda_r^\star\sqrt{k}}\Big)$, $\frac{\sigma_1}{mt}\Big(2\sqrt{rd} + 4\sqrt{\log rd}\Big) = \mathcal{O}\Big(\frac{1}{r\sqrt{r}\mu^\star\sqrt{\mu^\star}\sqrt{k}}\Big)$ and rearranging the terms in above gives

$$
\begin{aligned}
&= \Lambda\Big\{\tfrac{r\zeta}{t}\tfrac{\sqrt{\mu\lambda_r}}{\lambda_r\sqrt{k}} + \tfrac{c\sqrt{\mu}}{\sqrt{\lambda_r}}\sqrt{\tfrac{r^3\log(1/\delta_0)}{mt}} + \sqrt{\mu\lambda_r}\cdot\mathcal{O}\Big(\tfrac{1}{\sqrt{r}(\mu^\star)^2\sqrt{\mu^\star}\lambda_r^\star\sqrt{k}}\Big)\Big(\sqrt{\tfrac{4\zeta}{t}} + 4\sqrt{\tfrac{d\log(rd/\delta_0)}{mt}}\Big)\Big\} \\
&+ \sigma_2\Big\{\frac{2}{mt}\sqrt{\log(rd/\delta_0)}\cdot\frac{r\sqrt{r}}{\lambda_r} + \mathcal{O}\Big(\frac{1}{\sqrt{r}(\mu^\star)^2\sqrt{\mu^\star}\lambda_r^\star\sqrt{k}}\Big)\cdot\frac{1}{mt}6\sqrt{rd\log(rd)}\Big\} \\
&+ \sigma\Big\{\frac{2\sqrt{\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\frac{r}{\lambda_r} + \mathcal{O}\Big(\frac{1}{\sqrt{r}(\mu^\star)^2\sqrt{\mu^\star}\lambda_r^\star\sqrt{k}}\Big)\cdot\frac{2\sqrt{d\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big\}
\end{aligned}
$$

Using $\sqrt{\frac{r^2\zeta}{t}} = \mathcal{O}\Big(\frac{1}{\sqrt{\mu^\star}}\Big)$, $\sqrt{\frac{k}{d}} = \mathcal{O}(1)$, $\sqrt{\frac{r^3 d\log(1/\delta_0)}{mt}} = \mathcal{O}\Big(\frac{1}{\mu^\star}\Big)$, $\Lambda = \mathcal{O}(\sqrt{\lambda_r^\star})$, $mt = \Omega(\sigma_2 r(\mu^\star)^{1/2}\sqrt{rd\log d}/\lambda_r^\star)$ and $mt = \widetilde{\Omega}(\sigma^2 dr^3\mu^\star/\lambda_r^\star)$, $t = \widetilde{\Omega}(\zeta(\mu^\star\lambda_r^\star)\max(1,\lambda_r^\star/r))$, eigenvalue ratios and incoherence bounds for $\mathbf{H}^{(\ell)}$ and $\mathbf{W}^{(\ell)}$ from Corollaries 1 and 2 as well as $\lambda_r^\star \le \lambda_1^\star$, $r \ge 1$,$\mu^\star \ge 1$ for the bracketed terms, the above simplifies to

$$
J_2 = \frac{1}{\sqrt{k}}\Big\{\mathcal{O}\Big(\frac{\Lambda}{\sqrt{\mu^\star\lambda_r^\star}}\Big) + \mathcal{O}\Big(\frac{\sigma_2 r}{mt\lambda_r^\star}\sqrt{rd\log(rd)}\Big) + \mathcal{O}\Big(\sigma\sqrt{\frac{r^3 d\log^2(2rdmt/\delta_0)}{mt\lambda_r^\star}}\Big\}. \tag{211}
$$

Using equation 209 and equation 211 in equation 207 gives us the required bound for $\|\mathbf{U}^{(\ell)}\|_{2,\infty}$.

Similarly, using Assumption 4 to plug in values for $\mathbf{b}^{(i,\ell)}$ and $\mathbf{Q}^{(\ell-1)}$, the fact that $\mathbf{U}^\star$ is orthonormal and norm and incoherence bounds for $\mathbf{H}^{(\ell)}$ and $\mathbf{W}^{(\ell)}$ from Corollaries 1 and 2, the above becomes $\mathbf{U}_{2,\infty}^{(\ell)}$, we can simplify and express $\|\mathbf{U}^{(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_{2,\infty}$ as

$$
\begin{aligned}
&\le \sqrt{\frac{\nu^\star}{k}\frac{r}{t\lambda_r}}\cdot\mathcal{O}\Big(\sqrt{\frac{t}{r}}\sqrt{\lambda_1^\star}\Big)\cdot \\
&\quad\Big(\mathcal{O}\Big(\frac{\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\frac{\lambda_r^\star}{\lambda_1^\star}\sqrt{\frac{t}{r}\lambda_1^\star}}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\sqrt{\frac{t}{r}\lambda_r^\star}\frac{\Lambda'}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\frac{\Lambda}{\sqrt{r\mu^\star}}\sqrt{\frac{t}{r}}\Big) + \mathcal{O}\Big(\sigma\sqrt{\frac{r\log^2(r\delta^{-1})}{m}}\Big)\Big) \\
&\quad+ \sqrt{r}\mathcal{O}\Big(\frac{r}{\lambda_r}\sqrt{\mu\lambda_r}\Big)\sqrt{\frac{\log(1/\delta_0)}{mt}}\cdot \\
&\quad\Big(\mathcal{O}\Big(\frac{\sqrt{\mu^\star\lambda_r^\star}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\frac{\Lambda'\|\mathbf{w}^{\star(i)}\|_2}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\frac{\Lambda}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\sigma\sqrt{\frac{r\log^2(r\delta^{-1})}{m}}\Big)\Big) \\
&\quad+ \zeta\Big(c'\sqrt{\mu^\star\lambda_r^\star}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star k}} + \frac{\Lambda}{\sqrt{k}}\Big)\sqrt{\mu\lambda_r}\frac{r}{t\lambda_r} \\
&\quad+ c\frac{r\sqrt{r}}{\lambda_r}\Big(c'\sqrt{\mu^\star\lambda_r^\star}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} + \Lambda\Big)\sqrt{\mu\lambda_r}\sqrt{\frac{\log(1/\delta_0)}{mt}} \\
&\quad+ \frac{2\sigma_2}{mt}\sqrt{\log(rd/\delta_0)}\cdot\frac{r\sqrt{r}}{\lambda_r} + \frac{r^2\sigma_1\sqrt{rd\cdot 2\log(2r^2d^2/\delta_0)}}{mt\lambda_r} \\
&\quad+ \frac{2\sigma\sqrt{\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\frac{r}{\lambda_r}
\end{aligned}
$$

$$
\begin{aligned}
&+ r\sqrt{r}\Big(c\mu\sqrt{\frac{rd\log(rd/\delta_0)}{mt}} + \frac{\sigma_1}{mt\lambda_r}\Big(2\sqrt{rd} + 2\sqrt{2rd\log(2rd/\delta_0)}\Big)\Big)\cdot \\
&\Big\{\tfrac{2}{t}\sqrt{t\mu\lambda_r}\Big(\mathcal{O}\Big(\tfrac{\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\frac{\lambda_r^\star}{\lambda_1^\star}\sqrt{\frac{t}{r}\lambda_1^\star}}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\sqrt{\tfrac{t}{r}\lambda_r^\star}\tfrac{\Lambda'}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\tfrac{\Lambda}{\sqrt{r\mu^\star}}\sqrt{\tfrac{t}{r}}\Big) + \mathcal{O}\Big(\sigma\sqrt{\tfrac{r\log^2(r\delta^{-1})}{m}}\Big)\Big) \\
&+ \sqrt{\frac{4\zeta}{t}}\sqrt{\mu\lambda_r}\Big(c'\sqrt{\mu^\star\lambda_r^\star}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} + \Lambda\Big) \\
&+ \sqrt{\mu\lambda_r}\Big(\mathcal{O}\Big(\tfrac{\sqrt{\mu^\star\lambda_r^\star}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\tfrac{\Lambda'\|\mathbf{w}^{\star(i)}\|_2}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\tfrac{\Lambda}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}\Big(\sigma\sqrt{\tfrac{r\log^2(r\delta^{-1})}{m}}\Big) \\
&+ \Big(c'\sqrt{\mu^\star\lambda_r^\star}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} + \Lambda\Big)\Big)\sqrt{\frac{d\log(rd/\delta_0)}{mt}} \\
&+ \frac{\sigma_2}{mt}6\sqrt{rd\log(rd)} + \frac{\sigma_1}{mt}\Big(2\sqrt{rd} + 4\sqrt{\log rd}\Big)\sqrt{r} + \frac{2\sigma\sqrt{d\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big\} \\
= {}& J_1' + J_2'
\end{aligned}
\tag{212}
$$

where as before, J_1' denotes the terms which arise from analysing the problem in the noiseless setting and J_2' denotes the contribution of noise terms $(\sigma_1, \sigma_2, \sigma, \Lambda, \Lambda')$. Now, J_1'

$$
\begin{aligned}
&= \sqrt{\frac{\nu^\star}{k}}\frac{r}{t\lambda_r}\cdot\mathcal{O}\Big(\sqrt{\frac{t}{r}}\sqrt{\lambda_1^\star}\Big)\cdot\mathcal{O}\Big(\frac{\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\frac{\lambda_r^\star}{\lambda_1^\star}\sqrt{\frac{t}{r}\lambda_1^\star}}{\sqrt{r\mu^\star}}\Big) \\
&+ \sqrt{r}\mathcal{O}\Big(\frac{r}{\lambda_r}\sqrt{\mu\lambda_r}\Big)\sqrt{\frac{\log(1/\delta_0)}{mt}}\cdot\mathcal{O}\Big(\frac{\sqrt{\mu^\star\lambda_r^\star}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}}{\sqrt{r\mu^\star}}\Big) \\
&+ \zeta c'\sqrt{\mu^\star\lambda_r^\star}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star k}}\sqrt{\mu\lambda_r}\frac{r}{t\lambda_r} + c\frac{r\sqrt{r}}{\lambda_r}c'\sqrt{\mu^\star\lambda_r^\star}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\sqrt{\mu\lambda_r}\sqrt{\frac{\log(1/\delta_0)}{mt}} \\
&+ r\sqrt{r}\Big(c\mu\sqrt{\frac{rd\log(rd/\delta_0)}{mt}} + \frac{\sigma_1}{mt\lambda_r}\Big(2\sqrt{rd} + 2\sqrt{2rd\log(2rd/\delta_0)}\Big)\Big)\cdot \\
&\Big\{\frac{2}{t}\sqrt{t\mu\lambda_r}\cdot\mathcal{O}\Big(\frac{\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\frac{\lambda_r^\star}{\lambda_1^\star}\sqrt{\frac{t}{r}\lambda_1^\star}}{\sqrt{r\mu^\star}}\Big) + \sqrt{\frac{4\zeta}{t}}\sqrt{\mu\lambda_r}\cdot c'\sqrt{\mu^\star\lambda_r^\star}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} \\
&+ 4\sqrt{\mu\lambda_r}\cdot\Big(\mathcal{O}\Big(\frac{\sqrt{\mu^\star\lambda_r^\star}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}}{\sqrt{r\mu^\star}}\Big) + c'\sqrt{\mu^\star\lambda_r^\star}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big)\sqrt{\frac{d\log(rd/\delta_0)}{mt}}\Big\}.
\end{aligned}
$$

Taking $\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star k}}$ common across all terms, using eigenvalue ratio and incoherence bounds, $\mathsf{B}_{\mathbf{U}^{(\ell-1)}} = \mathcal{O}\Big(\frac{1}{\sqrt{r\mu^\star}}\Big)$, and rearranging the terms in the above we get J_1'

$$
\begin{aligned}
&= \mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star k}}\Big\{\sqrt{\nu^\star}\frac{1}{\lambda_r}\cdot\mathcal{O}(\sqrt{\lambda_1^\star})\cdot\mathcal{O}\Big(\frac{\sqrt{\lambda_r^\star}}{\sqrt{r\mu^\star}}\Big) \\
&+ c\sqrt{r}\mathcal{O}\Big(\frac{r}{\lambda_r}\sqrt{\mu\lambda_r}\Big)\cdot\mathcal{O}\Big(\frac{\sqrt{\mu^\star\lambda_r^\star}}{\sqrt{r\mu^\star}}\Big)\cdot\sqrt{\frac{k\log(1/\delta_0)}{mt}} \\
&+ \frac{\zeta}{t}c'\sqrt{\mu^\star\lambda_r^\star}\sqrt{\mu\lambda_r}\frac{r}{\lambda_r} + cc'\frac{r\sqrt{r}}{\lambda_r}\sqrt{\mu^\star\lambda_r^\star}\sqrt{\frac{k\log(1/\delta_0)}{mt}} \\
&+ r\sqrt{r}\sqrt{k}\Big(c\mu\sqrt{\frac{rd\log(rd/\delta_0)}{mt}} + \frac{\sigma_1}{mt\lambda_r}\Big(2\sqrt{rd} + 2\sqrt{2rd\log(2rd/\delta_0)}\Big)\Big)\cdot\mu^\star\lambda_r^\star\cdot \\
&\Big\{\frac{2}{r}\cdot\mathcal{O}\Big(\frac{1}{\mu^\star}\Big) + \sqrt{\frac{4\zeta}{t}} * \cdot\mathcal{O}(1) + 4\cdot\Big(\mathcal{O}\Big(\frac{1}{\sqrt{r\mu^\star}}\Big) + \mathcal{O}(1)\Big)\sqrt{\frac{d\log(rd/\delta_0)}{mt}}\Big\}.
\end{aligned}
$$

Using $\sqrt{\nu^\star} = \mathcal{O}\Big(\frac{1}{\sqrt{\mu^\star}}\frac{\lambda_r^\star}{\lambda_1^\star}\Big)$, $\sqrt{\frac{k}{d}} = \mathcal{O}(1)$, $\sqrt{\frac{r^3 d\log(1/\delta_0)}{mt}} = \mathcal{O}\Big(\frac{1}{\mu^\star}\Big)$, $\sqrt{\frac{r^2\zeta}{t}} = \mathcal{O}\Big(\frac{1}{\mu^\star}\Big)$, $\sqrt{\frac{r^3 d\log(1/\delta_0)}{mt}} = \mathcal{O}\Big(\frac{1}{\sqrt{r}(\mu^\star)^2\sqrt{\mu^\star}\lambda_r^\star\sqrt{k}}\Big)$, $\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big) = \mathcal{O}\Big(\frac{1}{r\sqrt{r}\mu^\star\sqrt{\mu^\star}\sqrt{k}}\Big)$ for the bracketed terms the above becomes

$$
\begin{aligned}
&= \mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star k}}\cdot\mathcal{O}(1) \\
&= \mathcal{O}\Big(\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star k}}\Big). \qquad (213)
\end{aligned}
$$

Similarly, we have J_2'

$$
\begin{aligned}
&= \sqrt{\frac{\nu^\star}{k}}\frac{r}{t\lambda_r}\cdot\mathcal{O}\Big(\sqrt{\frac{t}{r}}\sqrt{\lambda_1^\star}\Big)\Big(\mathcal{O}\Big(\sqrt{\frac{t}{r}\lambda_r^\star}\frac{\Lambda'}{\sqrt{r\mu^\star}}\Big)+\mathcal{O}\Big(\frac{\Lambda}{\sqrt{r\mu^\star}}\sqrt{\frac{t}{r}}\Big)+\mathcal{O}\Big(\sigma\sqrt{\frac{r\log^2(r\delta^{-1})}{m}}\Big)\Big) \\
&+\sqrt{r}\mathcal{O}\Big(\frac{r}{\lambda_r}\sqrt{\mu\lambda_r}\Big)\cdot\Big(\mathcal{O}\Big(\frac{\Lambda'\|\mathbf{w}^{\star(i)}\|_2}{\sqrt{r\mu^\star}}\Big)+\mathcal{O}\Big(\frac{\Lambda}{\sqrt{r\mu^\star}}\Big)+\mathcal{O}\Big(\sigma\sqrt{\frac{r\log^2(r\delta^{-1})}{m}}\Big)\Big)\sqrt{\frac{\log(1/\delta_0)}{mt}} \\
&+\zeta\frac{\Lambda}{\sqrt{k}}\sqrt{\mu\lambda_r}\frac{r}{t\lambda_r}+c\frac{r\sqrt{r}}{\lambda_r}\Lambda\sqrt{\mu\lambda_r}\sqrt{\frac{\log(1/\delta_0)}{mt}}+\frac{2\sigma_2}{mt}\sqrt{\log(rd/\delta_0)}\cdot\frac{r\sqrt{r}}{\lambda_r} \\
&+\frac{r^2\sigma_1\sqrt{rd\cdot 2\log(2r^2d^2/\delta_0)}}{mt\lambda_r}+\frac{2\sigma\sqrt{\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\frac{r}{\lambda_r} \\
&+r\sqrt{r}\Big(c\mu\sqrt{\frac{rd\log(rd/\delta_0)}{mt}}+\frac{\sigma_1}{mt\lambda_r}\Big(2\sqrt{rd}+2\sqrt{2rd\log(2rd/\delta_0)}\Big)\Big)\cdot \\
&\Big\{\frac{2}{t}\sqrt{t\mu\lambda_r}\Big(\mathcal{O}\Big(\sqrt{\frac{t}{r}\lambda_r^\star}\frac{\Lambda'}{\sqrt{r\mu^\star}}\Big)+\mathcal{O}\Big(\frac{\Lambda}{\sqrt{r\mu^\star}}\sqrt{\frac{t}{r}}\Big)+\mathcal{O}\Big(\sigma\sqrt{\frac{r\log^2(r\delta^{-1})}{m}}\Big)\Big) \\
&+\sqrt{\tfrac{4\zeta}{t}}\sqrt{\mu\lambda_r}\Lambda+4\sqrt{\mu\lambda_r}\Big(\mathcal{O}\Big(\tfrac{\Lambda'\|\mathbf{w}^{\star(i)}\|_2}{\sqrt{r\mu^\star}}\Big)+\mathcal{O}\Big(\tfrac{\Lambda}{\sqrt{r\mu^\star}}\Big)+\mathcal{O}\Big(\sigma\sqrt{\tfrac{r\log^2(r\delta^{-1})}{m}}\Big)+\Lambda\Big)\sqrt{\tfrac{d\log(rd/\delta_0)}{mt}} \\
&+\frac{\sigma_2}{mt}6\sqrt{rd\log(rd)}+\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)\sqrt{r}+\frac{2\sigma\sqrt{d\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big\}.
\end{aligned}
$$

Using $\sqrt{\frac{r^3 d\log(1/\delta_0)}{mt}} = \mathcal{O}\Big(\frac{1}{\sqrt{r}(\mu^\star)^2\sqrt{\mu^\star}\lambda_r^\star\sqrt{k}}\Big)$, $\frac{\sigma_1}{mt}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big) = \mathcal{O}\Big(\frac{1}{r\sqrt{r}\mu^\star\sqrt{\mu^\star}\sqrt{k}}\Big)$, eigenvalue ratio bounds and rearranging the terms in above gives

$$
\begin{aligned}
&= \Lambda\Big\{\sqrt{\frac{\nu^\star}{k}}\mathcal{O}\Big(\frac{\sqrt{\lambda_1^\star}}{\lambda_r^\star\sqrt{r\mu^\star}}\Big)+\mathcal{O}\Big(\frac{1}{\sqrt{r\lambda_r^\star}}\sqrt{\frac{r^3\log(1/\delta_0)}{mt}}+\frac{r\zeta}{t}\mathcal{O}\Big(\frac{1\sqrt{\mu^\star}}{\sqrt{k}\sqrt{\lambda_r^\star}}\Big) \\
&+\mathcal{O}\Big(\tfrac{1}{\mu^\star\sqrt{\mu^\star}\lambda_r^\star\sqrt{k}}\Big)\Big(\tfrac{2}{r}\sqrt{\mu\lambda_r}\mathcal{O}\Big(\tfrac{1}{\sqrt{r\mu^\star}}\Big)+\sqrt{\tfrac{4\zeta}{t}}\sqrt{\mu\lambda_r}+4\sqrt{\mu\lambda_r}\mathcal{O}\Big(\tfrac{1}{\sqrt{r\mu^\star}}+1\Big)\sqrt{\tfrac{d\log(rd/\delta_0)}{mt}}\Big)\Big\} \\
&+\Lambda'\Big\{\sqrt{\frac{\nu^\star}{k}}\cdot\mathcal{O}\Big(\sqrt{\frac{\lambda_1^\star}{\lambda_r^\star}}\frac{1}{\sqrt{r\mu^\star}}\Big)+c\sqrt{r}\frac{r}{\lambda_r}\sqrt{\mu\lambda_r}\cdot\mathcal{O}\Big(\frac{\|\mathbf{w}^{\star(i)}\|_2}{\sqrt{r\mu^\star}}\Big) \\
&+\mathcal{O}\Big(\frac{1}{\mu^\star\sqrt{\mu^\star}\lambda_r^\star\sqrt{k}}\Big)\Big(\frac{2}{t}\sqrt{t\mu\lambda_r}\sqrt{\frac{t}{r}\lambda_r^\star}\mathcal{O}\Big(\frac{1}{\sqrt{r\mu^\star}}\Big)+4\sqrt{\mu\lambda_r}\mathcal{O}\Big(\frac{\|\mathbf{w}^{\star(i)}\|_2}{\sqrt{r\mu^\star}}\Big)\Big)\Big\} \\
&+\sigma_1\Big\{\frac{r^2\sqrt{rd\cdot 2\log(2r^2d^2/\delta_0)}}{mt\lambda_r}+\mathcal{O}\Big(\frac{1}{\mu^\star\sqrt{\mu^\star}\lambda_r^\star\sqrt{k}}\Big)\cdot\frac{\sqrt{r}}{mt}\Big(2\sqrt{rd}+4\sqrt{\log rd}\Big)\Big\} \\
&+\sigma_2\Big\{\frac{2}{mt}\sqrt{\log(rd/\delta_0)}\cdot\frac{r\sqrt{r}}{\lambda_r}+\mathcal{O}\Big(\frac{1}{\mu^\star\sqrt{\mu^\star}\lambda_r^\star\sqrt{k}}\Big)\cdot\frac{1}{mt}6\sqrt{rd\log(rd)}\Big\}
\end{aligned}
$$

$$
\begin{aligned}
&+ \sigma\Big\{\sqrt{\frac{\nu^\star}{k}}\frac{r}{t\lambda_r}\cdot\mathcal{O}\Big(\sqrt{\frac{t}{r}}\sqrt{\lambda_1^\star}\Big)\cdot\Big(\sqrt{\frac{r\log^2(r\delta^{-1})}{m}}\Big)\\
&+ c\sqrt{r}\tfrac{r}{\lambda_r}\cdot\mathcal{O}\Big(\sqrt{\tfrac{r\log^2(r\delta^{-1})}{m}}\sqrt{\tfrac{\log(1/\delta_0)}{mt}}\Big) + \tfrac{2\sqrt{\mu^\star\lambda_r^\star}\log(2rdmt/\delta_0)}{\sqrt{mt}}\tfrac{r}{\lambda_r} + \mathcal{O}\Big(\tfrac{1}{\mu^\star\sqrt{\mu^\star}\lambda_r^\star\sqrt{k}}\Big)\cdot\sqrt{\mu^\star\lambda_r^\star}\\
&\quad\Big(\tfrac{2}{\sqrt{t}}\cdot\mathcal{O}\Big(\sqrt{\tfrac{r\log^2(r\delta^{-1})}{m}}\Big) + 4\mathcal{O}\Big(\sqrt{\tfrac{r\log^2(r\delta^{-1})}{m}}\sqrt{\tfrac{d\log(rd/\delta_0)}{mt}}\Big) + \tfrac{2\sigma\sqrt{d}\log(2rdmt/\delta_0)}{\sqrt{mt}}\Big)\Big\}.
\end{aligned}
$$

Taking $\sqrt{\frac{1}{k}}$ common and using $\sqrt{\nu^\star} = \mathcal{O}\Big(\frac{1}{\sqrt{\mu^\star}}\frac{\lambda_r^\star}{\lambda_1^\star}\Big)$, $\sqrt{\frac{r^3 d\log(1/\delta_0)}{mt}} = \mathcal{O}\Big(\frac{1}{\mu^\star}\Big)$, $\sqrt{\frac{k}{d}} = \mathcal{O}(1)$, $\sqrt{\frac{r^2\zeta}{t}} = \mathcal{O}\Big(\frac{1}{\sqrt{\mu^\star}}\Big)$, $\sqrt{\frac{k}{d}} = \mathcal{O}(1)$, $\sqrt{\frac{r^3 d\log(1/\delta_0)}{mt}} = \mathcal{O}\Big(\frac{1}{\mu^\star}\Big)$, $\Lambda = \mathcal{O}(\sqrt{\lambda_r^\star})$, $mt = \Omega(\sigma_2 r(\mu^\star)^{1/2}\sqrt{rd\log d}/\lambda_r^\star)$ and $mt = \widetilde{\Omega}(\sigma^2 dr^3\mu^\star/\lambda_r^\star)$, $t = \widetilde{\Omega}(\zeta(\mu^\star\lambda_r^\star)\max(1, \lambda_r^\star/r))$, eigenvalue ratios and incoherence bounds for $\mathbf{H}^{(\ell)}$ and $\mathbf{W}^{(\ell)}$ from Corollaries 1 and 2 as well as $\lambda_r^\star \le \lambda_1^\star$, $r \ge 1$,$\mu^\star \ge 1$ for the bracketed terms, the above simplifies to

$$
\begin{aligned}
J_2' = \frac{1}{\sqrt{k}}\Big\{&\mathcal{O}(\Lambda' + \mathcal{O}\Big(\frac{\Lambda}{\sqrt{\mu^\star\lambda_r^\star}}\Big) + \mathcal{O}\Big(\frac{\sigma_2 r}{mt\lambda_r^\star}\sqrt{rd\log(rd)}\Big) + \mathcal{O}\Big(\frac{\sigma_1 r^{3/2}}{mt\lambda_r^\star}\sqrt{rd\log rd}\Big)\\
&+ \mathcal{O}\Big(\sigma\sqrt{\frac{r^3 d\mu^\star\log^2(r\delta^{-1})}{mt\lambda_r^\star}}\Big)\Big\} \qquad (214)
\end{aligned}
$$

Using equation 213 and equation 214 in equation 212 gives us the required bound for $\|\mathbf{U}^{(\ell-1)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_{2,\infty}$.

□

Lemma 12. *For some constant $c > 0$ and for any iteration indexed by $\ell > 0, j > 0$, we set*

$$
\Delta^{(\ell-1,j)} := \alpha^{(\ell-1)} + c\sqrt{\frac{\log(d/\delta)}{m}}\Big(\gamma^{(\ell-1,j-1)} + \beta^{(\ell-1)}\Big) + \sigma\sqrt{\frac{\log(d\delta^{-1})}{m}}.
$$

, where $\alpha^{(\ell-1)}$, $\gamma^{(\ell-1,j-1)}$ and $\beta^{(\ell-1)}$ are known upper-bounds on $\|\mathbf{U}^{+(\ell-1)}\mathbf{w}^{(i,\ell-1)} - \mathbf{U}^\star\mathbf{w}^{\star(i)}\|_\infty$, $\|\mathbf{b}^{(i,\ell-1,j-1)} - \mathbf{b}^{\star(i)}\|_2$ and $\|\mathbf{U}^{+(\ell-1)}\mathbf{w}^{(i,\ell-1)} - \mathbf{U}^\star\mathbf{w}^{\star(i)}\|_2$ respectively. Subsequently, we have

$$
\begin{aligned}
&\Big|\Big|\mathbf{b}^{(i,\ell-1,j)} - \mathbf{b}^{\star(i)}\Big|\Big|_\infty \le 2\Delta^{(\ell-1,j)},\\
&\Big|\Big|\mathbf{b}^{(i,\ell-1,j)} - \mathbf{b}^{\star(i)}\Big|\Big|_2 \le 2\sqrt{k}\Delta^{(\ell-1,j)} := \gamma^{(\ell-1,j)},\\
&\textit{and } \mathsf{support}(\mathbf{b}^{(i,\ell-1,j)}) \subseteq \mathsf{support}(\mathbf{b}^{\star(i)})
\end{aligned}
$$

and with probability at least $1 - \delta$.

Proof. It is easy to see that update step of the algorithm gives us $\mathbf{c}^{(i,\ell-1,j)} - \mathbf{b}^{\star(i)}$

$$
\begin{aligned}
&= \Big(\mathbf{I} - \frac{1}{m}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big)\Big(\mathbf{b}^{(i,\ell-1,j-1)} - \mathbf{b}^{\star(i)}\Big)\\
&\quad + \frac{1}{m}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}(\mathbf{U}^\star\mathbf{w}^{\star(i)} - \mathbf{U}^{+(\ell-1)}\mathbf{w}^{(i,\ell-1)}) + \frac{1}{m}(\mathbf{X}^{(i)})^\mathsf{T}\xi^{(i)}\\
\implies &\mathbf{c}^{(i,\ell-1,j)} - \mathbf{b}^{\star(i)} - \mathbf{U}^\star\mathbf{w}^{\star(i)} + \mathbf{U}^{+(\ell-1)}\mathbf{w}^{(i,\ell-1)}\\
&\quad = \Big(\mathbf{I} - \frac{1}{m}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big)\Big(\mathbf{b}^{(i,\ell-1,j-1)} - \mathbf{b}^{\star(i)}\Big)\\
&\quad + \Big(\mathbf{I} - \frac{1}{m}(\mathbf{X}^{(i)})^\mathsf{T}\mathbf{X}^{(i)}\Big)(\mathbf{U}^{+(\ell-1)}\mathbf{w}^{(i,\ell-1)} - \mathbf{U}^\star\mathbf{w}^{\star(i)}) + \frac{1}{m}(\mathbf{X}^{(i)})^\mathsf{T}\xi^{(i)}. \qquad (215)
\end{aligned}
$$

Note that $\big|\big|\frac{1}{m}(\mathbf{X}^{(i)})^\mathsf{T}\xi^{(i)}\big|\big|_\infty \le \sqrt{\frac{\log(d\delta_0^{-1})}{m}}$ with probability at least $1 - \delta_0$. Let $\mathbf{e}_s \in \mathbb{R}^d$ denote the s^{th} basis vector for which the s^{th} coordinate entry is 1 and all other coordinate entries are 0. Then, note that $\left|\left(\mathbf{c}^{(i,\ell-1,j)} - \mathbf{b}^{\star(i)} - \mathbf{U}^{\star}\mathbf{w}^{\star(i)} + \mathbf{U}^{+(\ell-1)}\mathbf{w}^{(i,\ell-1)}\right)_s\right|$

$$\begin{aligned}
&= \left|\mathbf{e}_s^{\mathsf{T}}\Big(\mathbf{I} - \tfrac{1}{m}(\mathbf{X}^{(i)})^{\mathsf{T}}\mathbf{X}^{(i)}\Big)\Big(\mathbf{b}^{(i,\ell-1,j-1)} - \mathbf{b}^{\star(i)}\Big) + \mathbf{e}_s^{\mathsf{T}}\Big(\mathbf{I} - \tfrac{1}{m}(\mathbf{X}^{(i)})^{\mathsf{T}}\mathbf{X}^{(i)}\Big)(\mathbf{U}^{+(\ell-1)}\mathbf{w}^{(i,\ell-1)} - \mathbf{U}^{\star}\mathbf{w}^{\star(i)})\right| \\
&\qquad + \sigma\sqrt{\frac{\log(d\delta^{-1})}{m}} \\
&\le \left|\mathbf{e}_s^{\mathsf{T}}\Big(\mathbf{I} - \frac{1}{m}(\mathbf{X}^{(i)})^{\mathsf{T}}\mathbf{X}^{(i)}\Big)\Big(\mathbf{b}^{(i,\ell-1,j-1)} - \mathbf{b}^{\star(i)}\Big)\right| \\
&\qquad + \left|\mathbf{e}_s^{\mathsf{T}}\Big(\mathbf{I} - \frac{1}{m}(\mathbf{X}^{(i)})^{\mathsf{T}}\mathbf{X}^{(i)}\Big)(\mathbf{U}^{+(\ell-1)}\mathbf{w}^{(i,\ell-1)} - \mathbf{U}^{\star}\mathbf{w}^{\star(i)})\right| + \sigma\sqrt{\frac{\log(d\delta^{-1})}{m}} \\
&\le \left|\frac{1}{m}\mathbf{e}_s^{\mathsf{T}}(\mathbf{X}^{(i)})^{\mathsf{T}}\mathbf{X}^{(i)}(\mathbf{b}^{(i,\ell-1,j-1)} - \mathbf{b}^{\star(i)}) - \mathbf{e}_s^{\mathsf{T}}(\mathbf{b}^{(i,\ell-1,j-1)} - \mathbf{b}^{\star(i)})\right| + \sigma\sqrt{\frac{\log(d\delta^{-1})}{m}} \\
&\qquad + \left|\frac{1}{m}\mathbf{e}_s^{\mathsf{T}}(\mathbf{X}^{(i)})^{\mathsf{T}}\mathbf{X}^{(i)}(\mathbf{U}^{+(\ell-1)}\mathbf{w}^{(i,\ell-1)} - \mathbf{U}^{\star}\mathbf{w}^{\star(i)}) - \mathbf{e}_s^{\mathsf{T}}(\mathbf{U}^{+(\ell-1)}\mathbf{w}^{(i,\ell-1)} - \mathbf{U}^{\star}\mathbf{w}^{\star(i)})\right| \\
&\le c\sqrt{\frac{\log(1/\delta)}{m}}\Big(\|\mathbf{b}^{(i,\ell-1,j-1)} - \mathbf{b}^{\star(i)}\|_2 + \|\mathbf{U}^{+(\ell-1)}\mathbf{w}^{(i,\ell-1)} - \mathbf{U}^{\star}\mathbf{w}^{\star(i)}\|_2\Big) + \sigma\sqrt{\frac{\log(d\delta^{-1})}{m}},
\end{aligned}$$

w.p. $\ge 1-\delta$, where we invoke Lemma 17 in the last step and plugging $\mathbf{a} = \mathbf{e}_s$ and $\mathbf{b} = \mathbf{b}^{(i,\ell-1,j-1)} - \mathbf{b}^{\star(i)}$ and $w_i^{\star}\mathbf{u}^{\star} - w_i^{(\ell-1)}\widehat{\mathbf{u}}^{(\ell-1)}$ for the two terms respectively. Therefore, by taking a union bound over all entries $s \in [d]$, we can conclude that

$$\begin{aligned}
&\left\|\mathbf{c}^{(i,\ell-1,j)} - \mathbf{b}^{\star(i)} - \mathbf{U}^{\star}\mathbf{w}^{\star(i)} + \mathbf{U}^{+(\ell-1)}\mathbf{w}^{(i,\ell-1)}\right\|_{\infty} \\
&\quad \le c\sqrt{\frac{\log(d/\delta)}{m}}\Big(\|\mathbf{b}^{(i,\ell-1,j-1)} - \mathbf{b}^{\star(i)}\|_2 + \|\mathbf{U}^{+(\ell-1)}\mathbf{w}^{(i,\ell-1)} - \mathbf{U}^{\star}\mathbf{w}^{\star(i)}\|_2\Big) + \sigma\sqrt{\frac{\log(d\delta^{-1})}{m}}
\end{aligned}$$

$$\begin{aligned}
&\left\|\mathbf{c}^{(i,\ell-1,j)} - \mathbf{b}^{\star(i)}\right\|_{\infty} \\
&\le \|\mathbf{U}^{+(\ell-1)}\mathbf{w}^{(i,\ell-1)} - \mathbf{U}^{\star}\mathbf{w}^{\star(i)}\|_{\infty} \\
&\quad + c\sqrt{\frac{\log(d/\delta)}{m}}\Big(\|\mathbf{b}^{(i,\ell-1,j-1)} - \mathbf{b}^{\star(i)}\|_2 + \|\mathbf{U}^{+(\ell-1)}\mathbf{w}^{(i,\ell-1)} - \mathbf{U}^{\star}\mathbf{w}^{\star(i)}\|_2 + \sigma\sqrt{\frac{\log(d\delta^{-1})}{m}}\Big) \\
&\triangleq \widehat{\Delta}^{(\ell-1,j)} \\
&\le \underbrace{\alpha^{(\ell-1)} + c\sqrt{\frac{\log(d/\delta)}{m}}\Big(\gamma^{(\ell-1,j-1)} + \beta^{(\ell-1)}\Big) + \sigma\sqrt{\frac{\log(d\delta^{-1})}{m}}}_{\triangleq \Delta^{(\ell-1,j)}}
\end{aligned} \tag{216}$$

w.p. $\ge 1-\delta$. Now, we have

$$\mathbf{b}^{(i,\ell-1,j)} = \mathsf{HT}(\mathbf{c}^{(i,\ell-1,j)}, \Delta^{(\ell-1,j)})$$

$$\implies \mathbf{b}_s^{(i,\ell-1,j)} = \begin{cases} \mathbf{c}_s^{(i,\ell-1,j)} & \text{if } |\mathbf{c}_s^{(i,\ell-1,j)}| > \Delta^{(\ell-1,j)}, \\ 0 & \text{otherwise,} \end{cases} \tag{217}$$

$$\implies |\mathbf{b}_s^{(i,\ell-1,j)} - \mathbf{b}_s^{\star(i)}| = \begin{cases} |\mathbf{c}_s^{(i,\ell-1,j)} - \mathbf{b}_s^{\star(i)}| & \text{if } |\mathbf{c}_s^{(i,\ell-1,j)}| > \Delta^{(\ell-1,j)}, \\ |\mathbf{b}_s^{\star(i)}| & \text{otherwise.} \end{cases} \tag{218}$$

Therefore, by using equation 216 and equation 218, we have $\left\|\mathbf{b}^{(i,\ell-1,j)} - \mathbf{b}^{\star(i)}\right\|_{\infty} \le \Delta^{(\ell-1,j)}$ and $\left\|\mathbf{b}^{(i,\ell-1,j)} - \mathbf{b}^{\star(i)}\right\|_2 \le 2\sqrt{k}\Delta^{(\ell-1,j)}$. Further, from equation equation 216 we have for any coordinate s

$$\left|\Big(\mathbf{c}^{(i,\ell-1,j)} - \mathbf{b}^{\star(i)}\Big)_s\right| \le \Delta^{(\ell-1,j)}.$$

Thus, if $s \notin \mathsf{support}(\mathbf{b}^{\star(i)})$, then the above gives $|\mathbf{c}^{(i,\ell-1,j)}| \le \Delta^{(\ell-1,j)}$. Using this in equation 217 then gives $\mathbf{b}_s^{(i,\ell-1,j)} = 0$, i.e., $\forall\, s \notin \mathsf{support}(\mathbf{b}^{\star(i)}) \implies s \notin \mathsf{support}(\mathbf{b}^{\star(i,\ell-1,j)})$. Hence, $\mathsf{support}(\mathbf{b}^{(i,\ell-1,j)}) \subseteq \mathsf{support}(\mathbf{b}^{\star(i)})$. □

Lemma 13. *Suppose* $c > 0$ *and* $c_1 = c\sqrt{\frac{k\log(d/\delta)}{m}} \le \frac{1}{2}$ *be positive constants. For any iteration indexed by* $\ell > 0$*, after*

$$T^{(\ell)} = \Omega\Big(\ell \max_i \log\Big(\frac{\gamma^{(\ell-1)}}{\epsilon}\Big)\Big)$$

iterations of the inner loop at Step 3 in the ℓ^{th} *iteration of the outer loop, we have*

$$\Big|\Big|\mathbf{b}^{(i,\ell)} - \mathbf{b}^{\star(i)}\Big|\Big|_2 \le 2\varphi^{(i)} + \epsilon \text{ and } \Big|\Big|\mathbf{b}^{(i,\ell)} - \mathbf{b}^{\star(i)}\Big|\Big|_\infty \le \frac{1}{\sqrt{k}}\Big(2\varphi^{(i)} + \epsilon\Big)$$

with probability at least $1 - T^{(\ell)}\delta$*, where* $\varphi^{(i)}$ *is an upperbound on* $\widehat{\varphi}^{(i)}$ *s.t.*

$$\widehat{\varphi}^{(i)} = 2\Big(\sqrt{k}\|\mathbf{U}^{+(\ell-1)}\mathbf{w}^{(i,\ell-1)} - \mathbf{U}^\star\mathbf{w}^{\star(i)}\|_\infty + c_1\|\mathbf{U}^{+(\ell-1)}\mathbf{w}^{(i,\ell-1)} - \mathbf{U}^\star\mathbf{w}^{\star(i)}\|_2 + \sigma\sqrt{\tfrac{k\log(d\delta^{-1})}{m}}\Big)$$
$$\le \varphi^{(i)} = 2\Big(\sqrt{k}\alpha^{(\ell-1)} + c_1\beta^{(\ell-1)} + \sigma\sqrt{\frac{k\log(d\delta^{-1})}{m}}\Big).$$

and $\alpha^{(\ell-1)}$*,* $\gamma^{(\ell-1,j-1)}$ *and* $\beta^{(\ell-1)}$ *denote upperbounds on* $\|\mathbf{U}^{+(\ell-1)}\mathbf{w}^{(i,\ell-1)} - \mathbf{U}^\star\mathbf{w}^{\star(i)}\|_\infty$*,* $\|\mathbf{b}^{(i,\ell-1,j-1)} - \mathbf{b}^{\star(i)}\|_2$ *and* $\|\mathbf{U}^{+(\ell-1)}\mathbf{w}^{(i,\ell-1)} - \mathbf{U}^\star\mathbf{w}^{\star(i)}\|_2$ *respectively. Furthermore, we will also have that* $\mathsf{support}(\mathbf{b}^{(i,\ell)}) \subseteq \mathsf{support}(\mathbf{b}^{\star(i)})$.

Proof. Let $\varphi^{(i)}$ be an upperbound on $\widehat{\varphi}^{(i)}$ where,

$$\widehat{\varphi}^{(i)} = 2\Big(\sqrt{k}\|\mathbf{v}\|_\infty + c_1\|\mathbf{v}\|_2 + \sigma\sqrt{\frac{k\log(d\delta^{-1})}{m}}\Big) \le \varphi^{(i)}.$$

where $\mathbf{v} := \mathbf{U}^{+(\ell-1)}\mathbf{w}^{(i,\ell-1)} - \mathbf{U}^\star\mathbf{w}^{\star(i)}$. Then $\varphi^{(i)} := 2\Big(\sqrt{k}\alpha^{(\ell-1)} + c_1\beta^{(\ell-1)}\Big)$ From Lemma 12, we have for each iteration j,

$$\gamma^{(\ell-1,j)} := \Big|\Big|\mathbf{b}^{(i,\ell-1,j)} - \mathbf{b}^{\star(i)}\Big|\Big|_2 \le \varphi^{(i)} + 2c_1\gamma^{(\ell-1,j-1)} \tag{219}$$

$$\text{and } \Big|\Big|\mathbf{b}^{(i,\ell-1,j)} - \mathbf{b}^{\star(i)}\Big|\Big|_\infty \le \frac{\varphi^{(i)}}{\sqrt{k}} + \frac{2c_1}{\sqrt{k}}\gamma^{(\ell-1,j-1)} \tag{220}$$

with probability at least $1 - \delta_0$, where $c_1 = c\sqrt{\frac{k\log(d/\delta_0)}{m}}$. Therefore after $T^{(\ell)}$ iterations of Step 3 inner loop at the ℓ^{th} iteration of the outer loop, we have using equation 219:

$$\begin{aligned}\Big|\Big|\mathbf{b}^{(i,\ell)} - \mathbf{b}^{\star(i)}\Big|\Big|_2 &= \Big|\Big|\mathbf{b}^{(i,\ell-1,T^{(\ell)})} - \mathbf{b}^{\star(i)}\Big|\Big|_2 \\ &\le \varphi^{(i)} + 2c_1\gamma^{(\ell-1,T^{(\ell)}-1)} \\ &\le \varphi^{(i)} + 2c_1\varphi^{(i)} + (2c_1)^2\gamma^{(\ell-1,T^{(\ell)}-2)} \\ &\dots \\ &\le \varphi^{(i)}(1 + (2c_1)\varphi^{(i)} + (2c_1)^2 + \dots + (2c_1)^{T^{(\ell)}-1}) + (2c_1)^{T^{(\ell)}}\gamma^{(\ell-1,0)} \\ &= \varphi^{(i)}\frac{1-(2c_1)^{T^{(\ell)}}}{1-2c_1} + (2c_1)^{T^{(\ell)}}\gamma^{(\ell-1)} \\ &\le \varphi^{(i)}\frac{1}{1-2c_1} + (2c_1)^{T^{(\ell)}}\gamma^{(\ell-1)}, \end{aligned} \tag{221}$$

w.p. $\geq 1-T^{(\ell)}\delta_0$ where $\gamma^{(\ell-1)} = \gamma^{(\ell-1,0)}$ is the upper bound on $\|\mathbf{b}^{(i,\ell-1)} - \mathbf{b}^{\star(i)}\|_2 = \|\mathbf{b}^{(i,\ell-1,0)} - \mathbf{b}^{\star(i)}\|_2$. Similarly, unfolding equation 220 gives

$$\begin{aligned}
\left|\left|\mathbf{b}^{(i,\ell)} - \mathbf{b}^{\star(i)}\right|\right|_\infty &= \left|\left|\mathbf{b}^{(i,\ell-1,T^{(\ell)})} - \mathbf{b}^{\star(i)}\right|\right|_\infty \\
&\leq \frac{\varphi^{(i)}}{\sqrt{k}} + \frac{2c_1}{\sqrt{k}}\gamma^{(\ell-1,T^{(\ell)}-1)} \\
&\leq \frac{\varphi^{(i)}}{\sqrt{k}} + \frac{2c_1\varphi^{(i)}}{\sqrt{k}} + \frac{(2c_1)^2\gamma^{(\ell-1,T^{(\ell)}-2)}}{\sqrt{k}} \\
&\dots \\
&\leq \frac{\varphi^{(i)}}{\sqrt{k}}(1 + (2c_1)\varphi^{(i)} + (2c_1)^2 + \dots (2c_1)^{T^{(\ell)}-1}) + \frac{(2c_1)^{T^{(\ell)}}\gamma^{(\ell-1,0)}}{\sqrt{k}} \\
&\leq \frac{\varphi^{(i)}}{\sqrt{k}}\frac{1}{1-2c_1} + (2c_1)^{T^{(\ell)}}\frac{\gamma^{(\ell-1)}}{\sqrt{k}} \qquad (222)
\end{aligned}$$

w.p. $\geq 1-T^{(\ell)}\delta_0$. Therefore, if we set $T^{(\ell)} \geq \max_i \frac{1}{\log(1/2c_1)}\Big(\frac{\gamma^{(\ell-1)}}{\epsilon}\Big)$ and $c_1 < \frac{1}{2}$ is sufficiently small then equation 221 gives us

$$\left|\left|\mathbf{b}^{(i,\ell)} - \mathbf{b}^{\star(i)}\right|\right|_2 \leq 2\varphi^{(i)} + \epsilon \qquad (223)$$

and equation 222 gives us

$$\left|\left|\mathbf{b}^{(i,\ell)} - \mathbf{b}^{\star(i)}\right|\right|_\infty \leq \frac{1}{\sqrt{k}}\Big(2\varphi^{(i)} + \epsilon\Big) \qquad (224)$$

w.p. $\geq 1-T^{(\ell)}\delta_0$. Equations equation 223 equation 224 give us the required result. Also, note that we set

$$T^{(\ell)} = \Omega\Big(\ell \max_i \log\Big(\frac{\gamma^{(\ell-1)}}{\epsilon}\Big)\Big) \qquad (225)$$

□

Corollary 6. *Using Corollaries 1, 2, 3, 4 and 5, we have*

$$\begin{aligned}
\left|\left|\mathbf{b}^{(i,\ell+1)} - \mathbf{b}^{\star(i)}\right|\right|_2 &\leq c' \max\{\|\mathbf{w}^{\star(i)}\|_2, \epsilon\}\mathsf{B}_{\mathbf{U}^{(\ell)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} \\
\textit{and } \left|\left|\mathbf{b}^{(i,\ell+1)} - \mathbf{b}^{\star(i)}\right|\right|_\infty &\leq c' \max\{\|\mathbf{w}^{\star(i)}\|_2, \epsilon\}\mathsf{B}_{\mathbf{U}^{(\ell)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star k}}
\end{aligned}$$

with $c' = \max\Big(\mathcal{O}(1), O\Big(\frac{1}{\mathsf{B}_{\mathbf{U}^{(0)}}}\frac{\lambda_1^\star}{\lambda_r^\star}\Big)\Big)$, *and for sufficiently large constants* $\tilde{c}, \hat{c} > 0$

$$\begin{aligned}
\Lambda &= \tilde{c}\Big(\sqrt{\lambda_r^\star\mu^\star}\Big(\frac{\sigma_2 r}{mt\lambda_r^\star} + \frac{\sigma_1 r^{3/2}}{mt\lambda_r^\star}\sqrt{rd\log rd} + \sigma\sqrt{\frac{r^3 d\mu^\star\log^2(r\delta^{-1})}{mt\lambda_r^\star}}\Big) \\
&\quad + \sigma\Big(\sqrt{\frac{r^3\log^2(r\delta^{-1})}{m\lambda_r^\star}}\Big) + \sqrt{\frac{k\log(d\delta^{-1})}{m}}\Big)\Big) \\
\Lambda' &= \hat{c}\Big(\frac{\Lambda}{\sqrt{\mu^\star\lambda_r^\star}}\Big).
\end{aligned}$$

Proof. Using Corollary 5 we have $\|\mathbf{U}^{(\ell)}\mathbf{w}^{(i,\ell)} - \mathbf{U}^\star\mathbf{w}^{\star(i)}\|_\infty$

$$\begin{aligned}
&= \|\mathbf{U}^{(\ell)}\mathbf{w}^{(i,\ell)} - \mathbf{U}^{(\ell)}(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)} + \mathbf{U}^{(\ell)}(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)} - \mathbf{U}^\star\mathbf{w}^{\star(i)}\|_\infty \\
&\leq \|\mathbf{U}^{(\ell)}(\mathbf{w}^{(i,\ell)} - (\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)})\|_\infty + \|(\mathbf{U}^{(\ell)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\|_\infty \\
&\leq \|\mathbf{U}^{(\ell)}\|_{2,\infty}\|\mathbf{w}^{(i,\ell-1)} - (\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)})\|_2 + \|\mathbf{U}^{(\ell)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_{2,\infty}\|(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\|_2 \\
&= \mathcal{O}\Big(\frac{1}{\sqrt{k\mu^\star}}\Big)\mathcal{O}\Big(\frac{\max\{\epsilon, \|\mathbf{w}^{\star(i)}\|_2\}\mathsf{B}_{\mathbf{U}^{(\ell-\mathbf{1})}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}}{\sqrt{r\mu^\star}} + \frac{\Lambda'\|\mathbf{w}^{\star(i)}\|_2}{\sqrt{r\mu^\star}} + \frac{\Lambda}{\sqrt{r\mu^\star}} + \sigma\sqrt{\frac{r\log^2(r\delta^{-1})}{m}}\Big) \\
&\quad + \mathcal{O}\Big(\mathsf{B}_{\mathbf{U}^{(\ell-\mathbf{1})}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star k}} + \frac{1}{\sqrt{k}}\Big\{\Lambda' + \frac{\Lambda}{\sqrt{\mu^\star\lambda_r^\star}} \\
&\quad + \frac{\sigma_2 r}{mt\lambda_r^\star}\sqrt{rd\log(rd)} + \frac{\sigma_1 r^{3/2}}{mt\lambda_r^\star}\sqrt{rd\log rd} + \sigma\sqrt{\frac{r^3 d\mu^\star\log^2(r\delta^{-1})}{mt\lambda_r^\star}}\Big\}\Big)2\|\mathbf{w}^{\star(i)}\|_2 \\
&= \frac{1}{\sqrt{k}}\Big\{\mathcal{O}\Big(\max\{\epsilon, \|\mathbf{w}^{\star(i)}\|_2\}\mathsf{B}_{\mathbf{U}^{(\ell-\mathbf{1})}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big) + \mathcal{O}(\Lambda'\|\mathbf{w}^{\star(i)}\|_2) + \mathcal{O}(\Lambda) \\
&\quad + \mathcal{O}\Big(\frac{\|\mathbf{w}^{\star(i)}\|_2\sigma_2 r}{mt\lambda_r^\star}\sqrt{rd\log(rd)}\Big) + \mathcal{O}\Big(\frac{\|\mathbf{w}^{\star(i)}\|_2\sigma_1 r^{3/2}}{mt\lambda_r^\star}\sqrt{rd\log rd}\Big) \\
&\quad + \mathcal{O}\Big(\sigma\Big(\frac{1}{\sqrt{\mu^\star}}\sqrt{\frac{r\log^2(r\delta^{-1})}{m}} + \|\mathbf{w}^{\star(i)}\|_2\sqrt{\frac{r^3 d\mu^\star\log^2(r\delta^{-1})}{mt\lambda_r^\star}}\Big)\Big)\Big\} \\
&:= \alpha^{(\ell-1)}.
\end{aligned} \tag{226}$$

Similarly, Using Corollaries 3 and 4 we have $\|\mathbf{U}^{(\ell)}\mathbf{w}^{(i,\ell)} - \mathbf{U}^\star\mathbf{w}^{\star(i)}\|_2$

$$\begin{aligned}
&= \|\mathbf{U}^{(\ell)}\mathbf{w}^{(i,\ell)} - \mathbf{U}^{(\ell)}(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)} + \mathbf{U}^{(\ell)}(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)} - \mathbf{U}^\star\mathbf{w}^{\star(i)}\|_2 \\
&\leq \|\mathbf{U}^{(\ell)}(\mathbf{w}^{(i,\ell)} - (\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)})\|_2 + \|(\mathbf{U}^{(\ell)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)})(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\|_2 \\
&\leq \|\mathbf{U}^{(\ell)}\|_2\|\mathbf{w}^{(i,\ell)} - (\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)})\|_2 + \|\mathbf{U}^{(\ell)} - \mathbf{U}^\star\mathbf{Q}^{(\ell-1)}\|_{\mathsf{F}}\|(\mathbf{Q}^{(\ell-1)})^{-1}\mathbf{w}^{\star(i)}\|_2 \\
&\leq (1 + c'')\cdot\mathcal{O}\Big(\frac{\max\{\epsilon, \|\mathbf{w}^{\star(i)}\|_2\}\mathsf{B}_{\mathbf{U}^{(\ell-\mathbf{1})}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}}{\sqrt{r\mu^\star}} + \frac{\Lambda'\|\mathbf{w}^{\star(i)}\|_2}{\sqrt{r\mu^\star}} + \frac{\Lambda}{\sqrt{r\mu^\star}} + \sigma\sqrt{\frac{r\log^2(r\delta^{-1})}{m}}\Big) \\
&\quad + \mathcal{O}\Big(\mathsf{B}_{\mathbf{U}^{(\ell-\mathbf{1})}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} + \Lambda' + \frac{\Lambda}{\sqrt{\mu^\star\lambda_r^\star}} + \frac{\sigma_2 r}{mt\lambda_r^\star}\sqrt{rd\log(rd)} + \frac{\sigma_1 r\sqrt{r}}{mt\lambda_r^\star}\sqrt{rd\log rd} \\
&\quad + \sigma\Big(\sqrt{\frac{r^3 d\mu^\star\log^2(r\delta^{-1})}{mt\lambda_r^\star}} + \sqrt{\frac{r^3\log^2(r\delta^{-1})}{m\lambda_r^\star}}\Big)\Big)2\|\mathbf{w}^{\star(i)}\|_2\Big) \\
&= \mathcal{O}\Big(\max\{\epsilon, \|\mathbf{w}^{\star(i)}\|_2\}\mathsf{B}_{\mathbf{U}^{(\ell-\mathbf{1})}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big) + \mathcal{O}(\Lambda'\|\mathbf{w}^{\star(i)}\|_2) + \mathcal{O}(\Lambda) \\
&\quad + \mathcal{O}\Big(\frac{\|\mathbf{w}^{\star(i)}\|_2\sigma_2 r}{mt\lambda_r^\star}\sqrt{rd\log(rd)}\Big) + \mathcal{O}\Big(\frac{\|\mathbf{w}^{\star(i)}\|_2\sigma_1 r^{3/2}}{mt\lambda_r^\star}\sqrt{rd\log rd}\Big) \\
&\quad + \mathcal{O}\Big(\sigma\Big(\frac{1}{\sqrt{\mu^\star}}\sqrt{\frac{r\log^2(r\delta^{-1})}{m}} + \|\mathbf{w}^{\star(i)}\|_2\Big(\sqrt{\frac{r^3 d\mu^\star\log^2(r\delta^{-1})}{mt\lambda_r^\star}} + \sqrt{\frac{r^3\log^2(r\delta^{-1})}{m\lambda_r^\star}}\Big)\Big)\Big) \\
&:= \beta^{(\ell-1)}.
\end{aligned} \tag{227}$$

Using equation 226 and equation 227, we have:

$$
\begin{aligned}
\varphi^{(i)} &= 2\Big(\sqrt{k}\alpha^{(\ell-1)} + c_1\beta^{(\ell-1)} + \sigma\sqrt{\frac{k\log(d\delta^{-1})}{m}}\Big) && (228)\\
&\leq \mathcal{O}\Big(\max\{\epsilon, \|\mathbf{w}^{\star(i)}\|_2\}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big) + \mathcal{O}(\Lambda'\|\mathbf{w}^{\star(i)}\|_2) + \mathcal{O}(\Lambda)\\
&\quad + \mathcal{O}\Big(\frac{\|\mathbf{w}^{\star(i)}\|_2\sigma_2 r}{mt\lambda_r^\star}\sqrt{rd\log(rd)}\Big) + \mathcal{O}\Big(\frac{\|\mathbf{w}^{\star(i)}\|_2\sigma_1 r^{3/2}}{mt\lambda_r^\star}\sqrt{rd\log rd}\Big)\\
&\quad + \mathcal{O}\Big(\sigma\Big(\frac{1}{\sqrt{\mu^\star}}\sqrt{\frac{r\log^2(r\delta^{-1})}{m}} + \sqrt{\frac{k\log(d\delta^{-1})}{m}}\\
&\qquad + \|\mathbf{w}^{\star(i)}\|_2\Big(\sqrt{\frac{r^3 d\mu^\star\log^2(r\delta^{-1})}{mt\lambda_r^\star}} + \sqrt{\frac{r^3\log^2(r\delta^{-1})}{m\lambda_r^\star}}\Big)\Big)\Big). && (229)
\end{aligned}
$$

Using equation 229 in Lemma 13 and setting $\epsilon' \leftarrow \mathcal{O}\Big(\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\cdot\epsilon\Big)$, we have:

$$
\begin{aligned}
&\Big|\Big|\mathbf{b}^{(i,\ell+1)} - \mathbf{b}^{\star(i)}\Big|\Big|_2 \leq 2\varphi^{(i)} + \epsilon'\\
&= \mathcal{O}\Big(\max\{\epsilon, \|\mathbf{w}^{\star(i)}\|_2\}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big) + \mathcal{O}(\Lambda'\|\mathbf{w}^{\star(i)}\|_2) + \mathcal{O}(\Lambda)\\
&\quad + \mathcal{O}\Big(\frac{\|\mathbf{w}^{\star(i)}\|_2\sigma_2 r}{mt\lambda_r^\star}\sqrt{rd\log(rd)}\Big) + \mathcal{O}\Big(\frac{\|\mathbf{w}^{\star(i)}\|_2\sigma_1 r^{3/2}}{mt\lambda_r^\star}\sqrt{rd\log rd}\Big)\\
&\quad + \mathcal{O}\Big(\sigma\Big(\frac{1}{\sqrt{\mu^\star}}\sqrt{\frac{r\log^2(r\delta^{-1})}{m}} + \sqrt{\frac{k\log(d\delta^{-1})}{m}}\\
&\qquad + \|\mathbf{w}^{\star(i)}\|_2\Big(\sqrt{\frac{r^3 d\mu^\star\log^2(r\delta^{-1})}{mt\lambda_r^\star}} + \sqrt{\frac{r^3\log^2(r\delta^{-1})}{m\lambda_r^\star}}\Big)\Big)\Big). && (230)
\end{aligned}
$$

Recall that from Corollaries 3 and 4, we have

$$
\begin{aligned}
\Big|\Big|\Delta(\mathbf{U}^{+(\ell)}, \mathbf{U}^\star)\Big|\Big|_{\mathsf{F}} &\leq (1+c'')\mathcal{O}\Big(\Big\{\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} + \Lambda'\sqrt{\lambda_r^\star} + \frac{\Lambda}{r} + \frac{\sigma_2 r}{mt\lambda_r^\star}\sqrt{rd\log(rd)}\\
&\quad + \frac{\sigma_1 r\sqrt{r}}{mt\lambda_r^\star}\sqrt{rd\log rd} + \sigma\Big(\sqrt{\frac{r^3 d\mu^\star\log^2(rdmt/\delta_0)}{mt\lambda_r^\star}}\Big)\Big\}\Big) && (231)
\end{aligned}
$$

Therefore, it is sufficient to have for sufficiently large constants $\tilde{c}, \hat{c} > 0$

$$
\begin{aligned}
\Lambda &= \tilde{c}\Big(\sqrt{\lambda_r^\star\mu^\star}\Big(\frac{\sigma_2 r}{mt\lambda_r^\star} + \frac{\sigma_1 r^{3/2}}{mt\lambda_r^\star}\sqrt{rd\log rd} + \sigma\sqrt{\frac{r^3 d\mu^\star\log^2(r\delta^{-1})}{mt\lambda_r^\star}}\Big) && (232)\\
&\quad + \sigma\Big(\sqrt{\frac{r^3\log^2(r\delta^{-1})}{m\lambda_r^\star}}\Big) + \sqrt{\frac{k\log(d\delta^{-1})}{m}}\Big)\Big)\\
\Lambda' &= \hat{c}\Big(\frac{\Lambda}{\sqrt{\mu^\star\lambda_r^\star}}\Big). && (233)
\end{aligned}
$$

such that $\big|\big|\mathbf{b}^{(i,\ell+1)} - \mathbf{b}^{\star(i)}\big|\big|_2 \leq \frac{1}{10}\max\{\epsilon, \|\mathbf{w}^{\star(i)}\|_2\}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} + \Lambda$ and $\big|\big|\Delta(\mathbf{U}^{+(\ell)}, \mathbf{U}^\star)\big|\big|_{\mathsf{F}} \leq \frac{\mathsf{B}_{\mathbf{U}^{(\ell-1)}}}{100}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} + \Lambda'$ which satisfies the induction assumption and therefore completes the proof.

Comparing the contribution of noise-deficit terms on both sides for the next iteration, we also get the value of c' as

$$\begin{aligned}
c''' \max\{\|\mathbf{w}^{\star(i)}\|_2, \epsilon\}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} &:= c' \max\{\|\mathbf{w}^{\star(i)}\|_2, \epsilon\}\mathsf{B}_{\mathbf{U}^{(\ell)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} \\
&\leq \frac{51}{50 * 200} c' \max\{\|\mathbf{w}^{\star(i)}\|_2, \epsilon\}\mathsf{B}_{\mathbf{U}^{(\ell-1)}}\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} \\
\implies c' &:= \frac{50 * 200 * c'''}{51} < 5.
\end{aligned}$$

using sufficiently large m and t to pull down the value of c'''. Combining with the Base Case we have $c' = \max\left(\mathcal{O}(1), O\left(\frac{1}{\mathsf{B}_{\mathbf{U}^{(0)}}}\frac{\lambda_1^\star}{\lambda_r^\star}\right)\right)$. ☐

Theorem (Restatement of Theorem 3 (Parameter Estimation))**.** *Consider the LRS problem equation 2 with all parameters m, t, ζ obeying the bounds stated in Theorem 1 with $\zeta = O\Big(t(r^2(\mu^\star)^2)^{-1}(\frac{\lambda_r^\star}{\lambda_1^\star})^2\Big)$, $k = O\Big(d\cdot(\frac{\lambda_r^\star}{\lambda_1^\star})^2\Big)$, $m = \widetilde{\Omega}\Big(k + r^2\mu^\star\Big(\frac{\lambda_1^\star}{\lambda_r^\star}\Big)^2 + \frac{\sigma^2 r^3}{\lambda_r^\star}\Big)$, $mt = \widetilde{\Omega}\Big(r^3 d\mu^\star\Big(r(\mu^\star)^4(\lambda_r^\star)^2 k + \mu^\star\Big(\frac{\lambda_1^\star}{\lambda_r^\star}\Big)^2 + \mu^\star(\lambda_r^\star)^2 + \sigma^2\Big(1 + \frac{1}{\lambda_r^\star}\Big)\Big)\Big)$ and furthermore, $t = \widetilde{\Omega}\Big((rd)^{3/2}\mu^\star\Big(1 + \lambda_r^\star + \sqrt{rk}(\mu^\star)^{3/2}\lambda_r^\star + \sqrt{\mu^\star}\Big(1 + \frac{(\max_i \|\mathbf{b}^\star(i)\|_2}{\sqrt{\mu^\star\lambda_r^\star}} + \sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big)\Big)\frac{\sqrt{\log(1/\delta)+\epsilon}}{\epsilon}\Big)$. Suppose we run Algorithm 5 for $\mathsf{L} = \log\left(\frac{\lambda_r^\star}{\sigma\sqrt{\lambda_1^\star}}\cdot\sqrt{\frac{mt}{\mu^\star rd}}\right)$ iterations with parameters:*

$$\mathsf{A}_1 = \widetilde{O}(\sqrt{d}), \mathsf{A}_2 = \widetilde{O}(\sqrt{\mu^\star\lambda_r^\star} + (\max_i \|\mathbf{b}^{\star(i)}\|_2)), \mathsf{A}_3 = \widetilde{O}\Big(\lambda_r^\star\sqrt{\frac{\mu^\star}{\lambda_1^\star}}\Big), \mathsf{A}_w = \widetilde{O}(\sqrt{\mu^\star\lambda_r^\star}). \tag{234}$$

Then, w.p. $\geq 1 - O(\delta_0)$, the outputs $\mathbf{U}^{+(\mathsf{L})}, \{\mathbf{b}^{(i,\mathsf{L})}\}_{i\in[t]}$ satisfies:

$$\Big\|(\mathbf{I} - \mathbf{U}^\star(\mathbf{U}^\star)^\mathsf{T})\mathbf{U}^{+(\mathsf{L})}\Big\|_\mathsf{F} = \widetilde{O}\Big(\frac{\sigma}{\sqrt{\mu^\star\lambda_r^\star}}\Big(\mu^\star\sqrt{\frac{r^3 d}{mt}} + \sqrt{\frac{r^3}{m\lambda_r^\star}} + \sqrt{\frac{k}{m}}\Big)\Big) + \frac{\sqrt{k}\eta}{\sqrt{\mu^\star\lambda_r^\star}} \tag{235}$$

$$\Big\|\mathbf{b}^{(i,\mathsf{L})} - \mathbf{b}^{\star(i)}\Big\|_\infty \leq \widetilde{O}\Big(\frac{\sigma}{\sqrt{k}}\Big(\mu^\star\sqrt{\frac{r^3 d}{mt}} + \sqrt{\frac{r^3}{m\lambda_r^\star}} + \sqrt{\frac{k}{m}}\Big)\Big) + \eta, \textit{ for all } i \in [t], \tag{236}$$

where $\sqrt{k}\eta = \widetilde{O}\Big(t^{-1}(\mu^\star)^{3/2}\sqrt{\lambda_r^\star}r\sqrt{d}\Big(1 + \max_{i\in[t]}\frac{\|\mathbf{b}^{\star(i)}\|_2}{\sqrt{\mu^\star\lambda_r^\star}} + \sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}} + rd\Big)\frac{\sqrt{\log(1/\delta)+\epsilon}}{\epsilon}\Big)$.

Proof. We will denote the DP noise by $\sigma_{\mathsf{DP}} > 0$. Using standard gaussian concentration inequalities, we set $\mathsf{A}_1, \mathsf{A}_2, \mathsf{A}_3$ and A_w as written in the theorem statement which ensures that for all i, j, ℓ in $\mathbf{U}$ update of Algorithm, let $\|\mathbf{x}_j^{(i)}\|_2 \leq A_1$, $\|\mathbf{w}^{(i,\ell)}\|_2 \leq A_w$, $|y_j^{(i)}| \leq A_2$, and $\|(\mathbf{x}_j^{(i)})^\top\mathbf{b}^{(i,\ell)}\|_2 \leq A_3$ with probability $1 - \mathcal{O}(\frac{1}{\mathsf{Poly}(mt\mathsf{L})})$. Setting each entry of $\mathbf{N}_1$ independently according to $\mathcal{N}\left(0, m^2 \cdot A_1^4 \cdot A_w^4 \cdot \mathsf{L} \cdot \sigma_{\mathsf{DP}}^2\right)$ ($\sigma_1^2 = m^2 \cdot A_1^4 \cdot A_w^4 \cdot \mathsf{L} \cdot \sigma_{\mathsf{DP}}^2$) and each entry of $\mathbf{N}_2$ is independently set $\mathcal{N}\left(0, m^2 \cdot A_1^2(A_2 + A_3)^2 A_w^2 \cdot \mathsf{L} \cdot \sigma_{\mathsf{DP}}^2\right)$ ($\sigma_2^2 = m^2 \cdot A_1^2(A_2 + A_3)^2 A_w^2 \cdot \mathsf{L} \cdot \sigma_{\mathsf{DP}}^2$) ensures that the algorithm satisfies $\frac{1}{\sigma_{\mathsf{DP}}}^2$-zCDP and equivalently (ϵ, δ) Approximate Differential Privacy if $\sigma_{\mathsf{DP}} \geq \frac{\sqrt{\log(1/\delta)+\epsilon}}{\epsilon}$ (Theorem 2).

Using the bounds on m, t, mt, ζ, k in terms of the ground truth model parameters $\mu^\star, \lambda_1^\star, \lambda_r^\star$ expressed in the theorem statement, we invoke Corollaries 1, 2, 3, 4, 5 and 6 as well as the Base Case C.1 ($\ell = 1$) to show that our Inductive Assumption 4 holds for each iteration of ℓ and complete out proof using the Principle of Induction.

Now, note that the error bound guarantees in 4 have two terms in the upper bounds: the first one (a multiple of $\mathsf{B}_{\mathbf{U}^{(\ell-1)}}$, which stems from analysing the problem in the noiseless setting) decreases exponentially with the number of iterations the second unchanging one (Λ and Λ' depends on the inherent noise σ and DP noise σ_{DP}). Plugging $\mathsf{L} = \log\left(\frac{\lambda_r^\star}{\sigma\sqrt{\lambda_1^\star}} \cdot \sqrt{\frac{mt}{\mu^\star rd}}\right)$ in the geometric series expression, we obtain the guarantees as stated in the main theorem. $\square$

Corollary 7 (Restatement of Theorem 1 (Parameter Estimation))**.** *Consider the LRS problem equation 2 with t linear regression tasks and samples obtained by equation 1. Let model parameters satisfy assumptions A1, A2. Also, let the row sparsity of $\mathbf{B}^\star$ satisfy $\zeta = O\Big(t(r^2(\mu^\star)^2)^{-1}(\frac{\lambda_r^\star}{\lambda_1^\star})^2\Big)$, $k = O\Big(d \cdot (\frac{\lambda_r^\star}{\lambda_1^\star})^2\Big)$, $m = \widetilde{\Omega}\Big(k + r^2\mu^\star\left(\frac{\lambda_1^\star}{\lambda_r^\star}\right)^2 + \frac{\sigma^2 r^3}{\lambda_r^\star}\Big)$. Suppose Algorithm 1 is initialized with $\mathbf{U}^{+(0)}$ such that $\left|\left|(\mathbf{I} - \mathbf{U}^\star(\mathbf{U}^\star)^\mathsf{T})\mathbf{U}^{+(0)}\right|\right|_\mathsf{F} = O\Big(\sqrt{\frac{\lambda_r^\star}{\lambda_1^\star}}\Big)$ and $\left|\left|\mathbf{U}^{+(0)}\right|\right|_{2,\infty} = O(\sqrt{\mu^\star r/d})$, and is run for $\mathsf{L} = \log\left(\frac{\lambda_r^\star}{\sigma\sqrt{\lambda_1^\star}} \cdot \sqrt{\frac{mt}{\mu^\star rd}}\right)$ iterations. Then, w.p. $\geq 1 - O(\delta_0)$, the outputs $\mathbf{U}^{+(\mathsf{L})}, \{\mathbf{b}^{(i,\mathsf{L})}\}_{i\in[t]}$ satisfies:*

$$\left|\left|(\mathbf{I} - \mathbf{U}^\star(\mathbf{U}^\star)^\mathsf{T})\mathbf{U}^{+(\mathsf{L})}\right|\right|_\mathsf{F} = \widetilde{O}\Big(\frac{\sigma}{\sqrt{\mu^\star\lambda_r^\star}}\Big(\mu^\star\sqrt{\frac{r^3 d}{mt}} + \sqrt{\frac{r^3}{m\lambda_r^\star}} + \sqrt{\frac{k}{m}}\Big)\Big) \tag{237}$$

$$\left|\left|\mathbf{b}^{(i,\mathsf{L})} - \mathbf{b}^{\star(i)}\right|\right|_\infty \leq \widetilde{O}\Big(\frac{\sigma}{\sqrt{k}}\Big(\mu^\star\sqrt{\frac{r^3 d}{mt}} + \sqrt{\frac{r^3}{m\lambda_r^\star}} + \sqrt{\frac{k}{m}}\Big)\Big), \textit{ for all } i \in [t], \tag{238}$$

where, the total number of samples satisfies:

$$m = \widetilde{\Omega}\Big(k + r^2\mu^\star\Big(\frac{\lambda_1^\star}{\lambda_r^\star}\Big)^2 + \frac{\sigma^2 r^3}{\lambda_r^\star}\Big)$$
$$mt = \widetilde{\Omega}\Big(r^3 d\mu^\star\Big(r(\mu^\star)^4(\lambda_r^\star)^2 k + \mu^\star\Big(\frac{\lambda_1^\star}{\lambda_r^\star}\Big)^2 + \mu^\star(\lambda_r^\star)^2 + \sigma^2\Big(1 + \frac{1}{\lambda_r^\star}\Big)\Big)\Big).$$

Proof. The proof follows by substituting $\sigma_{\mathsf{DP}} = 0$ (hence $\sigma_1, \sigma_2 = 0$) in the proof of Theorem 3. $\square$

C.3 PROOF OF THEOREM 2

Following along similar lines of proof techniques used for privacy guarantees used in Varshney et al. (2022), our proof will broadly involve computing the Zero Mean Concentrated Differential Privacy (zCDP) parameters and then using them to prove the Approximate Differential Privacy. The Update Step for $\mathbf{U}^{(\ell)}$ without the additive DP Noise is:

$$\widehat{\mathbf{x}_j^{(i)}} \leftarrow \mathsf{clip}\left(\mathbf{x}_j^{(i)}, \mathsf{A}_1\right), \widehat{y_j^{(i)}} \leftarrow \mathsf{clip}\left(y_j^{(i)}, \mathsf{A}_2\right), \widehat{(\mathbf{x}_j^{(i)})^\mathsf{T}\mathbf{b}^{(i,\ell)}} \leftarrow \mathsf{clip}\left((\mathbf{x}_j^{(i)})^\mathsf{T}\mathbf{b}^{(i,\ell)}, \mathsf{A}_3\right)$$
$$\text{and } \widehat{\mathbf{w}^{(i,\ell)}} \leftarrow \mathsf{clip}\left(\mathbf{w}^{(i,\ell)}, \mathsf{A}_w\right) \tag{239}$$

$$\mathbf{A} := \frac{1}{mt}\sum_{i\in[t]}\Big(\widehat{\mathbf{w}^{(i,\ell)}}(\widehat{\mathbf{w}^{(i,\ell)}})^\mathsf{T} \otimes \Big(\sum_{j=1}^{m}\widehat{\mathbf{x}_j^{(i)}}(\widehat{\mathbf{x}_j^{(i)}})^\mathsf{T}\Big)\Big) \tag{240}$$

$$\mathbf{V} := \frac{1}{mt}\sum_{i\in[t]}\sum_{j\in[m]}\widehat{\mathbf{x}_j^{(i)}}\Big(\widehat{y_j^{(i)}} - \widehat{(\mathbf{x}_j^{(i)})^\mathsf{T}\mathbf{b}^{(i,\ell)}}\Big)(\widehat{\mathbf{w}^{(i,\ell)}})^\mathsf{T} \tag{241}$$

$$\mathbf{U}^{(\ell)} \leftarrow \mathsf{vec}_{d\times r}^{-1}(\mathbf{A}^{-1}\mathsf{vec}(\mathbf{V})). \tag{242}$$

where $\mathsf{clip}(\dot{;})$ denotes the clipping function. Therefore, the sensitivity of $\mathbf{A}$ and $\mathbf{V}$ due to samples from i^{th}-task (w.r.t. the Frobenius norm) is $\Gamma_1 = m\mathsf{A}_1^2\mathsf{A}_w^2$, and $\Gamma_2 = m\mathsf{A}_1(\mathsf{A}_2 + \mathsf{A}_3)\mathsf{A}_\mathsf{w}$ respectively. Now, since each entry of $\mathbf{N}^{(1)}$ is independently generated from $\mathcal{N}\left(0, m^2 \cdot \mathsf{A}_1^4 \cdot \mathsf{A}_w^4 \cdot \mathsf{L} \cdot \sigma_{\mathsf{DP}}^2\right)$ and each entry of $\mathbf{N}^{(2)}$ is independently generated from $\mathcal{N}\left(0, m^2 \cdot \mathsf{A}_1^2(\mathsf{A}_2 + \mathsf{A}_3)^2\mathsf{A}_w^2 \cdot \mathsf{L} \cdot \sigma_{\mathsf{DP}}^2\right)$, the update steps equation 240 and equation 241 are $\Big(\rho_{\ell,1} = \frac{\Gamma_1^2}{2\cdot m^2\cdot\mathsf{A}_1^4\cdot\mathsf{A}_w^4\cdot\mathsf{L}\cdot\sigma_{\mathsf{DP}}^2} = \frac{1}{2\mathsf{L}\cdot\sigma_{\mathsf{DP}}^2}\Big)$-zCDP and $\Big(\rho_{\ell,2} = \frac{\Gamma_2^2}{2\cdot m^2\cdot\mathsf{A}_1^4\cdot\mathsf{A}_w^4\cdot\mathsf{L}\cdot\sigma_{\mathsf{DP}}^2} = \frac{1}{2\mathsf{L}\cdot\sigma_{\mathsf{DP}}^2}\Big)$-zCDP respectively by virtue of the DP noise standard deviations Bun & Steinke (2016). Therefore by composition and robustness to post-processing, each iteration step is $\Big(\rho_\ell = \rho_{\ell,1} + \rho_{\ell,2} = \frac{1}{\mathsf{L}\cdot\sigma_{\mathsf{DP}}^2}\Big)$-zCDP. By composition of zCDPs, the overall ρ for the algorithm is given by $\rho = \sum_{\ell=1}^{\mathsf{L}} \rho_\ell = \frac{1}{\sigma_{\mathsf{DP}}^2}$.

Recall ρ-zCDP for an algorithm is equivalent to obtaining a $(\mu, \mu\rho)$-Renyi differential privacy (RDP) Mironov (2017) guarantee. Now, we will optimize for $\mu \in [1, \infty)$ and demonstrate that for the choice of the noise multiplier σ_{DP} mentioned in the theorem statement satisfies (ϵ, δ)-DP. Our analysis is similar to that of Theorem 1 of Chien et al. (2021).

Note that $(\mu, \mu\rho)$-(RDP) $\implies$ (ϵ, δ) Approximate Privacy where $\epsilon = \mu\rho + \frac{\log(1/\delta)}{\mu-1}$ $\forall \mu > 1$. Also note that $\epsilon_{\min} = \rho + 2\sqrt{\rho\log(1/\delta)}$ is attained at $\frac{\mathsf{d}\epsilon}{\mathsf{d}\mu} = 0 \implies \mu = 1 + \sqrt{\frac{\log(1/\delta)}{\rho}}$.

Consider a fixed ϵ. Since we want to minimize σ_{DP} (which scales as $1/\sqrt{\rho}$), we need to compute the maximum permissable ρ s.t. $\epsilon_{\min}(\rho) \leq \epsilon$. Since $\epsilon_{\min}(\rho)$ is an increasing function of ρ (thus an increasing function of σ_{DP}) and a second order polynomial in $\sqrt{\rho}$ with root at $\sqrt{\rho} = \sqrt{\log(1/\delta) + \epsilon_{\min}} - \sqrt{\log(1/\delta)}$, the maximum is achieved at $\epsilon_{\min}(\rho) = \epsilon$. Therefore,

$$\frac{1}{\sigma_{\mathsf{DP}}^2} = (\sqrt{\log(1/\delta) + \epsilon} - \sqrt{\log(1/\delta)})^2 = \frac{\epsilon^2}{(\sqrt{\log(1/\delta) + \epsilon} + \sqrt{\log(1/\delta)})^2}.$$

Since the above value of σ_{DP} satisfies (ϵ, δ)-DP and

$$\frac{\epsilon^2}{(\sqrt{\log(1/\delta) + \epsilon} + \sqrt{\log(1/\delta)})^2} \geq \frac{\epsilon^2}{4(\log(1/\delta) + \epsilon)},$$

choosing $\sigma_{\mathsf{DP}} \geq \frac{2\sqrt{(\log(1/\delta)+\epsilon)}}{\epsilon}$ ensures (ϵ, δ)-DP.

D ALGORITHM AND PROOF OF THEOREM 1 (GENERALIZATION GUARANTEES)

Consider a new task for which we get the samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^{m'}$ i.e. $y_i = \langle \mathbf{x}_i, \mathbf{U}^\star\mathbf{w}^\star + \mathbf{b}^\star \rangle$ for all $i \in [m']$. Suppose we have an estimate $\mathbf{U}^+$ of $\mathbf{U}^\star$ such that $(\mathbf{U}^+)^{\mathsf{T}}\mathbf{U}^+ = \mathbf{I}$ and

$$\Big|\Big|(\mathbf{I} - \mathbf{U}^\star(\mathbf{U}^\star)^{\mathsf{T}})\mathbf{U}^+\Big|\Big|_{\mathsf{F}} \leq \rho,\ \Big|\Big|(\mathbf{I} - \mathbf{U}^\star(\mathbf{U}^\star)^{\mathsf{T}})\mathbf{U}^+\Big|\Big|_{2,\infty} \leq \frac{\rho}{\sqrt{k}} \text{ and } \Big|\Big|\mathbf{U}^+\Big|\Big|_{2,\infty} \leq \sqrt{\frac{\nu}{k}}$$

Algorithm 6 AM-NEW TASK

```
Require: Data $\{(\mathbf{X} \in \mathbb{R}^{m' \times d}, \mathbf{y} \in \mathbb{R}^{m'})\}$, known bounds $||\mathbf{b}^\star||_\infty \leq \mathsf{C}$. Set parameter $\epsilon > 0$ appropriately. Estimate $\mathbf{U}^+$ of $\mathbf{U}^\star$ satisfying $\Big|\Big|(\mathbf{I} - \mathbf{U}^\star(\mathbf{U}^\star)^{\mathsf{T}})\mathbf{U}^+\Big|\Big|_{\mathsf{F}} \leq \rho$. Parameter $\mathsf{A}$.
for $\ell = 1, 2, \dots$ do
    Initialize $\mathbf{w}^{(0)}, \mathbf{b}^{(0)} = \mathbf{0}$. Set $\phi^{(0)} = 2$ since $\Big|\Big|\mathbf{w}^{(0)} - (\mathbf{U}^\star)^{\mathsf{T}}\mathbf{U}^+)^{-1}\mathbf{w}^\star\Big|\Big|_2 \leq \phi^{(0)}||\mathbf{w}^\star||_2 \leq 2||\mathbf{w}^\star||_2$. Set $\gamma^{(0)} \geq ||\mathbf{b}^\star||_\infty$.
    for $i = 1, 2, 3, \dots, t$ do
        Set $T^{(\ell)} = \Omega\Big(\ell \log\Big(\frac{\gamma^{(\ell-1)}}{\epsilon}\Big)\Big)$.
        $\mathbf{w}^{(\ell)} = \Big((\mathbf{X}^{(i)}\mathbf{U}^{+(\ell-1)})^{\mathsf{T}}(\mathbf{X}^{(i)}\mathbf{U}^{+(\ell-1)})\Big)^{-1}\Big((\mathbf{X}^{(i)}\mathbf{U}^{+(\ell-1)})^{\mathsf{T}}(\mathbf{y}^{(i)} - \mathbf{X}^{(i)}\mathbf{b}^{(i,\ell)})\Big)$ {Use a fresh batch of data samples}
        $\mathbf{b}^{(\ell)} \leftarrow \mathsf{OptimizeSparseVector}(\mathbf{X}, \mathbf{y}, \alpha = \mathsf{A} + c_1\phi^{(\ell-1)}||\mathbf{w}^\star||_2 + \frac{2\rho||\mathbf{w}^\star||_2}{\sqrt{k}}, \beta = \mathsf{A} + \phi^{(\ell-1)}||\mathbf{w}^\star||_2 + 2\rho||\mathbf{w}^\star||_2, \gamma = \mathsf{A} + \frac{||\mathbf{w}^\star||_2}{\sqrt{k}}\Big(\phi^{(\ell-1)}c' + ||\mathbf{w}^\star||_2\rho(1+c'')\Big), \mathsf{T} = T^{(\ell)})$ {Use a fresh batch of data samples, constants $c_1, c', c''$ set appropriately.}
        Set $\Phi^{(\ell)} \leftarrow ||\mathbf{w}^\star||_2\Phi^{(\ell-1)}c_3 + 2\rho||\mathbf{w}^\star||_2\Big(1 + c_4\Big) + \mathsf{A}$. {$c_3, c_4$ can be made arbitrarily small by increasing number of samples $m'$.}
    end for
end for
Return $\mathbf{w}^{(\ell)}$ and $\mathbf{b}^{(\ell)}$.
```

for some known parameters ν, ρ. Our goal is to recover the vectors $\mathbf{w}^\star \in \mathbb{R}^r$ and $\mathbf{b}^\star \in \mathbb{R}^d$ satisfying $||\mathbf{b}||_0 \le k$. We will again use an Alternating Minimization algorithm for recovery of $\mathbf{w}^\star, \mathbf{b}^\star$. In the ℓ^{th} iteration, with probability at least $1 - O(\delta/\mathsf{L})$ for $m = \Omega(k\log(d\mathsf{L}\delta^{-1}))$ we have the following updates for some constant $c > 0$, (note that the ℓ^{th} iterates of $\mathbf{w}^\star, \mathbf{b}^\star$ are given by $\mathbf{w}^{(\ell)}, \mathbf{b}^{(\ell)}$).

At the ℓ^{th} iteration, we will denote a known upper bound

$$\left|\left|\mathbf{w}^{(\ell-1)} - \mathbf{Q}^{-1}\mathbf{w}^\star\right|\right|_2 \le \phi^{(\ell-1)}||\mathbf{w}^\star||_2 + 2\sigma\frac{\sqrt{k\log(d\delta^{-1})}}{\sqrt{m}} + 2\sigma\frac{\sqrt{r\log^2(r\delta^{-1})}}{\sqrt{m}}$$

where $\phi^{(\ell)}$ is known. We can use Lemma 13 to have

$$\left|\left|\mathbf{b}^{(\ell)} - \mathbf{b}^\star\right|\right|_2 \le 2\varphi^{(i)} + \epsilon \text{ and } \left|\left|\mathbf{b}^{(\ell)} - \mathbf{b}\right|\right|_\infty \le \frac{1}{\sqrt{k}}\Big(2\varphi^{(i)} + \epsilon\Big)$$

with probability at least $1 - T^{(\ell)}\delta$, where φ is an upperbound on $\widehat{\varphi}$ s.t.

$$\begin{aligned}
&\widehat{\varphi} = 2\Big(\sqrt{k}\|\mathbf{U}^+\mathbf{w}^{(\ell-1)} - \mathbf{U}^\star\mathbf{w}^\star\|_\infty + c_1\|\mathbf{U}^+\mathbf{w}^{(\ell-1)} - \mathbf{U}^\star\mathbf{w}^\star\|_2 + \sigma\frac{\sqrt{k\log(d\delta^{-1})}}{\sqrt{m}}\Big) \\
&\le \varphi = 2\Big(\sqrt{k}\alpha^{(\ell-1)} + c_1\beta^{(\ell-1)} + \sigma\frac{\sqrt{k\log(d\delta^{-1})}}{\sqrt{m}} + 2\sigma\frac{\sqrt{r\log^2(r\delta^{-1})}}{\sqrt{m}}\Big).
\end{aligned}$$

and $\alpha^{(\ell-1)}, \beta^{(\ell-1)}$ denote upper bounds on $\|\mathbf{U}^+\mathbf{w}^{(\ell-1)} - \mathbf{U}^\star\mathbf{w}^\star\|_\infty$, and $\|\mathbf{U}^+\mathbf{w}^{(\ell-1)} - \mathbf{U}^\star\mathbf{w}^\star\|_2$ respectively. Furthermore, we will also have that $\mathsf{support}(\mathbf{b}^{(\ell)}) \subseteq \mathsf{support}(\mathbf{b}^\star)$. We denote $\mathbf{Q} = (\mathbf{U}^\star)^{\mathsf{T}}\mathbf{U}^+$. Using a similar analysis as in Corollary 6, we have:

$$\begin{aligned}
&\|\mathbf{U}^+\mathbf{w}^{(\ell-1)} - \mathbf{U}^\star\mathbf{w}^\star\|_\infty = \|\mathbf{U}^+\mathbf{w}^{(\ell-1)} - \mathbf{U}^{(\ell-1)}\mathbf{Q}^{-1}\mathbf{w}^\star + \mathbf{U}^{(\ell-1)}\mathbf{Q}^{-1}\mathbf{w}^\star - \mathbf{U}^\star\mathbf{w}^\star\|_\infty \\
&\le \|\mathbf{U}^+\mathbf{w}^{(\ell-1)} - \mathbf{Q}^{-1}\mathbf{w}^\star)\|_\infty + \|(\mathbf{U}^{(\ell-1)} - \mathbf{U}^\star\mathbf{Q})\mathbf{Q}^{-1}\mathbf{w}^\star\|_\infty \\
&\le \|\mathbf{U}^+\|_{2,\infty}\|\mathbf{w}^{(\ell-1)} - \mathbf{Q}^{-1}\mathbf{w}^\star\|_2 + \|\mathbf{U}^+ - \mathbf{U}^\star\mathbf{Q}\|_{2,\infty}\|\mathbf{Q}^{-1}\mathbf{w}^\star\|_2 \\
&\le \sqrt{\frac{\nu}{k}}\|\mathbf{w}^{(\ell-1)} - \mathbf{Q}^{-1}\mathbf{w}^\star\|_2 + \frac{2\rho||\mathbf{w}^\star||_2}{\sqrt{k}} \\
&\le \sqrt{\frac{\nu}{k}}\Big(\phi^{(\ell-1)}||\mathbf{w}^\star||_2 + 2\sigma\frac{\sqrt{k\log(d\delta^{-1})}}{\sqrt{m}} + 2\sigma\frac{\sqrt{r\log^2(r\delta^{-1})}}{\sqrt{m}}\Big) + \frac{2\rho||\mathbf{w}^\star||_2}{\sqrt{k}} \\
&:= \alpha^{(\ell-1)}.
\end{aligned} \tag{243}$$

Similarly, we have:

$$\begin{aligned}
&\|\mathbf{U}^{(\ell-1)}\mathbf{w}^{(\ell-1)} - \mathbf{U}^\star\mathbf{w}^\star\|_2 \\
&= \|\mathbf{U}^+\mathbf{w}^{(\ell-1)} - \mathbf{U}^{(\ell-1)}\mathbf{Q}^{-1}\mathbf{w}^\star + \mathbf{U}^{(\ell-1)}\mathbf{Q}^{-1}\mathbf{w}^\star - \mathbf{U}^\star\mathbf{w}^\star\|_2 \\
&\le \|\mathbf{U}^+\mathbf{w}^{(\ell-1)} - \mathbf{Q}^{-1}\mathbf{w}^\star)\|_2 + \|(\mathbf{U}^{(\ell-1)} - \mathbf{U}^\star\mathbf{Q})\mathbf{Q}^{-1}\mathbf{w}^\star\|_2 \\
&\le \|\mathbf{U}^+\|_2\|\mathbf{w}^{(\ell-1)} - \mathbf{Q}^{-1}\mathbf{w}^\star\|_2 + \|\mathbf{U}^+ - \mathbf{U}^\star\mathbf{Q}\|_2\|\mathbf{Q}^{-1}\mathbf{w}^\star\|_2 \\
&\le \|\mathbf{w}^{(\ell-1)} - \mathbf{Q}^{-1}\mathbf{w}^\star\|_2 + 2\rho||\mathbf{w}^\star||_2 \\
&\le \phi^{(\ell-1)}||\mathbf{w}^\star||_2 + 2\rho||\mathbf{w}^\star||_2 + 2\sigma\frac{\sqrt{k\log(d\delta^{-1})}}{\sqrt{m}} + 2\sigma\frac{\sqrt{r\log^2(r\delta^{-1})}}{\sqrt{m}} \\
&:= \beta^{(\ell-1)}.
\end{aligned} \tag{244}$$

Using equation 243 and equation 244, we have:

$$\varphi = \phi^{(\ell-1)}||\mathbf{w}^\star||_2(2\sqrt{\nu} + c_1) + ||\mathbf{w}^\star||_2(4\rho + 4c_1\rho) + \sigma\frac{\sqrt{k\log(d\delta^{-1})}}{\sqrt{m}}\Big)$$

since $c_1 < \frac{1}{40}$, $\sqrt{\nu} \le \frac{1}{40}$ and $\rho \le \frac{1}{80}$.

Using above in Lemma 6 and setting $\epsilon \leftarrow \varphi^{(i)}$, we have:

$$\Big|\Big|\mathbf{b}^{(\ell)} - \mathbf{b}^{\star}\Big|\Big|_2 \leq 3||\mathbf{w}^{\star}||_2\Big(\phi^{(\ell-1)}(2\sqrt{\nu} + c_1) + 4\rho(1+c_1)\Big) + 3\sigma\frac{\sqrt{k\log(d\delta^{-1})}}{\sqrt{m}} + 3\sigma\frac{\sqrt{r\log^2(r\delta^{-1})}}{\sqrt{m}}\Big)$$

Similarly, we will also have from our updates (with $\mathbf{S} = \frac{1}{m}\sum_{i=1}^{m'}\mathbf{x}_i(\mathbf{x}_i)^{\mathsf{T}}$).

$$\mathbf{w}^{(\ell)} - \mathbf{Q}^{-1}\mathbf{w}^{\star} = \Big(\mathbf{U}^{+\mathsf{T}}\mathbf{S}\mathbf{U}^{+}\Big)^{-1}\Big(\mathbf{U}^{+\mathsf{T}}\mathbf{S}(\mathbf{b}^{\star} - \mathbf{b}^{(\ell)}) + \mathbf{U}^{+\mathsf{T}}\mathbf{S}(\mathbf{U}^{\star}\mathbf{Q} - \mathbf{U}^{+})\mathbf{Q}^{-1}\mathbf{w}^{\star}\Big) + \Big(\mathbf{U}^{+\mathsf{T}}\mathbf{S}\mathbf{U}^{+}\Big)^{-1}\Big(\mathbf{U}^{\mathsf{T}}(\mathbf{X}^{(i)})^{\mathsf{T}}\mathbf{z}^{(i)}\Big)$$

We already know by using an ϵ-net argument that $\Big|\Big|\Big(\mathbf{U}^{+\mathsf{T}}\mathbf{S}\mathbf{U}^{+}\Big)^{-1}\Big|\Big|_2 \leq 2$. We also know that

$$\Big|\Big|\mathbb{E}\mathbf{U}^{+\mathsf{T}}\mathbf{S}(\mathbf{b}^{\star} - \widehat{\mathbf{b}}^{(\ell)})\Big|\Big|_2 \leq \sqrt{\nu}\Big|\Big|\mathbf{b}^{\star} - \mathbf{b}^{(\ell)}\Big|\Big|_2$$

$$\Big|\Big|\mathbb{E}\mathbf{U}^{+\mathsf{T}}\mathbf{S}(\mathbf{U}^{\star}\mathbf{Q} - \mathbf{U}^{+})\mathbf{Q}^{-1}\mathbf{w}^{\star}\Big|\Big|_2 = \Big|\Big|\mathbf{U}^{+\mathsf{T}}(\mathbf{U}^{\star}\mathbf{Q} - \mathbf{U}^{+})\mathbf{Q}^{-1}\mathbf{w}^{\star}\Big|\Big|_2 \leq 2\rho||\mathbf{w}^{\star}||_2$$

and moreover,

$$\Big|\Big|\mathbf{U}^{\star\mathsf{T}}(\mathbf{S}-\mathbf{I})(\mathbf{b}^{\star} - \mathbf{b}^{(\ell)})\Big|\Big|_2 \leq \Big|\Big|\mathbf{b}^{\star} - \mathbf{b}^{(\ell)}\Big|\Big|_2\sqrt{\frac{r\log\delta^{-1}}{m}}$$

$$\begin{aligned}\Big|\Big|\mathbf{U}^{\star\mathsf{T}}(\mathbf{S}-\mathbf{I})(\mathbf{U}^{\star}\mathbf{Q} - \mathbf{U}^{+})\mathbf{Q}^{-1}\mathbf{w}^{\star}\Big|\Big|_2 &\leq \Big|\Big|\mathbf{U}^{+\mathsf{T}}(\mathbf{U}^{\star}\mathbf{Q} - \mathbf{U}^{+})\mathbf{Q}^{-1}\mathbf{w}^{\star}\Big|\Big|_2\sqrt{\frac{r\log\delta^{-1}}{m}} \\ &\leq 2\rho\sqrt{\frac{r\log\delta^{-1}}{m}}||\mathbf{w}^{\star}||_2\end{aligned}$$

$$\Big|\Big|\Big(\mathbf{U}^{+\mathsf{T}}\mathbf{S}\mathbf{U}^{+}\Big)^{-1}\Big(\mathbf{U}^{\mathsf{T}}(\mathbf{X}^{(i)})^{\mathsf{T}}\mathbf{z}^{(i)}\Big)\Big|\Big|_2 \leq \frac{\sigma\sqrt{r\log^2(r\delta^{-1})}}{\sqrt{m}}$$

with probability at least $1 - \delta/\mathsf{L}$. Hence, we get that $\Big|\Big|\mathbf{w}^{(\ell)} - \mathbf{Q}^{-1}\mathbf{w}^{\star}\Big|\Big|_2 \leq 3||\mathbf{w}^{\star}||_2\Big(\phi^{(\ell-1)}(2\sqrt{\nu} + c_1) + 4\rho(1+c_1)\Big)\Big(\sqrt{\nu} + \sqrt{\frac{r\log\delta^{-1}}{m}}\Big) + 2\rho||\mathbf{w}^{\star}||_2\Big(1 + \sqrt{\frac{r\log\delta^{-1}}{m}}\Big) + 2\frac{\sigma\sqrt{r\log^2(r\delta^{-1})}}{\sqrt{m}} + 2\frac{\sigma\sqrt{k\log(d\delta^{-1})}}{\sqrt{m}}$. Therefore, for $m' = \Omega\Big(\max(k\log(d\mathsf{L}\delta^{-1}), r\log\delta^{-1})\Big)$, we get a decrease along with a bias term. We can have $\phi^{(0)} = 2||\mathbf{w}^{\star}||_2$ by using $\mathbf{w}^{(0)} = 0$. After L iterations, we will get $\Big|\Big|\mathbf{w}^{(\mathsf{L})} - \mathbf{Q}^{-1}\mathbf{w}^{\star}\Big|\Big|_2 = O\Big(\rho||\mathbf{w}^{\star}||_2 + c'^{\mathsf{L}-1}||\mathbf{w}^{\star}||_2 + \frac{\sigma\sqrt{k\log(d\delta^{-1})}}{m'}\Big)$; hence, we will have with $\mathsf{L} = O\Big(\log\rho^{-1}\Big)$ that $\Big|\Big|\mathbf{w}^{(\mathsf{L})} - \mathbf{Q}^{-1}\mathbf{w}^{\star}\Big|\Big|_2 = O\Big(\rho||\mathbf{w}^{\star}||_2 + \frac{\sigma\sqrt{k\log(d\delta^{-1})}}{\sqrt{m}} + \frac{\sigma\sqrt{r\log^2(r\delta^{-1})}}{\sqrt{m}}\Big)$. The generalization error is given by

$$\mathcal{L}(\mathbf{U}^{+}, \mathbf{w}^{(\mathsf{L})}, \mathbf{b}^{(\mathsf{L})}) - \mathcal{L}(\mathbf{U}^{\star}, \mathbf{w}^{\star}, \mathbf{b}^{\star})$$

where $\mathcal{L}(\mathbf{U}, \mathbf{w}, \mathbf{b}) \triangleq \mathbb{E}_{(\mathbf{x},y)}(y - \langle\mathbf{x}, \mathbf{U}\mathbf{w} + \mathbf{b}\rangle)^2$. Hence, we have that

$$\mathcal{L}(\mathbf{U}^{+}, \mathbf{w}^{(\mathsf{L})}, \mathbf{b}^{(\mathsf{L})}) - \mathcal{L}(\mathbf{U}^{\star}, \mathbf{w}^{\star}, \mathbf{b}^{\star}) \leq \widetilde{O}\Big(\rho^2||\mathbf{w}^{\star}||_2^2 + \frac{\sigma^2(r+k)}{m}\Big). \tag{245}$$

Theorem 5 (Restatement of Theorem 3 (Generalization properties in private setting))**.** *Generalization error for a new task scales as:*

$$\begin{aligned}&\mathcal{L}(\mathbf{U}, \mathbf{w}, \mathbf{b}) - \mathcal{L}(\mathbf{U}^{\star}, \mathbf{w}^{\star}, \mathbf{b}^{\star}) \\ &= \widetilde{O}\Big(\sigma^2\Big(\frac{r^3 d(\mu^{\star})^2}{mt} + \frac{r^3}{m\lambda_r^{\star}} + \frac{k+r}{m}\Big) + \frac{dr^2(\log(1/\delta)+\epsilon)(\lambda_r^{\star}\mu^{\star})^2}{\epsilon^2 t^2}\cdot(\kappa^2 + r^2 d^2)\Big)\end{aligned}$$

where $\kappa = 1 + \sqrt{\frac{\lambda_r^{\star}}{\lambda_1^{\star}}} + \max_{i\in[t]}\frac{\|\mathbf{b}^{\star(i)}\|_2}{\sqrt{\mu^{\star}\lambda_r^{\star}}}$.

Proof. We assume that $||\mathbf{w}^\star||_2 \leq \sqrt{\mu^\star \lambda_r^\star}$ due to the incoherence (see Assumption A2). We substitute ρ to be the guarantee that we had obtained in Theorem 3; hence we immediately obtain our desired guarantees by using equation 245. $\square$

Corollary 8 (Restatement of Theorem 1 (Generalization Properties in non-private setting))**.** *Furthermore, for a new task, Algorithm 6 ensures the following generalization error bound:*

$$\mathcal{L}(\mathbf{U}, \mathbf{w}, \mathbf{b}) - \mathcal{L}(\mathbf{U}^\star, \mathbf{w}^\star, \mathbf{b}^\star) = \widetilde{O}\Big(\sigma^2 \Big(\frac{r^3 d(\mu^\star)^2}{mt} + \frac{r^3}{m\lambda_r^\star} + \frac{k+r}{m}\Big)\Big).$$

Proof. The proof follows again by substituting $\sigma_{\mathsf{DP}} = 0$ (hence $\sigma_1, \sigma_2 = 0$) which removes the last term in the generalization properties in Theorem 3. $\square$

E DISCUSSION ON OBTAINING INITIAL ESTIMATES USING METHOD OF MOMENTS

Overview: Note that Algorithm 1 has local convergence properties as described in Theorem 1. In practice, typically we use random initialization for $\mathbf{U}^{+(0)}$. However, similar to the representation learning framework Tripuraneni et al. (2021), we can use the Method of Moments to obtain a good initialization. i.e. when the representation matrix $\mathbf{U}^\star$ is of rank r, we can compute the Singular Value Decomposition (SVD) of the matrix $(mt)^{-1}\sum_{i\in[t]}(y_j^{(i)})^2\mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}$. This is similar to the Method of Moments technique used in Tripuraneni et al. (2021) and has been used as an initialization technique in the AM framework of Thekumparampil et al. (2021) as well. Even in the presence of additional sparse vectors, the SVD decomposition is robust. Such a phenomenon has been also been characterized theoretically in the robust PCA setting Netrapalli et al. (2014).

Details for Rank-1: Assume $\|\mathbf{u}^\star\|_2 = \|\mathbf{w}^\star\|_2 = 1$ and $\|\mathbf{b}^{\star(i)}\|_0 \leq k$ for all $i \in [t]$. Moreover, for some constant $\mu > 0$, we will have $\|\mathbf{u}^\star\|_\infty \leq \sqrt{\mu/d}$, $\|\mathbf{w}^\star\|_\infty \leq \sqrt{\mu/t}$, $\max_{i\in[t]}\|\mathbf{B}\|_\infty \leq \mu/\sqrt{dt}$ where $\mathbf{B}$ is the matrix whose columns correspond to $\mathbf{b}^{\star(i)}$'s. Suppose, we obtain samples $(\mathbf{x}, y) \in \mathbb{R}^d \times \mathbb{R}$ where each sample is randomly generated from the t data generating models corresponding to each task. In order to generate the i^{th} sample we first draw a latent variable $j \sim_U [t]$ and subsequently generate the tuple according to the following process:

$$\mathbf{x}^{(i)} \mid j \sim \mathcal{N}(0, \mathbf{I}_d) \text{ and } y^{(i)} \mid \mathbf{x}^{(i)}, j \sim \mathcal{N}(\langle \mathbf{x}^{(i)}, w_j^\star \mathbf{u}^\star + \mathbf{b}^{\star(j)}\rangle, \sigma^2) \tag{246}$$

We look at the quantity $y^2\mathbf{x}\mathbf{x}^\mathsf{T}$. Our first result is the following lemma:

Lemma 14. *Suppose we obtain samples $\{(\mathbf{x}^{(i)}, y^{(i)})\}$ generated according to the model described in equation 246. In that case we have*

$$\mathbb{E}\left[y^2\mathbf{x}\mathbf{x}^\mathsf{T}\right] = \mathbf{I} + \frac{2}{t}\sum_j \Big(w_j^\star \mathbf{u}^\star + \mathbf{b}^{\star(j)}\Big)\Big(w_j^\star \mathbf{u}^\star + \mathbf{b}^{\star(j)}\Big)^\mathsf{T} \tag{247}$$

where $\mathbf{I}$ denotes the d-dimensional identity matrix.

The proof follows from simple calculations. From the data $\{(\mathbf{x}_j^{(i)}, y_j^{(i)})\}_{j=1}^m$ for the i^{th} task, we can compute an unbiased estimate $\mathbf{A} \triangleq \frac{1}{mt}\sum_{i=1}^t\sum_{j=1}^m (y_j^{(i)})^2\mathbf{x}_j^{(i)}(\mathbf{x}_j^{(i)})^\mathsf{T}$ of the matrix $\mathbb{E}\left[y^2\mathbf{x}\mathbf{x}^\mathsf{T}\right]$. Let us write $\mathbf{A} = \mathbb{E}[\mathbf{A}] + 2\mathbf{F}$ where $2\mathbf{F}$ is the error in estimating $\mathbb{E}[\mathbf{A}]$. Also, let us denote $0.5t(\mathbb{E}\left[y^2\mathbf{x}\mathbf{x}^\mathsf{T}\right] - \mathbf{I}) \triangleq \mathbf{L}$. In that case, we will have $0.5t(\mathbf{A} - \mathbf{I}) = 0.5t(\mathbf{A} - \mathbb{E}[\mathbf{A}] + \mathbb{E}[\mathbf{A}] - \mathbf{I}) = \mathbf{L} + \mathbf{F}$. We will also denote

$$\mathbf{E} \triangleq \underbrace{\sum_{j=1}^t \Big(w_j^\star \mathbf{b}^{\star(j)}(\mathbf{u}^\star)^\mathsf{T} + w_i^\star \mathbf{u}^\star(\mathbf{b}^{\star(j)})^\mathsf{T} + \mathbf{b}^{\star(j)}(\mathbf{b}^{\star(j)})^\mathsf{T}\Big)}_{\mathbf{G}} + \mathbf{F}.$$

Our goal is to show that any eigenvector of $\mathbf{L} + \mathbf{F}$ is close to $\mathbf{u}^\star$ in infinity norm. Note that $(\mathbf{L} + \mathbf{F})\mathbf{z} = (\mathbf{u}(\mathbf{u}^\star)^\mathsf{T} + \mathbf{E})\mathbf{z} = \lambda\mathbf{z}$. Hence, we have

$$\mathbf{z} = \Big(\mathbf{I} - \frac{\mathbf{E}}{\lambda}\Big)^{-1}\frac{\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\mathbf{z}}{\lambda}. \tag{248}$$

Fist, note that

$$\lambda \mathbf{z}\mathbf{z}^\mathsf{T} - \mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T} = \frac{\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\mathbf{z}\mathbf{z}^\mathsf{T}\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}}{\lambda} + \sum_{p,q:p+q\geq 1} \frac{\mathbf{E}^p\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\mathbf{z}\mathbf{z}^\mathsf{T}\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\mathbf{E}^q}{\lambda^{p+q+1}}$$

$$\|\lambda \mathbf{z}\mathbf{z}^\mathsf{T} - \mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\|_\infty = \|\tfrac{\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\mathbf{z}\mathbf{z}^\mathsf{T}\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}}{\lambda} - \mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\|_\infty + \|\textstyle\sum_{p,q:p+q\geq 1} \tfrac{\mathbf{E}^p\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\mathbf{z}\mathbf{z}^\mathsf{T}\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\mathbf{E}^q}{\lambda^{p+q+1}}\|_\infty.$$

We have that

$$\begin{aligned}
&\Big|\Big|\frac{\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\mathbf{z}\mathbf{z}^\mathsf{T}\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}}{\lambda} - \mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\Big|\Big|_\infty \\
=&\max_{ij} \mathbf{e}_i^\mathsf{T}\Big(\frac{\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\mathbf{z}\mathbf{z}^\mathsf{T}\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}}{\lambda} - \mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\Big)\mathbf{e}_j \\
=&\mathbf{e}_i^\mathsf{T}\Big(\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T} + \mathbf{U}_\perp^\star(\mathbf{U}_\perp^\star)^\mathsf{T}\Big)\Big(\frac{\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\mathbf{z}\mathbf{z}^\mathsf{T}\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}}{\lambda} - \mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\Big)\Big(\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T} + \mathbf{U}_\perp^\star(\mathbf{U}_\perp^\star)^\mathsf{T}\Big)\mathbf{e}_j \\
\leq&||\mathbf{u}^\star||_\infty^2\Big|\Big|\frac{\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\mathbf{z}\mathbf{z}^\mathsf{T}\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}}{\lambda} - \mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\Big|\Big|_2 \leq \frac{\mu^2}{d}\Big(\frac{((\mathbf{u}^\star)^\mathsf{T}\mathbf{z})^2}{\lambda} - 1\Big)
\end{aligned}$$

where $\mathbf{U}_\perp^\star$ is the subspace orthogonal to the vector $\mathbf{u}^\star$.

First, we will show an upper bound on $||\mathbf{F}||_\infty$. Recall that according to the data generating mechanism, each co-variate $\mathbf{x}$ is generated according to $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and given the co-variate, the response $y \mid \mathbf{x} \sim \mathcal{N}(\langle \mathbf{x}, w^\star\mathbf{u}^\star + \mathbf{b}^\star\rangle, \sigma^2)$ where $w^\star, \mathbf{b}^\star$ is uniformly chosen at random from the set $\{w_j^\star, \mathbf{b}^{\star(j)}\}_j$. Hence, we can bound the magnitude of y as follows: $Ey^2|\mathbf{x} = t^{-1}\sum_{j=1}^t \sigma^2 + \langle \mathbf{x}, w_j^\star\mathbf{u}^\star + \mathbf{b}^{\star(j)}\rangle^2$ and therefore $\mathbb{E}\left[y^2\right] = t^{-1}\sum_{j=1}^t \sigma^2 + \left|\left|w_j^\star\mathbf{u}^\star + \mathbf{b}^{\star(j)}\right|\right|_2^2$. Hence, $y \sim t^{-1}\sum_{j=1}^t \mathcal{N}(0, \sigma^2 + \left|\left|w_j^\star\mathbf{u}^\star + \mathbf{b}^{\star(j)}\right|\right|_2^2)$ and therefore, by using standard Gaussian concentration, we will have $|y| \leq \sqrt{\sigma^2 + \max_{j\in[t]}\left|\left|w_j^\star\mathbf{u}^\star + \mathbf{b}^{\star(j)}\right|\right|_2^2}\log(mt) \leq \sqrt{\sigma^2 + 4\mu t^{-1}}\log(mt)$ for all mt samples w.p. at least $1 - \mathsf{poly}((mt)^{-1})$. Moreover, $|\mathbf{x}_{p,j}^{(i)}| \leq \log(dmt)$ for all $i \in [t], p \in [m], j \in [d]$. Hence, with probability at least $1 - \mathsf{poly}((dmt)^{-1})$, by using standard concentration inequalities, we have $||\mathbf{F}||_\infty \leq \sqrt{\frac{\sigma^2+4\mu t^{-1}}{mt}}\log^3(dmt)$. We will now bound $||\mathbf{E}||_2 \leq ||\mathbf{G}||_2 + ||\mathbf{F}||_2$. In order to do so, fix unit vectors $\mathbf{x}, \mathbf{y}$ such that $||\mathbf{E}||_2 = \mathbf{x}^\mathsf{T}\mathbf{E}\mathbf{y} = \sum_{is} x_i y_s \mathbf{E}_{is} = \frac{1}{2}\sum_{is}(x_i^2 + y_s^2)\mathbf{E}_{is}$. We have the following:

$$\begin{aligned}
&\sum_i x_i^2 \sum_{j=1}^t \sum_{s=1}^d \mathbf{b}_i^{\star(j)}\mathbf{b}_s^{\star(j)} \leq \zeta k||\mathbf{B}||_\infty^2 \quad \text{and} \quad \sum_s y_s^2 \sum_{j=1}^t \mathbf{b}_s^{\star(j)} \sum_{i=1}^d \mathbf{b}_i^{\star(j)} \leq \zeta k||\mathbf{B}||_\infty^2 \leq \frac{\mu^2\zeta k}{dt} \\
\implies &\sum_i x_i^2 \sum_{j=1}^t w_j^\star \mathbf{b}_i^{\star(j)} \sum_{s=1}^d \mathbf{u}_s^\star \leq \zeta d||\mathbf{B}||_\infty||\mathbf{u}||_\infty||\mathbf{w}||_\infty \\
\text{and} \quad &\sum_s y_s^2 \sum_{j=1}^t w_j^\star \mathbf{u}_s^\star \sum_{i=1}^d \mathbf{b}_i^{\star(j)} \leq kt||\mathbf{B}||_\infty||\mathbf{u}||_\infty||\mathbf{w}||_\infty \leq \frac{\mu^2 k}{d} \\
\implies &\sum_i x_i^2 \sum_{j=1}^t w_j^\star \mathbf{u}_i^{\star(j)} \sum_{s=1}^d \mathbf{b}_s^{\star(j)} \leq kt||\mathbf{B}||_\infty||\mathbf{u}||_\infty||\mathbf{w}||_\infty \\
\text{and} \quad &\sum_s y_s^2 \sum_{j=1}^t w_j^\star \mathbf{b}_s^{\star(j)} \sum_{i=1}^d \mathbf{u}_i^\star \leq \zeta d||\mathbf{B}||_\infty||\mathbf{u}||_\infty||\mathbf{w}||_\infty \leq \frac{\mu^2\zeta}{t}.
\end{aligned}$$

and similarly $||\mathbf{F}||_2 \leq \sqrt{d}||\mathbf{F}||_\infty$. Hence $||\mathbf{F}||_2 \leq 1/800$ provided $mt = \Omega(d\sigma^2)$. In that case, we have $||\mathbf{E}||_2 \leq 1/400$ provided $\zeta \leq c_1 t$ and $k \leq c_2 d$ for appropriate constants $0 \leq c_1, c_2 \leq 1$. Therefore, λ must be at least $399/400$ (Weyl's inequality) and $(\langle \mathbf{u}^\star, \mathbf{z}\rangle^2 - 1) \leq 4||\mathbf{E}||_2$ (Davis

Kahan). Hence, we have the following inequality: $\Big(\frac{((\mathbf{u}^\star)^\mathsf{T}\mathbf{z})^2}{\lambda}-1\Big) \le 1/100$. Again, we have

$$
\begin{aligned}
&\Big|\Big|\frac{\mathbf{E}^p\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\mathbf{z}\mathbf{z}^\mathsf{T}\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\mathbf{E}^q}{\lambda^{p+q+1}}\Big|\Big|_\infty \\
&= \max_{ij}\mathbf{e}_i^\mathsf{T}\Big(\frac{\mathbf{E}^p\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\mathbf{z}\mathbf{z}^\mathsf{T}\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\mathbf{E}^q}{\lambda^{p+q+1}}\Big)\mathbf{e}_j \\
&= \max_{ij}\mathbf{e}_i^\mathsf{T}\mathbf{E}^p\Big(\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}+\mathbf{U}^\star_\perp(\mathbf{U}^\star_\perp)^\mathsf{T}\Big)\Big(\frac{\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\mathbf{z}\mathbf{z}^\mathsf{T}\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}}{\lambda^{p+q+1}}\Big)\Big(\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}+\mathbf{U}^\star_\perp(\mathbf{U}^\star_\perp)^\mathsf{T}\Big)\mathbf{E}^q\mathbf{e}_j \\
&\le ||\mathbf{E}^p\mathbf{u}^\star||_\infty||\mathbf{E}^q\mathbf{u}^\star||_\infty\Big|\Big|\frac{\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\mathbf{z}\mathbf{z}^\mathsf{T}\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}}{\lambda^{p+q+1}}\Big|\Big|_2 \le ||\mathbf{E}^p\mathbf{u}^\star||_\infty||\mathbf{E}^q\mathbf{u}^\star||_\infty\Big(\frac{((\mathbf{u}^\star)^\mathsf{T}\mathbf{z})^2}{\lambda^{p+q+1}}\Big)
\end{aligned}
$$

where $\mathbf{U}^\star_\perp$ is the subspace orthogonal to the vector $\mathbf{u}^\star$.

Lemma 15. *Let $\mathbf{e}_i \in \mathbb{R}^d$ denote the i^{th} standard basis vector. In that case, we will have*

$$\max_i \big|\big|\mathbf{e}_i^\mathsf{T}\mathbf{E}^p\mathbf{u}^\star\big|\big| \le \frac{\mu}{\sqrt{d}}\Big(\frac{\mu^2\zeta k}{dt}+\frac{\mu^2 k}{d}+\frac{\mu^2\zeta}{t}+\sqrt{\frac{d}{mt}}||\mathbf{F}||_\infty\Big)^p.$$

Proof. We can prove this statement via induction. For $p = 1$, the statement follows from the incoherence of $\mathbf{u}^\star$. Suppose the statement holds for $p = k$ for some $k > 1$. Under this induction hypothesis, we are going to show that the statement holds for $p = k + 1$. We will have

$$
\begin{aligned}
&\big|\big|\mathbf{e}_i^\mathsf{T}\mathbf{E}^{k+1}\mathbf{u}^\star\big|\big|_2^2 = \sum_\ell(\mathbf{e}_i^\mathsf{T}\mathbf{E}\mathbf{E}^k\mathbf{u}^\star\mathbf{e}_\ell)^2 = \sum_\ell(\sum_j\mathbf{E}_{ij}\mathbf{E}^k\mathbf{u}^\star\mathbf{e}_\ell)^2 \\
&= \sum_{j_1j_2}\mathbf{E}_{ij_1}\mathbf{E}_{ij_2}\big|\big|\mathbf{e}_{j_1}^\mathsf{T}\mathbf{E}^k\mathbf{u}^\star\big|\big|_2\big|\big|\mathbf{e}_{j_2}^\mathsf{T}\mathbf{E}^k\mathbf{u}^\star\big|\big|_2 \le \frac{\mu^2}{d}\Big(\frac{\mu^2\zeta k}{dt}+\frac{\mu^2 k}{d}+\frac{\mu^2\zeta}{t}+\sqrt{\frac{d}{mt}}||\mathbf{F}||_\infty\Big)^{2k+2}
\end{aligned}
$$

□

Hence, we must have $\|\sum_{p,q:p+q\ge1}\frac{\mathbf{E}^p\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\mathbf{z}\mathbf{z}^\mathsf{T}\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\mathbf{E}^q}{\lambda^{p+q+1}}\|_\infty$

$$
\begin{aligned}
&\le \sum_{p,q:p+q\ge1}\frac{\mu^2}{d}\Big(\frac{\mu^2\zeta k}{dt}+\frac{\mu^2 k}{d}+\frac{\mu^2\zeta}{t}\Big)^{p+q}\Big(\frac{1}{\lambda}\Big)^{p+q}\Big(\frac{((\mathbf{u}^\star)^\mathsf{T}\mathbf{z})^2}{\lambda}\Big) \\
&\le \frac{\mu^2}{d}\Big(\frac{((\mathbf{u}^\star)^\mathsf{T}\mathbf{z})^2}{\lambda}\Big)\sum_{p,q:p+q\ge1}\alpha^{p+q} = \frac{\mu^2}{d}\Big(\frac{((\mathbf{u}^\star)^\mathsf{T}\mathbf{z})^2}{\lambda}\Big)\Big(\Big(\frac{1}{1-\alpha}\Big)^2-1\Big)
\end{aligned}
$$

where $\alpha = \frac{\mu^2\zeta k}{dt\lambda}+\frac{\mu^2 k}{d\lambda}+\frac{\mu^2\zeta}{t\lambda}+\sqrt{\frac{d}{mt}}||\mathbf{F}||_\infty$. Again, if $\zeta \le c_1 t$ and $k \le c_2 d$ for appropriate constants $0 \le c_1, c_2 \le 1$, we will have $\big|\big|\lambda\mathbf{z}\mathbf{z}^\mathsf{T}-\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\big|\big|_\infty = O\Big(\frac{\mu^2}{d}\Big)$ and similarly, from our previous calculations on the operator norms, we will have $\big|\big|\lambda\mathbf{z}\mathbf{z}^\mathsf{T}-\mathbf{u}^\star(\mathbf{u}^\star)^\mathsf{T}\big|\big|_2 = O(1)\sigma\sqrt{\frac{d}{mt}}$. Hence, provided $mt = \Omega(d\sigma^2)$, by using Davis Kahan inequality, we obtain the initialization guarantees that we need for the rank-1 setting (see Theorem 1).

F USEFUL LEMMAS

Lemma 16 (Hanson-Wright lemma)**.** *Let $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(m)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d\times d})$ 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 $1-\delta_0$*

$$\Bigg|\frac{1}{m}\sum_{j=1}^m\mathbf{x}_j^\mathsf{T}\mathbf{A}_j\mathbf{x}_j-\frac{1}{m}\sum_{j=1}^m\mathsf{Tr}(\mathbf{A}_j)\Bigg| \le c\max\Bigg(\sqrt{\sum_{j=1}^m\|\mathbf{A}_j\|_\mathsf{F}^2\frac{\log\delta_0^{-1}}{m^2}}, \max_{j=1,\dots,m}\|\mathbf{A}_j\|_2\frac{\log\delta_0^{-1}}{m}\Bigg).$$

Lemma 17. *Let $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(m)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d\times d})$ be m i.i.d. standard isotropic Gaussian random vectors of dimension d. Then, for some universal constant $c \geq 0$, the following holds true with a probability of at least $1-\delta_0$.*

$$\left|\frac{1}{m}\sum_{j=1}^{m}\mathbf{a}^{\mathsf{T}}(\mathbf{x}^{(j)}(\mathbf{x}^{(j)})^{\mathsf{T}})\mathbf{b} - \mathbf{a}^{\mathsf{T}}\mathbf{b}\right| \leq c||\mathbf{a}||_2||\mathbf{b}||_2 \max\Big(\sqrt{\frac{\log\delta_0^{-1}}{m}}, \frac{\log\delta_0^{-1}}{m}\Big).$$

Lemma 18. *For three real r-rank matrices, satisfying $\mathbf{A} - \mathbf{B} = \mathbf{C}$, Weyl's inequality tells that*

$$\sigma_k(\mathbf{A}) - \sigma_k(\mathbf{B}) \leq \|\mathbf{C}\|$$

$\forall\ k \in [r]$ where $\sigma_k(\cdot)$ is the k-th largest singular value operator.

Lemma 19. *Let $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(m)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d\times d})$ be m i.i.d. standard isotropic Gaussian random vectors of dimension d. Then, for some universal constant $c \geq 0$, the following holds true with a probability of at least $1-\delta_0$, $\left|\left|\frac{1}{m}\sum_{j=1}^{m}\mathbf{a}_j(\mathbf{x}^{(j)}(\mathbf{x}^{(j)})^{\mathsf{T}}) - \frac{1}{m}\sum_{j=1}^{m}\mathbf{a}_j\mathbf{I}\right|\right|_2$*

$$\leq c\max\Big(\frac{||\mathbf{a}||_2}{\sqrt{m}}\sqrt{\frac{d\log 9 + \log\delta_0^{-1}}{m}}, ||\mathbf{a}||_\infty\frac{d\log 9 + \log\delta_0^{-1}}{m}\Big).$$

Lemma 20. *Let $\mathbf{a}_i, \mathbf{b}_i \in \mathbb{R}^d\ \forall\ i \in [t]$. Then,*

$$\|\sum_i \mathbf{a}_i\mathbf{b}_i^{\mathsf{T}}\|_p^2 \leq \|\sum_i \mathbf{a}_i\mathbf{a}_i^{\mathsf{T}}\|_p\|\sum_i \mathbf{b}_i\mathbf{b}_i^{\mathsf{T}}\|_p.$$

Lemma 21. *For a real matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and a real symmetric positive semi-definite (PSD) matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$, the following holds true: $\sigma_{\min}^2(\mathbf{A})\lambda_{\min}(\mathbf{B}) \leq \lambda_{\min}(\mathbf{A}\mathbf{B}\mathbf{A}^{\mathsf{T}})$, where $\sigma_{\min}(\cdot)$ and $\lambda_{\min}(\cdot)$ represents the minimum singular value and minimum eigenvalue operators respectively.*

Lemma 22. *For any three matrices $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$ for which the matrix product $\mathbf{ABC}$ is defined,*

$$\mathsf{vec}(\mathbf{ABC}) = (\mathbf{C}^{\mathsf{T}} \otimes \mathbf{A})\mathsf{vec}(\mathbf{B}).$$

Lemma 23. *For a (ν^2, α) sub-exponential random variable, we have the following tail bound*

$$\mathbb{P}(\|X - \mathbb{E}[X]\,| \geq t) \leq e^{-\frac{1}{2}\min\{t^2/\nu^2, t/\alpha\}}.$$