Exploiting Shared Representations for Personalized Federated Learning

Liam Collins[*], Hamed Hassani[†], Aryan Mokhtari[*], Sanjay Shakkottai[*]

Abstract

Deep neural networks have shown the ability to extract universal feature representations from data such as images and text that have been useful for a variety of learning tasks. However, the fruits of representation learning have yet to be fully-realized in federated settings. Although data in federated settings is often non-i.i.d. across clients, the success of centralized deep learning suggests that data often shares a global *feature representation*, while the statistical heterogeneity across clients or tasks is concentrated in the *labels*. Based on this intuition, we propose a novel federated learning framework and algorithm for learning a shared data representation across clients and unique local heads for each client. Our algorithm harnesses the distributed computational power across clients to perform many local-updates with respect to the low-dimensional local parameters for every update of the representation. We prove that this method obtains linear convergence to the ground-truth representation with near-optimal sample complexity in a linear setting, demonstrating that it can efficiently reduce the problem dimension for each client. Further, we provide extensive experimental results demonstrating the improvement of our method over alternative personalized federated learning approaches in heterogeneous settings.

[*]Department of Electrical and Computer Engineering, The University of Texas at Austin, Austin, TX, USA. {liamc@utexas.edu, mokhtari@austin.utexas.edu, sanjay.shakkottai@utexas.edu}.

[†]Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, USA. {hassani@seas.upenn.edu}.

1 Introduction

Many of the most heralded successes of modern machine learning have come in *centralized* settings, wherein a single model is trained on a large amount of centrally-stored data. The growing number of data-gathering devices, however, calls for a distributed architecture to train models. Federated learning aims at addressing this issue by providing a platform in which a group of clients collaborate to learn effective models for each client by leveraging the local computational power, memory, and data of all clients [McMahan et al., 2017]. The task of coordinating between the clients is fulfilled by a central server that combines the models received from the clients at each round and broadcasts the updated information to them. Importantly, the server and clients are restricted to methods that satisfy communication and privacy constraints, preventing them from directly applying centralized techniques.

However, one of the most important challenges in federated learning is the issue of *data heterogeneity*, where the underlying data distribution of client tasks could be substantially different from each other. In such settings, if the server and clients learn a single shared model (e.g., by minimizing average loss), the resulting model could perform poorly for many of the clients in the network (and also not generalize well across diverse data [Jiang et al., 2019]). In fact, for some clients, it might be better to simply use their own local data (even if it is small) to train a local model; see Figure 1. Finally, the (federated) trained model may not generalize well to unseen clients that have not participated in the training process. These issues raise this question:

> "*How can we exploit the data and computational power of all clients in data heterogeneous settings to learn a personalized model for each client?*"

We address this question by taking advantage of the common representation among clients. Specifically, we view the data heterogeneous federated learning problem as n parallel learning tasks that they possibly have some common structure, and *our goal is to learn and exploit this common representation to improve the quality of each client's model*. Indeed, this would be in line with our understanding from centralized learning, where we have witnessed success in training multiple tasks simultaneously by leveraging a common (low-dimensional) representation in popular machine learning tasks (e.g., image classification, next-word prediction) [Bengio et al., 2013, LeCun et al., 2015].

Main Contributions. We introduce a novel federated learning framework and an associated algorithm for data heterogeneous settings. Next, we present our main contributions.

(i) **FedRep Algorithm.** Federated Representation Learning (FedRep) leverages the full quantity of data stored across clients to learn a global low-dimensional representation using gradient-based updates. Further, it enables each client to compute a personalized, low-dimensional classifier, which we term as the client's local head, that accounts for the unique labeling of each client's local data.

(ii) **Convergence Rate.** We show that FedRep converges to the optimal representation at a *exponentially fast rate* with near-optimal sample complexity in the case that each client aims to solve a linear regression problem with a two-layer linear neural network. Our

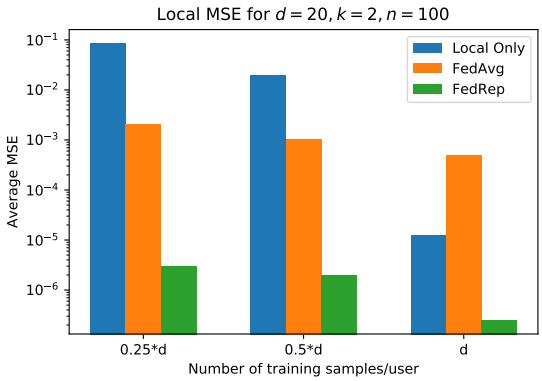

Figure 1: Local only training suffers in small-training data regimes, whereas training a single global model (FedAvg) cannot overcome client heterogeneity even when the number of training samples is large. FedRep exploits a common representation of the clients to achieve small error in all cases.

analysis further implies that we only need $\mathcal{O}(\kappa^4 k^2 (k \log(rn) + \frac{d}{rn}))$ samples per client, where n is the number of clients, d is the dimension of the data, k is the representation dimension, r is the participation rate and κ is the condition number of the ground-truth client-representation matrix.

(iii) **Empirical Results.** Through a combination of synthetic and real datasets (CIFAR10, CIFAR100, FEMNIST, Sent140) we show the benefits of FedRep in: (a) leveraging many local updates, (b) robustness to different levels of heterogeneity, and (c) generalization to new clients. We consider several important baselines including FedAvg [McMahan et al., 2017], Fed-MTL [Smith et al., 2017], LG-FedAvg [Liang et al., 2020], and Per-FedAvg [Fallah et al., 2020]. Our experiments indicate that FedRep outpeforms these baselines in heterogeneous settings that share a global representation.

Benefits of FedRep. Next, we list benefits of FedRep over standard federated learning (that learns a single model).

(I) More local updates. By reducing the problem dimension, each client can make many local updates at each communication round, which is beneficial in learning its own individual head. This is unlike standard federated learning where multiple local updates in a heterogeneous setting moves each client *away* from the best averaged representation, and thus *hurts* performance.

(II) Gains of cooperation. Denote d to be the data dimension and n the number of clients. From our sample complexity bounds, it follows that with FedRep, the sample complexity per client scales as $\Theta(\log(n) + d/n)$. On the other hand, local learning (without any collaboration) has a sample complexity that scales as $\Theta(d)$. Thus, if $1 \ll n \ll e^{\Theta(d)}$ (see Section 4.2 for details), we expect benefits of collaboration through federation. When d is large (as is typical in practice), $e^{\Theta(d)}$ is exponentially larger, and federation helps each client. *To the best of our knowledge, this is the first sample-complexity-based result for heterogeneous federated learning that demonstrates the benefit of cooperation.*

(III) Generalization to new clients. For a new client, since a ready-made representation is available, the client only needs to learn a head with a low-dimensional representation of dimension k. Thus, its sample complexity scales only as $\Theta(k \log(1/\epsilon))$ to have no more than ϵ error in accuracy.

Related Work. A variety of recent works have studied personalization in federated learning using, for example, local fine-tuning [Wang et al., 2019, Yu et al., 2020], meta-learning [Chen et al., 2018, Fallah et al., 2020, Jiang et al., 2019, Khodak et al., 2019], additive mixtures of local and global models [Deng et al., 2020, Hanzely and Richtárik, 2020, Mansour et al., 2020], and multi-task learning [Smith et al., 2017]. In all of these methods, each client's subproblem is still full-dimensional - there is no notion of learning a dimensionality-reduced set of local parameters. More recently, Liang et al. [2020] also proposed a representation learning method for federated learning, but their method attempts to learn many local representations and a single global head as opposed to a single global representation and many local heads. Earlier, Arivazhagan et al. [2019] presented an algorithm to learn local heads and a global network body, but their local procedure jointly updates the head and body (using the same number of updates), and they did not provide any theoretical justification for their proposed method. Meanwhile, another line of work has studied federated learning in heterogeneous settings [Haddadpour et al., 2020, Karimireddy et al., 2020, Pathak and Wainwright, 2020, Reddi et al., 2020, Wang et al., 2020], and the optimization-based insights from these works may be used to supplement our formulation and algorithm.

2 Problem Formulation

The generic form of federated learning with n clients is

$$\min_{(q_1,\ldots,q_n)\in\mathcal{Q}_n} \frac{1}{n}\sum_{i=1}^{n} f_i(q_i), \tag{1}$$

where f_i and q_i are the error function and learning model for the i-th client, respectively, and $\mathcal{Q}_n$ is the space of feasible sets of n models. We consider a supervised setting in which the data for the i-th client is generated by a distribution $(\mathbf{x}_i, y_i) \sim \mathcal{D}_i$. The learning model $q_i : \mathbb{R}^d \to \mathcal{Y}$ maps inputs $\mathbf{x}_i \in \mathbb{R}^d$ to predicted labels $q_i(\mathbf{x}_i) \in \mathcal{Y}$, which we would like to resemble the true labels y_i. The error f_i is in the form of an expected risk over $\mathcal{D}_i$, namely $f_i(q_i) \coloneqq \mathbb{E}_{(\mathbf{x}_i,y_i)\sim\mathcal{D}_i}[\ell(q_i(\mathbf{x}_i), y_i)]$, where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ is a loss function that penalizes the distance of $q_i(\mathbf{x}_i)$ from y_i.

In order to minimize f_i, the i-th client accesses a dataset of M_i labelled samples $\{(\mathbf{x}_i^j, y_i^j)\}_{j=1}^{M_i}$ from $\mathcal{D}_i$ for training. Federated learning addresses settings in which the M_i's are typically small relative to the problem dimension while the number of clients n is large. Thus, clients may not be able to obtain solutions q_i with small expected risk by training completely locally on *only* their M_i local samples. Instead, federated learning enables the clients to cooperate, by exchanging messages with a central server, in order to learn models using the cumulative data of all the clients.

Standard approaches to federated learning aim at learning a *single* shared model $q = q_1 = \cdots = q_n$ that performs well on average across the clients [Li et al., 2018, McMahan et al., 2017]. In this way, the clients aim to solve a special version of Problem (1), which is to minimize $(1/n)\sum_i f_i(q)$

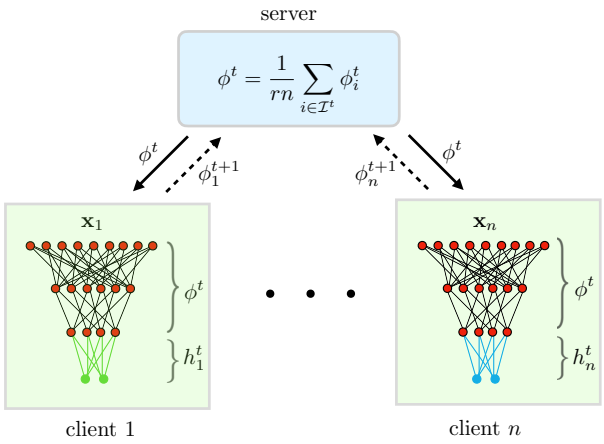

Figure 2: Federated representation learning structure where clients and the server aim at learning a global representation ϕ together, while each client i learns its unique head h_i locally.

over the choice of the shared model q. However, this approach may yield a solution that performs poorly in heterogeneous settings where the data distributions $\mathcal{D}_i$ vary across the clients. Indeed, in the presence of data heterogeneity, the error functions f_i will have different forms and their minimizers are not the same. Hence, learning a shared model q may not provide good solution to Problem (1). This necessities the search for more personalized solutions $\{q_i\}$ that can be learned in a federated manner using the clients' data.

Learning a Common Representation. We are motivated by insights from centralized machine learning that suggest that heterogeneous data distributed across tasks may share a common representation despite having different labels [Bengio et al., 2013, LeCun et al., 2015]; e.g., shared features across many types of images, or across word-prediction tasks. Using this common (low-dimensional) representation, the labels for each client can be simply learned using a linear classifier or a shallow neural network.

Formally, we consider a setting consisting of a global representation $\phi : \mathbb{R}^d \to \mathbb{R}^k$, which maps data points to a lower space of size k, and client-specific heads $h_i : \mathbb{R}^k \to \mathcal{Y}$. The model for the i-th client is the composition of the client's local parameters and the representation: $q_i(\mathbf{x}) = (h_i \circ \phi)(\mathbf{x})$. Critically, $k \ll d$, meaning that the number of parameters that must be learned locally by each client is small. Thus, we can assume that any client's optimal classifier for any *fixed representation* is easy to compute, which motivates the following re-written global objective:

$$\min_{\phi\in\Phi} \frac{1}{n}\sum_{i=1}^{n} \min_{h_i\in\mathcal{H}} f_i(h_i \circ \phi), \tag{2}$$

where Φ is the class of feasible representations and $\mathcal{H}$ is the class of feasible heads. In our proposed scheme, clients cooperate to learn the global model using all clients' data, while they use their local information to learn their personalized head. We discuss this in detail in Section 3.

2.1 Comparison with Standard Federated Learning

To formally demonstrate the advantage of our formulation over the standard (single-model) federated learning formulation in heterogeneous settings with a shared representation, we study a linear representation setting with quadratic loss. As we will see below, standard federated learning *cannot recover the underlying representation in the face of heterogeneity*, while our formulation does indeed recover it.

Consider a setting in which the functions f_i are quadratic losses, the representation ϕ is a projection onto a k-dimensional subspace of $\mathbb{R}^d$ given by matrix $\mathbf{B} \in \mathbb{R}^{d\times k}$, and the i-th client's local head h_i is a vector $\mathbf{w}_i \in \mathbb{R}^k$. In this setting, we model the local data of clients $\{\mathcal{D}_i\}_i$ such that $y_i = \mathbf{w}_i^{*\top}\mathbf{B}^{*\top}\mathbf{x}_i$ for some ground-truth representation $\mathbf{B}^* \in \mathbb{R}^{d\times k}$ and local heads $\mathbf{w}_i^* \in \mathbb{R}^k$. This setting will be described in detail in Section 4. In particular, one can show that the expected error over the data distribution $\mathcal{D}_i$ has the following form: $f_i(\mathbf{w}_i \circ \mathbf{B}) := \frac{1}{2}\|\mathbf{B}\mathbf{w}_i - \mathbf{B}^*\mathbf{w}_i^*\|_2^2$. Consequently, Problem (2) becomes

$$\min_{\mathbf{B}\in\mathbb{R}^{d\times k},\mathbf{w}_i,\ldots,\mathbf{w}_n\in\mathbb{R}^k} \frac{1}{2n}\sum_{i=1}^{n}\|\mathbf{B}\mathbf{w}_i - \mathbf{B}^*\mathbf{w}_i^*\|_2^2. \tag{3}$$

In contrast, standard federated learning methods, which aim to learn a shared model $(\mathbf{B}, \mathbf{w})$ for all the clients, solve

$$\min_{\mathbf{B}\in\mathbb{R}^{d\times k},\mathbf{w}\in\mathbb{R}^k} \frac{1}{2n}\sum_{i=1}^{n}\|\mathbf{B}\mathbf{w} - \mathbf{B}^*\mathbf{w}_i^*\|_2^2. \tag{4}$$

Let $(\hat{\mathbf{B}}, \{\hat{\mathbf{w}}_i\}_i)$ denote a global minimizer of (3). We thus have $\hat{\mathbf{B}}\hat{\mathbf{w}}_i = \mathbf{B}^*\mathbf{w}_i^*$ for all $i \in [n]$. Also, it is not hard to see that $(\mathbf{B}^\diamond, \mathbf{w}^\diamond)$ is a global minimizer of (4) if and only if $\mathbf{B}^\diamond\mathbf{w}^\diamond = \mathbf{B}^*(\frac{1}{n}\sum_{i=1}^n \mathbf{w}_i^*)$. Thus, our formulation finds an exact solution with zero global error, whereas standard federated learning has global error of $\frac{1}{2n}\sum_{i=1}^n \|\frac{1}{n}\mathbf{B}^*\sum_{i'=1}^n(\mathbf{w}_{i'}^* - \mathbf{w}_i^*)\|_2^2$, which grows with the heterogeneity of the $\mathbf{w}_i^*$. Moreover, since solving our formulation provides n matrix equations, we can fully recover the column space of $\mathbf{B}^*$ as long as $\mathbf{w}_i^*$'s span $\mathbb{R}^k$. In contrast, solving (4) yields only one matrix equation, so there is no hope to recover the column space of $\mathbf{B}^*$ for any $k > 1$.

3 FedRep Algorithm

FedRep solves Problem (2) by distributing the computation across clients. The server and clients aim to learn the global representation ϕ together, while the i-th client aims to learn its unique local head denoted by h_i locally (see Figure 2). To do so, FedRep alternates between client updates and a server update on each communication round.

Client Update. On each round, a constant fraction $r \in (0, 1]$ of the clients are selected to execute a client update. In the client update, client i makes τ local gradient-based updates to solve for its optimal head given the current global representation ϕ^t communicated by the server. Namely, for $s = 1, \ldots, \tau$, client i updates its head as follows:

$$h_i^{t,s+1} = \texttt{GRD}(f_i(h_i^{t,s} \circ \phi^t), h_i^{t,s}, \alpha),$$

Algorithm 1 `FedRep`

Parameters: Participation rate r, step sizes α, η; number of local updates τ; number of communication rounds T; initial $\phi^0, h_1^0, \ldots, h_n^0$.

for $t = 1, 2, \ldots, T$ **do**

  Server receives a batch of clients $\mathcal{I}^t$ of size rn

  Server sends current representation ϕ^t to these clients

  for each client i in $\mathcal{B}$ **do**

    Client i initializes $h_i^t \leftarrow h_i^{t-1,\tau}$

    Client i makes τ updates to its head h_i^t:

    for $s = 1$ **to** τ **do**

      $h_i^{t,s} \leftarrow \texttt{GRD}(f_i(h_i^{t,s} \circ \phi^t), h_i^{t,s}, \alpha)$

    end for

    Client i locally updates the representation as:

      $\phi_i^{t+1} \leftarrow \texttt{GRD}(f_i(h_i^{t,\tau} \circ \phi^t), \phi^t, \alpha)$

    Client i sends updated representation ϕ_i^{t+1} to server

  end for

  for each client i not in $\mathcal{B}$, **do**

    Set $h_i^{t,\tau} \leftarrow h_i^{t-1,\tau}$

  end for

  Server computes the new representation as

    $\phi^{t+1} = \frac{1}{rn} \sum_{i \in \mathcal{I}^t} \phi_i^{t+1}$

end for

where $\texttt{GRD}(f, h, \alpha)$ is generic notation for an update of the variable h using a gradient of function f with respect to h and the step size α. For example, $\texttt{GRD}(f_i(h_i^{t,s} \circ \phi^t), h_i^{t,s}, \alpha)$ can be a step of gradient descent, stochastic gradient descent (SGD), SGD with momentum, etc. The key is that client i makes many such local updates, i.e., τ is large, to find the optimal head based on its local data, given the most recent representation ϕ^t received from the server.

Server Update. Once the local updates with respect to the head h_i finish, the client participates in the server update by taking one local gradient-based update with respect to the current representation, i.e., computing

$$\phi_i^{t+1} \leftarrow \texttt{GRD}(f_i(h_i^{t,\tau} \circ \phi^t), \phi^t, \alpha).$$

It then sends ϕ_i^{t+1} to the server, which averages the local updates to compute the next representation ϕ^{t+1}. The entire procedure is outlined in Algorithm 1.

4 Low-Dimensional Linear Representation

In this section, we analyze an instance of Problem (2) with quadratic loss functions and linear models, as discussed in Section 2.1. Here, each client's problem is to solve a linear regression with a two-layer linear neural network. In particular, each client i attempts to find a shared global projection onto a low-dimension subspace $\mathbf{B} \in \mathbb{R}^{d \times k}$ and a unique regressor $\mathbf{w}_i \in \mathbb{R}^k$ that together accurately map its samples $\mathbf{x}_i \in \mathbb{R}^d$ to labels $y_i \in \mathbb{R}$. The matrix $\mathbf{B}$ corresponds

to the representation ϕ, and $\mathbf{w}_i$ corresponds to local head h_i for the i-th client. We thus have $h_i \circ \phi(\mathbf{x}_i) = \mathbf{w}_i^\top \mathbf{B}^\top \mathbf{x}_i$. Hence, the loss function for client i is given by:

$$f_i(\mathbf{w}_i \circ \mathbf{B}) := \tfrac{1}{2}\mathbb{E}_{(\mathbf{x}_i, y_i)\sim \mathcal{D}_i}\left[(y_i - \mathbf{w}_i^\top \mathbf{B}^\top \mathbf{x}_i)^2\right] \tag{5}$$

meaning that the global objective is:

$$\min_{\substack{\mathbf{B}\in\mathbb{R}^{d\times k}\\ \mathbf{W}\in\mathbb{R}^{n\times k}}} F(\mathbf{B}, \mathbf{W}) := \frac{1}{2n}\sum_{i=1}^{n} \mathbb{E}_{(\mathbf{x}_i, y_i)}\left[(y_i - \mathbf{w}_i^\top \mathbf{B}^\top \mathbf{x}_i)^2\right], \tag{6}$$

where $\mathbf{W} = [\mathbf{w}_1^\top, \ldots, \mathbf{w}_n^\top] \in \mathbb{R}^{n\times k}$ is the concatenation of client-specific heads. To evaluate the ability of FedRep to learn an accurate representation, we model the local datasets $\{\mathcal{D}_i\}_i$ such that, for $i = 1 \ldots, n$

$$y_i = \mathbf{w}_i^{*\top} \mathbf{B}^{*\top} \mathbf{x}_i,$$

for some ground-truth representation $\mathbf{B}^* \in \mathbb{R}^{d\times k}$ and local heads $\mathbf{w}_i^* \in \mathbb{R}^k$–i.e. a standard regression setting. In other words, all of the clients' optimal solutions live in the same k-dimensional subspace of $\mathbb{R}^d$, where k is assumed to be small. We consider the case in which y_i contains no noise for simplicity, as adding noise does not significantly change our analysis. Moreover, we make the following standard assumption on the samples $\mathbf{x}_i$.

Assumption 1. *(Sub-gaussian design) The samples $\mathbf{x}_i \in \mathbb{R}^d$ are i.i.d. with mean $\mathbf{0}$, covariance $\mathbf{I}_d$, and are $\mathbf{I}_d$-sub-gaussian, i.e. $\mathbb{E}[e^{\mathbf{v}^\top \mathbf{x}_i}] \le e^{\|\mathbf{v}\|_2^2/2}$ for all $\mathbf{v} \in \mathbb{R}^d$.*

4.1 FedRep

We next discuss how FedRep tries to recover the optimal representation in this setting.

Client Update. As in Algorithm 1, rn clients are selected on round t to update their current local head $\mathbf{w}_i^t$ and the global representation $\mathbf{B}^t$. Each selected client i samples a fresh batch $\{\mathbf{x}_i^{t,j}, y_i^{t,j}\}_{j=1}^m$ of m samples according to its local data distribution $\mathcal{D}_i$ to use for updating both its head and representation on each round t that it is selected. That is, within the round, client i considers the batch loss

$$\hat{f}_i^t(\mathbf{w}_i^t \circ \mathbf{B}^t) := \frac{1}{2m}\sum_{j=1}^{m}(y_i^{t,j} - \mathbf{w}_i^{t\top} \mathbf{B}^{t\top} \mathbf{x}_i^{t,j})^2. \tag{7}$$

Since $\hat{f}_i^t$ is strongly convex with respect to $\mathbf{w}_i^t$, the client can find an update for a local head that is ϵ-close to the global minimizer of (7) after at most $\mathcal{O}(\log(1/\epsilon))$ local gradient updates. Alternatively, since the function is also quadratic, the client can solve for the optimal $\mathbf{w}$ directly in only $\mathcal{O}(mk^2 + k^3)$ operations. Thus, to simplify the analysis we assume each selected client obtains $\mathbf{w}_i^{t+1} = \text{argmin}_{\mathbf{w}} \hat{f}_i^t(\mathbf{w} \circ \mathbf{B}^t)$ during each round of local updates.

Server Update. After updating its head, client i updates the global representation with one step of gradient descent using the same m samples and sends the update to the server, as outlined in Algorithm 2. Then, the server computes the new representation by averaging over received representations.

Algorithm 2 `FedRep` for linear regression

Input: Step size η; number of rounds T, participation rate r, initial $\mathbf{B}^0$.
for $t = 1, 2, \ldots, T$ **do**
  Server receives a subset $\mathcal{I}^t$ of clients of size rn
  Server sends current representation $\mathbf{B}^t$ to these clients
  for $i \in \mathcal{I}^t$ **do**
    Client i samples a fresh batch of m samples.
    Client i updates w_i:
      $\mathbf{w}_i^{t+1} \leftarrow \operatorname{argmin}_{\mathbf{w}} \hat{f}_i^t(\mathbf{w} \circ \mathbf{B}^t)$
    Client i updates representation:
      $\mathbf{B}_i^{t+1} \leftarrow \mathbf{B}^t - \eta \nabla_{\mathbf{B}} \hat{f}_i^t(\mathbf{w}_i^{t+1} \circ \mathbf{B}^t)$
    Client i sends $\mathbf{B}_i^{t+1}$ to the server.
  end for
  Server averages updates: $\mathbf{B}^{t+1} \leftarrow \frac{1}{rn} \sum_{i \in \mathcal{I}^t} \mathbf{B}_i^{t+1}$
end for

4.2 Analysis

As mentioned earlier, in FedRep, each client i performs an alternating minimization-descent method to solve its nonconvex objective in (7). This means the global loss over all clients at round t is given by

$$\frac{1}{n} \sum_{i=1}^{n} \hat{f}_i^t(\mathbf{w}_i^t \circ \mathbf{B}^t) := \frac{1}{2mn} \sum_{i=1}^{n} \sum_{j=1}^{m} (y_i^{t,j} - \mathbf{w}_i^{t\top} \mathbf{B}^{t\top} \mathbf{x}_i^{t,j})^2. \tag{8}$$

This objective has many global minima, including all pairs of matrices $(\mathbf{Q}^{-1}\mathbf{W}^*, \mathbf{B}^*\mathbf{Q}^\top)$ where $\mathbf{Q} \in \mathbb{R}^{k \times k}$ is invertible, eliminating the possibility of exactly recovering the ground-truth factors $(\mathbf{W}^*, \mathbf{B}^*)$. Instead, the ultimate goal of the server is to recover the ground-truth *representation*, i.e., the column space of $\mathbf{B}^*$. To evaluate how closely the column space is recovered, we define the distance between subspaces as follows.

Definition 1. *The principal angle distance between the column spaces of* $\mathbf{B}_1, \mathbf{B}_2 \in \mathbb{R}^{d \times k}$ *is given by*

$$dist(\mathbf{B}_1, \mathbf{B}_2) := \|\hat{\mathbf{B}}_{1,\perp}^\top \hat{\mathbf{B}}_2\|_2, \tag{9}$$

where $\hat{\mathbf{B}}_{1,\perp}$ *and* $\hat{\mathbf{B}}_2$ *are orthonormal matrices satisfying* $span(\hat{\mathbf{B}}_{1,\perp}) = span(\mathbf{B}_1)^\perp$ *and* $span(\hat{\mathbf{B}}_2) = span(\mathbf{B}_2)$.

Next, we define a key property required for our results.

Definition 2. *A rank-k matrix* $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$ *is* μ-row-wise incoherent *if* $\max_{i \in [d_1]} \|\mathbf{m}_i\|_2 \leq \frac{\mu\sqrt{d_2}}{\sqrt{d_1}} \|\mathbf{M}\|_F$, *where* $\mathbf{m}_i \in \mathbb{R}^{d_2}$ *is the i-th row of* $\mathbf{M}$.

Incoherence of the ground-truth matrices is a key property required for efficient matrix completion and other sensing problems with sparse measurements [Chi et al., 2019]. Since our measurement

matrices are row-wise sparse, we require the row-wise incoherence of $\mathbf{W}^*\mathbf{B}^{*\top}$, which the following assumption implies with $\mu = 1$.

Assumption 2. *(Client normalization) The ground-truth client-specific parameters satisfy* $\|\mathbf{w}_i^*\|_2 = \sqrt{k}$ *for all* $i \in [n]$*, and* $\mathbf{B}^*$ *has orthonormal columns.*

Assumption 3. *(Client diversity) Let* $\bar{\sigma}_{\min,*}$ *be the minimum singular value of any matrix* $\frac{1}{\sqrt{rn}}\overline{\mathbf{W}} \in \mathbb{R}^{rn\times k}$ *with rows being an* rn*-sized subset of ground-truth client-specific parameters* $\{\mathbf{w}_1^*, \ldots, \mathbf{w}_n^*\}$*. Then* $\bar{\sigma}_{\min,*} > 0$*.*

Assumption 3 states that if we select any rn clients, their optimal solutions span $\mathbb{R}^k$. Indeed, this assumption is weak as we expect the number of participating clients rn to be substantially larger than k. Note that if we do not have client solutions that span $\mathbb{R}^k$, recovering $\mathbf{B}^*$ would be impossible because the samples $(\mathbf{x}_i^j, y_i^j)$ may never contain any information about one or more features of $\mathbf{B}^*$.

Our main result shows that the iterates $\{\mathbf{B}^t\}_t$ generated by FedRep in this setting linearly converge to the optimal representation $\mathbf{B}_*$ in principal angle distance.

Theorem 1. *Suppose Assumptions 2 and 3 hold. Let* $\bar{\sigma}_{\min,*}$ *be defined as in Assumption 3, and similarly define* $\bar{\sigma}_{\max,*}$ *as the maximum singular value. Also define* $\kappa := \frac{\bar{\sigma}_{\max,*}}{\bar{\sigma}_{\min,*}}$ *and* $E_0 := 1 - dist^2(\mathbf{B}^0, \mathbf{B}^*)$*. 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* $\eta \leq 1/(4\bar{\sigma}_{\max,*}^2)$*, we have*

$$dist(\mathbf{B}^T, \mathbf{B}^*) \leq \left(1 - \eta E_0 \bar{\sigma}_{\min,*}^2/2\right)^{T/2} dist(\mathbf{B}^0, \mathbf{B}^*), \tag{10}$$

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

From Assumption 3, we have that $\bar{\sigma}_{\min,*}^2 > 0$, so the RHS of (10) strictly decreases with T for appropriate step size. Considering the complexity of m and the fact that the algorithm converges exponentially fast, the total number of samples required per client to reach an ϵ-accurate solution in principal angle distance is $\Theta\left(m \log\left(\kappa/\epsilon E_0\right)\right)$, which is

$$\Theta\left(\left[\kappa^4 k^2 \left(k\log(rn) + d/rn\right)\right] \log\left(\kappa/\epsilon E_0\right)\right). \tag{11}$$

Next, a few remarks about this sample complexity follow.

When and whom does federation help? Observe that for a single client with no collaboration, the sample complexity scales as $\Theta(d)$. With FedRep, however, the sample complexity scales as $\Theta(\log(rn) + d/rn)$. Thus, so long as $\log(rn) + d/rn \ll d$, federation helps. This indeed holds in several settings, for instance when $1 \ll n \ll e^{\Theta(d)}$. In practical scenarios, d (the data dimension) is large, and thus $e^{\Theta(d)}$ is exponentially larger; therefore collaboration helps *each individual client.* Furthermore, from the point of view of a new client who enters the system later, it has a representation available for free, and this new client's sample complexity for adapting to its task is only $k\log(1/\epsilon)$. Thus, both the overall system benefits (a representation has been learned, which is useful for the new client because it now only needs to learn a head), and each individual client that did take part in the federated training also benefits.

Connection to matrix sensing. We would like to mention that the problem in (6) has a close connection with the matrix sensing problem, and in fact it can be considered as an instance of that problem; see the proof in Appendix B for more details. Considering this connection, our theoretical results also contribute to the theoretical study of matrix sensing. Although matrix sensing is a well-studied problem, our setting presents two new analytical challenges: (i) due to row-wise sparsity in the measurements, the sensing operator does not satisfy the commonly-used Restricted Isometry Property within an efficient number of samples, i.e., it does not efficiently concentrate to an identity operation on all rank-k matrices, and (ii) FedRep executes a novel non-symmetric procedure. We further discuss these challenges in Appendix B.1. To the best of our knowledge, Theorem 1 provides the first convergence result for an alternating minimization-descent procedure to solve a matrix sensing problem. It is also the first result to show sample-efficient linear convergence of any solution to a matrix sensing with rank-one, row-wise sparse measurements. The state-of-the-art result for the closest matrix sensing setting to ours is given by Zhong et al. [2015] for rank-1, independent Gaussian measurements, which our result matches up to an $\mathcal{O}(\kappa^2)$ factor. However, our setting is more challenging as we have rank-1 *and* row-wise sparse measurements, and dependence on κ^4 has been previously observed in settings with sparse measurements, e.g. matrix completion [Jain et al., 2013].

New users and dimensionality reduction. Theorem 1 is related to works studying representation learning in the context of multi-task learning. [Tripuraneni et al., 2020] and [Du et al., 2020] provided upper bounds on the generalization error resulting from learning a low-dimensional representation of tasks assumed to share a common representation. They show that, if the common representation is learned, then excess risk bound on a new task is $\mathcal{O}(\frac{\mathcal{C}(\Phi)}{nm} + \frac{k}{m_{\text{new}}})$, where $\mathcal{C}(\Phi)$ is the complexity of the representation class Φ and m_{new} is the number of labelled samples from the new task that the learner can use for fine-tuning. Since the number of test samples must exceed only $\mathcal{O}(k)$, where k is assumed to small, these works demonstrate the dimensionality-reducing benefits of representation learning. Our work complements these results by showing how to provably and efficiently learn the representation in the linear case.

5 Experiments

We focus on three points in our experiments: (i) the effect of many local updates for the local head in FedRep (ii) the quality of the global representation learned by FedRep and (iii) the applicability of FedRep to a wide range of datasets. Further experimental details are provided in the appendix.

5.1 Synthetic Data

We start by experimenting with an instance of the multi-linear regression problem analyzed in Section 4. Consistent with this formulation, we generate synthetic samples $\mathbf{x}_i^j \sim \mathcal{N}(0, \mathbf{I}_d)$ and labels $y_i^j \sim \mathcal{N}(\mathbf{w}_i^{*\top} \hat{\mathbf{B}}^{*\top} \mathbf{x}_i^j, 10^{-3})$ (here we include an additive Gaussian noise). The ground-truth heads $\mathbf{w}_i^* \in \mathbb{R}^k$ for clients $i \in [n]$ and the ground-truth representation $\hat{\mathbf{B}}^* \in \mathbb{R}^{d \times k}$ are generated randomly by sampling and normalizing Gaussian matrices.

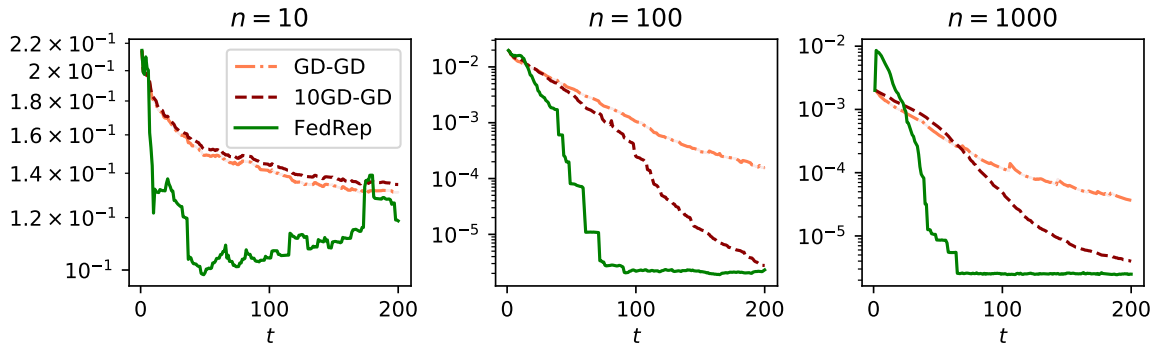

Figure 3: Comparison of (principal angle) distances between the ground-truth and estimated representations by FedRep and alternating gradient descent algorithms for different numbers of clients n. In all plots, $d = 10$, $k = 2$, $m = 5$, and $r = 0.1$.

Benefit of finding the optimal head. We first demonstrate that the convergence of FedRep improves with larger number of clients n, making it highly applicable to federated settings. Further, we give evidence showing that this improvement is augmented by the minimization step in FedRep, since methods that replace the minimization step in FedRep with 1 and 10 steps of gradient descent (GD-GD and 10GD-GD, respectively) do not scale properly with n. In Figure 3, we plot convergence trajectories for FedRep, GD-GD, and 10GD-GD for four different values of n and fixed m, d, k and r. As we observe in Figure 3, by increasing the number of nodes n, clients converge to the true representation faster. Also, running more local updates for finding the local head accelerates the convergence speed of FedRep. In particular, FedRep which exactly finds the optimal local head at each round has the fastest rate compared to GD-GD and 10GD-GD that only run 1 and 10 local updates, respectively, to learn the head.

Generalization to new clients. Next, we evaluate the effectiveness of the representation learned by FedRep in reducing the sample complexity for a new client which has not participated in training. We first train FedRep and FedAvg on a fixed set of $n = 100$ clients as in Figure 1. The new client has access to m_{new} labelled local samples. It will use the representation $\hat{\mathbf{B}}^* \in \mathbb{R}^{d \times k}$ learned by training clients, where $(d, k) = (20, 2)$, and trains a local head using this representation and its local training samples. For both FedRep and FedAvg, we solve for the optimal head given these samples and the representation learned during training. We compare the MSE of the resulting model on the new client's test data to that of a model trained by only using the m_{new} labelled samples from the new client (Local Only) in Figure 4. The large error for FedAvg demonstrates that it does not learn the ground-truth representation. Meanwhile, the representation learned by FedRep allows an accurate model to be found for the new client as long as $m_{\text{new}} \geq k$, which drastically improves over the complexity for Local Only ($m_{\text{new}} = \Omega(d)$).

5.2 Real Data

We next investigate whether these insights apply to nonlinear models and real datasets. Additional results and details of our experimental setup are provided in the appendix.

Datasets and Models. We use four real datasets: CIFAR10 and CIFAR100 [Krizhevsky et al.,

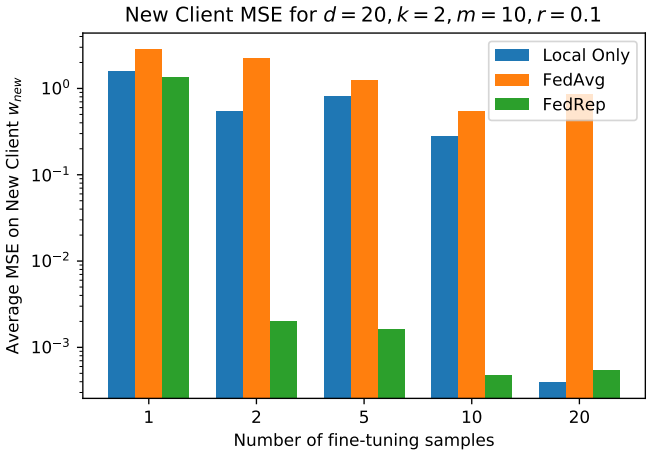

Figure 4: MSE on new clients sharing the representation after fine-tuning using various numbers of samples from the new client.

2009], FEMNIST [Caldas et al., 2018, Cohen et al., 2017] and Sent140 [Caldas et al., 2018]. The first three are image datasets and the last is a text dataset for which the goal is to classify the sentiment of a tweet as positive or negative. We control the heterogeneity of CIFAR10 and CIFAR100 by assigning different numbers S of classes per client, from among 10 and 100 classes, respectively. We use balanced versions of these datasets, meaning that each client has $50{,}000/n$ training samples. For FEMNIST, we first restrict the dataset to 10 handwritten letters and assign samples to clients according to a log-normal distribution as in [Li et al., 2019]. We consider a partition of $n = 150$ clients with an average of 148 samples/client. For Sent140, we use the natural assignment of tweets to their author, and use $n = 183$ clients with an average of 72 samples per client. We use a 5-layer CNN for the CIFAR datasets, a 2-hidden layer multilayer perceptron (MLP) for FEMNIST, and an RNN for Sent140 (details provided in the appendix).

Baselines. We compare against a variety of baselines: (1) GD-GD, which simultaneously updates a representation and unique local head in each local update as in [Arivazhagan et al., 2019]; (2) FedAvg [McMahan et al., 2017], which tries to learn a single global model; (3) Fed-MTL [Smith et al., 2017], which learns local models and a regularizer to encode relationships among the clients; (4) LG-FedAvg [Liang et al., 2020], which learns local representations and a global head; and (5) local only training with no global communication. Method (2) is the standard federated learning technique, (3) is a well-known method for achieving personalized models, and (1) and (4) are methods with similar intuitions as ours.

Implementation. In each experiment we sample 10% of the clients on every round. We initialize all models randomly and train for $T = 200$ communication rounds. Unless otherwise noted, FedRep executes multiple local epochs of SGD with momentum to train the local head, followed by one epoch for the representation, in each local update. All other methods use $E = 1$ local epoch (for all parameters) besides FedAvg, which uses $E = 10$.

Benefit of more local updates. As mentioned in Section 1, a key advantage of our formulation is that it enables clients to run many local updates without causing divergence from the global optimal solution. We demonstrate an example of this in Figure 5. Here, there are $n = 100$ clients

Table 1: Average local test accuracy over the final 10 rounds out of $T = 200$ total rounds, with 95% confidence intervals.

Dataset	FedRep	GD-GD	FedAvg	Fed-MTL	LG-FedAvg	Local
CIFAR10, $S = 2$, $n = 100$	$86.08 \pm .3$	$83.52 \pm .4$	44.29 ± 1.9	$70.44 \pm .2$	78.12 ± 1.1	$87.28 \pm .6$
CIFAR10, $S = 2$, $n = 1000$	$76.41 \pm .5$	$73.07 \pm .5$	41.11 ± 1.6	$56.90 \pm .1$	68.32 ± 1.4	$72.17 \pm .3$
CIFAR10, $S = 5$, $n = 100$	$72.19 \pm .9$	72.26 ± 1.0	58.14 ± 1.5	$54.21 \pm .3$	57.83 ± 1.2	$67.76 \pm .5$
CIFAR10, $S = 5$, $n = 1000$	$53.41 \pm .7$	$52.27 \pm .8$	$50.40 \pm .9$	$34.58 \pm .5$	41.05 ± 1.3	$47.56 \pm .4$
CIFAR100, $S = 20$, $n = 100$	$38.17 \pm .8$	$37.41 \pm .5$	$15.74 \pm .8$	$23.38 \pm .3$	$31.06 \pm .5$	$32.04 \pm .7$
CIFAR100, $S = 20$, $n = 500$	$21.88 \pm .6$	$21.92 \pm .6$	$19.43 \pm .7$	$10.56 \pm .3$	$14.83 \pm .4$	$21.06 \pm .6$
FEMNIST, $n = 150$	$65.52 \pm .8$	$39.21 \pm .6$	60.71 ± 1.8	$32.34 \pm .7$	$29.78 \pm .5$	$29.12 \pm .9$
Sent140, $n = 183$	$72.13 \pm .2$	$69.57 \pm .8$	56.82 ± 1.4	$54.82 \pm .5$	70.84 ± 1.4	$70.22 \pm .8$

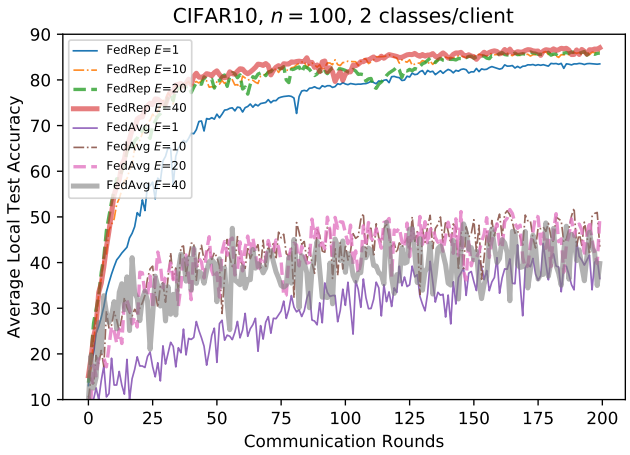

Figure 5: CIFAR10 local test errors for different numbers of local epochs E for FedRep and FedAvg.

where each has $S = 2$ classes of images. For FedAvg, we observe running more local updates does not necessarily improve the performance. In contrast, FedRep's performance is monotonically non-decreasing with E, i.e., *FedRep requires less tuning of E and is never hurt by more local computation.*

Robustness to varying numbers of samples/client and level of heterogeneity. We show the average local test errors for all of the algorithms for a variety of settings in Table 1. In all cases, FedRep performs at least as well as the best method, up to statistical equivalence. Not surprisingly, the largest gain for FedRep over FedAvg comes in heterogeneous settings with many samples/client: CIFAR10 and CIFAR100 with 2 and 20 classes/client, respectively, and $n = 100$. In these settings, FedRep matches, or nearly matches, the performance of local-only training. Meanwhile, even in homogeneous and small-data settings conducive to global-only training, i.e. $S = 5$ and $n = 1000$ on CIFAR10, FedRep still outperforms all the baselines, even without any regularization added to prevent over-fitting to the head. *This evidence supports our intuition that FedRep effectively interpolates between local and global learning.* Furthermore, FedRep outperforms all other baselines.

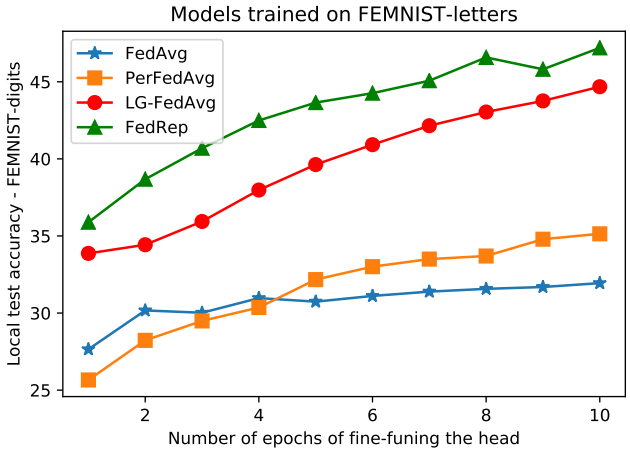

Figure 6: Test accuracy on handwritten digits from FEMNIST after fine-tuning the head of models trained on FEMNIST-letters.

Generalization to new clients. We evaluate the strength of the representation learned by FedRep in terms of adaptation for new users. To do so, we consider an additional baseline, PerFedAvg [Fallah et al., 2020], which aims to find a single global model that can quickly adapt to a new task. First, we train FedRep, FedAvg, PerFedAvg, and LG-FedAvg in the usual setting on the partition of FEMNIST containing images of 10 handwritten letters (FEMNIST-letters). Then, we encounter clients with data from a different partition of the FEMNIST dataset, containing images of handwritten digits. We assume we have access to a dataset of 500 samples at this new client to fine tune the head. Using these, with each of the algorithms, we fine tune the head over multiple epochs while keeping the representation fixed. In Figure 6, we repeatedly sweep over the same 500 samples over multiple epochs to further refine the head, and plot the corresponding local test accuracy. As is apparent, FedRep has significantly better performance than these baselines.

6 Discussion

We introduce a novel representation learning framework and algorithm for federated learning, and we provide both theoretical and empirical justification for its utility in federated settings. In particular, our proposed framework exploits the structure of federating learning by (i) leveraging all clients' data to learn a global representation that enhances each client's model and could possibly generalize to new users and (ii) leveraging the computational power of clients to run multiple local updates for learning their local heads. Future work remains to analyze the convergence properties of FedRep in non-linear settings.

Appendix

A Additional Experimental Results

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

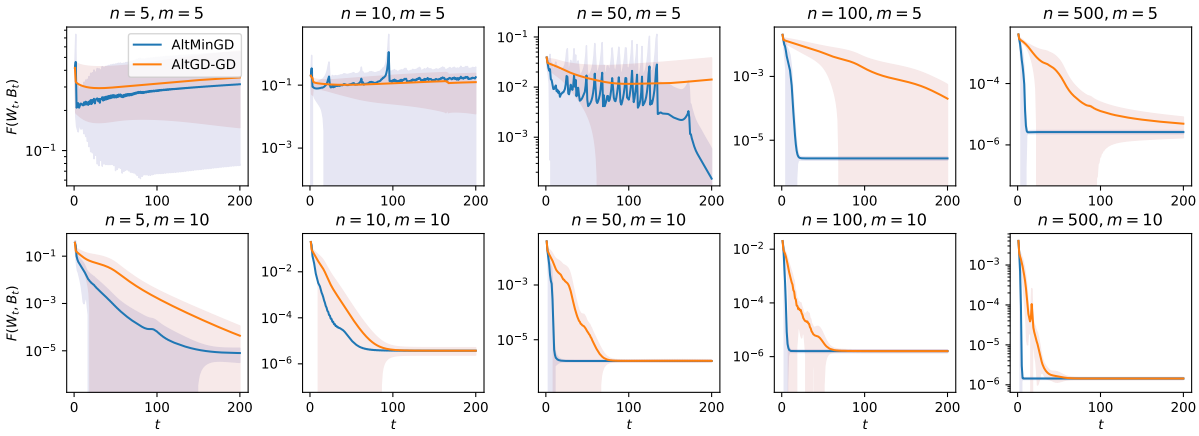

Figure 7: Function values for FedRep and GD-GD. The value of m is fixed in each row and n is fixed in each column. Here $r = 1$ (full participation) and the average trajectories over 10 trials are plotted along with 95% confidence intervals. Principal angle distances are not plotted as the results are 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: Comparison against additional baselines

Additional Results. Extensions of Table 1 and Figure 6 to include additional baselines are given in Table 2 and Figure 8. We compare against two recently proposed methods for personalized federated learning L2GD [Hanzely and Richtárik, 2020] and APFL [Deng et al., 2020], as well as FedProx [Li et al., 2018], a well-known method for federated learning in heterogeneous settings.

Table 2: Performance of additional baselines in the setting of Table 1.

Dataset	FedRep	L2GD	APFL	FedProx
CIFAR10,$S = 2$, $n = 100$	$86.08 \pm .3$	84.47 ± 1.3	$85.65 \pm .3$	$43.72 \pm .9$
CIFAR10, $S = 2$, $n = 1000$	$76.41 \pm .5$	60.34 ± 1.5	$67.84 \pm .6$	$40.10 \pm .4$
CIFAR10, $S = 5$, $n = 100$	$72.19 \pm .9$	68.36 ± 1.1	$68.38 \pm .2$	$54.94 \pm .2$
CIFAR10, $S = 5$, $n = 1000$	$53.41 \pm .7$	37.14 ± 1.2	$38.11 \pm .3$	$51.22 \pm .3$
CIFAR100, $S = 20$, $n = 100$	$38.17 \pm .8$	34.70 ± 1.2	$35.61 \pm .4$	$22.82 \pm .4$
CIFAR100, $S = 20$, $n = 500$	$21.88 \pm .6$	12.46 ± 1.1	$22.90 \pm .3$	$20.11 \pm .3$

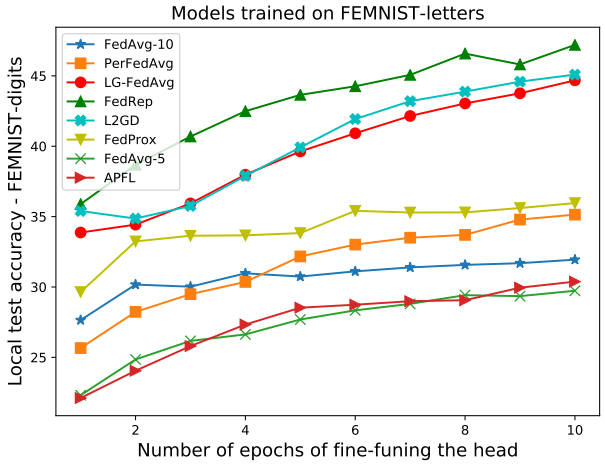

Figure 8: Test accuracy on handwritten digits from FEMNIST after fine-tuning the head of models trained on FEMNIST-letters.

We see from Figure 8 that L2GD and APFL nearly match the performance of FedRep in settings with a small number of clients and a large amount of data per client (recall that for the CIFAR datasets, the number of samples per client is $50,000/n$), but struggle in large-client, small-data-per-client environments. Meanwhile, in Figure 8, the L2GD global model is a relatively strong initial model for fine-tuning on new clients, but still does not match FedRep. Conversely, the APFL global model performs only as well as FedAvg with the same number (5) of local epochs (FedAvg-5), which is because the global model for APFL is essentially computed using FedAvg. Like FedAvg, FedProx performs best relative to FedRep in relatively homogeneous settings ($S = 5$ in Table 2) and learns an inferior representation on FEMNIST as suggested by Figure 8.

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 Tables 1 and 2 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 Figures 6 and 8, 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 3. For both FEMNIST and Sent140 we use the LEAF framework [Caldas et al., 2018].

Implementations. All experiments were performed on a 3.7GHz, 6-core Intel Corp i7-8700K CPU and the code was written in PyTorch. All methods use SGD with momentum with parameter 0.5. For CIFAR10 and CIFAR100, the sample batch size is 50, for FEMNIST it is 10, and for Sent140 it is 4. The participation rate r is always 0.1, besides in the fine-tuning phases in Figures 6 and 8, in which all clients are sampled in each round.

In Tables 1 and 2, we initialize all methods randomly and train for $T = 200$ communication rounds. The accuracy shown is the average local test accuracy over the final ten communication rounds. Intervals of 95% confidence are given over 3 random trials. Learning rates were adopted from published results [Liang et al., 2020, McMahan et al., 2017, Smith et al., 2017], or, in the case of Sent140, first tuned for FedAvg (resulting in a learning rate of 10^{-2}) then applied to all other methods. For FedRep, the same learning rate was used to update both local and global parameters.

We used the implementations of FedAvg, Fed-MTL and LG-FedAvg found at `https://github.com/pliang279/LG-FedAvg/` corresponding to the paper [Liang et al., 2020]. 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 [Pennington et al., 2014] and train RNN with an LSTM module followed by a fully-connected decoding layer, as in [Caldas et al., 2018].

For FedRep, we treated the head as the weights and biases in the final fully-connected layers in each of the models. For LG-FedAvg, we treated the first two convolutional layers and the first fully-connected layer of the model for CIFAR10 and CIFAR100 as the local representation, and the two final fully-connected layers as the global parameters. For FEMNIST, we set all parameters besides those in the final hidden layer and output layer to be 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 PerFedAvg [Fallah et al., 2020] we used the first-order version (FO) and 10 local epochs for training.

For L2GD [Hanzely and Richtárik, 2020] we tuned the step size η over $\{1, 0.5, 0.1\}$ (from which 0.5 was the best). For fair comparison with other methods, we set $p = 0.9$, thus the local parameters are trained on 10% of the communication rounds. We also set $\lambda = 1$ and $E = 5$. For APFL [Deng et al., 2020], we used the highest-performing constant value of α on CIFAR10 according to the experiments in [Deng et al., 2020], namely $\alpha = 0.75$. Like L2GD, we used $E = 5$ local epochs, and the global models produced by both methods are used as the initial models for fine-tuning in Figure 8. For FedProx, we use $E = 10$ local epochs when $n = 100$ and $E = 5$ local

epochs otherwise. We tune the parameter μ from among $\{1, 0.5, 0.1, 0.05\}$, and use $\mu = 0.1$ for all settings.

Note that allowing 5+ local epochs for L2GD, APFL, and FedProx, and 10 local epochs for FedAvg, means that these methods execute more computations in the local update than FedRep, since they update all parameters in each epoch. Moreover, these extra local epochs improved the test accuracy compared to the 1 local epoch case for all methods. Still, FedRep outperforms all of them.

Table 3: Dataset statistics.

Dataset	Number of users (n)	Avg samples/user	Min samples/user
FEMNIST	150	148	50
Sent140	183	72	50

B Proof of Main Theoretical Result

B.1 Preliminaries.

Definition 3. *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. $\mathbb{E}[\exp(\mathbf{y}^\top \mathbf{A}^\top \mathbf{x})] \leq \exp\left(\|\mathbf{y}\|_2^2 \|\mathbf{A}\|_2^2 / 2\right)$.*

We use hats to denote orthonormal matrices (a matrix is called orthonormal if its set of columns is an orthonormal set). By Assumption 2, 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 2, 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:

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

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

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 $\mathbf{W}^t$.

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

(i) Row-wise sparse measurements. Recall that the measurement matrices $\mathbf{A}_{i,j}^t$ 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. For background, the RIP is defined as follows:

Definition 4. *(RIP) 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*

$$(1 - \delta_k) \|\mathbf{M}\|_F^2 \leq \|\mathcal{B}(\mathbf{M})\|_2^2 \leq (1 + \delta_k) \|\mathbf{M}\|_F^2 \tag{15}$$

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

$$\begin{aligned} \|\frac{1}{\sqrt{m}} \mathcal{A}(\mathbf{M})\|_2^2 &= \frac{1}{m} \sum_{i=1}^{rn} \sum_{j=1}^{m} \langle \mathbf{e}_i(\mathbf{x}_i^j)^\top, \mathbf{e}_1(\mathbf{x}_1^1)^\top \rangle^2 \\ &= \frac{1}{m} \|\mathbf{x}_1^1\|_2^4 + \frac{1}{m} \sum_{j=2}^{m} \langle \mathbf{x}_1^j, \mathbf{x}_1^1 \rangle^2 \\ &\geq \frac{1}{m} \|\mathbf{x}_1^1\|_2^4 \end{aligned} \tag{16}$$

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

$$\begin{aligned}
\mathbb{P}\left(\frac{\left\|\frac{1}{\sqrt{m}}\mathcal{A}(\mathbf{M})\right\|_2^2}{\|\mathbf{M}\|_F^2} \geq \frac{d}{2m}\right) &\geq \mathbb{P}\left(\frac{\frac{1}{m}\|\mathbf{x}_1^1\|_2^4}{\|\mathbf{x}_1^1\|_2^2} \geq \frac{d}{2m}\right) \\
&= \mathbb{P}\left(\|\mathbf{x}_1^1\|_2^2 \geq \frac{d}{2}\right) \\
&= 1 - \mathbb{P}\left(\|\mathbf{x}_1^1\|_2^2 - d \leq \frac{-d}{2}\right) \\
&\geq 1 - e^{-cd} \qquad (17)
\end{aligned}$$

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 [Park et al., 2018, Tu et al., 2016, Zheng and Lafferty, 2016]. We combine aspects of both types of analysis in our proof.

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:*

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

$$\widetilde{\mathbf{B}}^{t+1} = \overline{\mathbf{B}}^t - \frac{\eta}{rnm}\left((\mathcal{A}^t)^\dagger \mathcal{A}^t(\widetilde{\mathbf{W}}^{t+1}(\overline{\mathbf{B}}^t)^\top - \mathbf{W}_*(\hat{\mathbf{B}}^*)^\top)\right)^\top \widetilde{\mathbf{W}}^{t+1} \qquad (19)$$

$$\overline{\mathbf{B}}^{t+1} = \widetilde{\mathbf{B}}^{t+1}(\widetilde{\mathbf{R}}^{t+1})^{-1} \qquad (20)$$

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

$$\widetilde{\mathbf{W}}^{t+1} = \arg\min_{\mathbf{W}} \|\mathcal{A}^t(\mathbf{W}\overline{\mathbf{B}}_t^\top - \mathbf{W}_*(\hat{\mathbf{B}}^*)^\top)\|_F^2 = \arg\min_{\mathbf{W}} \|\mathcal{A}^t((\mathbf{W}(\widetilde{\mathbf{R}}_t\mathbf{R}_B)^{-\top})\mathbf{B}_t^\top - \mathbf{W}_*(\hat{\mathbf{B}}^*)^\top)\|_F^2 \tag{21}$$

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 $\mathbf{B}^{t+1}$:

$$\mathbf{B}_{t+1} = \mathbf{B}_t - \frac{\eta}{rnm}\left((\mathcal{A})^\dagger\mathcal{A}^t(\mathbf{W}_{t+1}\mathbf{B}_t^\top - \mathbf{W}_*(\hat{\mathbf{B}}^*)^\top)\right)^\top \mathbf{W}_{t+1} \tag{22}$$

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

$$= \left[\overline{\mathbf{B}}^t,\ -\tfrac{\eta}{rnm}\left((\mathcal{A}^t)^\dagger\mathcal{A}^t(\widetilde{\mathbf{W}}^{t+1}(\overline{\mathbf{B}}^t)^\top - \mathbf{W}_*(\hat{\mathbf{B}}^*)^\top)\right)^\top \widetilde{\mathbf{W}}^{t+1}\right]\begin{bmatrix}\widetilde{\mathbf{R}}_t\mathbf{R}_B\\ (\widetilde{\mathbf{R}}_t\mathbf{R}_B)^{-\top}\end{bmatrix} \tag{23}$$

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:

$$\widetilde{\mathbf{B}}^{t+1} = \left[\overline{\mathbf{B}}^t,\ -\tfrac{\eta}{rnm}\left((\mathcal{A}^t)^\dagger\mathcal{A}^t(\widetilde{\mathbf{W}}^{t+1}(\overline{\mathbf{B}}^t)^\top - \mathbf{W}_*(\hat{\mathbf{B}}^*)^\top)\right)^\top \widetilde{\mathbf{W}}^{t+1}\right]\begin{bmatrix}\mathbf{I}_k\\ \mathbf{I}_k\end{bmatrix} \tag{24}$$

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

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

. □

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:

$$\mathbf{W}^{t+1} = \arg\min_{\mathbf{W}} \frac{1}{2rnm}\|\mathcal{A}^t(\mathbf{W}(\hat{\mathbf{B}}^t)^\top - \mathbf{W}_*(\hat{\mathbf{B}}^*)^\top)\|_F^2 \tag{25}$$

$$\mathbf{B}^{t+1} = \hat{\mathbf{B}}^t - \frac{\eta}{rnm}\left((\mathcal{A}^t)^\dagger\mathcal{A}^t(\mathbf{W}^{t+1}(\hat{\mathbf{B}}^t)^\top - \mathbf{W}^*(\hat{\mathbf{B}}^*)^\top)\right)^\top \mathbf{W}^{t+1} \tag{26}$$

$$\hat{\mathbf{B}}^{t+1} = \mathbf{B}^{t+1}(\mathbf{R}^{t+1})^{-1} \tag{27}$$

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:*

$$\mathbf{W}^{t+1} = \mathbf{W}^* {\hat{\mathbf{B}}^*}^\top \hat{\mathbf{B}}^t - \mathbf{F} \tag{28}$$

where $\mathbf{F}$ *is defined in equation* (33) *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 $\tilde{F}(\mathbf{W}, \hat{\mathbf{B}}^t) := \frac{1}{2rnm}\|\mathcal{A}^t(\mathbf{W}^*(\hat{\mathbf{B}}^*)^\top - \mathbf{W}(\mathbf{B}^t)^\top)\|_2^2$ 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

$$\frac{1}{m}\sum_{q=1}^{k}\left(\sum_{i=1}^{rn}\sum_{j=1}^{m}\mathbf{A}_{i,j}\hat{\mathbf{b}}_p^t(\hat{\mathbf{b}}_q^t)^\top\mathbf{A}_{i,j}^\top\right)\mathbf{w}_q^{t+1} = \frac{1}{m}\sum_{q=1}^{k}\left(\sum_{i=1}^{rn}\sum_{j=1}^{m}\mathbf{A}_{i,j}\hat{\mathbf{b}}_p^t(\hat{\mathbf{b}}_q^*)^\top\mathbf{A}_{i,j}^\top\right)\mathbf{w}_q^* \tag{29}$$

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

$$\mathbf{G} := \begin{bmatrix} \mathbf{G}_{11} & \cdots & \mathbf{G}_{1k} \\ \vdots & \ddots & \vdots \\ \mathbf{G}_{k1} & \cdots & \mathbf{G}_{kk} \end{bmatrix}, \mathbf{C} := \begin{bmatrix} \mathbf{C}_{11} & \cdots & \mathbf{C}_{1k} \\ \vdots & \ddots & \vdots \\ \mathbf{C}_{k1} & \cdots & \mathbf{C}_{kk} \end{bmatrix}, \mathbf{D} := \begin{bmatrix} \mathbf{D}_{11} & \cdots & \mathbf{D}_{1k} \\ \vdots & \ddots & \vdots \\ \mathbf{D}_{k1} & \cdots & \mathbf{D}_{kk} \end{bmatrix} \tag{30}$$

where, for $p, q \in [k]$: $\mathbf{G}_{pq} := \frac{1}{m}\sum_{i=1}^{rn}\sum_{j=1}^{m}\mathbf{A}_{i,j}\hat{\mathbf{b}}_p^t\hat{\mathbf{b}}_q^{t\top}\mathbf{A}_{i,j}^\top \in \mathbb{R}^{rn\times rn}$, $\mathbf{C}_{pq} := \frac{1}{m}\sum_{i=1}^{rn}\sum_{j=1}^{m}\mathbf{A}_{i,j}\hat{\mathbf{b}}_p^t(\hat{\mathbf{b}}_q^*)^\top\mathbf{A}_{i,j}^\top \in \mathbb{R}^{rn\times rn}$, and, $\mathbf{D}_{pq} := \langle\hat{\mathbf{b}}_p^t, \hat{\mathbf{b}}_q^*\rangle\mathbf{I}_{rn} \in \mathbb{R}^{rn\times rn}$. 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 (29), we have

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

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 $(\mathbf{GD}-\mathbf{C})\,\mathbf{w}^*)_p$ denote the p-th block of $(\mathbf{GD}-\mathbf{C})\,\mathbf{w}^*$. We have

$$\begin{aligned}\widetilde{\mathbf{w}}_p^{t+1} &= \sum_{q=1}^{k}\langle\hat{\mathbf{b}}_p^t,\hat{\mathbf{b}}_q^*\rangle\mathbf{w}_q^* - (\mathbf{G}^{-1}\,(\mathbf{GD}-\mathbf{C})\,\mathbf{w}^*)_p \\ &= \left(\sum_{q=1}^{k}\mathbf{w}_q^*(\hat{\mathbf{b}}_p^*)^\top\right)\hat{\mathbf{b}}_q^t - (\mathbf{G}^{-1}\,(\mathbf{GD}-\mathbf{C})\,\mathbf{w}^*)_p \\ &= \left(\mathbf{W}^*(\hat{\mathbf{B}}^*)^\top\right)\hat{\mathbf{b}}_q^t - (\mathbf{G}^{-1}\,(\mathbf{GD}-\mathbf{C})\,\mathbf{w}^*)_p \end{aligned} \tag{31}$$

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

$$\mathbf{W}^{t+1} = \mathbf{W}^*\hat{\mathbf{B}}^*(\hat{\mathbf{B}}^t)^\top - \mathbf{F} \tag{32}$$

where

$$\mathbf{F} = [(\mathbf{G}^{-1}(\mathbf{GD}-\mathbf{C})\widetilde{\mathbf{w}}^*)_1,\ldots,(\mathbf{G}^{-1}(\mathbf{GD}-\mathbf{C})\widetilde{\mathbf{w}}^*)_k] \tag{33}$$

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

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. $\mathrm{vec}(\mathbf{F}) = \mathbf{G}^{-1}(\mathbf{GD}-\mathbf{C})\mathbf{w}^*$. 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:

$$\begin{aligned}\mathbf{G}_{pq} &= \frac{1}{m}\sum_{i=1}^{rn}\sum_{j=1}^{m}\mathbf{A}_{i,j}\hat{\mathbf{B}}_p\hat{\mathbf{B}}_q^\top\mathbf{A}_{i,j}^\top \\ &= \frac{1}{m}\sum_{i=1}^{rn}\sum_{j=1}^{m}\mathbf{e}_i(\mathbf{x}_i^j)^\top\hat{\mathbf{B}}_p\hat{\mathbf{B}}_q^\top\mathbf{x}_i^j\mathbf{e}_i^\top \end{aligned} \tag{34}$$

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

$$(\mathbf{G}_{pq})_{ii} = \frac{1}{m}\sum_{j=1}^{m}(\mathbf{x}_i^j)^\top\hat{\mathbf{B}}_p\hat{\mathbf{B}}_q^\top\mathbf{x}_i^j = \hat{\mathbf{B}}_p^\top\left(\frac{1}{m}\sum_{j=1}^{m}\mathbf{x}_i^j(\mathbf{x}_i^j)^\top\right)\hat{\mathbf{B}}_q \tag{35}$$

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

$$(\mathbf{C}_{pq})_{ii} = \hat{\mathbf{B}}_p^\top\mathbf{\Pi}^i\hat{\mathbf{B}}_{*,q} \tag{36}$$

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

$$\mathbf{G}^i \coloneqq \left[\hat{\mathbf{B}}_p^\top \mathbf{\Pi}^i \hat{\mathbf{B}}_q\right]_{1\le p,q\le k} = \hat{\mathbf{B}}^\top \mathbf{\Pi}^i \hat{\mathbf{B}}, \quad \mathbf{C}^i \coloneqq \left[\hat{\mathbf{B}}_p^\top \mathbf{\Pi}^i \hat{\mathbf{B}}_{*,q}\right]_{1\le p,q\le k} = \hat{\mathbf{B}}^\top \mathbf{\Pi}^i \hat{\mathbf{B}}_* \tag{37}$$

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

$$\mathbf{f}^i = (\mathbf{G}^i)^{-1}(\mathbf{G}^i\mathbf{D}^i - \mathbf{C}^i)\mathbf{w}_i^* \tag{38}$$

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 $1 - e^{-110k^3\log(rn)}$.

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:

$$\hat{\mathbf{B}}^\top \mathbf{\Pi}^i \hat{\mathbf{B}} = \sum_{j=1}^{m} \frac{1}{\sqrt{m}} \hat{\mathbf{B}}^\top \mathbf{x}_i^j (\frac{1}{\sqrt{m}} \hat{\mathbf{B}}^\top \mathbf{x}_i^j)^\top \tag{39}$$

Let $\mathbf{v}_i^j \coloneqq \frac{1}{\sqrt{m}} \hat{\mathbf{B}}^\top \mathbf{x}_i^j$ 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

$$\sigma_{min}(\hat{\mathbf{B}}^\top \mathbf{\Pi}^i \hat{\mathbf{B}}) \ge 1 - C\left(\sqrt{\frac{k}{m}} + \frac{r}{\sqrt{m}}\right) \tag{40}$$

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

$$\sigma_{min}(\hat{\mathbf{B}}^\top \mathbf{\Pi}^i \hat{\mathbf{B}}) \geq 1 - \delta_k \tag{41}$$

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 (41) holds with probability at least

$$\begin{aligned} 1 - \exp\left(-\left(12k^{3/2}\sqrt{\log(rn)} - \sqrt{k}\right)^2\right) &\geq 1 - \exp\left(-k(12\sqrt{k}\sqrt{\log(rn)} - 1)^2\right) \\ &\geq 1 - \exp\left(121k^3\log(rn)\right) \end{aligned} \tag{42}$$

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

$$1 - rn\exp\left(-121k^3\log(rn)\right) \geq 1 - e^{-110k^3\log(rn)}, \tag{43}$$

completing the proof. □

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

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

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

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

$$\mathbf{H}^i \coloneqq \mathbf{G}^i\mathbf{D}^i - \mathbf{C}^i = \hat{\mathbf{B}}^\top \mathbf{\Pi}\hat{\mathbf{B}}\hat{\mathbf{B}}^\top\hat{\mathbf{B}}^* - \hat{\mathbf{B}}^\top\mathbf{\Pi}\hat{\mathbf{B}}^* = \hat{\mathbf{B}}^\top\left(\frac{1}{m}\mathbf{X}_i^\top\mathbf{X}_i\right)(\hat{\mathbf{B}}\hat{\mathbf{B}}^\top - \mathbf{I}_d)\hat{\mathbf{B}}^*, \tag{44}$$

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

$$\begin{aligned} \|(\mathbf{GD} - \mathbf{C})\mathbf{w}_*\|_2^2 &= \sum_{i=1}^{rn} \|\mathbf{H}^i\mathbf{w}_*^i\|_2^2 \\ &\leq \sum_{i=1}^{rn} \|\mathbf{H}^i\|_2^2\|\mathbf{w}_i^*\|_2^2 \\ &\leq \frac{k}{rn}\|\mathbf{W}^*\|_2^2 \sum_{i=1}^{rn}\|\mathbf{H}^i\|_2^2 \end{aligned} \tag{45}$$

where the last inequality follows almost surely from Assumption 2 (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 2 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 $\mathbf{V} := \frac{1}{\sqrt{m}}\mathbf{X}_i\hat{\mathbf{B}}$, 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 $\frac{1}{\sqrt{m}}\|\hat{\mathbf{B}}^{*\top}(\hat{\mathbf{B}}\hat{\mathbf{B}}^\top - \mathbf{I}_d)\|_2 = \frac{1}{\sqrt{m}}\mathrm{dist}(\hat{\mathbf{B}},\hat{\mathbf{B}}^*)$ 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 $\frac{c}{m}\mathrm{dist}(\hat{\mathbf{B}},\hat{\mathbf{B}}^*)$ 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 $\mathbf{z},\mathbf{y}\in\mathcal{N}_k$,

$$\mathbb{P}\left(\sum_{j=1}^m \langle\mathbf{z},\mathbf{u}_j\rangle\langle\mathbf{v}_j,\mathbf{y}\rangle \geq s\right) \leq \exp\left(-c'm\min\left(\frac{s^2}{\mathrm{dist}^2(\hat{\mathbf{B}},\hat{\mathbf{B}}^*)}, \frac{s}{\mathrm{dist}(\hat{\mathbf{B}},\hat{\mathbf{B}}^*)}\right)\right) \tag{46}$$

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

$$\mathbb{P}\left(\frac{1}{m}\|(\hat{\mathbf{B}}^*)^\top(\hat{\mathbf{B}}\hat{\mathbf{B}}^\top - \mathbf{I}_d)\mathbf{X}_i^\top\mathbf{X}_i\hat{\mathbf{B}}\|_2 \geq 2s\right) \leq 9^{2k}\exp\left(-c'm\min(s^2/\mathrm{dist}^2(\hat{\mathbf{B}},\hat{\mathbf{B}}^*), s/\mathrm{dist}(\hat{\mathbf{B}},\hat{\mathbf{B}}^*))\right) \tag{47}$$

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

$$\mathbb{P}\left(\frac{1}{m}\|(\hat{\mathbf{B}}^*)^\top(\hat{\mathbf{B}}\hat{\mathbf{B}}^\top - \mathbf{I}_d)\mathbf{X}_i^\top\mathbf{X}_i\hat{\mathbf{B}}\|_2 \geq 2\mathrm{dist}(\hat{\mathbf{B}},\hat{\mathbf{B}}^*)\max(\varepsilon,\varepsilon^2)\right) \leq 9^{2k}e^{-c'm\varepsilon^2} \tag{48}$$

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

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{m}\|(\hat{\mathbf{B}}^*)^\top(\hat{\mathbf{B}}\hat{\mathbf{B}}^\top - \mathbf{I}_d)\mathbf{X}_i^\top\mathbf{X}_i\hat{\mathbf{B}}\|_2 \geq \mathrm{dist}(\hat{\mathbf{B}},\hat{\mathbf{B}}^*)\sqrt{\frac{ck^2\log(rn)}{m}}\right) &\leq 9^{2k}e^{-c_1k^2\log(rn)} \\
&\leq e^{-111k^2\log(rn)}
\end{aligned} \tag{49}$$

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

$$\mathbb{P}\left(\|\mathbf{H}^i\|_2^2 \geq c\,\mathrm{dist}^2(\hat{\mathbf{B}},\hat{\mathbf{B}}^*)\frac{k^2\log(rn)}{m}\right) \leq e^{-111k^2\log(rn)} \tag{50}$$

Thus, using (45), 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*

$$\|\mathbf{F}\|_F \leq \frac{\delta_k}{1-\delta_k}\|\mathbf{W}^*\|_2\; dist(\hat{\mathbf{B}}_t,\hat{\mathbf{B}}_*) \tag{51}$$

with probability at least $1 - e^{-109k^2\log(n)}$.

Proof. By Cauchy-Schwarz, we have $\|\mathbf{F}\|_F \leq \|\mathbf{G}^{-1}\|_2\|(\mathbf{GD}-\mathbf{C})\mathbf{w}^*\|_2$. Combining the bound on $\|\mathbf{G}^{-1}\|_2$ from Lemma 3 and the bound on $\|(\mathbf{GD}-\mathbf{C})\mathbf{w}^*\|_2$ 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* $\delta_k' \leq k$,

$$\frac{1}{rn}\left\|\left(\frac{1}{m}\mathcal{A}^*\mathcal{A}(\mathbf{Q}^t)-\mathbf{Q}^t\right)^\top\mathbf{W}^{t+1}\right\|_2 \leq \delta_k'\; dist(\hat{\mathbf{B}}^t,\hat{\mathbf{B}}^*) \tag{52}$$

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

$$\begin{aligned}
\|\mathbf{q}_i\|_2^2 &= \|\hat{\mathbf{B}}\hat{\mathbf{B}}^\top \hat{\mathbf{B}}^* \hat{\mathbf{w}}_i^* - \hat{\mathbf{B}}\mathbf{f}_i - \hat{\mathbf{B}}^* \hat{\mathbf{w}}_i^*\|_2^2 \\
&= \|(\hat{\mathbf{B}}\hat{\mathbf{B}}^\top - \mathbf{I}_d)\hat{\mathbf{B}}^* \hat{\mathbf{w}}_i^* - \hat{\mathbf{B}}\mathbf{f}_i\|_2^2 \\
&\leq 2\|(\hat{\mathbf{B}}\hat{\mathbf{B}}^\top - \mathbf{I}_d)\hat{\mathbf{B}}^* \hat{\mathbf{w}}_i^*\|_2^2 + 2\|\hat{\mathbf{B}}\mathbf{f}_i\|_2^2 \\
&\leq 2\|(\hat{\mathbf{B}}\hat{\mathbf{B}}^\top - \mathbf{I}_d)\hat{\mathbf{B}}^*\|_2^2 \|\hat{\mathbf{w}}_i^*\|_2^2 + 2\|\mathbf{f}_i\|_2^2 \\
&= 2k\text{dist}^2(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*) + 2\|\mathbf{f}_i\|_2^2 \qquad (53)
\end{aligned}$$

Also recall that $\text{vec}(\mathbf{F}) = \mathbf{G}^{-1}(\mathbf{G}\mathbf{D} - \mathbf{C})\hat{\mathbf{w}}_*$ from Lemma 2. From equation (38), 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,

$$\begin{aligned}
\|\mathbf{f}_i\|_2^2 &\leq \|(\mathbf{G}^i)^{-1}\|_2^2 \, \|\mathbf{G}^i\mathbf{D}^i - \mathbf{C}^i\|_2^2 \;\; \|\mathbf{w}_i^*\|_2^2 \\
&\leq \|(\mathbf{G}^i)^{-1}\|_2^2 \, \|\mathbf{G}^i\mathbf{D}^i - \mathbf{C}^i\|_2^2 \, k \qquad (54)
\end{aligned}$$

where (54) follows by Assumption 2 From (50), we have that

$$\mathbb{P}\left(\|\mathbf{G}^i\mathbf{D}^i - \mathbf{C}^i\|_2^2 \geq \delta_k^2 \, \text{dist}^2(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*)\right) \leq e^{-101k^2 \log(rn)}$$

Similarly, from equations (41) and (42), we have that

$$\mathbb{P}\left(\|(\mathbf{G}^i)^{-1}\|_2^2 \geq \frac{1}{(1-\delta_k)^2}\right) \leq e^{-121k^3 \log(rn)} \qquad (55)$$

Now plugging this back into (54) and assuming $\delta_k \leq \frac{1}{2}$, we obtain

$$\|\mathbf{q}_i\|_2^2 \leq 2k \, \text{dist}^2(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*) \, \left(1 + \frac{\delta_k^2}{(1-\delta_k)^2}\right) \leq 4k \, \text{dist}^2(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*) \qquad (56)$$

with probability at least $1 - e^{-110k^2 \log(rn)}$. Likewise, to upper bound $\mathbf{w}_i$ we have

$$\begin{aligned}
\|\mathbf{w}_i\|_2^2 &\leq 2\|\hat{\mathbf{B}}^\top \hat{\mathbf{B}}^* \mathbf{w}_i^*\|_2^2 + 2\|\mathbf{f}_i\|_2^2 \\
&\leq 2\|\hat{\mathbf{B}}^\top \hat{\mathbf{B}}^*\|_2^2 \|\mathbf{w}_i^*\|_2^2 + 2\|\mathbf{f}_i\|_2^2 \\
&\leq 2k + 2\frac{\delta_k^2}{(1-\delta_k)^2}\text{dist}^2(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*)k \qquad (57) \\
&\leq 4k \qquad (58)
\end{aligned}$$

where (57) holds with probability at least $1 - e^{-110k^2 \log(n)}$ conditioning on the same event as in 56, and holds almost surely as long as $\delta_k \leq 1/2$. Observe that the matrix $\frac{1}{m}\mathcal{A}^*\mathcal{A}(\mathbf{Q}) - \mathbf{Q}$ can be re-written as

$$\begin{aligned}\frac{1}{m}\mathcal{A}^*\mathcal{A}(\mathbf{Q}) - \mathbf{Q} &= \frac{1}{m}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\langle \mathbf{e}_i(\mathbf{x}_i^j)^\top, \mathbf{Q}\rangle\ \mathbf{e}_i(\mathbf{x}_i^j)^\top - \mathbf{Q}\right) \\ &= \frac{1}{m}\sum_{i=1}^{rn}\sum_{j=1}^{m}\langle \mathbf{x}_i^j, \mathbf{q}_i\rangle\ \mathbf{e}_i(\mathbf{x}_i^j)^\top - \mathbf{Q} \end{aligned} \tag{59}$$

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

$$\frac{1}{rn}\left(\frac{1}{m}\mathcal{A}^*\mathcal{A}(\mathbf{Q}) - \mathbf{Q}\right)^\top \mathbf{W} = \frac{1}{rnm}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\langle \mathbf{x}_i^j,\ \mathbf{q}_i\rangle\ \mathbf{x}_i^j(\mathbf{w}_i)^\top - \mathbf{q}_i(\mathbf{w}_i)^\top\right) \tag{60}$$

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 (60).

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

$$\begin{aligned}&\left\|\frac{1}{rnm}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\langle \mathbf{x}_i^j, \mathbf{q}_i\rangle\ \mathbf{x}_i^j(\mathbf{w}_i)^\top - \mathbf{q}_i(\mathbf{w}_i)^\top\right)\right\|_2^2 \\ &\leq 2\max_{\mathbf{u}\in\mathcal{N}_d,\mathbf{v}\in\mathcal{N}_k}\mathbf{u}^\top\left(\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\frac{1}{rnm}\langle \mathbf{x}_i^j, \mathbf{q}_i\rangle\ \mathbf{x}_i^j(\mathbf{w}_i)^\top - \frac{1}{rnm}\mathbf{q}_i(\mathbf{w}_i)^\top\right)\right)\mathbf{v} \\ &= 2\max_{\mathbf{u}\in\mathcal{N}_d,\mathbf{v}\in\mathcal{N}_k}\sum_{i=1}^{rn}\sum_{j=1}^{m}\left(\frac{1}{rnm}\langle \mathbf{x}_i^j, \mathbf{q}_i\rangle\langle \mathbf{u}, \mathbf{x}_i^j\rangle\langle \mathbf{w}_i, \mathbf{v}\rangle - \frac{1}{rnm}\langle \mathbf{u}, \mathbf{q}_i\rangle\langle \mathbf{w}_i, \mathbf{v}\rangle\right)\end{aligned} \tag{61}$$

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 $\|\mathbf{q}_i\|_2 \leq 2c\sqrt{k}\ \mathrm{dist}(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*)$ with probability at least $1 - e^{-110k^2\log(rn)}$, using (56). 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

$$\frac{2c^2\sqrt{k}}{rnm}\ \mathrm{dist}(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*)\langle \mathbf{w}_i, \mathbf{v}\rangle \leq \frac{2c^2\sqrt{k}}{rnm}\ \mathrm{dist}(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*)\|\mathbf{w}_i\|_2 \leq \frac{4c'k}{rnm}\ \mathrm{dist}(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*)$$

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

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

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

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

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 $\epsilon^2 = \frac{110(d+k)}{c_1 rnm}$, then as long as $\epsilon^2 \leq 1$, we have

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

completing the proof along with a union bound over the event $\|\mathbf{F}\|_F \leq \frac{\delta_k}{1-\delta_k} k\mathrm{dist}(\hat{\mathbf{B}}, \hat{\mathbf{B}}^*)$.

□

Now we are finally ready to show the main result.

B.3 Main Result

Lemma 7. *Define* $E_0 := 1 - dist^2(\hat{\mathbf{B}}^0, \hat{\mathbf{B}}^*)$ *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*

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

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:

$$\mathbf{W}^{t+1} = \underset{\mathbf{W}\in\mathbb{R}^{rn\times k}}{\mathrm{argmin}} \frac{1}{2rnm}\|\mathcal{A}(\mathbf{W}^*\hat{\mathbf{B}}^{*\top} - \mathbf{W}\hat{\mathbf{B}}^{t\top})\|_2^2 \tag{62}$$

$$\mathbf{B}^{t+1} = \hat{\mathbf{B}}^t - \frac{\eta}{rnm}\left(\mathcal{A}^\dagger\mathcal{A}(\mathbf{W}^{t+1}\hat{\mathbf{B}}^{t\top} - \mathbf{W}^*\hat{\mathbf{B}}^{*\top})\right)^\top \mathbf{W}^{t+1} \tag{63}$$

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

$$\begin{aligned}\mathbf{B}^{t+1} &= \hat{\mathbf{B}}^t - \frac{\eta}{rnm}\left(\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^t)\right)^\top \mathbf{W}^{t+1} \\ &= \hat{\mathbf{B}}^t - \frac{\eta}{rn}\,\mathbf{Q}^{t\top}\mathbf{W}^{t+1} - \frac{\eta}{rn}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^t) - \mathbf{Q}^t\right)^\top \mathbf{W}^{t+1} \end{aligned} \tag{64}$$

Now, multiply both sides by $\hat{\mathbf{B}}_\perp^{*\top}$. We have

$$\begin{aligned}\hat{\mathbf{B}}_\perp^{*\top}\mathbf{B}_{t+1} &= \hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}^t - \frac{\eta}{rn}\hat{\mathbf{B}}_\perp^{*\top}\mathbf{Q}^{t\top}\mathbf{W}^{t+1} - \frac{\eta}{rn}\hat{\mathbf{B}}_\perp^{*\top}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^t) - \mathbf{Q}^t\right)^\top \mathbf{W}^{t+1} \\ &= \hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}^t(\mathbf{I}_k - \frac{\eta}{rn}\mathbf{W}^{t+1\top}\mathbf{W}^{t+1}) - \frac{\eta}{rn}\hat{\mathbf{B}}_\perp^{*\top}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^t) - \mathbf{Q}^t\right)^\top \mathbf{W}^{t+1} \end{aligned} \tag{65}$$

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 (65) from the right by $(\mathbf{R}^{t+1})^{-1}$ yields

$$\hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}_{t+1} = \left(\hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}^t(\mathbf{I}_k - \frac{\eta}{rn}(\mathbf{W}^{t+1})^\top\mathbf{W}^{t+1}) - \frac{\eta}{rn}\hat{\mathbf{B}}_\perp^{*\top}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^t) - \mathbf{Q}^t\right)^\top \mathbf{W}^{t+1}\right)(\mathbf{R}^{t+1})^{-1} \tag{66}$$

Hence,

$$\begin{aligned}&\operatorname{dist}(\hat{\mathbf{B}}^{t+1}, \hat{\mathbf{B}}^*) \\ &= \left\|\left(\hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}^t(\mathbf{I}_k - \frac{\eta}{rn}(\mathbf{W}^{t+1})^\top\mathbf{W}^{t+1}) - \frac{\eta}{rn}\hat{\mathbf{B}}_\perp^{*\top}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^t) - \mathbf{Q}^t\right)^\top \mathbf{W}^{t+1}\right)(\mathbf{R}^{t+1})^{-1}\right\|_2 \\ &\le \left\|\hat{\mathbf{B}}_\perp^{*\top}\hat{\mathbf{B}}^t(\mathbf{I}_k - \frac{\eta}{rn}(\mathbf{W}^{t+1})^\top\mathbf{W}^{t+1})\right\|_2 \left\|(\mathbf{R}^{t+1})^{-1}\right\|_2 \\ &\quad + \frac{\eta}{rn}\left\|\hat{\mathbf{B}}_\perp^{*\top}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^t) - \mathbf{Q}^t\right)^\top \mathbf{W}^{t+1}\right\|_2 \left\|(\mathbf{R}^{t+1})^{-1}\right\|_2 \qquad (67) \\ &=: A_1 + A_2. \qquad (68)\end{aligned}$$

where (67) 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:

$$\begin{aligned}A_2 &= \frac{\eta}{rn}\left\|\hat{\mathbf{B}}_\perp^{*\top}\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^t) - \mathbf{Q}^t\right)^\top \mathbf{W}^{t+1}\right\|_2 \left\|(\mathbf{R}^{t+1})^{-1}\right\|_2 \\ &\le \frac{\eta}{rn}\left\|\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^t) - \mathbf{Q}^t\right)^\top \mathbf{W}^{t+1}\right\|_2 \left\|(\mathbf{R}^{t+1})^{-1}\right\|_2 \qquad (69) \\ &\le \eta\delta_k' \operatorname{dist}(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*)\,\|(\mathbf{R}^{t+1})^{-1}\|_2 \qquad (70)\end{aligned}$$

where (69) follows by Cauchy-Schwarz and the fact that $\hat{\mathbf{B}}^*_\perp$ is normalized, and (70) follows with probability at least $1 - e^{-110d}$ by Lemma 6. Next we control A_1:

$$\begin{aligned} A_1 &= \left\| \hat{\mathbf{B}}^{*\top}_\perp \hat{\mathbf{B}}^t (\mathbf{I}_k - \frac{\eta}{rn}(\mathbf{W}^{t+1})^\top \mathbf{W}^{t+1}) \right\|_2 \|(\mathbf{R}^{t+1})^{-1}\|_2 \\ &\le \|\hat{\mathbf{B}}^{*\top}_\perp \hat{\mathbf{B}}^t\|_2 \left\| \mathbf{I} - \frac{\eta}{rn}(\mathbf{W}^{t+1})^\top \mathbf{W}^{t+1} \right\|_2 \|(\mathbf{R}^{t+1})^{-1}\|_2 \\ &= \mathrm{dist}(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*) \left\| \mathbf{I}_k - \frac{\eta}{rn}(\mathbf{W}^{t+1})^\top \mathbf{W}^{t+1} \right\|_2 \|(\mathbf{R}^{t+1})^{-1}\|_2 \end{aligned} \tag{71}$$

The middle factor gives us contraction, which we bound as follows. First 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

$$\|\mathbf{F}\|_2 \le \frac{\delta_k}{1-\delta_k} \|\mathbf{W}^*\|_2 \,\mathrm{dist}(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*) \tag{72}$$

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

$$\begin{aligned} \lambda_{\max}\left((\mathbf{W}^{t+1})^\top \mathbf{W}^{t+1}\right) &= \|\mathbf{W}^* \hat{\mathbf{B}}^{*\top} \hat{\mathbf{B}}^t - \mathbf{F}\|_2^2 \\ &\le 2\|\mathbf{W}^* \hat{\mathbf{B}}^{*\top} \hat{\mathbf{B}}^t\|_2^2 + 2\|\mathbf{F}\|_2^2 \\ &\le 2\|\mathbf{W}^*\|_2^2 + 2\frac{\delta_k^2}{(1-\delta_k)^2} \|\mathbf{W}^*\|_2^2 \,\mathrm{dist}^2(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*) \\ &\le 4\|\mathbf{W}^*\|_2^2 \end{aligned} \tag{73}$$

where (73) follows under the assumption that $\delta_k \le 1/2$. Thus, as long as $\eta \le 1/(4\bar{\sigma}^2_{\max,*})$, we have by Weyl's Inequality:

$$\begin{aligned} &\|\mathbf{I}_k - \frac{\eta}{rn}(\mathbf{W}^{t+1})^\top \mathbf{W}^{t+1}\|_2 \\ &\le 1 - \frac{\eta}{rn}\lambda_{\min}((\mathbf{W}^{t+1})^\top \mathbf{W}^{t+1}) && (74) \\ &= 1 - \frac{\eta}{rn}\lambda_{\min}((\mathbf{W}^* \hat{\mathbf{B}}^{*\top} \hat{\mathbf{B}}^t - \mathbf{F})^\top (\mathbf{W}^* \hat{\mathbf{B}}^{*\top} \hat{\mathbf{B}}^t - \mathbf{F})) \\ &\le 1 - \frac{\eta}{rn}\sigma^2_{\min}(\mathbf{W}^*(\hat{\mathbf{B}}^*)^\top \hat{\mathbf{B}}^t) + \frac{2\eta}{rn}\sigma_{\max}(\mathbf{F}^\top \mathbf{W}^*(\hat{\mathbf{B}}^*)^\top \hat{\mathbf{B}}^t) - \frac{\eta}{rn}\sigma^2_{\min}(\mathbf{F}) && (75) \\ &\le 1 - \frac{\eta}{rn}\sigma^2_{\min}(\mathbf{W}^*)\sigma^2_{\min}((\hat{\mathbf{B}}^*)^\top \hat{\mathbf{B}}^t) + \frac{2\eta}{rn}\|\mathbf{F}\|_2 \,\|\mathbf{W}^*(\hat{\mathbf{B}}^*)^\top \hat{\mathbf{B}}^t\|_2 && (76) \\ &\le 1 - \frac{\eta}{rn}\sigma^2_{\min}(\mathbf{W}^*)\sigma^2_{\min}((\hat{\mathbf{B}}^*)^\top \hat{\mathbf{B}}^t) + \frac{2\eta}{rn}\frac{\delta_k}{1-\delta_k}\|\mathbf{W}^*\|_2^2 && (77) \\ &= 1 - \eta\bar{\sigma}^2_{\min,*}\sigma^2_{\min}((\hat{\mathbf{B}}^*)^\top \hat{\mathbf{B}}^t) + 2\eta\frac{\delta_k}{1-\delta_k}\bar{\sigma}^2_{\max,*} && (78) \end{aligned}$$

where (75) follows by again applying Weyl's inequality, under the condition that $2\sigma_{\max}(\mathbf{F}^\top \mathbf{W}^*(\hat{\mathbf{B}}^*)^\top \hat{\mathbf{B}}^t) \le \sigma^2_{\min}(\mathbf{W}^*)\sigma^2_{\min}((\hat{\mathbf{B}}^*)^\top \hat{\mathbf{B}}^t)$, which we will enforce to be true (otherwise we would not have contraction). Also, (76) follows by the Cauchy-Schawrz inequality, and we use Lemma 5 to obtain (77). Lastly, (78) follows by the definitions of $\bar{\sigma}_{\min,*}$ and $\bar{\sigma}_{\max,*}$. In order to lower bound $\sigma^2_{\min}((\hat{\mathbf{B}}^*)^\top \hat{\mathbf{B}}^t)$, note that

$$\sigma^2_{\min}((\hat{\mathbf{B}}^*)^\top \hat{\mathbf{B}}^t) \ge 1 - \|(\hat{\mathbf{B}}^*_\perp)^\top \hat{\mathbf{B}}^t\|_2^2 = 1 - \mathrm{dist}^2(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*) \ge 1 - \mathrm{dist}^2(\hat{\mathbf{B}}^0, \hat{\mathbf{B}}^*) =: E_0 \tag{79}$$

As a result, defining $\bar{\delta}_k \coloneqq \delta_k + \delta'_k$ and combining (67), (70), (71), (78), and (79) yields

$$\begin{aligned}\operatorname{dist}(\hat{\mathbf{B}}^t, \hat{\mathbf{B}}^*) &\leq \|(\mathbf{R}^{t+1})^{-1}\|_2 \left(1 - \eta\bar{\sigma}^2_{\min,*}E_0 + 2\eta\frac{\delta_k}{1-\delta_k}\bar{\sigma}^2_{\max,*} + \eta\delta'_k\right) \operatorname{dist}(\hat{\mathbf{B}}^t, \mathbf{B}^*) \\ &\leq \|(\mathbf{R}^{t+1})^{-1}\|_2 \left(1 - \eta\bar{\sigma}^2_{\min,*}E_0 + 2\eta\frac{\bar{\delta}_k}{1-\bar{\delta}_k}\bar{\sigma}^2_{\max,*}\right) \operatorname{dist}(\hat{\mathbf{B}}^t, \mathbf{B}^*) \qquad (80)\end{aligned}$$

where (80) 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 $\mathbf{S}^t \coloneqq \frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^t)$ and observe that

$$\begin{aligned}(\mathbf{R}^{t+1})^\top\mathbf{R}^{t+1} &= (\mathbf{B}^{t+1})^\top\mathbf{B}^{t+1} \\ &= \hat{\mathbf{B}}^{t^\top}\hat{\mathbf{B}}^t - \frac{\eta}{rn}(\hat{\mathbf{B}}^{t^\top}\mathbf{S}^{t^\top}\mathbf{W}^{t+1} + (\mathbf{W}^{t+1})^\top\mathbf{S}^t\hat{\mathbf{B}}^t) + \frac{\eta^2}{(rn)^2}(\mathbf{W}^{t+1})^\top\mathbf{S}^t\mathbf{S}^{t^\top}\mathbf{W}^{t+1} \\ &= \mathbf{I}_k - \frac{\eta}{rn}(\hat{\mathbf{B}}^{t^\top}\mathbf{S}^{t^\top}\mathbf{W}^{t+1} + (\mathbf{W}^{t+1})^\top\mathbf{S}^t\hat{\mathbf{B}}^t) + \frac{\eta^2}{(rn)^2}(\mathbf{W}^{t+1})^\top\mathbf{S}^t\mathbf{S}^{t^\top}\mathbf{W}^{t+1} \qquad (81)\end{aligned}$$

thus, by Weyl's Inequality, we have

$$\begin{aligned}\sigma^2_{\min}(\mathbf{R}_{t+1}) &\geq 1 - \frac{\eta}{rn}\lambda_{\max}(\hat{\mathbf{B}}^{t^\top}\mathbf{S}^{t^\top}\mathbf{W}^{t+1} + (\mathbf{W}^{t+1})^\top\mathbf{S}^t\hat{\mathbf{B}}^t) + \frac{\eta^2}{(rn)^2}\lambda_{\min}((\mathbf{W}^{t+1})^\top\mathbf{S}^t\mathbf{S}^{t^\top}\mathbf{W}^{t+1}) \\ &\geq 1 - \frac{\eta}{rn}\lambda_{\max}(\hat{\mathbf{B}}^{t^\top}\mathbf{S}^{t^\top}\mathbf{W}^{t+1} + (\mathbf{W}^{t+1})^\top\mathbf{S}^t\hat{\mathbf{B}}^t) \qquad (82)\end{aligned}$$

where (82) 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

$$\begin{aligned}&\frac{\eta}{rn}\lambda_{\max}(\hat{\mathbf{B}}^{t^\top}\mathbf{S}^{t^\top}\mathbf{W}^{t+1} + (\mathbf{W}^{t+1})^\top\mathbf{S}^t\hat{\mathbf{B}}^t) \\ &\quad= \max_{\mathbf{x}:\|\mathbf{x}\|_2=1} \frac{\eta}{rn}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}(\mathbf{S}^t)^\top\mathbf{W}^{t+1}\mathbf{x} + \mathbf{x}^\top(\mathbf{W}^{t+1})^\top\mathbf{S}^t\hat{\mathbf{B}}^t\mathbf{x} \\ &\quad= \max_{\mathbf{x}:\|\mathbf{x}\|_2=1} \frac{2\eta}{rn}\mathbf{x}^\top(\mathbf{W}^{t+1})^\top\mathbf{S}^t\hat{\mathbf{B}}^t\mathbf{x} \\ &\quad= \max_{\mathbf{x}:\|\mathbf{x}\|_2=1} \frac{2\eta}{rn}\mathbf{x}^\top(\mathbf{W}^{t+1})^\top\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^t) - \mathbf{Q}^t\right)\hat{\mathbf{B}}^t\mathbf{x} + \frac{2\eta}{rn}\mathbf{x}^\top(\mathbf{W}^{t+1})^\top\mathbf{Q}^t\hat{\mathbf{B}}^t\mathbf{x} \qquad (83)\end{aligned}$$

We first consider the first term. We have

$$\max_{\mathbf{x}:\|\mathbf{x}\|_2=1} \frac{2\eta}{rn}\mathbf{x}^\top(\mathbf{W}^{t+1})^\top\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^t) - \mathbf{Q}^t\right)\hat{\mathbf{B}}^t\mathbf{x} \leq \frac{2\eta}{rn}\left\|(\mathbf{W}^{t+1})^\top\left(\frac{1}{m}\mathcal{A}^\dagger\mathcal{A}(\mathbf{Q}^t) - \mathbf{Q}^t\right)\right\|_2 \left\|\hat{\mathbf{B}}^t\right\|_2 \leq 2\eta\delta'_k \qquad (84)$$

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 (83). We have

$$\begin{aligned}\max_{\mathbf{x}:\|\mathbf{x}\|_2=1} \frac{2\eta}{rn}\mathbf{x}^\top(\mathbf{W}^{t+1})^\top\mathbf{Q}^t\hat{\mathbf{B}}^t\mathbf{x} &= \max_{\mathbf{x}:\|\mathbf{x}\|_2=1} \frac{2\eta}{rn}\left\langle\mathbf{Q}^t, \mathbf{W}^{t+1}\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}\right\rangle \\ &= \max_{\mathbf{x}:\|\mathbf{x}\|_2=1} \frac{2\eta}{rn}\langle\mathbf{Q}^t, \mathbf{W}^*\hat{\mathbf{B}}^{*^\top}\hat{\mathbf{B}}^t\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}\rangle - \frac{2\eta}{rn}\langle\mathbf{Q}^t, \mathbf{F}\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}\rangle \qquad (85)\end{aligned}$$

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}\mathrm{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}\mathrm{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}\mathrm{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}\mathrm{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}\mathrm{tr}(\hat{\mathbf{B}}_\perp^t\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}\mathrm{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}\mathrm{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}}_\perp^t) - \frac{2\eta}{rn}\mathrm{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}\mathrm{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) && (86) \\
&= -\frac{2\eta}{rn}\mathrm{tr}(\mathbf{F}^\top\mathbf{W}^*\hat{\mathbf{B}}^{*^\top}\hat{\mathbf{B}}^t\mathbf{x}\mathbf{x}^\top) && (87) \\
&\le \frac{2\eta}{rn}\|\mathbf{F}\|_F\,\|\mathbf{W}^*\hat{\mathbf{B}}^{*^\top}\hat{\mathbf{B}}^t\mathbf{x}\mathbf{x}^\top\|_F && (88) \\
&\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 && (89) \\
&\le \frac{2\eta}{rn}\|\mathbf{F}\|_F\,\|\mathbf{W}^*\|_2 && (90) \\
&\le 2\eta\frac{\delta_k}{1-\delta_k}\bar{\sigma}^2_{\max,*} && (91)
\end{aligned}
$$

where (86) follows since $\hat{\mathbf{B}}^{t^\top}\hat{\mathbf{B}}_\perp^t = \mathbf{0}$, (87) follows since $\hat{\mathbf{B}}^{t^\top}\hat{\mathbf{B}}_\perp^t = \mathbf{I}_k$, (88) and (89) follow by the Cauchy-Schwarz inequality, (90) follows by the orthonormality of $\hat{\mathbf{B}}^t$ and $\hat{\mathbf{B}}^*$ and (91) follows by Lemma 5 and the definition of $\bar{\sigma}_{\max,*}$. Next, again for any $\mathbf{x} \in \mathbb{R}^k : \|\mathbf{x}\|_2 = 1$,

$$
\begin{aligned}
-\frac{2\eta}{rn}\langle \mathbf{Q}^t, \mathbf{F}\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}\rangle &= -\frac{2\eta}{rn}\mathrm{tr}((\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t^\top}\hat{\mathbf{B}}^*\mathbf{W}^{*^\top} - \hat{\mathbf{B}}^t\mathbf{F}^\top - \hat{\mathbf{B}}^*\mathbf{W}^{*^\top})\mathbf{F}\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}) \\
&= -\frac{2\eta}{rn}\mathrm{tr}((\hat{\mathbf{B}}^t\hat{\mathbf{B}}^{t^\top} - \mathbf{I}_d)\hat{\mathbf{B}}^*\mathbf{W}^{*^\top}\mathbf{F}\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}) + \frac{2\eta}{rn}\mathrm{tr}(\mathbf{F}\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}\hat{\mathbf{B}}^t\mathbf{F}^\top) \\
&= -\frac{2\eta}{rn}\mathrm{tr}(\hat{\mathbf{B}}^*\mathbf{W}^{*^\top}\mathbf{F}\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}\mathbf{B}_\perp^t) + \frac{2\eta}{rn}\mathbf{x}^\top\mathbf{F}^\top\mathbf{F}\mathbf{x} \\
&= \frac{2\eta}{rn}\mathbf{x}^\top\mathbf{F}^\top\mathbf{F}\mathbf{x} \\
&\le \frac{2\eta}{rn}\|\mathbf{F}\|_2^2 \\
&\le 2\eta\frac{\delta_k^2}{(1-\delta_k)^2}\bar{\sigma}^2_{\max,*} && (92)
\end{aligned}
$$

Thus, we have the following bound on the second term of (83):

$$
\max_{\mathbf{x}:\|\mathbf{x}\|_2=1}\frac{2\eta}{rn}\langle \mathbf{Q}^t, \mathbf{W}^{t+1}\mathbf{x}\mathbf{x}^\top\hat{\mathbf{B}}^{t^\top}\rangle \le 2\eta\bar{\sigma}^2_{\max,*}\left(\frac{\delta_k}{1-\delta_k} + \frac{\delta_k^2}{(1-\delta_k)^2}\right) \le 4\eta\frac{\delta_k}{(1-\delta_k)^2}\bar{\sigma}^2_{\max,*} \tag{93}
$$

since $\delta_k \le 1 \implies \delta_k^2 \le \delta_k$. Therefore, using (82), (83), (84) and (93), we have

$$\sigma_{\min}^2(\mathbf{R}_{t+1}) \ge 1 - 2\eta\delta_k' - 4\eta\frac{\delta_k}{(1-\delta_k)^2}\bar{\sigma}_{\max,*}^2 \ge 1 - 4\eta\frac{\bar{\delta}_k}{(1-\bar{\delta}_k)^2}\bar{\sigma}_{\max,*}^2 \tag{94}$$

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

$$\|(\mathbf{R}^{t+1})^{-1}\|_2 \le \left(1 - 4\eta\frac{\bar{\delta}_k}{(1-\bar{\delta}_k)^2}\bar{\sigma}_{\max,*}^2\right)^{-1/2} \tag{95}$$

Note that $1 - 4\eta\frac{\bar{\delta}_k}{(1-\bar{\delta}_k)^2}\bar{\sigma}_{\max,*}^2$ 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 $\eta \le 1/(4\bar{\sigma}_{\max,*}^2)$. Therefore, from (80), we have

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

Next, let $\bar{\delta}_k < 16E_0/(25\cdot 5\kappa^2)$. This implies that $\bar{\delta}_k < 1/5$. Then $\bar{\delta}_k/(1-\bar{\delta}_k)^2 < 25\bar{\delta}_k/16 \le E_0/(5\kappa^2) \le 1$, validating (95). Further, it is easily seen that

$$\begin{aligned} 1 - \eta E_0\bar{\sigma}_{\min,*}^2 + \eta\frac{\bar{\delta}_k}{(1-\bar{\delta}_k)^2}\bar{\sigma}_{\max,*}^2 &\le 1 - 4\eta\frac{\bar{\delta}_k}{(1-\bar{\delta}_k)^2}\bar{\sigma}_{\max}^2 \\ &\le 1 - \eta E_0\bar{\sigma}_{\min,*}^2/2 \end{aligned} \tag{96}$$

Thus

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

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 $m \ge c'(\kappa^4 k^3\log(rn)/E_0^2 + \kappa^4 k^2 d/(E_0^2 rn))$ for another absolute constant c' satisfies $\bar{\delta}_k \le 16E_0/(25\cdot 5\kappa^2)$. Also, we have conditioned on two events, described in Lemmas 5 and 6, which occur with probability at least $1 - e^{-110d} - e^{-110k^2\log(rn)} \ge 1 - e^{-100\min(k^2\log(rn),d)}$, completing the proof.

□

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

Theorem 1. *Define* $E_0 := 1 - dist^2(\mathbf{B}^0, \mathbf{B}^*)$ *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* $m \ge c(\kappa^4 k^3\log(rn)/E_0^2 + \kappa^4 k^2 d/(E_0^2 rn))$ *for some absolute constant* c*. Then for any* t *and any* $\eta \le 1/(4\bar{\sigma}_{\max,*}^2)$*, we have*

$$dist(\mathbf{B}^T, \mathbf{B}^*) \le \left(1 - \eta E_0\bar{\sigma}_{\min,*}^2/2\right)^{T/2} dist(\mathbf{B}^0, \mathbf{B}^*),$$

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

C Further Discussion of Related Works

The majority of methods for personalized federated learning do not involve dimensionality-reduction at the local level. For instance, Jiang et al. [2019] and Fallah et al. [2020] obtain full-dimensional models advantageous for local fine-tuning via SGD by applying model-agnostic meta-learning (MAML) methods to federated learning. On the other hand, many works propose to learn multiple full-dimensional models, typically a single global model and unique local models for each client. Hanzely and Richtárik [2020], T Dinh et al. [2020], and Li et al. [2020] add a regularization term to the standard federated learning objective which penalizes local models from being too far from a global model. Huang et al. [2020] take a similar approach but use a regularizer with an attention-inducing function to encode relationships between the local models, similar to [Smith et al., 2017]. Meanwhile, Deng et al. [2020] and Mansour et al. [2020] also use full-dimensional local and global models, and propose to linearly combine them to obtain the model for each client, a framework for which they are able to provide generalization bounds on new client data. Mansour et al. [2020] additionally analyze two other methods for personalization, but these methods require combining client data and hence are not applicable to federated learning.

References

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

Fei Chen, Mi Luo, Zhenhua Dong, Zhenguo Li, and Xiuqiang He. Federated meta-learning with fast convergence and efficient communication. *arXiv preprint arXiv:1802.07876*, 2018.

Yuejie Chi, Yue M Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926. IEEE, 2017.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

Simon S. Du, Wei Hu, Sham M. Kakade, Jason D. Lee, and Qi Lei. Few-shot learning via learning the representation, provably, 2020.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning: A meta-learning approach, 2020.

Farzin Haddadpour, Mohammad Mahdi Kamani, Aryan Mokhtari, and Mehrdad Mahdavi. Federated learning with compression: Unified analysis and sharp guarantees. *arXiv preprint arXiv:2007.01154*, 2020.

Filip Hanzely and Peter Richtárik. Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*, 2020.

Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized federated learning: An attentive collaboration approach. *arXiv preprint arXiv:2007.03797*, 2020.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. *Proceedings of the 45th annual ACM symposium on Symposium on theory of computing - STOC '13*, 2013.

Yihan Jiang, Jakub Konečnỳ, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*, 2019.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

Mikhail Khodak, Maria-Florina F Balcan, and Ameet S Talwalkar. Adaptive gradient-based meta-learning methods. In *Advances in Neural Information Processing Systems*, pages 5915–5926, 2019.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Feddane: A federated newton-type method. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pages 1227–1231. IEEE, 2019.

Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. Federated multi-task learning for competing constraints. *arXiv preprint arXiv:2012.04221*, 2020.

Paul Pu Liang, Terrance Liu, Liu Ziyin, Ruslan Salakhutdinov, and Louis-Philippe Morency. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020.

Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

Dohyung Park, Anastasios Kyrillidis, Constantine Caramanis, and Sujay Sanghavi. Finding low-rank solutions via nonconvex matrix factorization, efficiently and provably. *SIAM Journal on Imaging Sciences*, 11(4):2165–2204, 2018.

Reese Pathak and Martin J Wainwright. Fedsplit: An algorithmic framework for fast federated optimization. *arXiv preprint arXiv:2005.05238*, 2020.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. In *Advances in neural information processing systems*, pages 4424–4434, 2017.

Canh T Dinh, Nguyen Tran, and Tuan Dung Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33, 2020.

Nilesh Tripuraneni, Chi Jin, and Michael I. Jordan. Provable meta-learning of linear representations, 2020.

Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pages 964–973. PMLR, 2016.

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

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *arXiv preprint arXiv:2007.07481*, 2020.

Kangkang Wang, Rajiv Mathews, Chloé Kiddon, Hubert Eichner, Françoise Beaufays, and Daniel Ramage. Federated evaluation of on-device personalization. *arXiv preprint arXiv:1910.10252*, 2019.

Tao Yu, Eugene Bagdasaryan, and Vitaly Shmatikov. Salvaging federated learning by local adaptation. *arXiv preprint arXiv:2002.04758*, 2020.

Qinqing Zheng and John Lafferty. Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

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