A. Additional Experimental Results

A.1. Synthetic Data: Further comparison with GD-GD

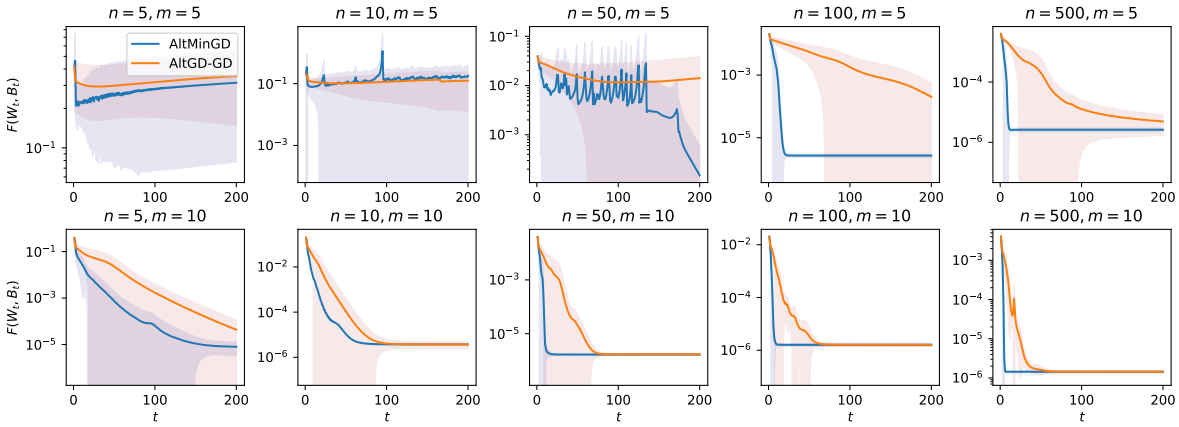

Figure 7. Function values for FedRep and GD-GD. The value of m is fixed in each row and n is fixed in each column. Here $r = 1$ (full participation) and the average trajectories over 10 trials are plotted along with 95% confidence intervals. Principal angle distances are not plotted as the results are very similar. *We see that the relative improvement of FedRep over GD-GD increases with n, highlighting the advantage of FedRep in settings with many clients.*

Further experimental details. In the synthetic data experiments, the ground-truth matrices $\mathbf{W}^*$ and $\mathbf{B}^*$ were generated by first sampling each element as an i.i.d. standard normal variable, then taking the QR factorization of the resulting matrix, and scaling it by $\sqrt{k}$ in the case of $\mathbf{W}^*$. The clients each trained on the same m samples throughout the entire training process. Test samples were generated identically as the training samples but without noise. Both the iterates of FedRep and GD-GD were initialized with the SVD of the result of 10 rounds of projected gradient descent on the unfactorized matrix sensing objective as in Algorithm 1 in (Tu et al., 2016). We would like to note that FedRep exhibited the same convergence trajectories regardless of whether its iterates were initialized with random Gaussian samples or with the projected gradient descent procedure, whereas GD-GD was highly sensitive to its initialization, often not converging when initialized randomly.

A.2. Real Data: Further experimental details

Datasets. The CIFAR10 and CIFAR100 datasets (Krizhevsky et al., 2009) were generated by randomly splitting the training data into Sn shards with $50,000/(Sn)$ images of a single class in each shard, as in (McMahan et al., 2017). The full Federated-EMNIST (FEMNIST) dataset contains 62 classes of handwritten letters, but in Table 1 we use a subset with only 10 classes of handwritten letters. In particular, we followed the same dataset generation procedure as in (Li et al., 2019), but used 150 clients instead of 200. When testing on new clients as in Figure 6, we use samples from 10 classes of handwritten digits from FEMNIST, i.e., the MNIST dataset. In this phase there are 100 new clients, each with 500 samples from 5 different classes for fine-tuning. The fine-tuned models are then evaluated on 100 testing samples from these same 5 classes. For Sent140, we randomly sample 183 clients (Twitter users) that each have at least 50 samples (tweets). Each tweet is either positive sentiment or negative sentiment. Statistics of both the FEMNIST and Sent140 datasets we use are given in Table 2. For both FEMNIST and Sent140 we use the LEAF framework (Caldas et al., 2018).

Hyperparameters. As in (Liang et al., 2020), all methods use SGD with momentum with parameter equal to 0.5. In Table 1, for CIFAR10, CIFAR100, and FEMNIST the local sample batch size is 10 and for Sent140 it is 4. The participation rate r is always 0.1, besides in the fine-tuning phases in Figure 6, in which all clients are sampled in each round. For each dataset learning rates were tuned in $\{0.001, 0.01, 0.1\}$. We observed that the optimal learning rates for FedAvg were also typically the optimal base learning rates for the other methods, so we used the same base learning rates for all methods for each dataset, which was 0.01 in all cases, unless stated otherwise. Note that the batch size and learning rate for CIFAR10 used in Table 1 differs from the standard setting of a batch size of 50 and learning rate of 0.1 (McMahan et al., 2017), but

we observed improved performance for all methods by using $(10, 0.01)$ instead. In particular, the simulation in Figure 5, the standard setting of $(50, 0.1)$ is used, but the accuracies are worse than those reported in Table 1 for both FedAvg and FedRep. Additionally, in Table 1, for CIFAR10 with $(n, S) = (100, 2)$ and $(n, S) = (100, 5)$, we executed 1 local epoch of SGD with momentum for the representation for FedRep and 1 local epoch for all other methods. For all other datasets we executed 5 local epochs for the representation for FedRep and for the local updates for all other methods.

Evaluation. As mentioned in the main body, in Table 1, we initialize all methods randomly and train for $T = 100$ communication rounds for the CIFAR datasets, $T = 200$ for FEMNIST, and $T = 50$ for Sent140. The accuracy shown is the average local test accuracy over all users over the final ten communication rounds, besides for the fine-tuning results, in which case we report the average local test accuracies of the locally fine-tuned models over all users, after the global model has been fully trained. We repeat the entire training and evaluation process five times for each model and dataset and report the averages in Table 1.

Implementations. Our code is an adaptation of the repository from (Liang et al., 2020), written in Pytorch and available at `https://github.com/pliang279/LG-FedAvg/`. In particular, we used the implementations of FedAvg, Fed-MTL and LG-FedAvg given in this repository. For consistency we use this same codebase to implement FedRep, FedPer, SCAFFOLD, FedProx, APFL, Ditto, L2GD, and PerFedAvg. As in the experiments in (Liang et al., 2020), we used a 5-layer CNN with two convolutional layers for CIFAR10 and CIFAR100 followed by three fully-connected layers. For FEMNIST, we use an MLP with two hidden layers, and for Sent140 we use a pre-trained 300-dimensional GloVe embedding[1] and train RNN with an LSTM module followed by two fully-connected decoding layers.

For FedRep, we treated the head as the weights and biases of the final fully-connected layer in each of the models. For LG-FedAvg, we treated the first two convolutional layers of the model for CIFAR10 and CIFAR100 as the local representation, and the fully-connected layers as the global parameters, and the input layer and hidden layers as the global parameters. For FEMNIST, we set all parameters besides those in the output layer as the local representation parameters. For Sent140, we set the RNN module to be the local representation and the decoder to be the global parameters. Unlike in the paper introducing LG-FedAvg (Liang et al., 2020), we did not initialize the models for all methods with the solution of many rounds of FedAvg (instead, we initialized randomly) and we computed the local test accuracy as the average local test accuracy over the final ten communication rounds, rather than the average of the maximum local test accuracy for each client over the entire training procedure.

For L2GD (Hanzely & Richtárik, 2020) we executed multiple epochs of local SGD (discussed above) instead of one step of GD in the local update in order for reasonable comparison with the other methods. We also set $p = 0.9$, thus the local parameters are trained on 10% of the communication rounds. We tuned α in $\{0.05, 0.1, 0.25, 0.5, 0.75\}$ and we tuned λ over $\{1, 0.5\}$. We used $(\alpha, \lambda) = (0.25, 1)$ in all cases besides the $(n, S) = (100, 5)$ case for CIFAR100, for which we used $\alpha = 0.1$. Also, for FEMNIST we improved performance by using a learning rate of 0.001 instead of 0.01. For APFL, we used a fixed α that we tuned in $\{0.1, 0.25, 0.5, 0.75\}$, and chose $\alpha = 0.25$ for all cases besides the most heterogeneous CIFAR versions, namely $(n, S) = (100, 2)$ for CIFAR10 and $(n, S) = (100, 25)$ for CIFAR100. For Ditto we tuned λ among $\{0.25, 0.5, 0.75, 1\}$, and used $\lambda = 0.75$ for all cases besides CIFAR100, for which we used $\lambda = 1$. For PerFedAvg, we used an inner learning rate of 10^{-4} and 8 samples as the support set and 2 samples as the target set in each local meta-gradient update. We used the Hessian-free version. For FedProx we tuned μ among $\{0.05, 0.1, 0.25, 0.5\}$, and used $\mu = 0.1$ for CIFAR and $\mu = 0.25$ for FEMNIST and Sent140. For SCAFFOLD we used a global learning rate of 1 in all cases besides FEMNIST, for which 0.5 was superior.

Table 2. Dataset statistics.

DATASET	NUMBER OF USERS (n)	AVG SAMPLES/USER	MIN SAMPLES/USER
FEMNIST	150	148	50
SENT140	183	72	50

[1] Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

B. Proof of Main Result

B.1. Preliminaries.

Definition 2. *For a random vector $\mathbf{x} \in \mathbb{R}^d$ and a fixed matrix $\mathbf{A} \in \mathbb{R}^{d_1 \times d_2}$, the vector $\mathbf{A}^\top \mathbf{x}$ is called $\|\mathbf{A}\|_2$-sub-gaussian if $\mathbf{y}^\top \mathbf{A}^\top \mathbf{x}$ is sub-gaussian with sub-gaussian norm $\|\mathbf{A}\|_2 \|\mathbf{y}\|_2$ for all $\mathbf{y} \in \mathbb{R}^{d_2}$, i.e. $\mathbb{E}[\exp(\mathbf{y}^\top \mathbf{A}^\top \mathbf{x})] \leq \exp\left(\|\mathbf{y}\|_2^2 \|\mathbf{A}\|_2^2 / 2\right)$.*

Definition 3. *A rank-k matrix $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$ is μ-*row-wise incoherent *if $\max_{i \in [d_1]} \|\mathbf{m}_i\|_2 \leq (\mu\sqrt{d_2}/\sqrt{d_1})\|\mathbf{M}\|_F$, where $\mathbf{m}_i \in \mathbb{R}^{d_2}$ is the i-th row of $\mathbf{M}$.*

We use hats to denote orthonormal matrices (a matrix is called orthonormal if its set of columns is an orthonormal set). By Assumption 3, the ground truth representation $\mathbf{B}^*$ is orthonormal, so from now on we will write it as $\hat{\mathbf{B}}^*$.

For a matrix $\mathbf{W} \in \mathbb{R}^{n \times k}$ and a random set of indices $\mathcal{I} \in [n]$ of cardinality rn, define $\mathbf{W}_{\mathcal{I}} \in \mathbb{R}^{rn \times k}$ as the matrix formed by taking the rows of $\mathbf{W}$ indexed by $\mathcal{I}$. Define $\bar{\sigma}_{\max,*} := \max_{\mathcal{I} \in [n], |\mathcal{I}| = rn} \sigma_{\max}(\frac{1}{\sqrt{rn}} \mathbf{W}^*_{\mathcal{I}})$ and $\bar{\sigma}_{\min,*} := \min_{\mathcal{I} \in [n], |\mathcal{I}| = rn} \sigma_{\min}(\frac{1}{\sqrt{rn}} \mathbf{W}^*_{\mathcal{I}})$, i.e. the maximum and minimum singular values of any matrix that can be obtained by taking rn rows of $\frac{1}{\sqrt{rn}} \mathbf{W}^*$. Note that by Assumption 3, each row of $\mathbf{W}^*$ has norm $\sqrt{k}$, so $\frac{1}{\sqrt{rn}}$ acts as a normalizing factor such that $\|\frac{1}{\sqrt{rn}} \mathbf{W}^*_{\mathcal{I}}\|_F = \sqrt{k}$. In addition, define $\kappa = \bar{\sigma}_{\max,*} / \bar{\sigma}_{\min,*}$.

Let i now be an index over $[rn]$, and let i' be an index over $[n]$. For random batches of samples $\{\{(\mathbf{x}_i^j, y_i^j)\}_{j=1}^m\}_{i=1}^{rn}$, define the random linear operator $\mathcal{A} : \mathbb{R}^{rn \times d} \to \mathbb{R}^{rnm}$ as $\mathcal{A}(\mathbf{M}) = [\langle \mathbf{A}_i^j, \mathbf{M} \rangle]_{1 \leq i \leq rn, 1 \leq j \leq m} \in \mathbb{R}^{rnm}$. Here, $\mathbf{A}_i^j := \mathbf{e}_i (\mathbf{x}_i^j)^\top$, where $\mathbf{e}_i$ is the i-th standard vector in $\mathbb{R}^{rn}$, and $\mathbf{M} \in \mathbb{R}^{rn \times d}$. Then, the loss function in (6) is equivalent to

$$\min_{\mathbf{B} \in \mathbb{R}^{d \times k}, \mathbf{W} \in \mathbb{R}^{n \times k}} \{F(\mathbf{B}, \mathbf{W}) := \frac{1}{2rnm} \mathbb{E}_{\mathcal{A}, \mathcal{I}} \left[\|\mathbf{Y} - \mathcal{A}(\mathbf{W}_{\mathcal{I}} \mathbf{B}^\top)\|_2^2\right]\}, \tag{12}$$

where $\mathbf{Y} = \mathcal{A}(\mathbf{W}^*_{\mathcal{I}} \hat{\mathbf{B}}^{*\top}) \in \mathbb{R}^{rnm}$ is a concatenated vector of labels. It is now easily seen that the problem of recovering $\mathbf{W}_* \hat{\mathbf{B}}^{*\top}$ from finitely-many measurements $\mathcal{A}(\mathbf{W}^*_{\mathcal{I}} \hat{\mathbf{B}}^{*\top})$ is an instance of matrix sensing. Moreover, the updates of FedRep satisfy the following recursion:

$$\mathbf{W}^{t+1}_{\mathcal{I}^t} = \underset{\mathbf{W}_{\mathcal{I}^t} \in \mathbb{R}^{rn \times k}}{\operatorname{argmin}} \frac{1}{2rnm} \|\mathcal{A}^t(\mathbf{W}^*_{\mathcal{I}^t} \hat{\mathbf{B}}^{*\top} - \mathbf{W}_{\mathcal{I}^t} \mathbf{B}^{t\top})\|_2^2 \tag{13}$$

$$\mathbf{B}^{t+1} = \mathbf{B}^t - \frac{\eta}{rnm} \left((\mathcal{A}^t)^\dagger \mathcal{A}^t(\mathbf{W}^{t+1}_{\mathcal{I}^t} \mathbf{B}^{t\top} - \mathbf{W}^*_{\mathcal{I}^t} \hat{\mathbf{B}}^{*\top})\right)^\top \mathbf{W}^{t+1}_{\mathcal{I}^t} \tag{14}$$

where $\mathcal{A}^t$ is an instance of $\mathcal{A}$, and $(\mathcal{A}^t)^\dagger$ is the adjoint operator of $\mathcal{A}^t$, i.e. $(\mathcal{A}^t)^\dagger \mathcal{A}(\mathbf{M}) = \sum_{i=1}^{rn} \sum_{j=1}^m (\langle \mathbf{A}_i^{t,j}, \mathbf{M} \rangle) \mathbf{A}_i^{t,j}$. Note that for the purposes of analysis, it does not matter how $\mathbf{w}_{i'}^{t+1}$ is computed for all $i' \notin \mathcal{I}^t$, as these vectors do not affect the computation of $\mathbf{B}^{t+1}$. Moreover, our analysis does not rely on any particular properties of the batches $\mathcal{I}^1, \ldots, \mathcal{I}^T$ other than the fact that they have cardinality rn, so without loss of generality we assume $\mathcal{I}^t = [rn]$ for all $t = 1, \ldots T$ and drop the subscripts $\mathcal{I}^t$ on $\mathbf{W}^t$.

B.2. Auxilliary Lemmas

We start by showing that we can assume without loss of generality that $\mathbf{B}^t$ is orthonormalized at the end of every communication round.

Lemma 1. *Let $\mathbf{W}^t \in \mathbb{R}^{rn \times k}$ and $\mathbf{B}^t \in \mathbb{R}^{d \times k}$ denote the iterates of Algorithm 2 as outlined in (13) and (14) (with the subscript $\mathcal{I}^t$ dropped). Now consider the modified algorithm given by the following recursion:*

$$\widetilde{\mathbf{W}}^{t+1} = \arg\min_{\mathbf{W}} \|\mathcal{A}^t(\mathbf{W}(\overline{\mathbf{B}}^t)^\top - \mathbf{W}_*(\hat{\mathbf{B}}^*)^\top)\|_F^2 \tag{15}$$

$$\widetilde{\mathbf{B}}^{t+1} = \overline{\mathbf{B}}^t - \frac{\eta}{rnm} \left((\mathcal{A}^t)^\dagger \mathcal{A}^t(\widetilde{\mathbf{W}}^{t+1}(\overline{\mathbf{B}}^t)^\top - \mathbf{W}_*(\hat{\mathbf{B}}^*)^\top)\right)^\top \widetilde{\mathbf{W}}^{t+1} \tag{16}$$

$$\overline{\mathbf{B}}^{t+1} = \widetilde{\mathbf{B}}^{t+1} (\widetilde{\mathbf{R}}^{t+1})^{-1} \tag{17}$$

where $\overline{\mathbf{B}}^{t+1} \widetilde{\mathbf{R}}^{t+1}$ is the QR factorization of $\widetilde{\mathbf{B}}^{t+1}$. Then the column spaces of $\mathbf{B}^t$ and $\widetilde{\mathbf{B}}^t$ are equivalent for all t.

Proof. The proof follows a similar argument as Lemma 4.4 in (Jain et al., 2013). Assume that the claim holds for iteration t. Then there is some full-rank $\mathbf{R}_B \in \mathbb{R}^{k\times k}$ such that $\widetilde{\mathbf{B}}^t\mathbf{R}_B = \mathbf{B}^t$. Then $\overline{\mathbf{B}}^t\widetilde{\mathbf{R}}^t\mathbf{R}_B = \mathbf{B}^t$, where $\widetilde{\mathbf{R}}^t\mathbf{R}_B$ is full rank. Since

$$\widetilde{\mathbf{W}}^{t+1} = \arg\min_{\mathbf{W}} \|\mathcal{A}^t(\mathbf{W}(\overline{\mathbf{B}}^t)^\top - \mathbf{W}_*(\hat{\mathbf{B}}^*)^\top)\|_F^2 = \arg\min_{\mathbf{W}} \|\mathcal{A}^t((\mathbf{W}(\widetilde{\mathbf{R}}^t\mathbf{R}_B)^{-\top})(\mathbf{B}^t)^\top - \mathbf{W}^*(\hat{\mathbf{B}}^*)^\top)\|_F^2 \tag{18}$$

we have that $\widetilde{\mathbf{W}}^{t+1}(\widetilde{\mathbf{R}}^t\mathbf{R}_B)^{-\top}$ minimizes $\|\mathcal{A}^t(\mathbf{W}(\mathbf{B}^t)^\top - \mathbf{W}_*(\hat{\mathbf{B}}^*)^\top)\|_F^2$ over $\mathbf{W}$ since $(\widetilde{\mathbf{R}}^t\mathbf{R}_B)^\top$ is full rank. So $\mathbf{W}^{t+1} = \widetilde{\mathbf{W}}^{t+1}(\widetilde{\mathbf{R}}^t\mathbf{R}_B)^{-\top}$ and the column spaces of $\widetilde{\mathbf{W}}^{t+1}$ and $\mathbf{W}^{t+1}$ are equivalent. Next, recall the definition of $\mathbf{B}^{t+1}$:

$$\begin{aligned}
\mathbf{B}^{t+1} &= \mathbf{B}^t - \frac{\eta}{rnm}\left((\mathcal{A}^t)^\dagger\mathcal{A}^t(\mathbf{W}^{t+1}(\mathbf{B}^t)^\top - \mathbf{W}^*(\hat{\mathbf{B}}^*)^\top)\right)^\top \mathbf{W}^{t+1} &(19)\\
&= \overline{\mathbf{B}}^t\widetilde{\mathbf{R}}^t\mathbf{R}_B - \frac{\eta}{rnm}\left((\mathcal{A}^t)^\dagger\mathcal{A}^t(\widetilde{\mathbf{W}}^{t+1}(\widetilde{\mathbf{R}}^t\mathbf{R}_B)^{-\top}(\widetilde{\mathbf{R}}^t\mathbf{R}_B)^\top(\overline{\mathbf{B}}^t)^\top - \mathbf{W}^*(\hat{\mathbf{B}}^*)^\top)\right)^\top \widetilde{\mathbf{W}}^{t+1}(\widetilde{\mathbf{R}}^t\mathbf{R}_B)^{-\top}\\
&= \left[\overline{\mathbf{B}}^t,\ -\tfrac{\eta}{rnm}\left((\mathcal{A}^t)^\dagger\mathcal{A}^t(\widetilde{\mathbf{W}}^{t+1}(\overline{\mathbf{B}}^t)^\top - \mathbf{W}^*(\hat{\mathbf{B}}^*)^\top)\right)^\top \widetilde{\mathbf{W}}^{t+1}\right]\begin{bmatrix}\widetilde{\mathbf{R}}^t\mathbf{R}_B\\ (\widetilde{\mathbf{R}}^t\mathbf{R}_B)^{-\top}\end{bmatrix} &(20)
\end{aligned}$$

so the column space of $\mathbf{B}^{t+1}$ is equal to the column space of

$$\left[\overline{\mathbf{B}}^t,\ -\tfrac{\eta}{rnm}\left((\mathcal{A}^t)^\dagger\mathcal{A}^t(\widetilde{\mathbf{W}}^{t+1}(\overline{\mathbf{B}}^t)^\top - \mathbf{W}^*(\hat{\mathbf{B}}^*)^\top)\right)^\top \widetilde{\mathbf{W}}^{t+1}\right].$$

Finally, note that $\widetilde{\mathbf{B}}^{t+1}$ can be written as:

$$\widetilde{\mathbf{B}}^{t+1} = \left[\overline{\mathbf{B}}^t,\ -\tfrac{\eta}{rnm}\left((\mathcal{A}^t)^\dagger\mathcal{A}^t(\widetilde{\mathbf{W}}^{t+1}(\overline{\mathbf{B}}^t)^\top - \mathbf{W}_*(\hat{\mathbf{B}}^*)^\top)\right)^\top \widetilde{\mathbf{W}}^{t+1}\right]\begin{bmatrix}\mathbf{I}_k\\ \mathbf{I}_k\end{bmatrix} \tag{21}$$

so $\widetilde{\mathbf{B}}^{t+1}$ has column space that is also equal to the column space of

$$\left[\overline{\mathbf{B}}^t,\ -\tfrac{\eta}{rnm}\left((\mathcal{A}^t)^\dagger\mathcal{A}^t(\widetilde{\mathbf{W}}^{t+1}(\overline{\mathbf{B}}^t)^\top - \mathbf{W}_*(\hat{\mathbf{B}}^*)^\top)\right)^\top \widetilde{\mathbf{W}}^{t+1}\right]$$

. □

Note that we cannot orthonormalize $\mathbf{W}^t$, neither in practice (due to privacy constraints) nor for analysis only.

In light of Lemma 1, we now analyze the modified algorithm in Lemma 1 in which $\mathbf{B}^t$ is orthonormalized after each iteration. We will use our standard notation $\mathbf{W}^t$, $\mathbf{B}^t$ to denote the iterates of this algorithm, with $\hat{\mathbf{B}}^t$ being the orthonormalized version of $\mathbf{B}^t$. For clarity we restate this modified algorithm with the standard notation here:

$$\mathbf{W}^{t+1} = \arg\min_{\mathbf{W}} \frac{1}{2rnm}\|\mathcal{A}^t(\mathbf{W}(\hat{\mathbf{B}}^t)^\top - \mathbf{W}_*(\hat{\mathbf{B}}^*)^\top)\|_F^2 \tag{22}$$

$$\mathbf{B}^{t+1} = \hat{\mathbf{B}}^t - \frac{\eta}{rnm}\left((\mathcal{A}^t)^\dagger\mathcal{A}^t(\mathbf{W}^{t+1}(\hat{\mathbf{B}}^t)^\top - \mathbf{W}^*(\hat{\mathbf{B}}^*)^\top)\right)^\top \mathbf{W}^{t+1} \tag{23}$$

$$\hat{\mathbf{B}}^{t+1} = \mathbf{B}^{t+1}(\mathbf{R}^{t+1})^{-1} \tag{24}$$

We next explicitly compute $\mathbf{W}^{t+1}$. Since the rest of the proof analyzes a particular communication round t, we drop superscripts t on the measurement operators $\mathcal{A}^t$ and matrices $\mathbf{A}_{i,j}^t$ for ease of notation.

Lemma 2. *In the modified algorithm, where $\mathbf{B}$ is orthonormalized after each update, the update for $\mathbf{W}$ is:*

$$\mathbf{W}^{t+1} = \mathbf{W}^*\hat{\mathbf{B}}^{*\top}\hat{\mathbf{B}}^t - \mathbf{F} \tag{25}$$

where $\mathbf{F}$ is defined in equation (30) below.

Proof. We adapt the argument from Lemma 4.5 in (Jain et al., 2013) to compute the update for $\mathbf{W}^{t+1}$, and borrow heavily from their notation.

Let $\mathbf{w}_p^{t+1}$ (respectively $\hat{\mathbf{b}}_p^{t+1}$) be the p-th column of $\mathbf{W}^t$ (respectively $\hat{\mathbf{B}}^t$). Since $\mathbf{W}^{t+1}$ minimizes $\tilde{F}(\mathbf{W}, \hat{\mathbf{B}}^t) \coloneqq \frac{1}{2rnm}\|\mathcal{A}^t(\mathbf{W}^*(\hat{\mathbf{B}}^*)^\top - \mathbf{W}(\mathbf{B}^t)^\top)\|_2^2$ with respect to $\mathbf{W}$, we have $\nabla_{\mathbf{w}_p}\tilde{F}(\mathbf{W}^{t+1}, \hat{\mathbf{B}}^t) = \mathbf{0}$ for all $p \in [k]$. Thus, for any $p \in [k]$, we have

$$\begin{aligned}
\mathbf{0} &= \nabla_{\mathbf{w}_p}\tilde{F}(\mathbf{W}^{t+1}, \hat{\mathbf{B}}^t) \\
&= \frac{1}{rnm}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\langle \mathbf{A}_{i,j}, \mathbf{W}^{t+1}(\hat{\mathbf{B}}^t)^\top - \mathbf{W}^*(\hat{\mathbf{B}}^*)^\top\rangle\right)\mathbf{A}_{i,j}\hat{\mathbf{b}}_p^t \\
&= \frac{1}{rnm}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\sum_{q=1}^{k}(\hat{\mathbf{b}}_q^t)^\top\mathbf{A}_{i,j}^\top\mathbf{w}_q^{t+1} - \sum_{q=1}^{k}(\hat{\mathbf{b}}_q^*)^\top\mathbf{A}_{i,j}^\top\mathbf{w}_q^*\right)\mathbf{A}_{i,j}\hat{\mathbf{b}}_p^t
\end{aligned}$$

This implies

$$\frac{1}{m}\sum_{q=1}^{k}\left(\sum_{i=1}^{rn}\sum_{j=1}^{m}\mathbf{A}_{i,j}\hat{\mathbf{b}}_p^t(\hat{\mathbf{b}}_q^t)^\top\mathbf{A}_{i,j}^\top\right)\mathbf{w}_q^{t+1} = \frac{1}{m}\sum_{q=1}^{k}\left(\sum_{i=1}^{rn}\sum_{j=1}^{m}\mathbf{A}_{i,j}\hat{\mathbf{b}}_p^t(\hat{\mathbf{b}}_q^*)^\top\mathbf{A}_{i,j}^\top\right)\mathbf{w}_q^* \tag{26}$$

To solve for $\mathbf{w}^{t+1}$, we define $\mathbf{G}$, $\mathbf{C}$, and $\mathbf{D}$ as rnk-by-rnk block matrices, as follows:

$$\mathbf{G} \coloneqq \begin{bmatrix} \mathbf{G}_{11} & \cdots & \mathbf{G}_{1k} \\ \vdots & \ddots & \vdots \\ \mathbf{G}_{k1} & \cdots & \mathbf{G}_{kk} \end{bmatrix}, \mathbf{C} \coloneqq \begin{bmatrix} \mathbf{C}_{11} & \cdots & \mathbf{C}_{1k} \\ \vdots & \ddots & \vdots \\ \mathbf{C}_{k1} & \cdots & \mathbf{C}_{kk} \end{bmatrix}, \mathbf{D} \coloneqq \begin{bmatrix} \mathbf{D}_{11} & \cdots & \mathbf{D}_{1k} \\ \vdots & \ddots & \vdots \\ \mathbf{D}_{k1} & \cdots & \mathbf{D}_{kk} \end{bmatrix} \tag{27}$$

where, for $p, q \in [k]$: $\mathbf{G}_{pq} \coloneqq \frac{1}{m}\sum_{i=1}^{rn}\sum_{j=1}^{m}\mathbf{A}_{i,j}\hat{\mathbf{b}}_p^t\hat{\mathbf{b}}_q^{t\top}\mathbf{A}_{i,j}^\top \in \mathbb{R}^{rn\times rn}$, $\mathbf{C}_{pq} \coloneqq \frac{1}{m}\sum_{i=1}^{rn}\sum_{j=1}^{m}\mathbf{A}_{i,j}\hat{\mathbf{b}}_p^t(\hat{\mathbf{b}}_q^*)^\top\mathbf{A}_{i,j}^\top \in \mathbb{R}^{rn\times rn}$, and, $\mathbf{D}_{pq} \coloneqq \langle\hat{\mathbf{b}}_p^t, \hat{\mathbf{b}}_q^*\rangle\mathbf{I}_{rn} \in \mathbb{R}^{rn\times rn}$. Recall that $\hat{\mathbf{b}}_p^t$ is the p-th column of $\hat{\mathbf{B}}^t$ and $\hat{\mathbf{b}}_q^*$ is the q-th column of $\hat{\mathbf{B}}^*$. Further, define

$$\widetilde{\mathbf{w}}^{t+1} = \begin{bmatrix}\mathbf{w}_1^{t+1}\\ \vdots \\ \mathbf{w}_k^{t+1}\end{bmatrix} \in \mathbb{R}^{rnk}, \quad \widetilde{\mathbf{w}}^* = \begin{bmatrix}\mathbf{w}_1^{*}\\ \vdots \\ \mathbf{w}_k^{*}\end{bmatrix} \in \mathbb{R}^{rnk}.$$

Then, by (26), we have

$$\begin{aligned}
\widetilde{\mathbf{w}}^{t+1} &= \mathbf{G}^{-1}\mathbf{C}\widetilde{\mathbf{w}}^* \\
&= \mathbf{D}\widetilde{\mathbf{w}}^* - \mathbf{G}^{-1}\left(\mathbf{G}\mathbf{D} - \mathbf{C}\right)\widetilde{\mathbf{w}}^*
\end{aligned}$$

where we can invert $\mathbf{G}$ conditioned on the event that its minimum singular value is strictly positive, which Lemma 3 shows holds with high probability. Now consider the p-th block of $\widetilde{\mathbf{w}}^{t+1}$, and let $\left(\left(\mathbf{G}\mathbf{D} - \mathbf{C}\right)\mathbf{w}^*\right)_p$ denote the p-th block of $\left(\mathbf{G}\mathbf{D} - \mathbf{C}\right)\mathbf{w}^*$. We have

$$\begin{aligned}
\widetilde{\mathbf{w}}_p^{t+1} &= \sum_{q=1}^{k}\langle\hat{\mathbf{b}}_p^t, \hat{\mathbf{b}}_q^*\rangle\mathbf{w}_q^* - \left(\mathbf{G}^{-1}\left(\mathbf{G}\mathbf{D} - \mathbf{C}\right)\mathbf{w}^*\right)_p \\
&= \left(\sum_{q=1}^{k}\mathbf{w}_q^*(\hat{\mathbf{b}}_p^*)^\top\right)\hat{\mathbf{b}}_q^t - \left(\mathbf{G}^{-1}\left(\mathbf{G}\mathbf{D} - \mathbf{C}\right)\mathbf{w}^*\right)_p \\
&= \left(\mathbf{W}^*(\hat{\mathbf{B}}^*)^\top\right)\hat{\mathbf{b}}_q^t - \left(\mathbf{G}^{-1}\left(\mathbf{G}\mathbf{D} - \mathbf{C}\right)\mathbf{w}^*\right)_p
\end{aligned} \tag{28}$$

By constructing $\mathbf{W}^{t+1}$ such that the p-th column of $\mathbf{W}^{t+1}$ is $\mathbf{w}_p^{t+1}$ for all $p \in [k]$, we obtain

$$\mathbf{W}^{t+1} = \mathbf{W}^*\hat{\mathbf{B}}^*(\hat{\mathbf{B}}^t)^\top - \mathbf{F} \tag{29}$$

where

$$\mathbf{F} = [(\mathbf{G}^{-1}(\mathbf{G}\mathbf{D} - \mathbf{C})\widetilde{\mathbf{w}}^*)_1, \ldots, (\mathbf{G}^{-1}(\mathbf{G}\mathbf{D} - \mathbf{C})\widetilde{\mathbf{w}}^*)_k] \tag{30}$$

and $(\mathbf{G}^{-1}(\mathbf{G}\mathbf{D} - \mathbf{C})\widetilde{\mathbf{w}}^*)_p$ is the p-th n-dimensional block of the rnk-dimensional vector $\mathbf{G}^{-1}(\mathbf{G}\mathbf{D} - \mathbf{C})\widetilde{\mathbf{w}}^*$. □

Next we bound the Frobenius norm of the matrix $\mathbf{F}$, which requires multiple steps. First, we establish some helpful notations. We drop superscripts indicating the iteration number t for simplicity.

Again let $\mathbf{w}^*$ be the rnk-dimensional vector formed by stacking the columns of $\mathbf{W}^*$, and let $\hat{\mathbf{b}}_p$ (respectively $\hat{\mathbf{b}}_q^*$) be the p-th column of $\hat{\mathbf{B}}$ (respectively the q-th column of $\hat{\mathbf{B}}_*$). Recall that $\mathbf{F}$ can be obtained by stacking $\mathbf{G}^{-1}(\mathbf{GD}-\mathbf{C})\mathbf{w}^*$ into k columns of length n, i.e. $\text{vec}(\mathbf{F}) = \mathbf{G}^{-1}(\mathbf{GD}-\mathbf{C})\mathbf{w}^*$. Further, $\mathbf{G} \in \mathbb{R}^{rnk \times rnk}$ is a block matrix whose blocks $\mathbf{G}_{pq} \in \mathbf{R}^{rn \times rn}$ for $p, q \in [k]$ are given by:

$$\begin{aligned}\mathbf{G}_{pq} &= \frac{1}{m}\sum_{i=1}^{rn}\sum_{j=1}^{m}\mathbf{A}_{i,j}\hat{\mathbf{b}}_p\hat{\mathbf{b}}_q^\top\mathbf{A}_{i,j}^\top \\ &= \frac{1}{m}\sum_{i=1}^{rn}\sum_{j=1}^{m}\mathbf{e}_i(\mathbf{x}_i^j)^\top\hat{\mathbf{b}}_p\hat{\mathbf{b}}_q^\top\mathbf{x}_i^j\mathbf{e}_i^\top \end{aligned} \tag{31}$$

So, each $\mathbf{G}_{pq}$ is diagonal with diagonal entries

$$(\mathbf{G}_{pq})_{ii} = \frac{1}{m}\sum_{j=1}^{m}(\mathbf{x}_i^j)^\top\hat{\mathbf{b}}_p\hat{\mathbf{b}}_q^\top\mathbf{x}_i^j = \hat{\mathbf{b}}_p^\top\left(\frac{1}{m}\sum_{j=1}^{m}\mathbf{x}_i^j(\mathbf{x}_i^j)^\top\right)\hat{\mathbf{b}}_q \tag{32}$$

Define $\mathbf{\Pi}^i \coloneqq \frac{1}{m}\sum_{j=1}^{m}\mathbf{x}_i^j(\mathbf{x}_i^j)^\top$ for all $i \in [rn]$. Similarly as above, each block $\mathbf{C}_{pq}$ of $\mathbf{C}$ is diagonal with entries

$$(\mathbf{C}_{pq})_{ii} = \hat{\mathbf{b}}_p^\top\mathbf{\Pi}^i\hat{\mathbf{b}}_{*,q} \tag{33}$$

Analogously to the matrix completion analysis in (Jain et al., 2013), we define the following matrices, for all $i \in [rn]$:

$$\mathbf{G}^i \coloneqq \left[\hat{\mathbf{b}}_p^\top\mathbf{\Pi}^i\hat{\mathbf{b}}_q\right]_{1\le p,q\le k} = \hat{\mathbf{B}}^\top\mathbf{\Pi}^i\hat{\mathbf{B}}, \quad \mathbf{C}^i \coloneqq \left[\hat{\mathbf{b}}_p^\top\mathbf{\Pi}^i\hat{\mathbf{b}}_{*,q}\right]_{1\le p,q\le k} = \hat{\mathbf{B}}^\top\mathbf{\Pi}^i\hat{\mathbf{B}}_* \tag{34}$$

In words, $\mathbf{G}^i$ is the $k \times k$ matrix formed by taking the i-th diagonal entry of each block $\mathbf{G}_{pq}$, and likewise for $\mathbf{C}^i$. Recall that $\mathbf{D}$ also has diagonal blocks, in particular $\mathbf{D}_{pq} = \langle\hat{\mathbf{B}}_p, \hat{\mathbf{B}}_q^*\rangle\mathbf{I}_d$, thus we also define $\mathbf{D}^i \coloneqq [\langle\hat{\mathbf{B}}_p, \hat{\mathbf{B}}_q^*\rangle]_{1\le p,q\le k} = \hat{\mathbf{B}}^\top\hat{\mathbf{B}}_*$.

Using this notation we can decouple $\mathbf{G}^{-1}(\mathbf{GD}-\mathbf{C})\mathbf{w}^*$ into i subvectors. Namely, let $\mathbf{w}_i^* \in \mathbb{R}^k$ be the vector formed by taking the $((p-1)rn+i)$-th elements of $\mathbf{w}^*$ for $p = 0, ..., k-1$, and similarly, let $\mathbf{f}_i$ be the vector formed by taking the $((p-1)rn+i)$-th elements of $\mathbf{G}^{-1}(\mathbf{GD}-\mathbf{C})\mathbf{w}^*$ for $p = 0, ..., k-1$. Then

$$\mathbf{f}_i = (\mathbf{G}^i)^{-1}(\mathbf{G}^i\mathbf{D}^i - \mathbf{C}^i)\mathbf{w}_i^* \tag{35}$$

is the i-th row of $\mathbf{F}$. Now we control $\|\mathbf{F}\|_F$.

Lemma 3. *Let $\delta_k = c\frac{k^{3/2}\sqrt{\log(rn)}}{\sqrt{m}}$ for some absolute constant c, then*

$$\|\mathbf{G}^{-1}\|_2 \le \frac{1}{1-\delta_k}$$

with probability at least $1 - e^{-111k^3\log(rn)}$.

Proof. We must lower bound $\sigma_{\min}(\mathbf{G})$. For some vector $\mathbf{z} \in \mathbb{R}^{rnk}$, let $\mathbf{z}^i \in \mathbb{R}^k$ denote the vector formed by taking the $((p-1)rn+i)$-th elements of $\mathbf{z}$ for $p = 0, ..., k-1$. Since $\mathbf{G}$ is symmetric, we have

$$\begin{aligned}\sigma_{\min}(\mathbf{G}) &= \min_{\mathbf{z}:\|\mathbf{z}\|_2=1}\mathbf{z}^\top\mathbf{G}\mathbf{z} \\ &= \min_{\mathbf{z}:\|\mathbf{z}\|_2=1}\sum_{i=1}^{rn}(\mathbf{z}^i)^\top\mathbf{G}^i\mathbf{z}^i \\ &= \min_{\mathbf{z}:\|\mathbf{z}\|_2=1}\sum_{i=1}^{rn}(\mathbf{z}^i)^\top\hat{\mathbf{B}}^\top\mathbf{\Pi}^i\hat{\mathbf{B}}\mathbf{z}^i \\ &\ge \min_{i\in[rn]}\sigma_{\min}(\hat{\mathbf{B}}^\top\mathbf{\Pi}^i\hat{\mathbf{B}})\end{aligned}$$

Note that the matrix $\hat{\mathbf{B}}^\top \mathbf{\Pi}^i \hat{\mathbf{B}}$ can be written as follows:

$$\hat{\mathbf{B}}^\top \mathbf{\Pi}^i \hat{\mathbf{B}} = \sum_{j=1}^{m} \frac{1}{\sqrt{m}} \hat{\mathbf{B}}^\top \mathbf{x}_i^j \left(\frac{1}{\sqrt{m}} \hat{\mathbf{B}}^\top \mathbf{x}_i^j \right)^\top \tag{36}$$

Let $\mathbf{v}_i^j := \frac{1}{\sqrt{m}} \hat{\mathbf{B}}^\top \mathbf{x}_i^j$ for all $i \in [rn]$ and $j \in [m]$, and note that each $\mathbf{v}_i^j$ is i.i.d. $\frac{1}{\sqrt{m}} \hat{\mathbf{B}}$-sub-gaussian. Thus using the one-sided version of equation (4.22) (Theorem 4.6.1) in (Vershynin, 2018), we have

$$\sigma_{min}(\hat{\mathbf{B}}^\top \mathbf{\Pi}^i \hat{\mathbf{B}}) \geq 1 - C \left(\sqrt{\frac{k}{m}} + \frac{r}{\sqrt{m}} \right) \tag{37}$$

with probability at least $1 - e^{-r^2}$ for $m \geq k$ and some absolute constant C. Choosing r such that $\delta_k = C \left(\sqrt{\frac{k}{m}} + \frac{r}{\sqrt{m}} \right)$ yields

$$\sigma_{min}(\hat{\mathbf{B}}^\top \mathbf{\Pi}^i \hat{\mathbf{B}}) \geq 1 - \delta_k \tag{38}$$

with probability at least $1 - e^{-(\delta_k \sqrt{m}/C - \sqrt{k})^2}$ for $m > k$. Now, letting $\delta_k = \frac{12Ck^{3/2}\sqrt{\log(rn)}}{\sqrt{m}}$, we have that (38) holds with probability at least

$$\begin{aligned} 1 - \exp\left(- \left(12k^{3/2}\sqrt{\log(rn)} - \sqrt{k} \right)^2 \right) &\geq 1 - \exp\left(-k(12\sqrt{k}\sqrt{\log(rn)} - 1)^2 \right) \\ &\geq 1 - \exp\left(121k^3 \log(rn) \right) \end{aligned} \tag{39}$$

Finally, taking a union bound over $i \in [n]$ yields $\sigma_{\min}(\mathbf{G}) \geq 1 - \delta_k$ with probability at least

$$1 - rn \exp\left(-121k^3 \log(rn) \right) \geq 1 - e^{-110k^3 \log(rn)}, \tag{40}$$

completing the proof. □

Lemma 4. *Let $\delta_k = c\frac{k^{3/2}\sqrt{\log(rn)}}{\sqrt{m}}$ for some absolute constant c, then*

$$\|(\mathbf{GD} - \mathbf{C})\mathbf{w}^*\|_2 \leq \delta_k \|\mathbf{W}^*\|_2 \ \operatorname{dist}(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*)$$

with probability at least $1 - e^{-111k^2 \log(rn)}$.

Proof. For ease of notation we drop superscripts t. We define $\mathbf{H} = \mathbf{GD} - \mathbf{C}$ and

$$\mathbf{H}^i := \mathbf{G}^i \mathbf{D}^i - \mathbf{C}^i = \hat{\mathbf{B}}^\top \mathbf{\Pi} \hat{\mathbf{B}} \hat{\mathbf{B}}^\top \hat{\mathbf{B}}^* - \hat{\mathbf{B}}^\top \mathbf{\Pi} \hat{\mathbf{B}}^* = \hat{\mathbf{B}}^\top \left(\frac{1}{m} \mathbf{X}_i^\top \mathbf{X}_i \right) (\hat{\mathbf{B}} \hat{\mathbf{B}}^\top - \mathbf{I}_d) \hat{\mathbf{B}}^*, \tag{41}$$

for all $i \in [rn]$. Then we have

$$\begin{aligned} \|(\mathbf{GD} - \mathbf{C})\mathbf{w}_*\|_2^2 &= \sum_{i=1}^{rn} \|\mathbf{H}^i \mathbf{w}_*^i\|_2^2 \\ &\leq \sum_{i=1}^{rn} \|\mathbf{H}^i\|_2^2 \|\mathbf{w}_i^*\|_2^2 \\ &\leq \frac{k}{rn} \|\mathbf{W}^*\|_2^2 \sum_{i=1}^{rn} \|\mathbf{H}^i\|_2^2 \end{aligned} \tag{42}$$

where the last inequality follows almost surely from Assumption 3 (the 1-row-wise incoherence of $\mathbf{W}^*$) and the fact that $krn = \|\mathbf{W}^*\|_F^2 \leq k\|\mathbf{W}^*\|_2^2$ by Assumption 3 and the fact that $\mathbf{W}^*$ has rank k. It remains to bound $\frac{1}{rn}\sum_{i=1}^{rn} \|\mathbf{H}^i\|_2^2$. Although $\|\mathbf{H}^i\|_2$ is sub-exponential, as we will show, $\|\mathbf{H}^i\|_2^2$ is not sub-exponential, so we cannot directly apply standard

concentration results. Instead, we compute a tail bound for each $\|\mathbf{H}^i\|_2^2$ individually, then then union bound over $i \in [rn]$. Let $\mathbf{U} \coloneqq \frac{1}{\sqrt{m}}\mathbf{X}_i(\hat{\mathbf{B}}\hat{\mathbf{B}}^\top - \mathbf{I}_d)\hat{\mathbf{B}}^*$, then the j-th row of $\mathbf{U}$ is given by

$$\mathbf{u}_j = \frac{1}{\sqrt{m}}\hat{\mathbf{B}}^{*\top}(\hat{\mathbf{B}}\hat{\mathbf{B}}^\top - \mathbf{I}_d)\mathbf{x}_i^j,$$

and is $\frac{1}{\sqrt{m}}\hat{\mathbf{B}}^{*\top}(\hat{\mathbf{B}}\hat{\mathbf{B}}^\top - \mathbf{I}_d)$-sub-gaussian. Likewise, define $\mathbf{V} \coloneqq \frac{1}{\sqrt{m}}\mathbf{X}_i\hat{\mathbf{B}}$, then the j-th row of $\mathbf{V}$ is

$$\mathbf{v}_j = \frac{1}{\sqrt{m}}\hat{\mathbf{B}}^\top\mathbf{x}_i^j,$$

therefore is $\frac{1}{\sqrt{m}}\hat{\mathbf{B}}$-sub-gaussian. We leverage the sub-gaussianity of the rows of $\mathbf{U}$ and $\mathbf{V}$ to make a similar concentration argument as in Proposition 4.4.5 in (Vershynin, 2018). First, let $\mathcal{S}^{k-1}$ denote the unit sphere in k dimensions, and let $\mathcal{N}_k$ be a $\frac{1}{4}$-th net of cardinality $|\mathcal{N}_k| \leq 9^k$, which exists by Corollary 4.2.13 in (Vershynin, 2018). Next, using equation 4.13 in (Vershynin, 2018), we obtain

$$\begin{aligned}
\|(\hat{\mathbf{B}}^*)^\top(\hat{\mathbf{B}}\hat{\mathbf{B}}^\top - \mathbf{I}_d)\mathbf{X}_i^\top\mathbf{X}_i\mathbf{B}\|_2 = \left\|\mathbf{U}^\top\mathbf{V}\right\|_2 &\leq 2\max_{\mathbf{z},\mathbf{y}\in\mathcal{N}_k}\mathbf{z}^\top\left(\mathbf{U}^\top\mathbf{V}\right)\mathbf{y} \\
&= 2\max_{\mathbf{z},\mathbf{y}\in\mathcal{N}_k}\mathbf{z}^\top\left(\sum_{j=1}^m\mathbf{u}_j\mathbf{v}_j^\top\right)\mathbf{y} \\
&= 2\max_{\mathbf{z},\mathbf{y}\in\mathcal{N}_k}\sum_{j=1}^m\langle\mathbf{z},\mathbf{u}_j\rangle\langle\mathbf{v}_j,\mathbf{y}\rangle
\end{aligned}$$

By definition of sub-gaussianity, $\langle\mathbf{z},\mathbf{u}_j\rangle$ and $\langle\mathbf{v}_j,\mathbf{y}\rangle$ are sub-gaussian with norms $\frac{1}{\sqrt{m}}\|\hat{\mathbf{B}}^{*\top}(\hat{\mathbf{B}}\hat{\mathbf{B}}^\top - \mathbf{I}_d)\|_2 = \frac{1}{\sqrt{m}}\text{dist}(\hat{\mathbf{B}},\hat{\mathbf{B}}^*)$ and $\frac{1}{\sqrt{m}}\|\hat{\mathbf{B}}\|_2 = \frac{1}{\sqrt{m}}$, respectively. Thus for all $j \in [m]$, $\langle\mathbf{z},\mathbf{u}_j\rangle\langle\mathbf{v}_j,\mathbf{z}\rangle$ is sub-exponential with norm $\frac{c}{m}\text{dist}(\hat{\mathbf{B}},\hat{\mathbf{B}}^*)$ for some absolute constant c. Note that for any $j \in [m]$ and any $\mathbf{z}$, $\mathbb{E}[\langle\mathbf{z},\mathbf{u}_j\rangle\langle\mathbf{v}_j,\mathbf{y}\rangle] = \mathbf{z}^\top((\hat{\mathbf{B}}^*)^\top(\hat{\mathbf{B}}\hat{\mathbf{B}}^\top - \mathbf{I}_d)\mathbf{B})\mathbf{y} = 0$. Thus we have a sum of m mean-zero, independent sub-exponential random variables. We can now use Bernstein's inequality to obtain, for any fixed $\mathbf{z},\mathbf{y} \in \mathcal{N}_k$,

$$\mathbb{P}\left(\sum_{j=1}^m\langle\mathbf{z},\mathbf{u}_j\rangle\langle\mathbf{v}_j,\mathbf{y}\rangle \geq s\right) \leq \exp\left(-c'm\min\left(\frac{s^2}{\text{dist}^2(\hat{\mathbf{B}},\hat{\mathbf{B}}^*)}, \frac{s}{\text{dist}(\hat{\mathbf{B}},\hat{\mathbf{B}}^*)}\right)\right) \tag{43}$$

Now union bound over all $\mathbf{z},\mathbf{y} \in \mathcal{N}_k$ to obtain

$$\mathbb{P}\left(\frac{1}{m}\|(\hat{\mathbf{B}}^*)^\top(\hat{\mathbf{B}}\hat{\mathbf{B}}^\top - \mathbf{I}_d)\mathbf{X}_i^\top\mathbf{X}_i\hat{\mathbf{B}}\|_2 \geq 2s\right) \leq 9^{2k}\exp\left(-c'm\min(s^2/\text{dist}^2(\hat{\mathbf{B}},\hat{\mathbf{B}}^*), s/\text{dist}(\hat{\mathbf{B}},\hat{\mathbf{B}}^*))\right) \tag{44}$$

Let $\frac{s}{\text{dist}(\hat{\mathbf{B}},\hat{\mathbf{B}}^*)} = \max(\varepsilon,\varepsilon^2)$ for some $\epsilon > 0$, then it follows that $\min(s^2/\text{dist}^2(\hat{\mathbf{B}},\hat{\mathbf{B}}^*), s/\text{dist}(\hat{\mathbf{B}},\hat{\mathbf{B}}^*)) = \varepsilon^2$. So we have

$$\mathbb{P}\left(\frac{1}{m}\|(\hat{\mathbf{B}}^*)^\top(\hat{\mathbf{B}}\hat{\mathbf{B}}^\top - \mathbf{I}_d)\mathbf{X}_i^\top\mathbf{X}_i\hat{\mathbf{B}}\|_2 \geq 2\text{dist}(\hat{\mathbf{B}},\hat{\mathbf{B}}^*)\max(\varepsilon,\varepsilon^2)\right) \leq 9^{2k}e^{-c'm\varepsilon^2} \tag{45}$$

Moreover, letting $\varepsilon^2 = \frac{ck^2\log(rn)}{4m}$ for some constant c, and $m \geq ck^2\log(rn)$, we have

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{m}\|(\hat{\mathbf{B}}^*)^\top(\hat{\mathbf{B}}\hat{\mathbf{B}}^\top - \mathbf{I}_d)\mathbf{X}_i^\top\mathbf{X}_i\hat{\mathbf{B}}\|_2 \geq \text{dist}(\hat{\mathbf{B}},\hat{\mathbf{B}}^*)\sqrt{\frac{ck^2\log(rn)}{m}}\right) &\leq 9^{2k}e^{-c_1k^2\log(rn)} \\
&\leq e^{-111k^2\log(rn)}
\end{aligned} \tag{46}$$

for large enough constant c_1. Thus, noting that $\|\mathbf{H}^i\|_2^2 = \|\frac{1}{m}(\hat{\mathbf{B}}^*)^\top(\hat{\mathbf{B}}\hat{\mathbf{B}}^\top - \mathbf{I}_d)\mathbf{X}_i^\top\mathbf{X}_i\hat{\mathbf{B}}\|_2^2$, we obtain

$$\mathbb{P}\left(\|\mathbf{H}^i\|_2^2 \geq c\,\text{dist}^2(\hat{\mathbf{B}},\hat{\mathbf{B}}^*)\frac{k^2\log(rn)}{m}\right) \leq e^{-111k^2\log(rn)} \tag{47}$$

Thus, using (42), we have

$$\begin{aligned}
&\mathbb{P}\left(\|(\mathbf{GD}-\mathbf{C})\mathbf{w}_*\|_2^2 \geq c\|\mathbf{W}^*\|_2^2\ \mathrm{dist}^2(\hat{\mathbf{B}},\hat{\mathbf{B}}^*)\frac{k^3\log(rn)}{m}\right)\\
&\leq \mathbb{P}\left(\frac{k}{rn}\|\mathbf{W}^*\|_2^2\sum_{i=1}^{rn}\|\mathbf{H}^i\|_2^2 \geq c\|\mathbf{W}^*\|_2^2\,\mathrm{dist}^2(\hat{\mathbf{B}},\hat{\mathbf{B}}^*)\frac{k^3\log(rn)}{m}\right)\\
&= \mathbb{P}\left(\frac{1}{rn}\sum_{i=1}^{rn}\|\mathbf{H}^i\|_2^2 \geq c\,\mathrm{dist}^2(\hat{\mathbf{B}},\hat{\mathbf{B}}^*)\frac{k^2\log(rn)}{m}\right)\\
&\leq rn\mathbb{P}\left(\|\mathbf{H}^1\|_2^2 \geq c\,\mathrm{dist}^2(\hat{\mathbf{B}},\hat{\mathbf{B}}^*)\frac{k^2\log(rn)}{m}\right)\\
&\leq e^{-110k^2\log(rn)}
\end{aligned}$$

completing the proof.

□

Lemma 5. *Let* $\delta_k = \frac{ck^{3/2}\sqrt{\log(rn)}}{\sqrt{m}}$*, then*

$$\|\mathbf{F}\|_F \leq \frac{\delta_k}{1-\delta_k}\|\mathbf{W}^*\|_2\ \mathrm{dist}(\hat{\mathbf{B}}_t,\hat{\mathbf{B}}_*) \tag{48}$$

with probability at least $1-e^{-110`k^2\log(n)}$.

Proof. By the definition of $\mathbf{F}$ and the Cauchy-Schwarz inequality, we have $\|\mathbf{F}\|_F = \|\mathbf{G}^{-1}(\mathbf{GD}-\mathbf{C})\widetilde{\mathbf{w}}^*\|_2 \leq \|\mathbf{G}^{-1}\|_2\|(\mathbf{GD}-\mathbf{C})\widetilde{\mathbf{w}}^*\|_2$. Combining the bound on $\|\mathbf{G}^{-1}\|_2$ from Lemma 3 and the bound on $\|(\mathbf{GD}-\mathbf{C})\widetilde{\mathbf{w}}^*\|_2$ from Lemma 4 via a union bound yields the result.

□

We next focus on showing concentration of the operator $\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}$ to the identity operator.

Lemma 6. *Let* $\delta_k' = ck\frac{\sqrt{d}}{\sqrt{rnm}}$ *for some absolute constant c. Then for any t, if* $\delta_k' \leq k$*,*

$$\frac{1}{rn}\left\|\left(\frac{1}{m}\mathcal{A}^*\mathcal{A}(\mathbf{Q}^t)-\mathbf{Q}^t\right)^\top\mathbf{W}^{t+1}\right\|_2 \leq \delta_k'\ \mathrm{dist}(\hat{\mathbf{B}}^t,\hat{\mathbf{B}}^*) \tag{49}$$

with probability at least $1-e^{-110d}-e^{-110k^2\log(rn)}$.

Proof. We drop superscripts t for simplicity. We first bound the norms of the rows of $\mathbf{Q}$ and $\mathbf{W}$. Let $\mathbf{q}_i \in \mathbb{R}^d$ be the i-th row of $\mathbf{Q}$ and let $\mathbf{w}_i \in \mathbb{R}^k$ be the i-th row of $\mathbf{W}$. Recall the computation of $\mathbf{W}$ from Lemma 2:

$$\mathbf{W} = \mathbf{W}_*\hat{\mathbf{B}}_*^\top\hat{\mathbf{B}} - \mathbf{F} \implies \mathbf{w}_i^\top = (\hat{\mathbf{w}}_i^*)^\top\hat{\mathbf{B}}_*^\top\hat{\mathbf{B}} - \mathbf{f}_i^\top$$

Thus

$$\begin{aligned}
\|\mathbf{q}_i\|_2^2 &= \|\hat{\mathbf{B}}\hat{\mathbf{B}}^\top\hat{\mathbf{B}}^*\hat{\mathbf{w}}_i^* - \hat{\mathbf{B}}\mathbf{f}_i - \hat{\mathbf{B}}^*\hat{\mathbf{w}}_i^*\|_2^2\\
&= \|(\hat{\mathbf{B}}\hat{\mathbf{B}}^\top - \mathbf{I}_d)\hat{\mathbf{B}}^*\hat{\mathbf{w}}_i^* - \hat{\mathbf{B}}\mathbf{f}_i\|_2^2\\
&\leq 2\|(\hat{\mathbf{B}}\hat{\mathbf{B}}^\top - \mathbf{I}_d)\hat{\mathbf{B}}^*\hat{\mathbf{w}}_i^*\|_2^2 + 2\|\hat{\mathbf{B}}\mathbf{f}_i\|_2^2\\
&\leq 2\|(\hat{\mathbf{B}}\hat{\mathbf{B}}^\top - \mathbf{I}_d)\hat{\mathbf{B}}^*\|_2^2\|\hat{\mathbf{w}}_i^*\|_2^2 + 2\|\mathbf{f}_i\|_2^2\\
&= 2k\mathrm{dist}^2(\hat{\mathbf{B}},\hat{\mathbf{B}}^*) + 2\|\mathbf{f}_i\|_2^2
\end{aligned} \tag{50}$$

Also recall that $\text{vec}(\mathbf{F}) = \mathbf{G}^{-1}(\mathbf{GD} - \mathbf{C})\hat{\mathbf{w}}_*$ from Lemma 2. From equation (35), the i-th row of $\mathbf{F}$ is given by:

$$\mathbf{f}_i = (\mathbf{G}^i)^{-1}(\mathbf{G}^i\mathbf{D}^i - \mathbf{C}^i)\mathbf{w}_i^*$$

Thus, using the Cauchy-Schwarz inequality and our previous bounds,

$$\begin{aligned}\|\mathbf{f}_i\|_2^2 &\leq \|(\mathbf{G}^i)^{-1}\|_2^2 \, \|\mathbf{G}^i\mathbf{D}^i - \mathbf{C}^i\|_2^2 \;\; \|\mathbf{w}_i^*\|_2^2 \\ &\leq \|(\mathbf{G}^i)^{-1}\|_2^2 \, \|\mathbf{G}^i\mathbf{D}^i - \mathbf{C}^i\|_2^2 \; k \end{aligned} \tag{51}$$

where (51) follows by Assumption 3. From (47), we have that

$$\mathbb{P}\left(\|\mathbf{G}^i\mathbf{D}^i - \mathbf{C}^i\|_2^2 \geq \delta_k^2 \, \text{dist}^2(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*)\right) \leq e^{-112k^2\log(rn)}$$

where δ_k is defined in 3. Similarly, from equations (38) and (39), we have that

$$\mathbb{P}\left(\|(\mathbf{G}^i)^{-1}\|_2^2 \geq \frac{1}{(1-\delta_k)^2}\right) \leq e^{-121k^3\log(rn)} \tag{52}$$

Now plugging this back into (51) and assuming $\delta_k \leq \frac{1}{2}$, we obtain

$$\|\mathbf{q}_i\|_2^2 \leq 2k \, \text{dist}^2(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*) \;\; \left(1 + \frac{\delta_k^2}{(1-\delta_k)^2}\right) \leq 4k \, \text{dist}^2(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*) \tag{53}$$

with probability at least $1 - e^{-111k^2\log(rn)}$. Likewise, to upper bound $\|\mathbf{w}_i\|_2$ we have

$$\begin{aligned}\|\mathbf{w}_i\|_2^2 &\leq 2\|\hat{\mathbf{B}}^\top\hat{\mathbf{B}}^*\mathbf{w}_i^*\|_2^2 + 2\|\mathbf{f}_i\|_2^2 \\ &\leq 2\|\hat{\mathbf{B}}^\top\hat{\mathbf{B}}^*\|_2^2\|\mathbf{w}_i^*\|_2^2 + 2\|\mathbf{f}_i\|_2^2 \\ &\leq 2k + 2\frac{\delta_k^2}{(1-\delta_k)^2}\text{dist}^2(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*)k \qquad (54) \\ &\leq 4k \qquad (55)\end{aligned}$$

where (54) holds with probability at least $1 - e^{-111k^2\log(rn)}$ conditioning on the same event as in (53), and (55) holds almost surely as long as $\delta_k \leq 1/2$. For the rest of the proof we condition on the event $\mathcal{E} := \cap_{i=1}^{rn}\left\{\|\mathbf{q}_i\|_2^2 \leq 4k \, \text{dist}^2(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*) \cap \|\mathbf{w}_i\|_2^2 \leq 4k\right\}$, which holds with probability at least $1 - e^{-110k^2\log(rn)}$ by a union bound over $i \in [rn]$. Observe that the matrix $\frac{1}{m}\mathcal{A}^*\mathcal{A}(\mathbf{Q}) - \mathbf{Q}$ can be re-written as

$$\begin{aligned}\frac{1}{m}\mathcal{A}^*\mathcal{A}(\mathbf{Q}) - \mathbf{Q} &= \frac{1}{m}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\langle \mathbf{e}_i(\mathbf{x}_i^j)^\top, \mathbf{Q}\rangle \; \mathbf{e}_i(\mathbf{x}_i^j)^\top - \mathbf{Q}\right) \\ &= \frac{1}{m}\sum_{i=1}^{rn}\sum_{j=1}^{m}\langle \mathbf{x}_i^j, \mathbf{q}_i\rangle \; \mathbf{e}_i(\mathbf{x}_i^j)^\top - \mathbf{Q} \end{aligned} \tag{56}$$

Multiplying the transpose by $\frac{1}{rn}\mathbf{W}$ yields

$$\frac{1}{rn}\left(\frac{1}{m}\mathcal{A}^*\mathcal{A}(\mathbf{Q}) - \mathbf{Q}\right)^\top \mathbf{W} = \frac{1}{rnm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\langle \mathbf{x}_i^j, \, \mathbf{q}_i\rangle \; \mathbf{x}_i^j(\mathbf{w}_i)^\top - \mathbf{q}_i(\mathbf{w}_i)^\top\right) \tag{57}$$

where we have used the fact that $(\mathbf{Q})^\top\mathbf{W} = \sum_{i=1}^n \mathbf{q}_i(\mathbf{w}_i)^\top$. We will argue similarly as in Proposition 4.4.5 in (Vershynin, 2018) to bound the spectral norm of the d-by-k matrix in the RHS of (57).

First, let $\mathcal{S}^{d-1}$ and $\mathcal{S}^{k-1}$ denote the unit spheres in d and k dimensions, respectively. Construct $\frac{1}{4}$-nets $\mathcal{N}_d$ and $\mathcal{N}_k$ over $\mathcal{S}^{d-1}$ and $\mathcal{S}^{k-1}$, respectively, such that $|\mathcal{N}_d| \leq 9^d$ and $|\mathcal{N}_k| \leq 9^k$ (which is possible by Corollary 4.2.13 in (Vershynin,

2018)). Then, using equation 4.13 in (Vershynin, 2018), we have

$$\begin{aligned}
&\left\| \frac{1}{rnm} \sum_{i=1}^{rn} \sum_{j=1}^{m} \left(\langle \mathbf{x}_i^j, \mathbf{q}_i \rangle \, \mathbf{x}_i^j (\mathbf{w}_i)^\top - \mathbf{q}_i (\mathbf{w}_i)^\top \right) \right\|_2^2 \\
&\leq 2 \max_{\mathbf{u} \in \mathcal{N}_d, \mathbf{v} \in \mathcal{N}_k} \mathbf{u}^\top \left(\sum_{i=1}^{rn} \sum_{j=1}^{m} \left(\frac{1}{rnm} \langle \mathbf{x}_i^j, \mathbf{q}_i \rangle \, \mathbf{x}_i^j (\mathbf{w}_i)^\top - \frac{1}{rnm} \mathbf{q}_i (\mathbf{w}_i)^\top \right) \right) \mathbf{v} \\
&= 2 \max_{\mathbf{u} \in \mathcal{N}_d, \mathbf{v} \in \mathcal{N}_k} \sum_{i=1}^{rn} \sum_{j=1}^{m} \left(\frac{1}{rnm} \langle \mathbf{x}_i^j, \mathbf{q}_i \rangle \langle \mathbf{u}, \mathbf{x}_i^j \rangle \langle \mathbf{w}_i, \mathbf{v} \rangle - \frac{1}{rnm} \langle \mathbf{u}, \mathbf{q}_i \rangle \langle \mathbf{w}_i, \mathbf{v} \rangle \right) \qquad (58)
\end{aligned}$$

By the $\mathbf{I}_d$-sub-gaussianity of $\mathbf{x}_i^j$, the inner product $\langle \mathbf{u}, \mathbf{x}_i^j \rangle$ is sub-gaussian with norm at most $c\|\mathbf{u}\|_2 = c$ for some absolute constant c for any fixed $\mathbf{u} \in \mathcal{N}_d$. Similarly, $\langle \mathbf{x}_i^j, \mathbf{q}_i \rangle$ is sub-gaussian with norm at most $\|\mathbf{q}_i\|_2 \leq 2c\sqrt{k}\ \mathrm{dist}(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*)$ using (53). Further, since the sub-exponential norm of the product of two sub-gaussian random variables is at most the product of the sub-gaussian norms of the two random variables (Lemma 2.7.7 in (Vershynin, 2018)), we have that $\langle \mathbf{x}_i^j, \mathbf{q}_i \rangle \langle \mathbf{u}, \mathbf{x}_i^j \rangle$ is sub-exponential with norm at most $2c^2\sqrt{k}\ \mathrm{dist}(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*)$. Further, $\frac{1}{rnm} \langle \mathbf{x}_i^j, \mathbf{q}_i \rangle \langle \mathbf{u}, \mathbf{x}_i^j \rangle \langle \mathbf{w}_i, \mathbf{v} \rangle$ is sub-exponential with norm at most

$$\frac{2c^2\sqrt{k}}{rnm}\ \mathrm{dist}(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*) \langle \mathbf{w}_i, \mathbf{v} \rangle \leq \frac{2c^2\sqrt{k}}{rnm}\ \mathrm{dist}(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*) \|\mathbf{w}_i\|_2 \leq \frac{c_1 k}{rnm}\ \mathrm{dist}(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*).$$

Finally, note that $\mathbb{E}[\frac{1}{rnm} \langle \mathbf{x}_i^j, \mathbf{q}_i \rangle \langle \mathbf{u}, \mathbf{x}_i^j \rangle \langle \mathbf{w}_i, \mathbf{v} \rangle - \frac{1}{rnm} \langle \mathbf{u}, \mathbf{q}_i \rangle \langle \mathbf{w}_i, \mathbf{v} \rangle] = 0$. Thus, we have a sum of rnm independent, mean zero sub-exponential random variables, so we apply Bernstein's inequality.

$$\begin{aligned}
&\mathbb{P}\left(\sum_{i=1}^{rn} \sum_{j=1}^{m} \left(\frac{1}{rnm} \langle \mathbf{x}_i^j, \mathbf{q}_i \rangle \langle \mathbf{u}, \mathbf{x}_i^j \rangle \langle \mathbf{w}_i, \mathbf{v} \rangle - \frac{1}{rnm} \langle \mathbf{u}, \mathbf{q}_i \rangle \langle \mathbf{w}_i, \mathbf{v} \rangle \right) \geq s \right) \\
&\leq \exp\left(-c_1 rnm \min\left(\frac{s^2}{k^2\ \mathrm{dist}^2(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*)}, \frac{s}{k \mathrm{dist}(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*)} \right) \right)
\end{aligned}$$

Union bounding over all $\mathbf{u} \in \mathcal{N}_d$ and $\mathbf{v} \in \mathcal{N}_k$, we obtain

$$\mathbb{P}\left(\left\| \frac{1}{rn} \left(\frac{1}{m} \mathcal{A}^* \mathcal{A}(\mathbf{Q}) - \mathbf{Q} \right)^\top \mathbf{W} \right\|_2 \geq 2s \;\middle|\; \mathcal{E} \right) \leq 9^{d+k} \exp\left(-c_1 rnm \min\left(\frac{s^2}{k^2\ \mathrm{dist}^2(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*)}, \frac{s}{k \mathrm{dist}(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*)} \right) \right)$$

Let $\frac{s}{k\ \mathrm{dist}(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*)} = \max(\epsilon, \epsilon^2)$ for some $\epsilon > 0$, then $\epsilon^2 = \min\left(\frac{s^2}{k^2\ \mathrm{dist}^2(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*)}, \frac{s}{k\ \mathrm{dist}(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*)} \right)$. Further, let $\epsilon^2 = \frac{112(d+k)}{c_1 rnm}$, then as long as $\epsilon^2 \leq 1$, we have

$$\mathbb{P}\left(\left\| \frac{1}{rn} \left(\frac{1}{m} \mathcal{A}^* \mathcal{A}(\mathbf{Q}) - \mathbf{Q} \right)^\top \mathbf{W} \right\|_2 \geq c_2 k\ \mathrm{dist}(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*) \sqrt{d/(rnm)} \;\middle|\; \mathcal{E}^c \right) \leq e^{-110(d+k)} \leq e^{-110d}.$$

Finally, we use $\mathbb{P}(A \mid \mathcal{E}^c) \leq \mathbb{P}(A \mid \mathcal{E}^c) + \mathbb{P}(\mathcal{E}^c)$, where $A := \left\{ \left\| \frac{1}{rn} \left(\frac{1}{m} \mathcal{A}^* \mathcal{A}(\mathbf{Q}) - \mathbf{Q} \right)^\top \mathbf{W} \right\|_2 \geq c_2 k\ \mathrm{dist}(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*) \sqrt{d/(rnm)} \right\}$, to complete the proof.

□

B.3. Main Result

Now we are ready to show Theorem 1, which follows immediately from the following descent lemma.

Lemma 7. *Define* $E_0 := 1 - \mathrm{dist}^2(\hat{\mathbf{B}}^0, \hat{\mathbf{B}}^*)$ *and* $\bar{\sigma}_{\max,*} := \max_{\mathcal{I} \in [n], |\mathcal{I}| = rn} \sigma_{\max}(\frac{1}{\sqrt{rn}} \mathbf{W}_{\mathcal{I}}^*)$ *and* $\bar{\sigma}_{\min,*} := \min_{\mathcal{I} \in [n], |\mathcal{I}| = rn} \sigma_{\min}(\frac{1}{\sqrt{rn}} \mathbf{W}_{\mathcal{I}}^*)$*, i.e. the maximum and minimum singular values of any matrix that can be obtained by taking* rn *rows of* $\frac{1}{\sqrt{rn}} \mathbf{W}^*$.

Suppose that $m \geq c(\kappa^4 k^3 \log(rn)/E_0^2 + \kappa^4 k^2 d/(E_0^2 rn))$ *for some absolute constant* c*. Then for any* t *and any* $\eta \leq 1/(4\bar{\sigma}^2_{\max,*})$*, we have*

$$\operatorname{dist}(\hat{\mathbf{B}}^{t+1}, \hat{\mathbf{B}}^*) \leq \left(1 - \eta E_0 \bar{\sigma}^2_{\min,*}/2\right)^{1/2} \operatorname{dist}(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*),$$

with probability at least $1 - e^{-100 \min(k^2 \log(rn), d)}$.

Proof. Recall that $\mathbf{W}^{t+1} \in \mathbb{R}^{rn \times k}$ and $\mathbf{B}^{t+1} \in \mathbb{R}^{d \times k}$ are computed as follows:

$$\mathbf{W}^{t+1} = \operatorname*{argmin}_{\mathbf{W} \in \mathbb{R}^{rn \times k}} \frac{1}{2rnm} \|\mathcal{A}(\mathbf{W}^* \hat{\mathbf{B}}^{*\top} - \mathbf{W} \hat{\mathbf{B}}^{t\top})\|_2^2 \tag{59}$$

$$\mathbf{B}^{t+1} = \hat{\mathbf{B}}^t - \frac{\eta}{rnm} \left(\mathcal{A}^\dagger \mathcal{A}(\mathbf{W}^{t+1} \hat{\mathbf{B}}^{t\top} - \mathbf{W}^* \hat{\mathbf{B}}^{*\top})\right)^\top \mathbf{W}^{t+1} \tag{60}$$

Let $\mathbf{Q}^t = \mathbf{W}^{t+1} \hat{\mathbf{B}}^{t\top} - \mathbf{W}^* \hat{\mathbf{B}}^{*\top}$. We have

$$\begin{aligned}
\mathbf{B}^{t+1} &= \hat{\mathbf{B}}^t - \frac{\eta}{rnm} \left(\mathcal{A}^\dagger \mathcal{A}(\mathbf{Q}^t)\right)^\top \mathbf{W}^{t+1} \\
&= \hat{\mathbf{B}}^t - \frac{\eta}{rn} \mathbf{Q}^{t\top} \mathbf{W}^{t+1} - \frac{\eta}{rn} \left(\frac{1}{m} \mathcal{A}^\dagger \mathcal{A}(\mathbf{Q}^t) - \mathbf{Q}^t\right)^\top \mathbf{W}^{t+1}
\end{aligned} \tag{61}$$

Now, multiply both sides by $\hat{\mathbf{B}}_\perp^{*\top}$ to obtain

$$\begin{aligned}
\hat{\mathbf{B}}_\perp^{*\top} \mathbf{B}_{t+1} &= \hat{\mathbf{B}}_\perp^{*\top} \hat{\mathbf{B}}^t - \frac{\eta}{rn} \hat{\mathbf{B}}_\perp^{*\top} \mathbf{Q}^{t\top} \mathbf{W}^{t+1} - \frac{\eta}{rn} \hat{\mathbf{B}}_\perp^{*\top} \left(\frac{1}{m} \mathcal{A}^\dagger \mathcal{A}(\mathbf{Q}^t) - \mathbf{Q}^t\right)^\top \mathbf{W}^{t+1} \\
&= \hat{\mathbf{B}}_\perp^{*\top} \hat{\mathbf{B}}^t (\mathbf{I}_k - \frac{\eta}{rn} \mathbf{W}^{t+1\top} \mathbf{W}^{t+1}) - \frac{\eta}{rn} \hat{\mathbf{B}}_\perp^{*\top} \left(\frac{1}{m} \mathcal{A}^\dagger \mathcal{A}(\mathbf{Q}^t) - \mathbf{Q}^t\right)^\top \mathbf{W}^{t+1}
\end{aligned} \tag{62}$$

where the second equality follows because $\hat{\mathbf{B}}_\perp^{*\top} \mathbf{Q}^{t\top} = \hat{\mathbf{B}}_\perp^{*\top} \hat{\mathbf{B}}^t \mathbf{W}^{t+1\top} - \hat{\mathbf{B}}_\perp^{*\top} \hat{\mathbf{B}}^* \mathbf{W}^{*\top} = \hat{\mathbf{B}}_\perp^{*\top} \hat{\mathbf{B}}^t \mathbf{W}^{t+1\top}$. Then, writing the QR decomposition of $\mathbf{B}^{t+1}$ as $\mathbf{B}^{t+1} = \hat{\mathbf{B}}^{t+1} \mathbf{R}^{t+1}$ and multiplying both sides of (62) from the right by $(\mathbf{R}^{t+1})^{-1}$ yields

$$\hat{\mathbf{B}}_\perp^{*\top} \hat{\mathbf{B}}_{t+1} = \left(\hat{\mathbf{B}}_\perp^{*\top} \hat{\mathbf{B}}^t (\mathbf{I}_k - \frac{\eta}{rn} (\mathbf{W}^{t+1})^\top \mathbf{W}^{t+1}) - \frac{\eta}{rn} \hat{\mathbf{B}}_\perp^{*\top} \left(\frac{1}{m} \mathcal{A}^\dagger \mathcal{A}(\mathbf{Q}^t) - \mathbf{Q}^t\right)^\top \mathbf{W}^{t+1}\right) (\mathbf{R}^{t+1})^{-1} \tag{63}$$

Hence,

$$\begin{aligned}
&\operatorname{dist}(\hat{\mathbf{B}}^{t+1}, \hat{\mathbf{B}}^*) \\
&= \left\| \left(\hat{\mathbf{B}}_\perp^{*\top} \hat{\mathbf{B}}^t (\mathbf{I}_k - \frac{\eta}{rn} (\mathbf{W}^{t+1})^\top \mathbf{W}^{t+1}) - \frac{\eta}{rn} \hat{\mathbf{B}}_\perp^{*\top} \left(\frac{1}{m} \mathcal{A}^\dagger \mathcal{A}(\mathbf{Q}^t) - \mathbf{Q}^t\right)^\top \mathbf{W}^{t+1}\right) (\mathbf{R}^{t+1})^{-1} \right\|_2 \\
&\leq \left\| \hat{\mathbf{B}}_\perp^{*\top} \hat{\mathbf{B}}^t (\mathbf{I}_k - \frac{\eta}{rn} (\mathbf{W}^{t+1})^\top \mathbf{W}^{t+1}) \right\|_2 \left\| (\mathbf{R}^{t+1})^{-1} \right\|_2 \\
&\quad + \frac{\eta}{rn} \left\| \hat{\mathbf{B}}_\perp^{*\top} \left(\frac{1}{m} \mathcal{A}^\dagger \mathcal{A}(\mathbf{Q}^t) - \mathbf{Q}^t\right)^\top \mathbf{W}^{t+1} \right\|_2 \left\| (\mathbf{R}^{t+1})^{-1} \right\|_2 \qquad (64) \\
&=: A_1 + A_2. \qquad (65)
\end{aligned}$$

where (64) follows by applying the triangle and Cauchy-Schwarz inequalities. We have thus split the upper bound on $\operatorname{dist}(\mathbf{B}_{t+1}, \hat{\mathbf{B}}_*)$ into two terms, A_1 and A_2. The second term, A_2, is small due to the concentration of $\frac{1}{m} \mathcal{A}^\dagger \mathcal{A}$ to the identity operator, and the first term is strictly smaller than $\operatorname{dist}(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*)$. We start by controlling A_2:

$$\begin{aligned}
A_2 &= \frac{\eta}{rn} \left\| \hat{\mathbf{B}}_\perp^{*\top} \left(\frac{1}{m} \mathcal{A}^\dagger \mathcal{A}(\mathbf{Q}^t) - \mathbf{Q}^t\right)^\top \mathbf{W}^{t+1} \right\|_2 \left\| (\mathbf{R}^{t+1})^{-1} \right\|_2 \\
&\leq \frac{\eta}{rn} \left\| \left(\frac{1}{m} \mathcal{A}^\dagger \mathcal{A}(\mathbf{Q}^t) - \mathbf{Q}^t\right)^\top \mathbf{W}^{t+1} \right\|_2 \left\| (\mathbf{R}^{t+1})^{-1} \right\|_2 \qquad (66) \\
&\leq \eta \delta_k' \operatorname{dist}(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*) \left\| (\mathbf{R}^{t+1})^{-1} \right\|_2 \qquad (67)
\end{aligned}$$

where (66) follows almost surely by Cauchy-Schwarz and the fact that $\hat{\mathbf{B}}_\perp^*$ is normalized, and (67) follows with probability at least $1 - e^{-110d}$ by Lemma 6. Next we control A_1:

$$\begin{aligned} A_1 &= \left\| \hat{\mathbf{B}}_\perp^{*\top} \hat{\mathbf{B}}^t (\mathbf{I}_k - \frac{\eta}{rn} (\mathbf{W}^{t+1})^\top \mathbf{W}^{t+1}) \right\|_2 \|(\mathbf{R}^{t+1})^{-1}\|_2 \\ &\leq \|\hat{\mathbf{B}}_\perp^{*\top} \hat{\mathbf{B}}^t\|_2 \left\| \mathbf{I} - \frac{\eta}{rn} (\mathbf{W}^{t+1})^\top \mathbf{W}^{t+1} \right\|_2 \|(\mathbf{R}^{t+1})^{-1}\|_2 \\ &= \text{dist}(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*) \left\| \mathbf{I}_k - \frac{\eta}{rn} (\mathbf{W}^{t+1})^\top \mathbf{W}^{t+1} \right\|_2 \|(\mathbf{R}^{t+1})^{-1}\|_2 \end{aligned} \tag{68}$$

The middle factor gives us contraction. To see this, recall that $\mathbf{W}^{t+1} = \mathbf{W}^* \hat{\mathbf{B}}^{*\top} \hat{\mathbf{B}}^t - \mathbf{F}$ where $\mathbf{F}$ is defined in Lemma 2. By Lemma 5, we have that

$$\|\mathbf{F}\|_2 \leq \frac{\delta_k}{1 - \delta_k} \|\mathbf{W}^*\|_2 \, \text{dist}(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*) \tag{69}$$

with probability at least $1 - e^{-110k^2 \log(rn)}$, which we will use throughout the proof. Conditioning on this event, we have

$$\begin{aligned} \lambda_{\max}\left((\mathbf{W}^{t+1})^\top \mathbf{W}^{t+1}\right) &= \|\mathbf{W}^* \hat{\mathbf{B}}^{*\top} \hat{\mathbf{B}}^t - \mathbf{F}\|_2^2 \\ &\leq 2\|\mathbf{W}^* \hat{\mathbf{B}}^{*\top} \hat{\mathbf{B}}^t\|_2^2 + 2\|\mathbf{F}\|_2^2 \\ &\leq 2\|\mathbf{W}^*\|_2^2 + 2\frac{\delta_k^2}{(1-\delta_k)^2} \|\mathbf{W}^*\|_2^2 \, \text{dist}^2(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*) \\ &\leq 4\|\mathbf{W}^*\|_2^2 \end{aligned} \tag{70}$$

where (70) follows under the assumption that $\delta_k \leq 1/2$. Thus, as long as $\eta \leq 1/(4\bar{\sigma}_{\max,*}^2)$, we have by Weyl's Inequality:

$$\begin{aligned} &\|\mathbf{I}_k - \frac{\eta}{rn} (\mathbf{W}^{t+1})^\top \mathbf{W}^{t+1}\|_2 \\ &\leq 1 - \frac{\eta}{rn} \lambda_{\min}((\mathbf{W}^{t+1})^\top \mathbf{W}^{t+1}) && (71) \\ &= 1 - \frac{\eta}{rn} \lambda_{\min}((\mathbf{W}^* \hat{\mathbf{B}}^{*\top} \hat{\mathbf{B}}^t - \mathbf{F})^\top (\mathbf{W}^* \hat{\mathbf{B}}^{*\top} \hat{\mathbf{B}}^t - \mathbf{F})) \\ &\leq 1 - \frac{\eta}{rn} \sigma_{\min}^2(\mathbf{W}^* (\hat{\mathbf{B}}^*)^\top \hat{\mathbf{B}}^t) + \frac{2\eta}{rn} \sigma_{\max}(\mathbf{F}^\top \mathbf{W}^* (\hat{\mathbf{B}}^*)^\top \hat{\mathbf{B}}^t) - \frac{\eta}{rn} \sigma_{\min}^2(\mathbf{F}) && (72) \\ &\leq 1 - \frac{\eta}{rn} \sigma_{\min}^2(\mathbf{W}^*) \sigma_{\min}^2((\hat{\mathbf{B}}^*)^\top \hat{\mathbf{B}}^t) + \frac{2\eta}{rn} \|\mathbf{F}\|_2 \, \|\mathbf{W}^* (\hat{\mathbf{B}}^*)^\top \hat{\mathbf{B}}^t\|_2 && (73) \\ &\leq 1 - \frac{\eta}{rn} \sigma_{\min}^2(\mathbf{W}^*) \sigma_{\min}^2((\hat{\mathbf{B}}^*)^\top \hat{\mathbf{B}}^t) + \frac{2\eta}{rn} \frac{\delta_k}{1-\delta_k} \|\mathbf{W}^*\|_2^2 && (74) \\ &= 1 - \eta \bar{\sigma}_{\min,*}^2 \sigma_{\min}^2((\hat{\mathbf{B}}^*)^\top \hat{\mathbf{B}}^t) + 2\eta \frac{\delta_k}{1-\delta_k} \bar{\sigma}_{\max,*}^2 && (75) \end{aligned}$$

where (72) follows by again applying Weyl's inequality, under the condition that $2\sigma_{\max}(\mathbf{F}^\top \mathbf{W}^* (\hat{\mathbf{B}}^*)^\top \hat{\mathbf{B}}^t) \leq \sigma_{\min}^2(\mathbf{W}^*) \sigma_{\min}^2((\hat{\mathbf{B}}^*)^\top \hat{\mathbf{B}}^t)$, which we will enforce to be true (otherwise we would not have contraction). Also, (73) follows by the Cauchy-Schwarz inequality, and we use Lemma 5 to obtain (74). Lastly, (75) follows by the definitions of $\bar{\sigma}_{\min,*}$ and $\bar{\sigma}_{\max,*}$. In order to lower bound $\sigma_{\min}^2((\hat{\mathbf{B}}^*)^\top \hat{\mathbf{B}}^t)$, note that

$$\sigma_{\min}^2((\hat{\mathbf{B}}^*)^\top \hat{\mathbf{B}}^t) \geq 1 - \|(\hat{\mathbf{B}}_\perp^*)^\top \hat{\mathbf{B}}^t\|_2^2 = 1 - \text{dist}^2(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*) \geq 1 - \text{dist}^2(\hat{\mathbf{B}}^0, \hat{\mathbf{B}}^*) =: E_0 \tag{76}$$

As a result, defining $\bar{\delta}_k := \delta_k + \delta_k'$ and combining (64), (67), (68), (75), and (76) yields

$$\begin{aligned} \text{dist}(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*) &\leq \|(\mathbf{R}^{t+1})^{-1}\|_2 \, (1 - \eta \bar{\sigma}_{\min,*}^2 E_0 + 2\eta \frac{\delta_k}{1-\delta_k} \bar{\sigma}_{\max,*}^2 + \eta \delta_k') \, \text{dist}(\hat{\mathbf{B}}^t, \mathbf{B}^*) \\ &\leq \|(\mathbf{R}^{t+1})^{-1}\|_2 \, (1 - \eta \bar{\sigma}_{\min,*}^2 E_0 + 2\eta \frac{\bar{\delta}_k}{1-\bar{\delta}_k} \bar{\sigma}_{\max,*}^2) \, \text{dist}(\hat{\mathbf{B}}^t, \mathbf{B}^*) \end{aligned} \tag{77}$$

where (77) follows from the fact that $krn = \|\mathbf{W}^*\|_F^2 \leq k\|\mathbf{W}^*\|_2^2 \implies 1 \leq \|\mathbf{W}^*\|_2^2/rn \leq \bar{\sigma}^2_{\max,*}$. All that remains to bound is $\|(\mathbf{R}^{t+1})^{-1}\|_2$. Define $\mathbf{S}^t := \frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^t)$ and observe that

$$\begin{aligned}(\mathbf{R}^{t+1})^\top\mathbf{R}^{t+1} &= (\mathbf{B}^{t+1})^\top\mathbf{B}^{t+1} \\ &= \hat{\mathbf{B}}^{t^\top}\hat{\mathbf{B}}^t - \frac{\eta}{rn}(\hat{\mathbf{B}}^{t^\top}\mathbf{S}^{t^\top}\mathbf{W}^{t+1} + (\mathbf{W}^{t+1})^\top\mathbf{S}^t\hat{\mathbf{B}}^t) + \frac{\eta^2}{(rn)^2}(\mathbf{W}^{t+1})^\top\mathbf{S}^t\mathbf{S}^{t^\top}\mathbf{W}^{t+1} \\ &= \mathbf{I}_k - \frac{\eta}{rn}(\hat{\mathbf{B}}^{t^\top}\mathbf{S}^{t^\top}\mathbf{W}^{t+1} + (\mathbf{W}^{t+1})^\top\mathbf{S}^t\hat{\mathbf{B}}^t) + \frac{\eta^2}{(rn)^2}(\mathbf{W}^{t+1})^\top\mathbf{S}^t\mathbf{S}^{t^\top}\mathbf{W}^{t+1} \end{aligned} \tag{78}$$

thus, by Weyl's Inequality, we have

$$\begin{aligned}\sigma^2_{\min}(\mathbf{R}_{t+1}) &\geq 1 - \frac{\eta}{rn}\lambda_{\max}(\hat{\mathbf{B}}^{t^\top}\mathbf{S}^{t^\top}\mathbf{W}^{t+1} + (\mathbf{W}^{t+1})^\top\mathbf{S}^t\hat{\mathbf{B}}^t) + \frac{\eta^2}{(rn)^2}\lambda_{\min}((\mathbf{W}^{t+1})^\top\mathbf{S}^t\mathbf{S}^{t^\top}\mathbf{W}^{t+1}) \\ &\geq 1 - \frac{\eta}{rn}\lambda_{\max}(\hat{\mathbf{B}}^{t^\top}\mathbf{S}^{t^\top}\mathbf{W}^{t+1} + (\mathbf{W}^{t+1})^\top\mathbf{S}^t\hat{\mathbf{B}}^t) \end{aligned} \tag{79}$$

where (79) follows because $(\mathbf{W}^{t+1})^\top\mathbf{S}^t\mathbf{S}^{t^\top}\mathbf{W}^{t+1}$ is positive semi-definite. Next, note that

$$\begin{aligned}&\frac{\eta}{rn}\lambda_{\max}(\hat{\mathbf{B}}^{t^\top}\mathbf{S}^{t^\top}\mathbf{W}^{t+1} + (\mathbf{W}^{t+1})^\top\mathbf{S}^t\hat{\mathbf{B}}^t) \\ &\quad= \max_{\mathbf{x}:\|\mathbf{x}\|_2=1}\frac{\eta}{rn}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}(\mathbf{S}^t)^\top\mathbf{W}^{t+1}\mathbf{x} + \mathbf{x}^\top(\mathbf{W}^{t+1})^\top\mathbf{S}^t\hat{\mathbf{B}}^t\mathbf{x} \\ &\quad= \max_{\mathbf{x}:\|\mathbf{x}\|_2=1}\frac{2\eta}{rn}\mathbf{x}^\top(\mathbf{W}^{t+1})^\top\mathbf{S}^t\hat{\mathbf{B}}^t\mathbf{x} \\ &\quad= \max_{\mathbf{x}:\|\mathbf{x}\|_2=1}\frac{2\eta}{rn}\mathbf{x}^\top(\mathbf{W}^{t+1})^\top\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^t) - \mathbf{Q}^t\right)\hat{\mathbf{B}}^t\mathbf{x} + \frac{2\eta}{rn}\mathbf{x}^\top(\mathbf{W}^{t+1})^\top\mathbf{Q}^t\hat{\mathbf{B}}^t\mathbf{x} \end{aligned} \tag{80}$$

We first consider the first term. We have

$$\max_{\mathbf{x}:\|\mathbf{x}\|_2=1}\frac{2\eta}{rn}\mathbf{x}^\top(\mathbf{W}^{t+1})^\top\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^t) - \mathbf{Q}^t\right)\hat{\mathbf{B}}^t\mathbf{x} \leq \frac{2\eta}{rn}\left\|(\mathbf{W}^{t+1})^\top\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^t) - \mathbf{Q}^t\right)\right\|_2\left\|\hat{\mathbf{B}}^t\right\|_2 \leq 2\eta\delta'_k \tag{81}$$

where the last inequality follows with probability at least $1 - e^{-110d} - e^{-110k^2\log(rn)}$ from Lemma 6. Next we turn to the second term in (80). We have

$$\begin{aligned}\max_{\mathbf{x}:\|\mathbf{x}\|_2=1}\frac{2\eta}{rn}\mathbf{x}^\top(\mathbf{W}^{t+1})^\top\mathbf{Q}^t\hat{\mathbf{B}}^t\mathbf{x} &= \max_{\mathbf{x}:\|\mathbf{x}\|_2=1}\frac{2\eta}{rn}\left\langle\mathbf{Q}^t, \mathbf{W}^{t+1}\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}\right\rangle \\ &= \max_{\mathbf{x}:\|\mathbf{x}\|_2=1}\frac{2\eta}{rn}\langle\mathbf{Q}^t, \mathbf{W}^*\hat{\mathbf{B}}^{*^\top}\hat{\mathbf{B}}^t\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}\rangle - \frac{2\eta}{rn}\langle\mathbf{Q}^t, \mathbf{F}\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}\rangle \end{aligned} \tag{82}$$

For any $\mathbf{x} \in \mathbb{R}^k : \|\mathbf{x}\|_2 = 1$, we have

$$\begin{aligned}
&\frac{2\eta}{rn}\langle \mathbf{Q}^t, \mathbf{W}^*(\hat{\mathbf{B}}^*)^\top \hat{\mathbf{B}}^t \mathbf{x}\mathbf{x}^\top \hat{\mathbf{B}}^{t^\top}\rangle \\
&= \frac{2\eta}{rn}\text{tr}((\hat{\mathbf{B}}^t(\mathbf{W}^{t+1})^\top - \hat{\mathbf{B}}^*(\mathbf{W}^*)^\top)\mathbf{W}^*\hat{\mathbf{B}}^{*^\top}\hat{\mathbf{B}}^t\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}) \\
&= \frac{2\eta}{rn}\text{tr}((\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t^\top}\hat{\mathbf{B}}^*\mathbf{W}^{*^\top} - \hat{\mathbf{B}}^t\mathbf{F}^\top - \hat{\mathbf{B}}^*\mathbf{W}^{*^\top})\mathbf{W}^*\hat{\mathbf{B}}^{*^\top}\hat{\mathbf{B}}^t\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}) \\
&= \frac{2\eta}{rn}\text{tr}((\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t^\top} - \mathbf{I})\hat{\mathbf{B}}^{*^\top}\mathbf{W}^{*^\top}\mathbf{W}^*\hat{\mathbf{B}}^{*^\top}\hat{\mathbf{B}}^t\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}) - \frac{2\eta}{rn}\text{tr}(\hat{\mathbf{B}}^t\mathbf{F}^\top\mathbf{W}^*\hat{\mathbf{B}}^{*^\top}\hat{\mathbf{B}}^t\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}) \\
&= \frac{2\eta}{rn}\text{tr}(\hat{\mathbf{B}}^t_\perp\hat{\mathbf{B}}^{*^\top}\mathbf{W}^{*^\top}\mathbf{W}^*\hat{\mathbf{B}}^{*^\top}\hat{\mathbf{B}}^t\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}) - \frac{2\eta}{rn}\text{tr}(\hat{\mathbf{B}}^t\mathbf{F}^\top\mathbf{W}^*\hat{\mathbf{B}}^{*^\top}\hat{\mathbf{B}}^t\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}) \\
&= \frac{2\eta}{rn}\text{tr}(\hat{\mathbf{B}}^{*^\top}\mathbf{W}^{*^\top}\mathbf{W}^*\hat{\mathbf{B}}^{*^\top}\hat{\mathbf{B}}^t\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}\hat{\mathbf{B}}^t_\perp) - \frac{2\eta}{rn}\text{tr}(\hat{\mathbf{B}}^t\mathbf{F}^\top\mathbf{W}^*\hat{\mathbf{B}}^{*^\top}\hat{\mathbf{B}}^t\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}) \\
&= -\frac{2\eta}{rn}\text{tr}(\mathbf{F}^\top\mathbf{W}^*\hat{\mathbf{B}}^{*^\top}\hat{\mathbf{B}}^t\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}\hat{\mathbf{B}}^t) && (83)\\
&= -\frac{2\eta}{rn}\text{tr}(\mathbf{F}^\top\mathbf{W}^*\hat{\mathbf{B}}^{*^\top}\hat{\mathbf{B}}^t\mathbf{x}\mathbf{x}^\top) && (84)\\
&\le \frac{2\eta}{rn}\|\mathbf{F}\|_F\ \|\mathbf{W}^*\hat{\mathbf{B}}^{*^\top}\hat{\mathbf{B}}^t\mathbf{x}\mathbf{x}^\top\|_F && (85)\\
&\le \frac{2\eta}{rn}\|\mathbf{F}\|_F\|\mathbf{W}^*\|_2\|\hat{\mathbf{B}}^{*^\top}\|_2\|\hat{\mathbf{B}}^t\|_2\|\mathbf{x}\mathbf{x}^\top\|_F && (86)\\
&\le \frac{2\eta}{rn}\|\mathbf{F}\|_F\ \|\mathbf{W}^*\|_2 && (87)\\
&\le 2\eta\frac{\delta_k}{1-\delta_k}\bar{\sigma}^2_{\max,*} && (88)
\end{aligned}$$

where (83) follows since $\hat{\mathbf{B}}^{t^\top}\hat{\mathbf{B}}^t_\perp = \mathbf{0}$, (84) follows since $\hat{\mathbf{B}}^{t^\top}\hat{\mathbf{B}}^t = \mathbf{I}_k$, (85) and (86) follow by the Cauchy-Schwarz inequality, (87) follows by the orthonormality of $\hat{\mathbf{B}}^t$ and $\hat{\mathbf{B}}^*$ and (88) follows by Lemma 5 and the definition of $\bar{\sigma}_{\max,*}$. Next, again for any $\mathbf{x} \in \mathbb{R}^k : \|\mathbf{x}\|_2 = 1$,

$$\begin{aligned}
-\frac{2\eta}{rn}\langle \mathbf{Q}^t, \mathbf{F}\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}\rangle &= -\frac{2\eta}{rn}\text{tr}((\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t^\top}\hat{\mathbf{B}}^*\mathbf{W}^{*^\top} - \hat{\mathbf{B}}^t\mathbf{F}^\top - \hat{\mathbf{B}}^*\mathbf{W}^{*^\top})\mathbf{F}\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}) \\
&= -\frac{2\eta}{rn}\text{tr}((\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t^\top} - \mathbf{I}_d)\hat{\mathbf{B}}^*\mathbf{W}^{*^\top}\mathbf{F}\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}) + \frac{2\eta}{rn}\text{tr}(\mathbf{F}\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}\hat{\mathbf{B}}^t\mathbf{F}^\top) \\
&= -\frac{2\eta}{rn}\text{tr}(\hat{\mathbf{B}}^*\mathbf{W}^{*^\top}\mathbf{F}\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}\mathbf{B}^t_\perp) + \frac{2\eta}{rn}\mathbf{x}^\top\mathbf{F}^\top\mathbf{F}\mathbf{x} \\
&= \frac{2\eta}{rn}\mathbf{x}^\top\mathbf{F}^\top\mathbf{F}\mathbf{x} \\
&\le \frac{2\eta}{rn}\|\mathbf{F}\|_2^2 \\
&\le 2\eta\frac{\delta_k^2}{(1-\delta_k)^2}\bar{\sigma}^2_{\max,*} && (89)
\end{aligned}$$

Thus, we have the following bound on the second term of (80):

$$\max_{\mathbf{x}:\|\mathbf{x}\|_2=1}\frac{2\eta}{rn}\langle \mathbf{Q}^t, \mathbf{W}^{t+1}\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}\rangle \le 2\eta\bar{\sigma}^2_{\max,*}\left(\frac{\delta_k}{1-\delta_k} + \frac{\delta_k^2}{(1-\delta_k)^2}\right) \le 4\eta\frac{\delta_k}{(1-\delta_k)^2}\bar{\sigma}^2_{\max,*} \tag{90}$$

since $0 \ge \delta_k \le 1 \implies \delta_k^2 \le \delta_k$. Therefore, using (79), (80), (81) and (90), we have

$$\sigma^2_{\min}(\mathbf{R}_{t+1}) \ge 1 - 2\eta\delta'_k - 4\eta\frac{\delta_k}{(1-\delta_k)^2}\bar{\sigma}^2_{\max,*} \ge 1 - 4\eta\frac{\bar{\delta}_k}{(1-\bar{\delta}_k)^2}\bar{\sigma}^2_{\max,*} \tag{91}$$

where $\bar{\delta}_k = \delta'_k + \delta_k$. This means that

$$\|(\mathbf{R}^{t+1})^{-1}\|_2 \le \left(1 - 4\eta\frac{\bar{\delta}_k}{(1-\bar{\delta}_k)^2}\bar{\sigma}^2_{\max,*}\right)^{-1/2} \tag{92}$$

Note that $1 - 4\eta\frac{\bar{\delta}_k}{(1-\bar{\delta}_k)^2}\bar{\sigma}^2_{\max,*}$ is strictly positive as long as $\frac{\bar{\delta}_k}{(1-\bar{\delta}_k)^2} < 1$, which we will verify shortly, due to our earlier assumption that $\eta \leq 1/(4\bar{\sigma}^2_{\max,*})$. Therefore, from (77), we have

$$\text{dist}(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*) \leq \frac{1}{\sqrt{1 - 4\eta\frac{\bar{\delta}_k}{(1-\bar{\delta}_k)^2}\bar{\sigma}^2_{\max,*}}} \left(1 - \eta\bar{\sigma}^2_{\min,*}E_0 + 2\eta\frac{\bar{\delta}_k}{(1-\bar{\delta}_k)^2}\bar{\sigma}^2_{\max,*}\right) \text{dist}(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*)$$

Next, let $\bar{\delta}_k < 16E_0/(25 \cdot 5\kappa^2)$. This implies that $\bar{\delta}_k < 1/5$. Then $\bar{\delta}_k/(1-\bar{\delta}_k)^2 < 25\bar{\delta}_k/16 \leq E_0/(5\kappa^2) \leq 1$, validating (92). Further, it is easily seen that

$$\begin{aligned} 1 - \eta E_0\bar{\sigma}^2_{\min,*} + \eta\frac{\bar{\delta}_k}{(1-\bar{\delta}_k)^2}\bar{\sigma}^2_{\max,*} &\leq 1 - 4\eta\frac{\bar{\delta}_k}{(1-\bar{\delta}_k)^2}\bar{\sigma}^2_{\max} \\ &\leq 1 - \eta E_0\bar{\sigma}^2_{\min,*}/2 \end{aligned} \tag{93}$$

Thus

$$\text{dist}(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*) \leq \left(1 - \eta E_0\bar{\sigma}^2_{\min,*}/2\right)^{1/2} \text{dist}(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*).$$

Finally, recall that $\bar{\delta}_k = \delta_k + \delta'_k = c\left(\frac{k^{3/2}\sqrt{\log(rn)}}{\sqrt{m}} + \frac{k\sqrt{d}}{\sqrt{rnm}}\right)$ for some absolute constant c. Choosing $m \geq c'(\kappa^4 k^3 \log(rn)/E_0^2 + \kappa^4 k^2 d/(E_0^2 rn))$ for another absolute constant c' satisfies $\bar{\delta}_k \leq 16E_0/(25 \cdot 5\kappa^2)$. Also, we have conditioned on two events, described in Lemmas 5 and 6, which occur with probability at least $1 - e^{-110d} - e^{-110k^2\log(rn)} \geq 1 - e^{-100\min(k^2\log(rn), d)}$, completing the proof.

□

Finally, Theorem 1 follows by recursively applying Lemma 7 and taking a union bound over all $t \in [T]$.

B.4. Initialization

As mentioned in the main body, our interpretation of Theorem 1 assumes that the initial distance is bounded above by a constant less than one, i.e., E_0 is bounded below by a constant greater than zero. We can achieve such an initialization without increasing the overall sample complexity via the Method-of-Moments algorithm, ignoring log factors. To show this, we adapt a result from (Tripuraneni et al., 2020).

Theorem 2 (Theorem 3, (Tripuraneni et al., 2020))**.** *Suppose that each client $i \in [n]$ sends the server $\mathbf{Z}_i := \frac{1}{m}\sum_{j=1}^m (y_i^{0,j})^2 \mathbf{x}_i^{0,j}(\mathbf{x}_i^{0,j})^\top$ and the server computes $\hat{\mathbf{U}}\mathbf{D}\hat{\mathbf{U}}^\top \leftarrow$ rank-k SVD$(\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i)$ and sets $\mathbf{B}^0 = \hat{\mathbf{U}}$. Then, if $m \geq c\,\text{polylog}(d, mn)kd/(\sigma^4_{\min,*}n)$,*

$$\text{dist}\left(\mathbf{B}, \hat{\mathbf{B}}^*\right) \leq \tilde{O}\left(\frac{kd}{\sigma^4_{\min,*}mn}\right) \tag{94}$$

with probability at least $1 - O((mn)^{-100})$ for some absolute constant c, where $\sigma_{\min,*} := \frac{1}{\sqrt{n}}\mathbf{W}^*$ and $\tilde{O}(\cdot)$ hides log factors.

The above result is a direct adaptation of Theorem 3 in (Tripuraneni et al., 2020) so we omit the proof. This result shows that $m = \tilde{\Omega}(\frac{kd}{\sigma^4_{\min,*}n})$ are required for proper initialization. Since $\kappa \gtrsim 1/\sigma_{\min,*}$, the overall sample complexity is not increased up to log factors.

B.5. Proof Challenges

We next discuss two analytical challenges involved in proving Theorem 1.

(i) Row-wise sparse measurements. Recall that the measurement matrices $\mathbf{A}^t_{i,j}$ have non-zero elements only in the i-th row. This property is beneficial in the sense that it allows for distributing the sensing computation across the n clients. However, it also means that the operators $\{\frac{1}{\sqrt{m}}\mathcal{A}^t\}_t$ *do not* satisfy Restricted Isometry Property (RIP), which therefore prevents us from using standard RIP-based analysis. The RIP is defined as follows:

Definition 4 (Restricted Isometry Property). *An operator $\mathcal{B} : \mathbb{R}^{n\times d} \to \mathbb{R}^{nm}$ satisfies the k-RIP with parameter $\delta_k \in [0, 1)$ if and only if*

$$(1 - \delta_k)\|\mathbf{M}\|_F^2 \leq \|\mathcal{B}(\mathbf{M})\|_2^2 \leq (1 + \delta_k)\|\mathbf{M}\|_F^2 \tag{95}$$

holds simultaneously for all $\mathbf{M} \in \mathbb{R}^{n\times d}$ of rank at most k.

Claim 1. *Let $\mathcal{A} : \mathbb{R}^{rn\times d} \to \mathbb{R}^{rnm}$ such that $\mathcal{A}(\mathbf{M}) = [\langle \mathbf{e}_i(\mathbf{x}_i^j)^\top, \mathbf{M}\rangle]_{1\leq i\leq rn, 1\leq j\leq m}$, and let the samples $\mathbf{x}_i^j$ be i.i.d. sub-gaussian random vectors with mean $\mathbf{0}_d$ and covariance $\mathbf{I}_d$. Then if $m \leq d/2$, with probability at least $1 - e^{-cd}$ for some absolute constant c, $\frac{1}{\sqrt{m}}\mathcal{A}$ does not satisfy 1-RIP for any constant $\delta_1 \in [0, 1)$.*

Proof. Let $\mathbf{M} = \mathbf{e}_1(\mathbf{x}_1^1)^\top$. Then

$$\begin{aligned}
\|\frac{1}{\sqrt{m}}\mathcal{A}(\mathbf{M})\|_2^2 &= \frac{1}{m}\sum_{i=1}^{rn}\sum_{j=1}^{m}\langle \mathbf{e}_i(\mathbf{x}_i^j)^\top, \mathbf{e}_1(\mathbf{x}_1^1)^\top\rangle^2 \\
&= \frac{1}{m}\|\mathbf{x}_1^1\|_2^4 + \frac{1}{m}\sum_{j=2}^{m}\langle \mathbf{x}_1^j, \mathbf{x}_1^1\rangle^2 \\
&\geq \frac{1}{m}\|\mathbf{x}_1^1\|_2^4
\end{aligned} \tag{96}$$

Also observe that $\|\mathbf{M}\|_F^2 = \|\mathbf{x}_1^1\|_2^2$. Therefore, we have

$$\begin{aligned}
\mathbb{P}\left(\frac{\left\|\frac{1}{\sqrt{m}}\mathcal{A}(\mathbf{M})\right\|_2^2}{\|\mathbf{M}\|_F^2} \geq \frac{d}{2m}\right) &\geq \mathbb{P}\left(\frac{\frac{1}{m}\left\|\mathbf{x}_1^1\right\|_2^4}{\|\mathbf{x}_1^1\|_2^2} \geq \frac{d}{2m}\right) \\
&= \mathbb{P}\left(\left\|\mathbf{x}_1^1\right\|_2^2 \geq \frac{d}{2}\right) \\
&= 1 - \mathbb{P}\left(\|\mathbf{x}_1^1\|_2^2 - d \leq \frac{-d}{2}\right) \\
&\geq 1 - e^{-cd}
\end{aligned} \tag{97}$$

where the last inequality follows for some absolute constant c by the sub-exponential property of $\|\mathbf{x}_1^1\|_2^2$ and the fact that $\mathbb{E}[\|\mathbf{x}_1^1\|_2^2] = d$. Thus, with probability at least $1 - e^{-cd}$, $\left\|\frac{1}{\sqrt{m}}\mathcal{A}(\mathbf{M})\right\|_2^2 \geq \frac{d}{2m}\|\mathbf{M}\|_2^2$, meaning that $\frac{1}{\sqrt{m}}\mathcal{A}$ does not satisfy 1-RIP with high probability if $m \leq \frac{d}{2}$. □

Claim 1 shows that we cannot use the RIP to show $\mathcal{O}(d/(rn))$ sample complexity for m - instead, this approach would require $m = \Omega(d)$. Fortunately, we do not need concentration of the measurements for all rank-k matrices $\mathbf{M}$, but only a particular class of rank-k matrices that are *row-wise incoherent*. Leveraging the row-wise incoherence of the matrices being measured allows us to show that we only require $m = \Omega(k^3\log(rn) + k^2 d/(rn))$ samples per user (ignoring dimension-independent constants).

(ii) Non-symmetric updates. Existing analyses for nonconvex matrix sensing study algorithms with symmetric update schemes for the factors $\mathbf{W}$ and $\mathbf{B}$, either alternating minimization, e.g. (Jain et al., 2013), or alternating gradient descent, e.g. (Tu et al., 2016). Here we show contraction due to the gradient descent step in principal angle distance, differing from the standard result for gradient descent using Procrustes distance (Tu et al., 2016; Zheng & Lafferty, 2016; Park et al., 2018). We combine aspects of both types of analysis in our proof.