SHARE YOUR REPRESENTATION ONLY: GUARANTEED IMPROVEMENT OF THE PRIVACY-UTILITY TRADEOFF IN FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Repeated parameter sharing in federated learning causes significant information leakage about private data, thus defeating its main purpose: data privacy. Mitigating the risk of this information leakage, using state of the art differentially private algorithms, also does not come for free. Randomized mechanisms can prevent convergence of models on learning even the useful representation functions, especially if there is more disagreement between local models on the classification functions (due to data heterogeneity). In this paper, we consider a representation federated learning objective that encourages various parties to collaboratively refine the consensus part of the model, with differential privacy guarantees, while separately allowing sufficient freedom for local personalization (without releasing it). We prove that in the linear representation setting, while the objective is non-convex, `CENTAUR` converges to a ball centered around the *global optimal* solution at a linear rate, and the radius of the ball is proportional to the reciprocal of the privacy budget. With this novel utility analysis, we improve the SOTA utility-privacy trade-off for this problem by a factor of $\sqrt{d}$, where d is the input dimension. We empirically evaluate our method with the image classification task on CIFAR10, CIFAR100, and EMNIST, and observe a significant performance improvement over the prior work under the same small privacy budget.

1 INTRODUCTION

In federated learning (FL), multiple parties cooperate to learn a model under the orchestration of a central server while keeping the data local. However, this paradigm alone is insufficient to provide rigorous privacy guarantees, even when local parties only share partial information (e.g. gradients) about their data. An adversary (e.g. one of the parties) can infer whether a particular record is in the training data set of other parties (Nasr et al., 2019), or even precisely reconstruct their training data (Zhu et al., 2019). To formally mitigate these privacy risks, we need to guarantee that any information *shared between the parties* during the training phase has *bounded information leakage* about the local data. This can be achieved using FL under differential privacy (DP) guarantees.

FL and DP are relatively well-studied separately. However, their challenges multiply when conducting FL under a DP constraint, in real-world settings where the data distributions can vary substantially across the clients (Li et al., 2020b; Acar et al., 2021; Shen et al., 2022). A direct consequence of such data heterogeneity is that the optimal local models might vary significantly across clients and differ drastically from the global solution. This results in large local gradients (Jiang et al., 2019). However, these large signals leak information about the local training data, and cannot be communicated as such when we need to guarantee DP. We require clipping gradient values (usually by a small threshold (De et al., 2022)) before sending them to the server, to bound the sensitivity of the gradient function with respect to changes in training data Abadi et al. (2016). As the local per-sample gradients (due to data heterogeneity) tend to be large even at the global optimum, clipping per-example gradient by a small threshold and then randomizing it, will result in a high error in the overall gradient computation, and thus degrading the accuracy of the model learned via FL.

Contributions. In order to achieve high-accuracy FL under a tight privacy budget, we propose a novel framework that is based on learning information that is agreed on among most parties in a differentially private manner, while also allowing each client to personalize its learning on information that others disagree with. It is shown that in many FL classification scenarios, participants have minimal disagreement on data representations (Bengio et al., 2013; Chen et al., 2020; Collins et al., 2021), but possibly have very different classifier heads (e.g., the last layer of the neural network). Therefore, instead of solving the standard classification problem, we view the neural network model as a composition of a representation extractor and a small classifier head, and optimize these two components in different manners. In our scheme, we train a single differentially private global representation extractor while allowing each participant to have a different *personalized* classifier head. Due to low heterogeneity in data representation (compared to the whole model), the DP learned representation in our new scheme outperforms prior schemes that perform DP optimization over the entire model.

In the setting where both the representation function and the classifier heads are linear w.r.t. their parameters, we prove a novel utility-privacy trade-off for an instance of `CENTAUR`, yielding a significant $O(\sqrt{d})$ improvement over previous art, where d is the input dimension (Corollary 5.1). A major algorithmic novelty of our proposed approach is a cross-validation scheme for boosting the success probability of the classic noisy power method for privacy-preserving spectral analysis.

We present strong empirical evidence for the superior performance of `CENTAUR` over the prior work, under the small DP budget of $(1, 10^{-5})$ in a variety of data-heterogeneity settings on benchmark datasets CIFAR10, CIFAR100, and EMNIST. Our method outperforms the prior work in all settings. Moreover, we showcase that `CENTAUR` *uniformly* enjoys a better utility-privacy trade-off over its competitors on the EMNIST dataset across different privacy budget ϵ (Figure 1). Importantly, `CENTAUR` ($\epsilon > 0.5$) is **the only method** that could **outperform the local stand-alone training baseline**, and therefore justifies the choice of collaborative learning as opposed to stand-alone training.

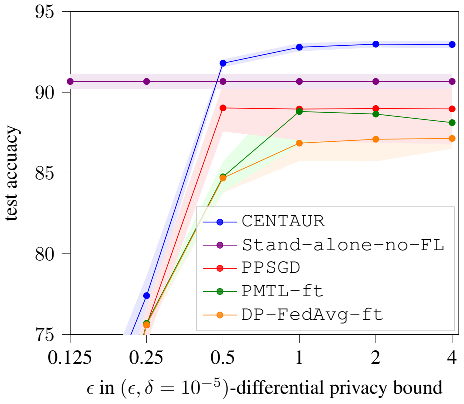

Figure 1: Privacy utility trade-off for models trained under `CENTAUR` and other algorithms on EMNIST dataset (2000 clients, 5 shards per user). Error bar denotes standard deviation across 3 runs of the algorithms.

1.1 RELATED WORK

Federated learning with differential privacy (DP) guarantee has been extensively studied since its emergence (Shokri & Shmatikov, 2015; McMahan et al., 2017a). Without assuming any trusted central party, the model of local DP requires each client to randomize its messages before sending them to any other (malicious) parties. Consequently, the trade-off between local DP and accuracy is significantly worse than that for centralized setting and requires huge amount of data for learning even simple statistics (Duchi et al., 2014; Erlingsson et al., 2014; Ding et al., 2017). By using secure aggregation protocol as an equivalent for trusted central party, recent works (McMahan et al., 2017b; Agarwal et al., 2018; Levy et al., 2021; Kairouz et al., 2021) study user-level DP under Billboard model to enable utility. We also focus on such user-level DP setting.

Model personalization approaches (Smith et al., 2017; Fallah et al., 2020; Li et al., 2020b; Arivazhagan et al., 2019; Collins et al., 2021; Pillutla et al., 2022) enable each client to learn a different (while related) model, thus alleviating the model drifting issue due to data heterogeneity. Recent works further investigate whether model personalization approaches enable improved privacy accuracy trade-off for federated learning. Hu et al. (2021) propose a private federated multi-task learning algorithm by adding task-specific regularization to each client's optimization objective. Hence, the algorithm is similar to the regularized `FedAvg-ft` algorithm, and has limited ability to deal with data heterogeneity. Bietti et al. (2022) propose the `PPSGD` algorithm that enables training additive personalized models with user-level DP. However, their generalization guarantees crucially rely on the convexity of loss functions. On the contrary, we study the convergence of `CENTAUR` algorithm under more general non-convex objectives.

The closest work to our approach in the literature is Jain et al. (2021), who also propose differentially privately learning shared low-dimensional linear representation with individualized classification head. However, their algorithm relies on solving the least square problem exactly at server side,

by performing SVD on perturbed noisy representation matrix. Hence, the generalization guarantee of their algorithm intrinsically has an expensive $d^{1.5}$ dependency on the data dimension d. By contrast, we perform noisy gradient descent at server side and improve upon this error dependency by a factor of $\sqrt{d}$. Their algorithm is also limited to the linear representation learning problem, unlike our `CENTAUR` algorithm which enables training multiple layers of shared representations.
Our work builds on the `FedRep` algorithm (Collins et al., 2021), which also relies on learning shared representations between clients but does not consider privacy. In contrast, our work provides a novel private and federated representation learning framework. Moreover, a major different ingredient of our algorithm is the initialization procedure, which requires performing *differentially private* SVD to the data matrix. We use the noisy power method (Hardt & Price, 2014) as a crucial tool to enable a constant probability for utility guarantee. We then perform cross-validation to further boost success probability to arbitrarily large (inspired by Liang et al. (2014)).

2 NOTATIONS AND BACKGROUND ON PRIVACY

Notations. We adhere to the notations suggested by the ICLR template, with the following additional ones. We denote the clipping operation by $\mathrm{clip}(\boldsymbol{x};\zeta) \doteq \boldsymbol{x} \cdot \min\{1, \zeta/\|\boldsymbol{x}\|\}$. We denote the Gaussian mechanism as $\mathrm{GM}_{\zeta,\sigma}(\{\boldsymbol{x}_i\}_{i=1}^s) \doteq \frac{1}{s}(\sum_{i=1}^s \mathrm{clip}(\boldsymbol{x}_i;\zeta) + \sigma\zeta W)$ where $W \sim \mathcal{N}(0, \mathbb{I})$.

Define Rényi Differential Privacy (RDP) on a dataset space $\mathcal{D}$ equipped with a distance d as follows.
Definition 2.1 (Rényi Differential Privacy (RDP))**.** *For a pair of measures ν and ν' over the same space with $\nu \ll \nu'$, the Rényi divergence $R_\alpha(\nu, \nu')$ is defined as*

$$R_\alpha(\nu,\nu') = \frac{1}{\alpha-1}\log E_\alpha(\nu,\nu'), \text{ where } E_\alpha(\nu,\nu') = \int \left(\frac{\mathrm{d}\nu}{\mathrm{d}\nu'}\right)^\alpha \mathrm{d}\nu'.$$

Let Θ be an output space. A randomized algorithm $\mathcal{M} : \mathcal{D} \to \Theta$ satisfies (α, ϵ)-RDP, if for all $\mathbb{D}, \mathbb{D}' \in \mathcal{D}$ with $\mathrm{d}(\mathbb{D}, \mathbb{D}') \le 1$, we have $R_\alpha(\mathcal{M}(\mathbb{D}), \mathcal{M}(\mathbb{D}')) \le \epsilon$.

User-level-RDP. Let $\mathcal{D}$ be the space of all of n tuples of local datasets $\{\mathbb{S}_i\}_{i=1}^n$, where each local dataset consists of m data points, i.e. $\mathcal{D} = \{\{\mathbb{S}_i\}_{i=1}^n \mid \mathbb{S}_i = \{\boldsymbol{z}_{ij}\}_{j=1}^m\}$. The distance d is the Hamming distance in the dataset level, i.e. $\mathrm{d}(\{\mathbb{S}_i\}_{i=1}^n, \{\mathbb{S}_i'\}_{i=1}^n) = \sum_{i=1}^n \mathbf{1}_{\mathbb{S}_i \neq \mathbb{S}_i'}$. We refer to the privacy guarantee recovered by this choice of dataset space as user-level-RDP.

Our privacy analysis relies on the Gaussian Mechanism and the Composition of RDP which we review in Appendix A.3. To facilitate the comparison of our result with previous works, we define the standard notion of Differential Privacy (DP) and the billboard model. Note that one can convert RDP to DP via the conversion lemma (see Appendix A.3).
Definition 2.2 (Differential Privacy (DP))**.** *Let Θ be an abstract output space. A randomized algorithm $\mathcal{M} : \mathcal{D} \to \Theta$ is (ϵ, δ)-differential private if for all $\mathbb{D}, \mathbb{D}' \in \mathcal{D}$ with $\mathrm{d}(\mathbb{D}, \mathbb{D}') \le 1$, we have that for all subset of the range, $S \subseteq \Theta$, the algorithm $\mathcal{M}$ satisfies:*

$$\Pr\{\mathcal{M}(\mathbb{D}) \in S\} \le \exp(\epsilon)\Pr\{\mathcal{M}(\mathbb{D}') \in S\} + \delta.$$

The Billboard model is a communication protocol that is particularly compatible with algorithms in the federated setting and has been adopted in many previous works (Jain et al., 2021; Hu et al., 2021; Bietti et al., 2022). In this model, a trusted server (either a trusted party or one that uses cryptographic techniques like multiparty computation) aggregates information subject to a DP constraint, which is then shared as public messages with all the n users. Then, each user computes its own personalized model based on the public messages and its own private data (Hsu et al., 2014).
Threat Models. In this work, we aim to ensure the privacy protection against potential adversarial clients so that any eavesdropper will not be able to tell whether one of the users has participated in the collaborative learning procedure, given the information released during the training phase. Consequently, we will establish user-level RDP guarantees under the billboard model, which is a setting commonly considered in the literature (Jain et al., 2021; Bietti et al., 2022; Hu et al., 2021).

3 PROBLEM FORMULATION

Consider a Federated Learning (FL) setting where there are n agents interacting with a central server and the ith agent holds m_i local data points $\mathbb{S}_i = \{(\boldsymbol{x}_{ij}, y_{ij})\}_{j=1}^{m_i}$. Here $(\boldsymbol{x}_{ij}, y_{ij}) \in \mathbb{R}^d \times \mathbb{R}$ denotes a pair of an input vector and the corresponding label. Suppose that each local dataset $\mathbb{S}_i$ is sampled from an underlying joint distribution $p_i(\boldsymbol{x}, y)$ of the input and response pair. We consider the *data-heterogenous* setting where potentially $p_i \neq p_j$ for $i \neq j$. We further consider a hypothesis model $f(\cdot; \boldsymbol{\theta}) : \mathbb{R}^d \to \mathbb{R}$ which maps an input $\boldsymbol{x}$ to the predicted label y. Here $\boldsymbol{\theta} \in \mathbb{R}^q$ is the trainable parameter of the model f. Let $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a loss function that penalizes the error of the prediction $f(\boldsymbol{x}; \boldsymbol{\theta})$ given the true label y. The local objective on client i is defined as

$$l(\boldsymbol{\theta}; \mathbb{S}_i) \doteq \frac{1}{m_i} \sum_{j=1}^{m_i} \ell(f(\boldsymbol{x}_{ij}; \boldsymbol{\theta}), y_{ij}). \tag{1}$$

The standard FL formulation seeks a global consensus solution, whose objective is defined as

$$\arg\min_{\theta} \mathcal{L}(\boldsymbol{\theta}) \doteq \frac{1}{n} \sum_{i=1}^{n} l(\boldsymbol{\theta}; \mathbb{S}_i). \tag{2}$$

While this formulation is reasonable when the local data distributions are similar, the obtained global solution may be far from the local optima $\arg\min l(\boldsymbol{\theta}; \mathbb{S}_i)$ under diverse local data distributions, a phenomenon known as *statistical heterogeneity* in the FL literature (Li et al., 2020a; Wang et al., 2019; Malinovskiy et al., 2020; Mitra et al., 2021; Charles & Konečnỳ, 2021; Acar et al., 2020; Karimireddy et al., 2020). Such a (potentially significant) mismatch between local and global optimal solutions limits the incentive for collaboration, and cause extra difficulties when DP constraints are imposed (Remark 3.1). These considerations motivate us to search for personalized solutions that can be learned in a federated fashion, with less utility loss due to the DP constraint.

Federated representation learning with differential privacy. It is now well-documented that in some common and real-world FL tasks, such as image classification and word prediction, clients have minimal disagreement on data representations (Chen et al., 2020; Collins et al., 2021). Based on this observation, a more reasonable alternative to the FL objective in equation 2 should focus on learning the data representation, which is the information that is agreed on among most parties, while also allowing each client to personalize its learning on information that the other clients disagree on. To formalize this, suppose that the variable $\boldsymbol{\theta} \in \mathbb{R}^q$ can be partitioned into a pair $[\boldsymbol{w}, \boldsymbol{b}] \in \mathbb{R}^{q_1} \times \mathbb{R}^{q_2}$ with $q = q_1 + q_2$ and the parameterized model admits the composition $f = h \circ \phi$ where $\phi : \mathbb{R}^d \to \mathbb{R}^k$ is the *representation extractor* that maps d-dimensional data points to a lower dimensional space of size k and $h : \mathbb{R}^k \to \mathbb{R}$ is a *classifier head* that maps from the lower dimensional subspace to the space of labels. An important example is bottom and top layers of the neural network model. We use $\boldsymbol{w}$ and $\boldsymbol{b}$ to denote the parameters that determine h and ϕ, respectively. With the above notation, we consider the following federated representation learning (FRL) problem

$$\min_{\boldsymbol{b} \in \mathbb{B}} \frac{1}{n} \sum_{i=1}^{n} \min_{\boldsymbol{w}_i} \left\{ l([\boldsymbol{w}_i, \boldsymbol{b}]; \mathbb{S}_i) := \frac{1}{m_i} \sum_{j=1}^{m_i} \ell(h(\phi(\boldsymbol{x}_{ij}; \boldsymbol{b}); \boldsymbol{w}_i), y_{ij}) \right\}, \tag{3}$$

where we maintain a single global representation extraction function $\phi(\cdot; \boldsymbol{b})$ subject to the constraint $\boldsymbol{b} \in \mathbb{B} \subseteq \mathbb{R}^{q_2}$ while allowing each client to use its personalized classification head $h(\cdot; \boldsymbol{w}_i)$ locally. The constraint $\mathbb{B}$ is included so that equation 3 also covers the linear case studied in section 5.

The choice of the FRL formulation in equation 3 entails considerations from both DP and optimization perspectives: From the DP standpoint, the phenomenon of statistical heterogeneity introduces additional difficulties for federated learning under DP constraint (see Remark 3.1 below). If the clients collaborate to train only a shared representation function, then the aforementioned disadvantages can be alleviated; From the optimization standpoint, we typically have $k \ll d$, i.e. the dimension of the extracted features is much smaller than that of the original input. Hence, for a fixed representation function $\phi(\cdot; \boldsymbol{b})$, the client specific heads $h(\cdot; \boldsymbol{w}_i)$ are in general easy to optimize locally as the number of parameters, k, is typically small.

Remark 3.1 (Statistical heterogeneity makes DP guarantee harder to establish.)**.** *To establish DP guarantees for gradient based methods, e.g. DP-SGD, a common choice is the Gaussian mechanism, which is comprised of the gradient clipping step and the noise injection step. It is empirically observed that to achieve a better privacy-utility trade-off, a small clipping threshold is preferred, since it limits the large variance due to the injected noise (De et al., 2022). Moreover, the effect of the bias (due to clipping) subsides as the per-sample gradient norm diminishes during the centralized*

Algorithm 1 SERVER procedure of CENTAUR

```
procedure SERVER($\boldsymbol{b}^0$, $p_g$, $T_g$, $\eta_g$, $\sigma_g$, $\zeta_g$)
    // $\boldsymbol{b}^0$ is obtained from the INITIALIZATION procedure.
    for $t \leftarrow 0$ to $T_g - 1$ do
        Sample set $\mathbb{C}^t$ of active clients using Poisson sampling with parameter $p_c$.
        Broadcast the current global representation parameter $\boldsymbol{b}^t$ to active clients.
        Receive the local update directions $\{\boldsymbol{g}_i^t\}_{i\in\mathbb{C}^t}$ from the CLIENT procedures.
        Compute the update direction $\boldsymbol{g}^t = \mathrm{GM}_{\zeta_g,\sigma_g}(\{\boldsymbol{g}_i^t\}_{i\in\mathbb{C}^t})$
        Update the global representation function $\boldsymbol{b}^{t+1} := \mathrm{AGGREGATION}(\boldsymbol{b}^t, \boldsymbol{g}^t, \eta_g)$.
        // The AGGREGATION procedure depends on the feasible set $\mathbb{B}$ in equation 3.
```

Algorithm 2 CLIENT procedure of CENTAUR in the general case (for client i)

```
procedure CLIENT($\boldsymbol{b}^t$, $\bar{m}$, $T_l$, $\eta_l$)
    [Phase 1: Local classifier update.] $\boldsymbol{w}_i^{t+1} = \arg\min_{\boldsymbol{w}} l([\boldsymbol{b}^t, \boldsymbol{w}]; \mathbb{S}_i)$.
    [Phase 2: Local representation function update.] Set $\boldsymbol{b}_i^{t,0} = \boldsymbol{b}^t$;
    for $s \leftarrow 0$ to $T_l - 1$ do
        Sample a subset $\mathbb{S}_i^s$ of size $\bar{m}$ from the local dataset $\mathbb{S}_i$ without replacement
        Update the local representation function $\boldsymbol{b}_i^{t,s+1} := \boldsymbol{b}_i^{t,s} - \eta_l \cdot \partial_{\boldsymbol{b}} l([\boldsymbol{b}_i^{t,s}, \boldsymbol{w}_i^{t+1}]; \mathbb{S}_i^s)$.
    [Phase 3: Summarize the local update direction.] return $\boldsymbol{g}_i^t := \boldsymbol{b}_i^{t,T_l} - \boldsymbol{b}^t$.
```

training, a phenomenon known as benign overfitting in deep learning (Bartlett et al., 2020; Li et al., 2021; Bartlett et al., 2021). However, due to the phenomenon of distribution shift, the local (per-sample) gradients in the standard FL setting (described in equation 2) remain large even at the global optimal solution, and hence setting a small (per-sample) gradient clipping threshold will result in a large and non-diminishing bias in the overall gradient computation.

In contrast, for tasks where the representation extracting functions are approximately homogeneous, the local and global optimal of the FRL formulation 3 are close and hence the gradients w.r.t. the representation function vanishes at the optimal, which amiable to small clipping threshold.

4 DIFFERENTIAL PRIVATE FEDERATED REPRESENTATION LEARNING

In this section we present the proposed CENTAUR method to solve the FRL problem in equation 3.

SERVER procedure (Algorithm 1) takes the following quantities as inputs: $\boldsymbol{b}^0$ denotes the initializer for the parameter of the global representation function, obtained from a procedure INITIALIZATION; p_g denotes the portion of the clients that will participate in training per global communication round; T_g denotes the total number of global communication rounds; η_g denotes the global update step size; (σ_g, ζ_g) stand for the noise multiplier and the clipping threshold of the Gaussian mechanism (that ensures user-level RDP). Note that in this section, we consider random INITIALIZATION over unconstrained space $\mathbb{B} = \mathbb{R}^{q_2}$, and the procedure $\mathrm{AGGREGATION}(\boldsymbol{b}^t, \boldsymbol{g}^t, \eta_g) = \boldsymbol{b}^t + \eta_g \cdot \boldsymbol{g}^t$. Under these configurations, SERVER follows the standard FL protocol: After broadcasting the current global representation function to the activate clients, it aggregates the information returned from the CLIENT procedure to update the global representation function.

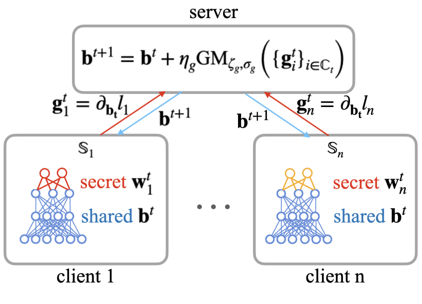
server
$\mathbf{b}^{t+1} = \mathbf{b}^t + \eta_g \mathrm{GM}_{\zeta_g,\sigma_g}\left(\{\mathbf{g}_i^t\}_{i\in\mathbb{C}_t}\right)$
$\mathbf{g}_1^t = \partial_{\mathbf{b}_t} l_1$
$\mathbf{b}^{t+1}$
$\mathbf{b}^{t+1}$
$\mathbf{g}_n^t = \partial_{\mathbf{b}_t} l_n$
$\mathbb{S}_1$
secret $\mathbf{w}_1^t$
shared $\mathbf{b}^t$
. . .
$\mathbb{S}_n$
secret $\mathbf{w}_n^t$
shared $\mathbf{b}^t$
client 1
client n

Figure 2: The t-th global round of the CENTAUR algorithm, where clients keep their classification head $\boldsymbol{w}_i^t$ secret while updating shared representation $\boldsymbol{b}^t \to \boldsymbol{b}^{t+1}$ based on perturbed gradients $\boldsymbol{g}_i^t$ from sampled clients $i \in \mathbb{C}_t$.

CLIENT procedure (Algorithm 2) takes the following quantities as inputs: $\boldsymbol{b}^t$ denotes the parameter of the global representation function received from the server; $\bar{m}$ denotes the number of local data points used as mini-batch to update the local representation function; T_l denotes the number of local update iterations; η_l denotes the local update step size. CLIENT can be divided into three phases: 1. After receiving the current global parameter $\boldsymbol{b}^t$ of the representation function from the server, the client update the local classifier head to the minimizer of the local objective $\boldsymbol{w}_i^{t+1} = \arg\min_{\boldsymbol{w}} l([\boldsymbol{b}^t, \boldsymbol{w}]; \mathbb{S}_i)$. This is possible since the local objective l usually admits very simple structure, e.g. it is convex w.r.t. $\boldsymbol{w}$, once the representation function is fixed. In practice, we would run multiple SGD epochs on $\boldsymbol{w}$ to approximately minimize l with $\boldsymbol{b}^t$ fixed. This is computationally cheap since the dimension of $\boldsymbol{w}$ is much smaller compared to the whole variable $\boldsymbol{\theta} = [\boldsymbol{w}, \boldsymbol{b}]$. 2. Once the local classifier head is updated to $\boldsymbol{w}_i^{t+1}$, the client optimizes its local representation function with multiple SGD steps, starting from the current global consensus $\boldsymbol{b}^t$. 3. Finally, each client calculate its local update direction for representation via the difference between its latest local representation function and the previous global consensus, $\boldsymbol{b}_i^{t,T_l} - \boldsymbol{b}^t$.

Remark 4.1 (Privacy guarantee)**.** *By the composition theorem for subsampled Gaussian Mechanism Mironov et al. (2019), we prove that Algorithm 1 and Algorithm 2 satisfies user-level (α, ϵ)-Rényi differential privacy for $\epsilon = T_g \cdot S_\alpha(p_c, \sigma_g)$, where $S_\alpha(p_c, \sigma_g) = R_\alpha(\mathcal{N}(0, \sigma_g^2) \| \mathcal{N}(0, \sigma_g^2) + p_c \cdot \mathcal{N}(1, \sigma_g^2))$. In the special case of full gradient descent with $p_c = 1$, we have that $\epsilon = \alpha \cdot T_g/(2\sigma_g^2)$.*

5 GUARANTEED IMPROVEMENT OF THE UTILITY-PRIVACY TRADE-OFF

In the previous section, we present `CENTAUR` for the general FRL problems (3). Due to the lack of structure information, for a general non-convex optimization problem we cannot expect any utility guarantee beyond the convergence to a stationary point. In this section, we consider a specific instance of the FRL problems where both the representation function ϕ and the local classifiers h are *linear* w.r.t. their parameters $\boldsymbol{b}$ and $\boldsymbol{w}_i$. This model has been commonly used in the analysis of representation learning (Collins et al., 2021; 2022), meta learning (Tripuraneni et al., 2021; Du et al., 2020; Thekumparampil et al., 2021; Sun et al., 2021), model personalization (Jain et al., 2021), multi-task learning (Maurer et al., 2016). For this (still nonconvex) instance, we prove that `CENTAUR` converges to a ball centered around the *global minimizer* in a linear rate where the size of the ball depends on the required RDP parameters (α, ϵ). Let ϵ_a be the utility and let ϵ_{dp} be the DP-parameter (`CENTAUR` is an (ϵ_{dp}, δ)-DP mechanism). We obtain the improved privacy-utility tradeoff $\epsilon_a \cdot \epsilon_{dp} \geq O(d/n)$, which is $O(\sqrt{d})$ better than the current SOTA result, $\epsilon_a \cdot \epsilon_{dp} \geq O(d^{1.5}/n)$, by (Jain et al., 2021). In the following, we will first review objective (3) in the linear representation setting, and analyse `CENTAUR` to show the improved utility-privacy tradeoff.

Federated Linear Representation Learning (LRL) Recall the problem formulation in Section 3. For simplicity, we assume $m_i = m$ for all $i \in \{1, \ldots, n\}$ for some constant m. Consider the LRL setting, where given the input $\boldsymbol{x}_{ij}$ the response $y_{ij} \in \mathbb{R}$ satisfies $y_{ij} = \boldsymbol{w}_i^{*\top} \boldsymbol{B}^{*\top} \boldsymbol{x}_{ij}$. Here $\boldsymbol{B}^* \in \mathbb{R}^{d\times k}$ is a *column orthonormal* global representation matrix and $\boldsymbol{w}_i^* \in \mathbb{R}^k$ is an agent-specific optimal local linear classifiers. In terms of the notations, $\boldsymbol{B}$ corresponds to the parameter $\boldsymbol{b}$ in the general formulation (3), but is capitalized since it is now regarded as a matrix. The feasible domain $\mathbb{B}$ is the set of column orthonormal matrices $\mathbb{O}_{d,k} = \{\boldsymbol{B} \in \mathbb{R}^{d\times k} | \boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{I}_k\}$.
Given a local dataset $\mathbb{S}$ (could be $\mathbb{S}_i$ or its subset), define

$$l([\boldsymbol{w}, \boldsymbol{B}]; \mathbb{S}) = \frac{1}{|\mathbb{S}|} \sum_{(\boldsymbol{x}, y)\in\mathbb{S}} \frac{1}{2} (\boldsymbol{w}^\top \boldsymbol{B}^\top \boldsymbol{x} - y)^2. \tag{4}$$

Our goal is to recover the ground truth representation matrix $\boldsymbol{B}^*$ using the collection of the local datasets $\{\mathbb{S}_i\}_{i=1}^n$ via solving the following linear instance of the FRL problem equation 3

$$\min_{\boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{I}_k} \frac{1}{n} \sum_{i=1}^n \min_{\boldsymbol{w}_i} l([\boldsymbol{w}_i, \boldsymbol{B}]; \mathbb{S}_i) \tag{L-FRL}$$

in a federated and DP manner. Here, equation L-FRL is an instance of the general FRL problem (3) with h and ϕ set to linear functions and ℓ set to the least square loss. Note that, despite the (relatively) simple structure of equation L-FRL, it is still *non-convex* w.r.t. the variable $\boldsymbol{B}$.

Methodology for the LRL case. To establish our novel privacy and utility guarantees, we need to specify INITIALIZATION and AGGREGATION in the procedures SERVER and we also need to slightly modify the CLIENT procedure, which are elaborated as follows.

1. To obtain the novel guarantee of converging to the global minimizer for the LRL case, the initializer $\boldsymbol{B}^0$ needs to be within a constant distance to the global optimal $\boldsymbol{B}^*$ which requires a more

Algorithm 3 INITIALIZATION procedure for `CENTAUR` in the LRL case.

```
procedure INITILIZATION($T_0, \epsilon_i, L_0, \sigma_0, \zeta_0$)
    Run $T_0$ independent copies of PPM($L, \sigma_0, \zeta_0$) to obtain $T_0$ candidates $\{\boldsymbol{B}_c\}_{c=1}^{T_0}$;
    // PPM stands for private power method and is presented in Algorithm 5 in the appendix.
    // Boost this probability to $1 - n^{-k}$ via cross validation.
    Find $\hat{c}$ in $[T_0]$ such that for at least half of $c \in [T_0]$, $s_i(\boldsymbol{B}_c^\top \boldsymbol{B}_{\hat{c}}) \geq 1 - 2\epsilon_i^2, \forall i \in [k]$.
    return $\boldsymbol{B}_{\hat{c}}$.
```

Algorithm 4 CLIENT procedure for `CENTAUR` in the LRL case

```
1: procedure CLIENT($\boldsymbol{B}^t$, $\bar{m}$)
2:     Sample without replacement two subsets $\mathbb{S}_i^{t,1}$ and $\mathbb{S}_i^{t,2}$ from $\mathbb{S}_i$ both with cardinality $\bar{m}$.
3:     [Phase 1:] Update the local head $\boldsymbol{w}_i^{t+1} = \arg\min_{\boldsymbol{w} \in \mathbb{R}^k} l(\boldsymbol{w}, \boldsymbol{B}^t; \mathbb{S}_i^{t,1})$.
4:     [Phase 2:] Compute the local gradient of the representation $\boldsymbol{G}_i^t = \partial_{\boldsymbol{B}} l([\boldsymbol{w}_i^{t+1}, \boldsymbol{B}^t]; \mathbb{S}_i^{t,2})$;
5:     [Phase 3:] return $\boldsymbol{G}_i^t$
```

sophisticated procedure than the simple random initialization. We show that this requirement can be ensured by running a modified instance of the private power method (`PPM`) by (Hardt & Price, 2014), but the utility guarantee only holds *with a constant probability* (Lemma F.1). A key contribution of our work is to propose a novel *cross-validation* scheme to boost the success probability of the `PPM` with a small extra cost. Our scheme only takes as input the outputs of independent `PPM` trials, and hence can be treated as post-processing, which is free of privacy risk (Lemma F.3). The proposed INITIALIZATION procedure is presented in Algorithm 3. The analyses are deferred to Appendix F.
2. As discussed earlier, the feasible domain $\mathbb{B}$ is the set of column orthonormal matrices. In order to ensure the feasibility of $\boldsymbol{B}^{t+1}$, we set AGGREGATION($\boldsymbol{B}^t, \{\boldsymbol{G}_i^t\}_{i \in \mathbb{C}^t}, \eta_g$) $= \mathcal{QR}(\boldsymbol{B}^t - \eta_g \cdot \boldsymbol{G}^t)$, where $\mathcal{QR}(\cdot)$ denotes the QR decomposition and only returns the Q matrix.
3. We make a small modification in line 2 where two subsets $\mathbb{S}_i^{t,1}$ and $\mathbb{S}_i^{t,2}$ of the local dataset are sampled *without replacement* from $\mathbb{S}_i$ and are used to replace $\mathbb{S}_i$ in Phases 1 and 2. This modification is required to establish our novel utility result, which ensures that clipping threshold of the Gaussian mechanism (line 5) in the CLIENT procedure is never reached with a high probability (Lemma C.1).

5.1 ANALYSIS OF `CENTAUR` IN THE LRL SETTING

Use $s_i(\boldsymbol{A})$ to denote the ith largest singular value of a matrix $\boldsymbol{A}$. Let $\boldsymbol{W}^* \in \mathbb{R}^{n \times k}$ be the collection of the local optimal classifier heads with $\boldsymbol{W}_{i,:}^* = \boldsymbol{w}_i^*$. We use s_i as a shorthand for $s_i(\boldsymbol{W}^*/\sqrt{n})$ and use $\kappa = s_1/s_k$ to denote the condition number of the problem. We choose the scaling $1/\sqrt{n}$ such that s_i remains meaningful as $n \to \infty$. We make the following assumptions.

Assumption 5.1. *$\boldsymbol{x}_{ij}$ is zero mean, $\boldsymbol{I}_d$-covariance, 1-subgaussian(defined in Appendix A.1).*

Assumption 5.2. *There exists a constant $\mu > 0$ such that $\max_{i \in [n]} \|\boldsymbol{w}_i^*\|_2 \leq \mu\sqrt{k}s_k$.*

Assumption 5.3. *The number of local data points is sufficiently large: $m \geq \tilde{c}_{ld} \max(k^2, d/n)$, and the number of global data points is sufficiently large: $mn \geq \tilde{c}_{gd} k^4 d$. Here $\tilde{c}_{ld}$ and $\tilde{c}_{gd}$ hide dependence on κ, μ, and the log factors.*

These assumptions are standard in literature (Li et al., 2020b; Collins et al., 2021; Jain et al., 2021). While Problem (L-FRL) is non-convex, we show that the consensus representation $\boldsymbol{B}^t$ converges to a ball centered around the ground truth solution $\boldsymbol{B}^*$ in a linear rate. The size of the ball is controlled by the noise multiplier σ_g, which will be the free parameter that controls the utility-privacy tradeoff.

High level idea. For Problem (L-FRL), we find an initial point close to the ground truth solution via the method of moments. Given this strong initialization, `CENTAUR` converges linearly to the vicinity of the ground truth since it can be interpreted as an inexact gradient descent method with a fast decreasing bias term. One caveat that requires particular attention is the clipping step in the Gaussian mechanism (line 5 in Algorithm 4) will destroy the above interpretation if the threshold parameter ζ_g is set too small. To resolve this, we set ζ_g to be a high probability upper bound of $\|\boldsymbol{G}_i^t\|_F$ so that the clipping step only takes effect with a negligible probability.

In the utility analysis of the LRL case, we use the principal angle distance to measure the convergence of the column-orthonormal variable $\boldsymbol{B}$ towards the ground truth $\boldsymbol{B}^*$. We also refer to this quantity as the *utility* of the algorithm. Let $\boldsymbol{B}_\perp$ be the orthogonal complement of $\boldsymbol{B}$. We define

$$\operatorname{dist}(\boldsymbol{B}, \boldsymbol{B}^*) := \|(\boldsymbol{I}_d - \boldsymbol{B}\boldsymbol{B}^\top)\boldsymbol{B}^*\|_2 = \|\boldsymbol{B}_\perp^\top \boldsymbol{B}^*\|_2.$$

To simplify the presentation of the result, we make the following assumptions: The dimension of the input is sufficiently large $d \geq \kappa^8 k^3 \log n$ and the number of clients is sufficiently large $n \geq m$. The proof of the following theorem can be find in Appendix C.

Theorem 5.1 (Utility Analysis). *Consider the instance of* `CENTAUR` *for the LRL setting with its* CLIENT *and* INITIALIZATION *procedures defined in Algorithms 4 and 3 respectively. Suppose that the matrix* $\boldsymbol{B}^0$ *returned by* INITIALIZATION *satisfies* $\operatorname{dist}(\boldsymbol{B}^0, \boldsymbol{B}^*) \leq \epsilon_0 = 0.2$*, and suppose that the mini-batch size parameter* $\bar{m}$ *satisfies* $\bar{m} \geq c_m \max\{\kappa^2 k^2 \log n, k^2 d/n\}$*. Set the clipping threshold of the Gaussian mechanism* $\zeta_g = c_\zeta \mu^2 k s_k^2 \sqrt{dk \log n}$*, the global step size* $\eta_g = 1/4s_1^2$*, the number of global rounds* $T_g = c_T \kappa^2 \log(\kappa \eta_g \zeta_g \sigma_g d/n)$*. Assuming that the noise multiplier in the Gaussian mechanism is sufficiently small:* $\sigma_g \leq c_\sigma n \kappa^4/(\mu^2 d \sqrt{k \log n})$*. We have with probability at least* $1 - c_p \bar{m} T_g \cdot n^{-k}$,

$$\operatorname{dist}(\boldsymbol{B}^{T_g}, \boldsymbol{B}^*) \leq c_d \kappa^2 \eta_g \sigma_g \zeta_g \sqrt{d}/n = \tilde{c}_d \sigma_g \mu^2 k^{1.5} d/n. \tag{5}$$

Here c_m, c_ζ, c_T, c_σ, c_p *and* c_d *are universal constants.*

As will be discussed in Theorem 5.2, a larger noise multiplier σ_g leads to a smaller privacy risk. On the other hand, in order for the above utility guarantee to hold, we cannot have arbitrarily good privacy protection as that would violate the limitation on σ_g. The intuition behind this requirement is that our convergence analysis requires the iterates to stay within a ball centered around the ground truth, with a constant radius (measured in terms of the principal angle distance). Adding a large Gaussian noise will break this argument. Similar requirements are made in Tripuraneni et al. (2021).

Since the SERVER procedure remains exactly the same as Algorithm 1 in the LRL case, the main body (anything after INITIALIZATION) of the resulting `CENTAUR` instance has the same privacy guarantee as described in remark 4.1. However, we still need to account for the privacy leakage of the INITIALIZATION procedure in Algorithm 3 as it is data-dependent. This will be deferred to Appendix F, where we show that Algorithm 3 is an $(\alpha, \epsilon_{init})$-RDP mechanism, with ϵ_{init} defined in Corollary F.1. Combining this fact with the RDP analysis for the main body leads to the following privacy guarantee for `CENTAUR` in the LRL case (see Appendix A.4).

Theorem 5.2 (Privacy Bound). *Consider the instance of* `CENTAUR` *with its* CLIENT *procedure defined in Algorithm 4 and its* INITIALIZATION *procedure defined in Algorithm 3. Suppose that the* INITIALIZATION *procedure is an* $(\alpha, \epsilon_{init})$*-RDP mechanism (proved in Corollary F.1). Let* σ_g*, the noise multiple in the* CLIENT *procedure, be a free parameter that controls the privacy-utility trade-off. By setting the inputs to* SERVER *as* $p_g = 1$, $T_g = c_T \kappa^2 \log(\kappa \eta_g \zeta_g \sigma_g d/n)$*, the instance of* `CENTAUR` *under consideration is an* $(\alpha, \epsilon_{init} + \epsilon_{rdp}/2)$*-RDP mechanism, where* $\epsilon_{rdp} = 4\alpha T_g/\sigma_g^2$*. Moreover, when* $\sigma_g = \tilde{O}(n/(k^4 \mu^2 d))$*, we have* $\epsilon_{init} \leq \epsilon_{rdp}/2$*, in which case* `CENTAUR` *is an* (α, ϵ_{rdp})*-RDP mechanism and is also an* (ϵ_{dp}, δ)*-DP mechanism with* $\epsilon_{dp} = 2\sqrt{T_g \log(1/\delta)}/\sigma_g$.

5.2 OVERALL UTILITY-PRIVACY TRADE-OFF

We now combine the utility and privacy analyses of `CENTAUR` in the LRL setting to obtain the overall utility-privacy trade-off in the following sense: According to Theorem 5.1, to achieve a high utility, i.e. a small ϵ_a, we need to choose a small noise multiplier σ_g while Theorem 5.2 states that the smaller σ_g is, the larger the privacy cost. For the simplicity of the presentation, we focus on the regime where the trade-off parameter satisfies $\sigma_g = \tilde{O}(n/k^4 \mu^2 d)$.

Corollary 5.1. *Use* ϵ_a *to denote the target utility, i.e.* $\operatorname{dist}(\boldsymbol{B}^T, \boldsymbol{B}^*) \leq \epsilon_a$ *where* $\boldsymbol{B}^T$ *is the output of* `CENTAUR` *and use* ϵ_{dp} *to denote the privacy budget, i.e.* `CENTAUR` *is an* (ϵ_{dp}, δ)*-DP mechanism. We have the following tradeoff where* $\tilde{c}_t$ *hides the constants and* log *terms:*

$$\tilde{c}_t \kappa k^{1.5} \mu^2 d/n \leq \epsilon_a \cdot \epsilon_{dp}.$$

When focusing on the input dimension d and the number of clients n and treating other factors as constants, the above trade-off can be simplified to $O(d/n) \leq \epsilon_a \cdot \epsilon_{dp}$. Recall that the previous SOTA result of the LRL setting by Jain et al. (2021) (Lemma 4.4 therein) can be interpreted as $O(d^{1.5}/n) \leq \epsilon_a \cdot \epsilon_{dp}$. Hence, we obtain an $O(\sqrt{d})$ improvement.

Table 1: Testing accuracy (%) on CIFAR10/CIFAR100/EMNIST with various data allocation settings. No data augmentation is used. n stands for the number clients and S stands for the number classes per client. The δ parameter of DP is fixed to 10^{-5} as a common choice in the literature. The budget parameter of DP is fixed to a small value of 1 for results in this table, i.e. $\epsilon_{dp} = 1$.

Methods	`Stand-alone-no-FL`	`DP-FedAvg-ft`	`PMTL-ft`	`PPSGD`	`CENTAUR`
CIFAR10 n=500, S=5	50.03 (1.59)	38.04 (1.97)	42.06 (0.94)	47.94 (1.87)	**52.55 (0.69)**
CIFAR10 n=1000,S=2	74.06 (0.45)	57.02 (1.36)	67.79 (1.77)	71.52 (2.30)	**75.75 (0.20)**
CIFAR10 n=1000,S=5	44.60 (1.30)	37.15 (1.22)	35.99 (1.79)	44.61 (2.68)	**47.67 (1.62)**
CIFAR100 n=1000,S=5	39.20 (1.12)	22.17 (2.12)	24.17 (1.43)	38.40 (0.66)	**41.59 (1.17)**
EMNIST n=1000,S=5	93.47 (0.14)	91.32 (0.22)	92.32 (0.22)	93.44 (0.20)	**94.17 (0.19)**
EMNIST n=2000,S=5	90.67 (0.46)	86.85 (1.11)	88.81 (2.08)	88.96 (1.93)	**92.79 (0.25)**

6 EXPERIMENTS

In this section, we present the empirical results that show the significant advantage of the proposed `CENTAUR` over previous arts. Four baselines are included: `Stand-alone-no-FL` which stands for local stand-alone training; `DP-FedAvg-ft` which stands for DP-FedAvg with local fine tuning (Yu et al., 2020); `PPSGD` proposed by Bietti et al. (2022); and `PMTL-ft` which stands for PMTL proposed by Hu et al. (2021) with local fine tuning. Note that `Stand-alone-no-FL` does not involve any global communications, therefore no privacy mechanism is added to its implementation. This makes `Stand-alone-no-FL` a strong baseline as the utility of all included differentially private competing methods are affected by gradient clipping and noise injection, especially when the DP budget is small, e.g. $\epsilon_{dp} = 1$. Another advantage of `Stand-alone-no-FL` setting is that, the local stand-alone models are highly flexible, i.e. the model on one client and be completely different from the one on others. On the contrary, while the models of all other non-local methods share a common representation part, which takes up the major portion of the whole model.

We focus on the task of image classification and conduct experiments on three representative datasets, namely CIFAR10, CIFAR100, and EMNIST. In terms of architecture of the neural network, we use LeNet for CIFAR10/CIFAR100 and use MLP for EMNIST, as commonly used in the federated learning literature, the details of which are discussed in Appendix. In terms of data augmentation, we do not perform any data augmentation, as we observe that naive data augmentation for DP training leads to worse accuracy, as also reported in De et al. (2022).

`CENTAUR` Has the Best Privacy Utility Trade-off. We first present the utility (testing accuracy) of models trained with `CENTAUR` and other baselines algorithms under a fixed small privacy budget $\epsilon_{dp} = 1$, for a variety of heterogeneous FL settings. To simulate the data-heterogeneity phenomenon ubiquitous in the research of federated learning, we follow the data allocation scheme of (Collins et al., 2021): Specifically, we first split the original dataset into a training part (90%) and a validation part (10%) and we then allocate the training part equally to n clients while ensuring that each client has at most data from S classes. In Table 1, we observe that, under this small privacy budget, our proposed `CENTAUR` enjoy better performance than all the included baseline algorithms. Importantly, `CENTAUR` is **the only method** that consistently **outperforms the strong local-only baseline**, and therefore justifies the choice of collaborative learning as opposed of local stand-alone training. Finally, we further demonstrate that `CENTAUR` enables superior privacy utility trade-off uniformly across different privacy budget ϵ, for the setting of EMNIST dataset ($n = 2000, S = 5$) in Figure 1.

Conclusion. In this work, we point out that the phenomenon of statistical heterogeneity, one of the major challenges of federated learning, introduces extra difficulty when DP constraints are imposed. To alleviate this difficulty, we consider the federated representation learning where only the representation function is to be globally shared and trained. We provide a rigorous guarantee for the utility-privacy trade-off of the proposed `CENTAUR` method in the linear representation setting, which is $O(\sqrt{d})$ better than the SOTA result. We also empirically show that `CENTAUR` provides better utility uniformly on several vision datasets under various data heterogeneous settings.

REFERENCES

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2020.

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021.

Naman Agarwal, Ananda Theertha Suresh, Felix Xinnan X Yu, Sanjiv Kumar, and Brendan McMahan. cpsgd: Communication-efficient and differentially-private distributed sgd. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *Acta numerica*, 30:87–201, 2021.

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.

Alberto Bietti, Chen-Yu Wei, Miroslav Dudik, John Langford, and Steven Wu. Personalization improves privacy-accuracy tradeoffs in federated learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1945–1962. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/bietti22a.html`.

Zachary Charles and Jakub Konečný. Convergence and accuracy trade-offs in federated learning and meta-learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2575–2583. PMLR, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International Conference on Machine Learning*, pp. 2089–2099. PMLR, 2021.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Fedavg with fine tuning: Local updates lead to representation learning. *NeurIPS*, 2022.

Soham De, Leonard Berrada, Jamie Hayes, Samuel L Smith, and Borja Balle. Unlocking high-accuracy differentially private image classification through scale. *arXiv preprint arXiv:2204.13650*, 2022.

Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. *Advances in Neural Information Processing Systems*, 30, 2017.

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

John C Duchi, Michael I Jordan, and Martin J Wainwright. Privacy aware learning. *Journal of the ACM (JACM)*, 61(6):1–57, 2014.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, pp. 1054–1067, 2014.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33:3557–3568, 2020.

Moritz Hardt and Eric Price. The noisy power method: A meta algorithm with applications. *Advances in neural information processing systems*, 27, 2014.

Justin Hsu, Zhiyi Huang, Aaron Roth, Tim Roughgarden, and Zhiwei Steven Wu. Private matchings and allocations. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, STOC '14, pp. 21–30, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450327107. doi: 10.1145/2591796.2591826. URL `https://doi.org/10.1145/2591796.2591826`.

Shengyuan Hu, Zhiwei Steven Wu, and Virginia Smith. Private multi-task learning: Formulation and applications to federated learning. *arXiv preprint arXiv:2108.12978*, 2021.

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

Prateek Jain, John Rush, Adam Smith, Shuang Song, and Abhradeep Guha Thakurta. Differentially private model personalization. *Advances in Neural Information Processing Systems*, 34, 2021.

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

Peter Kairouz, Brendan McMahan, Shuang Song, Om Thakkar, Abhradeep Thakurta, and Zheng Xu. Practical and private (deep) learning without sampling or shuffling. In *International Conference on Machine Learning*, pp. 5213–5225. PMLR, 2021.

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*, pp. 5132–5143. PMLR, 2020.

Daniel Levy, Ziteng Sun, Kareem Amin, Satyen Kale, Alex Kulesza, Mehryar Mohri, and Ananda Theertha Suresh. Learning with user-level privacy. *Advances in Neural Information Processing Systems*, 34:12466–12479, 2021.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020b.

Zhu Li, Zhi-Hua Zhou, and Arthur Gretton. Towards an understanding of benign overfitting in neural networks. *arXiv preprint arXiv:2106.03212*, 2021.

Yingyu Liang, Maria-Florina F Balcan, Vandana Kanchanapally, and David Woodruff. Improved distributed principal component analysis. *Advances in neural information processing systems*, 27, 2014.

Grigory Malinovskiy, Dmitry Kovalev, Elnur Gasanov, Laurent Condat, and Peter Richtarik. From local sgd to local fixed-point methods for federated learning. In *International Conference on Machine Learning*, pp. 6692–6701. PMLR, 2020.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

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*, pp. 1273–1282. PMLR, 2017a.

H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*, 2017b.

Ilya Mironov, Kunal Talwar, and Li Zhang. R\'enyi differential privacy of the sampled gaussian mechanism. *arXiv preprint arXiv:1908.10530*, 2019.

Aritra Mitra, Rayana Jaafar, George J Pappas, and Hamed Hassani. Achieving linear convergence in federated learning under objective and systems heterogeneity. *NeurIPS*, 2021.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pp. 739–753. IEEE, 2019.

Krishna Pillutla, Kshitiz Malik, Abdel-Rahman Mohamed, Mike Rabbat, Maziar Sanjabi, and Lin Xiao. Federated learning with partial model personalization. In *International Conference on Machine Learning*, pp. 17716–17758. PMLR, 2022.

Mark Rudelson and Roman Vershynin. The littlewood–offord problem and invertibility of random matrices. *Advances in Mathematics*, 218(2):600–633, 2008.

Zebang Shen, Hamed Hassani, Satyen Kale, and Amin Karbasi. Federated functional gradient boosting. In *International Conference on Artificial Intelligence and Statistics*, pp. 7814–7840. PMLR, 2022.

Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1310–1321, 2015.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. *Advances in neural information processing systems*, 30, 2017.

Yue Sun, Adhyyan Narang, Ibrahim Gulluk, Samet Oymak, and Maryam Fazel. Towards sample-efficient overparameterized meta-learning. *Advances in Neural Information Processing Systems*, 34:28156–28168, 2021.

Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Statistically and computationally efficient linear meta-representation learning. *Advances in Neural Information Processing Systems*, 34:18487–18500, 2021.

Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pp. 10434–10443. PMLR, 2021.

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

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.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.

A APPENDIX

A.1 SUBGAUSSIAN RANDOM VARIABLE

Let $\mathrm{x} \in \mathbb{R}$ be a subgaussian random variable. We use $\|\mathrm{x}\|_{\psi_2} = \inf\{s > 0 | \mathbb{E}[\exp(\mathrm{x}^2/s^2)] \leq 2\}$ to denote subgaussian norm and say x is $\|\mathrm{x}\|_{\psi_2}$-subgaussian. For a random vector $\mathbf{x} \in \mathbb{R}^d$, we say $\mathbf{x}$ is subgaussian if $\langle \boldsymbol{x}, \mathbf{x} \rangle$ is subgaussian for any weight vector $\boldsymbol{x} \in \mathbb{R}^d$ and the subgaussian norm of the random vector $\mathbf{x}$ is defined as $\|\mathbf{x}\|_{\psi_2} = \sup_{\boldsymbol{x} \in \mathbb{R}^d, \|\boldsymbol{x}\|_2 = 1} \|\langle \boldsymbol{x}, \mathrm{x} \rangle\|_{\psi_2}$. We refer to x as $\|\mathrm{x}\|_{\psi_2}$-subgaussian.

A.2 USEFUL FACTS

Note that $\boldsymbol{I}_k = \boldsymbol{B}^\top [\boldsymbol{B}^* \ \boldsymbol{B}^*_\perp][\boldsymbol{B}^* \ \boldsymbol{B}^*_\perp]^\top \boldsymbol{B} = \boldsymbol{B}^\top \boldsymbol{B}^* \boldsymbol{B}^{*\top} \boldsymbol{B} + \boldsymbol{B}^\top \boldsymbol{B}^*_\perp \boldsymbol{B}^{*\ \top}_\perp \boldsymbol{B}$ and hence

$$\begin{aligned} s^2_{\min}(\boldsymbol{B}^\top \boldsymbol{B}^*) &= s_{\min}(\boldsymbol{B}^\top \boldsymbol{B}^* \boldsymbol{B}^{*\top} \boldsymbol{B}) = \min_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^\top \boldsymbol{B}^\top \boldsymbol{B}^* \boldsymbol{B}^{*\top} \boldsymbol{B} \boldsymbol{a} \\ &= \min_{\|\boldsymbol{a}\|=1} \boldsymbol{a}^\top \left(\boldsymbol{I}_k - \boldsymbol{B}^\top \boldsymbol{B}^*_\perp \boldsymbol{B}^{*\ \top}_\perp \boldsymbol{B} \right) \boldsymbol{a} \\ &= 1 - s_{\max}(\boldsymbol{B}^\top \boldsymbol{B}^*_\perp \boldsymbol{B}^{*\ \top}_\perp \boldsymbol{B}) = 1 - (\mathrm{dist}(\boldsymbol{B}^*, \boldsymbol{B}))^2. \end{aligned} \tag{6}$$

A.3 MORE ON PRIVACY PRELIMINARIES

Lemma A.1 (Gaussian Mechanism of RDP). *Let R_α be the Rényi divergence defined in Definition 2.1, we have $R_\alpha(\mathcal{N}(0, \sigma^2), \mathcal{N}(\mu, \sigma^2)) = \alpha\mu^2/(2\sigma^2)$. Here $\mathcal{N}$ stands for the standard Gaussian distribution.*

Lemma A.2 (Composition of RDP). *Recall the definition of $\mathcal{D}$ in Definition 2.1 and let R_1 and R_2 be some abstract space. Let $\mathcal{M}_1 : \mathcal{D} \to R_1$ and $\mathcal{M}_1 : \mathcal{D} \times R_1 \to R_2$ be (α, ϵ_1)-RDP and (α, ϵ_2)-RDP respectively, then the mechanism defined as (X, Y), where $X \sim \mathcal{M}_1(\mathcal{S})$ and $Y \sim \mathcal{M}_2(\mathcal{S}, X)$, satisfies $(\alpha, \epsilon_1 + \epsilon_2)$-RDP.*

Theorem A.1 (Conversion from RDP to DP). *If $\mathcal{M}$ is an (α, ϵ)-RDP mechanism, it also satisfies $(\epsilon + \frac{\log 1/\delta}{\alpha - 1}, \delta)$-differential privacy for any $0 < \delta < 1$.*

A.4 PROOF OF THEOREM 5.2

Proof. Recall the definition of the Gaussian mechanism

$$\mathrm{GM}_{\zeta,\sigma}(\{\boldsymbol{x}_i\}_{i=1}^s) \doteq \frac{1}{s}(\sum_{i=1}^s \mathrm{clip}(\boldsymbol{x}_i; \zeta) + \sigma\zeta W)$$

where $W \sim \mathcal{N}(0, \boldsymbol{I})$. Lemma A.1 states that $\mathrm{GM}_{\zeta_g,\sigma_g}$ is a $(\alpha, 2\alpha/\sigma_g^2)$-RDP mechanism for $\alpha \geq 1$ (the sensitivity is $(2\zeta_g/n)^2$ while the variance of the noise is $(\sigma_g\zeta_g/n)^2$). Using the composition of RDP again over all the iterates $t \in [T]$, we obtain that Algorithm 1 is an $(\alpha, \epsilon_{init} + \frac{2\alpha T_g}{\sigma_g^2})$-RDP mechanism.

□

B DETAILS ON EXPERIMENTS SETUP

Models. We use LeNet-5 for the datasets CIFAR10 and CIFAR100. LeNet-5 consists of two convolution layers with (64, 64) channels and two hidden fully-connected layers. For CIFAR10, the number of hidden neurons are (384, 192) while for CIFAR100, the number of hidden neurons are (128, 32). We use ReLU for activation. No batch normalization or dropout layer is used.
We use MLP for experiments on EMNIST. It consists of three hidden layers with size (256, 128, 16). We use ReLU for activation. No batch normalization or dropout layer is used.

Hyperparameters. All of our experiments are conducted in the fully participating setting, i.e. $p_c = 1$. According to our experiments, the following hyperparameters are most important to the performance of `CENTAUR`: the clipping threshold of the Gaussian mechanism ζ_g, the global step size η_g, the local step size η_l, the number of global rounds T_g.
For CIFAR10 and CIFAR100, we uniformly set $\zeta_g = 0.02$, $\eta_l = 0.01$, $T_g = 100$, $\eta_g = 5$. Note that once the privacy budget ϵ_{dp} is given ($\epsilon_{dp} = 1$ in our experiments), we use the privacy engine from the package of Opacus to determine the noise multiplier σ_g, given T_g.
For EMNIST, we uniformly set $\zeta_g = 0.25$, $\eta_l = 0.01$, $T_g = 40$, $\eta_g = 1$.

There are also some algorithm-specific parameters: For `PPSGD`, we set the step size for the local correction to $\eta = 0.1$ and set ratio between the global and the local step size to $\alpha = 0.1$. For `PMTL-ft`, we set λ, the regularization parameter to 1.

For baselines that require local fine tuning, we perform 15 local epochs to fine tune the local head with a fixed step size of 0.01.

C UTILITY ANALYSIS OF THE `CENTAUR` INSTANCE FOR THE LRL CASE

To present the utility analysis of the `CENTAUR` instance for the LRL case, Problem (L-FRL) is equivalently formulated as a standard matrix sensing problem. By setting the clipping threshold ζ_g to a high probability upper bound of the norm of the local gradient $\boldsymbol{G}_i^t$ (see Lemma C.1), we show that `CENTAUR` can be regarded as an inexact gradient descent method. Given that $\bar{m}$, the mini-batch size parameter, is sufficiently large and that the initializer $\boldsymbol{B}^0$ is sufficiently close the ground truth $\boldsymbol{B}^*$, we establish a high probability one-step contraction lemma that controls the utility $\mathrm{dist}(\boldsymbol{B}^t, \boldsymbol{B}^*)$, which directly leads to the main utility theorem 5.1.

Matrix Sensing Formulation Consider a reparameterization[1] of the local classifier $\boldsymbol{w}_i = \sqrt{n}\boldsymbol{v}_i$. Problem (L-FRL) can be written as

$$\min_{B^\top B = \boldsymbol{I}_k} \frac{1}{n}\sum_{i=1}^{n} \min_{\boldsymbol{v}_i \in \mathbb{R}^k} \left(\frac{1}{m}\sum_{j=1}^{m} \frac{1}{2}(\sqrt{n}\langle \boldsymbol{v}_i, \boldsymbol{B}^\top \boldsymbol{x}_{ij}\rangle - y_{ij})^2 \right). \tag{7}$$

Further, denote $\boldsymbol{V} \in \mathbb{R}^{n\times k}$ the collection of local classifiers, $\boldsymbol{V}_{i,:} = \boldsymbol{v}_i$. The collection of optimal local classifier heads $\boldsymbol{W}^*$ can also be rescaled as $\boldsymbol{W}^* = \sqrt{n}\boldsymbol{V}^*$ and the responses $y_{ij} \in \mathbb{R}$ satisfy

$$y_{ij} = \sqrt{n}\langle \boldsymbol{B}^{*\top}\boldsymbol{x}_{ij}, \boldsymbol{V}^*_{i,:}\rangle. \tag{8}$$

Define the rank-1 matrices $\boldsymbol{A}_{ij} = \boldsymbol{x}_{ij}\boldsymbol{e}^{(i)\top} \in \mathbb{R}^{d\times n}$ ($\boldsymbol{e}^{(i)} \in \mathbb{R}^n$) and define the operators $\mathcal{A}_i : \mathbb{R}^{d\times n} \to \mathbb{R}^m$ and $\mathcal{A} : \mathbb{R}^{d\times n} \to \mathbb{R}^N$

$$\mathcal{A}_i(\boldsymbol{X}) = \{\sqrt{n}\langle \boldsymbol{A}_{ij}, \boldsymbol{X}\rangle\}_{j\in[m]} \in \mathbb{R}^m, \mathcal{A}(\boldsymbol{X}) = \{\mathcal{A}_i(\boldsymbol{X})\}_{i\in[n]} \in \mathbb{R}^N, \tag{9}$$

where we use $N = nm$ to denote the total number of data points globally. Note that $\frac{1}{\sqrt{N}}\mathcal{A}$ is an isometric operator in expectation w.r.t. the randomness of $\{\boldsymbol{x}_{ij}\}$, i.e. for any $\boldsymbol{X} \in \mathbb{R}^{d\times n}$

$$\begin{aligned}\mathbb{E}_{\{\boldsymbol{x}_{ij}\}}[\langle \mathcal{A}(\boldsymbol{X}), \mathcal{A}(\boldsymbol{X})\rangle] &= \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m} n\boldsymbol{e}^{(i)\top}\boldsymbol{X}^\top \mathbb{E}_{\boldsymbol{x}_{ij}}[\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top]\boldsymbol{X}\boldsymbol{e}^{(i)} \\ &= \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m} n\boldsymbol{X}_{:,i}^\top \boldsymbol{X}_{:,i} = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m} n\|\boldsymbol{X}_{:,i}\|^2 = \|\boldsymbol{X}\|_F^2,\end{aligned}$$

where we use Assumption 5.1 in the second equality. With the notations defined above, we can rewrite Problem (L-FRL) as a standard matrix sensing problem with the operator $\mathcal{A}$

$$\min_{B^\top B = \boldsymbol{I}_k} \min_{\boldsymbol{V}\in\mathbb{R}^{n\times k}} \mathcal{F}(\boldsymbol{B}, \boldsymbol{V}; \mathcal{A}) = \frac{1}{2N}\|\mathcal{A}(\boldsymbol{B}\boldsymbol{V}^\top) - \mathcal{A}(\boldsymbol{B}^*\boldsymbol{V}^{*\top})\|^2. \tag{MSP}$$

---

[1] We consider this rescaling of $\boldsymbol{w}_i$ so that the corresponding linear operator in equation 9 is an isometric operator in expectation (see more discussion below).

Since `CENTAUR` only uses a portion of the data points from $\mathbb{S}_i$ to compute the local gradient $\boldsymbol{G}_i^t$ (see line 2 in Algorithm 4), it is useful to define the operators $\mathcal{A}_i^{t,1}$ and $\mathcal{A}_i^{t,2}$ that corresponds to $\mathbb{S}_i^{t,1}$ and $\mathbb{S}_i^{t,2}$ respectively, and their globally aggregated versions $\mathcal{A}^{t,1}$ and $\mathcal{A}^{t,2}$:

$$\mathcal{A}_i^{t,l}(\boldsymbol{X}) = \{\sqrt{n}\langle \boldsymbol{x}_{ij}\boldsymbol{e}^{(i)^\top}, \boldsymbol{X}\rangle\}_{j\in\mathbb{S}_i^{t,l}} \in \mathbb{R}^{\bar{m}}, \mathcal{A}^{t,l}(\boldsymbol{X}) = \{\mathcal{A}_i^{t,l}(\boldsymbol{X})\}_{i\in[n]} \in \mathbb{R}^{\bar{N}}, l = 1, 2, \quad (10)$$

where we denote $\bar{N} \doteq \bar{m}n$.

Clippings are inactive with a high probability The following lemma shows that by properly setting the clipping thresholds ζ_g, the clipping step of the Gaussian mechanism in Algorithm 4 takes no effect with a high probability.

Lemma C.1. *Consider the LRL setting. Under the assumptions 5.2 and 5.1, we have with a probability at least* $1 - \bar{m}n^{-k}$,

$$\|\boldsymbol{G}_i^t\|_F \le \zeta_g \doteq c_\zeta \mu^2 k s_k^2 \sqrt{dk\log n},$$

where $\boldsymbol{G}_i^t$ *is computed in line 4 of Algorithm 4 and* ζ_g *is some universal constant.*

Proof. The detailed expression of $\boldsymbol{G}_i^t$ in line 4 of Algorithm 4 can be calculated as follows:

$$\boldsymbol{G}_i^t = \partial_{\boldsymbol{B}} l([\boldsymbol{w}_i^{t+1}, \boldsymbol{B}^t]; \mathbb{S}_i^{t,2}) = \frac{1}{\bar{m}} \sum_{(\boldsymbol{x}_{ij}, y_{ij})\in\mathbb{S}_i^{t,2}} \Big(\langle {\boldsymbol{B}^t}^\top \boldsymbol{x}_{ij}, \boldsymbol{w}_i^{t+1}\rangle - y_{ij}\Big) \boldsymbol{x}_{ij} {\boldsymbol{w}_i^{t+1}}^\top. \quad (11)$$

Using the triangle inequality of the matrix norm, ζ is a high probability upper bound of $\|\boldsymbol{G}_i^t\|_F$ if the inequality

$$\| \Big(\langle {\boldsymbol{B}^t}^\top \boldsymbol{x}_{ij}, \boldsymbol{w}\rangle - y_{ij}\Big) \boldsymbol{x}_{ij} {\boldsymbol{w}_i^{t+1}}^\top \|_F \le \zeta \quad (12)$$

holds with a high probability. In the following, we show that the inequalities $|\langle {\boldsymbol{B}^t}^\top \boldsymbol{x}_{ij}, \boldsymbol{w}\rangle| \le \zeta_1$, $|y_{ij}| \le \zeta_y$, $\|\boldsymbol{x}_{ij}\|_2 \le \zeta_x$ and $\|\boldsymbol{w}_i^{t+1}\| \le \zeta_w$ hold jointly with probability at least $1 - 5n^{-k}$, which together with $(\zeta_y + \zeta_1)\zeta_x\zeta_w \le \zeta_g$ and the union bound leads to the result of the lemma.

Choice of ζ_y Recall that in Assumption 5.1, we assume that $\boldsymbol{x}_{ij}$ is a sub-Gaussian random vector with $\|\boldsymbol{x}_{ij}\|_{\psi_2} = 1$. Using the definition of a sub-Gaussian random vector, we have

$$\mathbb{P}\{|y_{ij}| \ge \zeta_y\} \le 2\exp(-c_s\zeta_y^2/\|\boldsymbol{w}_i^*\|^2) \le \exp(-k\log n), \quad (13)$$

with the choice $\zeta_y = \mu\sqrt{k}s_k \cdot \sqrt{(k\log n + \log 2)/c_s} = O(\mu s_k k\sqrt{\log n})$ since $\|\boldsymbol{w}_i^*\|_2 \le \mu\sqrt{k}s_k$. Here c_s is some constant and we recall that s_k is a shorthand for $s_k(\boldsymbol{W}^*/\sqrt{n})$.

Choice of ζ_x Recall that $\boldsymbol{x}_{ij}$ is a sub-Gaussian random vector with $\|\boldsymbol{x}_{ij}\|_{\psi_2} = 1$ and therefore with probability at least $1 - \delta$,

$$\|\boldsymbol{x}_{ij}\|_2 \le 4\sqrt{d} + 2\sqrt{\log\frac{1}{\delta}}. \quad (14)$$

Therefore by taking $\delta = \exp(-k\log n)$, we have that $\zeta_x = 4\sqrt{d} + 2\sqrt{\log\frac{1}{\delta}} = O(\sqrt{d})$.

Choice of ζ_w We can show that $\zeta_w = 2\mu\sqrt{k}s_k$ is a high probability upper bound of $\|\boldsymbol{w}_i^{t+1}$. Proving this bound requires detailed analysis of FedRep and is discussed later in equation 64.

Choice of ζ_1 The following event is conditioned on the event $\|\boldsymbol{w}_i^{t+1}\|_2 \le 2\mu\sqrt{k}s_k$. We will then bound the probability of both events happen simultaneously using the union bound. Since $\boldsymbol{x}_{ij}$ is a sub-Gaussian random vector with $\|\boldsymbol{x}_{ij}\|_{\psi_2} = 1$, using the definition of a sub-Gaussian random vector, we have

$$\mathbb{P}\{|\langle \boldsymbol{B}^t\boldsymbol{w}_i^{t+1}, \boldsymbol{x}_{ij}\rangle| \ge \zeta_1\} \le 2\exp(-c_s\zeta_1^2/\|\boldsymbol{w}_i^{t+1}\|^2) \le \exp(-k\log n), \quad (15)$$

with the choice $\zeta_1 = 2\mu\sqrt{k}s_k \cdot \sqrt{(k\log n + \log 2)/c_s} = O(\mu s_k k\sqrt{\log n})$ since $\|\boldsymbol{w}_i^{t+1}\|_2 \le 2\mu\sqrt{k}s_k$. Here c_s is some constant and we recall that s_k is a shorthand for $s_k(\boldsymbol{W}^*/\sqrt{n})$. Using the union bound, we have that w.p. at least $1 - 2\exp(-k\log n)$, the upper bound ζ_1 is valid. $\square$

The idea behind the proof To present the intuition of the utility analysis of `CENTAUR`, define

$$V(\boldsymbol{B};\mathcal{A}) = \operatorname*{arg\,min}_{\boldsymbol{V}\in\mathbb{R}^{n\times k}} \mathcal{F}(\boldsymbol{B},\boldsymbol{V};\mathcal{A}), \tag{16}$$

where $\mathcal{A}$ is some matrix sensing operator and $\mathcal{F}$ is defined in Problem (MSP). Under the event that the clipping steps in the Gaussian mechanism in Algorithm 4 takes no effect, the average gradient $\boldsymbol{G}^t \doteq \frac{1}{n}\sum_{i=1}^n \boldsymbol{G}_i^t$ admits the following compact form

$$\boldsymbol{G}^t = \partial_{\boldsymbol{B}}\mathcal{F}(\boldsymbol{B}^t, V(\boldsymbol{B}^t;\mathcal{A}^{t,1});\mathcal{A}^{t,2}). \tag{17}$$

Suppose that $\mathcal{A}^{t,1} \simeq \mathcal{A}^{t,2} \simeq \mathcal{A}$ (recall that all these linear operators are comprised of i.i.d. data points $\boldsymbol{x}_{ij}$ and are hence similar when m is large). Further define the objective

$$\mathcal{G}(\boldsymbol{B};\mathcal{A}) = \min_{\boldsymbol{V}\in\mathbb{R}^{n\times k}} \mathcal{F}(\boldsymbol{B},\boldsymbol{V};\mathcal{A}). \tag{18}$$

We have $\boldsymbol{G}^t \simeq \nabla\mathcal{G}(\boldsymbol{B}^t;\mathcal{A})$ since

$$\nabla\mathcal{G}(\boldsymbol{B}^t;\mathcal{A}) = \partial_{\boldsymbol{B}}\mathcal{F}(\boldsymbol{B}^t, V(\boldsymbol{B}^t);\mathcal{A}) + \mathcal{J}_V(\boldsymbol{B}^t)^\top \partial_{\boldsymbol{V}}\mathcal{F}(\boldsymbol{B}^t, V(\boldsymbol{B}^t);\mathcal{A}) = \partial_{\boldsymbol{B}}\mathcal{F}(\boldsymbol{B}^t, V(\boldsymbol{B}^t;\mathcal{A});\mathcal{A}),$$

where $\mathcal{J}_V(\boldsymbol{B})$ denotes the Jacobian matrix of $V(\boldsymbol{B};\mathcal{A})$ with respect to $\boldsymbol{B}$ and the second equality holds due to the optimality of $V(\boldsymbol{B}^t;\mathcal{A})$. Consequently, conditioned on the event that all the clipping operation are inactive, `CENTAUR` behaves similar to the noisy gradient descent on the objective $\mathcal{G}(\boldsymbol{B};\mathcal{A})$ (up to the difference between $\mathcal{A}^{t,1}$, $\mathcal{A}^{t,1}$, and $\mathcal{A}$). In the following, we show that while the objective $\mathcal{G}(\boldsymbol{B};\mathcal{A})$ is non-convex globally, but we can show that `CENTAUR` converges locally within a region around the underlying ground truth.

One-step contraction To present our theory, we first establish the following properties that the operators $\mathcal{A}^{t,1}$ and $\mathcal{A}^{t,2}$ (defined in equation 10) satisfy. Recall that $\bar{N} = \bar{m}n = \sum_{i=1}^n |\mathcal{S}_i^{t,1}|$. The proofs are deferred to Appendix E.

Lemma C.2. *Under Assumption 5.1, the linear operator $\mathcal{A}^{t,1}$ satisfies the following property with probability at least $1-\exp(-c_1 k\log n)$:*

$$\sup_{\boldsymbol{V}\in\mathbb{R}^{n\times k}, \|\boldsymbol{V}\|_F=1} \left|\frac{1}{\bar{N}}\langle \mathcal{A}^{t,1}(\boldsymbol{B}^t\boldsymbol{V}^\top), \mathcal{A}^{t,1}(\boldsymbol{B}^t\boldsymbol{V}^\top)\rangle - 1\right| \le \delta^{(1)}.$$

Here, the factor $\delta^{(1)} = \sqrt{k\log n}/\sqrt{\bar{m}}$.

Lemma C.3. *Under Assumption 5.1, the linear operator $\mathcal{A}^{t,1}$ satisfies the following property with probability at least $1-\exp(-c_2 k\log n)$: For $\boldsymbol{V}_1, \boldsymbol{V}_2 \in \mathbb{R}^{n\times k}$,*

$$\sup_{\|\boldsymbol{V}_1\|_F=\|\boldsymbol{V}_2\|_F=1} \left|\frac{1}{\bar{N}}\langle \mathcal{A}^{t,1}((\boldsymbol{B}^t{\boldsymbol{B}^t}^\top - \boldsymbol{I}_d)\boldsymbol{B}^*\boldsymbol{V}_1^\top), \mathcal{A}^{t,1}(\boldsymbol{B}^t\boldsymbol{V}_2^\top)\rangle\right| \le \delta^{(2)}\mathrm{dist}(\boldsymbol{B}^t,\boldsymbol{B}^*).$$

Here, the factor $\delta^{(2)} = \sqrt{k\log n}/\sqrt{\bar{m}}$.

Lemma C.4. *Under Assumptions 5.1 and 5.2, the linear operator $\mathcal{A}^{t,2}$ satisfies the following property with probability at least $1-\exp(-c_3 k\log n)$: For $\boldsymbol{a}\in\mathbb{R}^d, \boldsymbol{b}\in\mathbb{R}^k$*

$$\sup_{\|\boldsymbol{a}\|=\|\boldsymbol{b}\|=1} \left|\frac{1}{\bar{N}}\langle \mathcal{A}^{t,2}(\boldsymbol{B}^t{\boldsymbol{V}^{t+1}}^\top - \boldsymbol{B}^*{\boldsymbol{V}^*}^\top), \mathcal{A}^{t,2}(\boldsymbol{a}\boldsymbol{b}^\top{\boldsymbol{V}^{t+1}}^\top)\rangle\right| \le \delta^{(3)}s_1^2 k\,\mathrm{dist}(\boldsymbol{B}^t,\boldsymbol{B}^*).$$

Here, the factor $\delta^{(3)} = 4(\sqrt{d}+\sqrt{k\log n})/\left(\sqrt{\bar{m}n}\kappa^2\right)$.

[remark why is the factor $\sqrt{d}$ not avoidable? We need to have the union bound over the epsilon net $\mathcal{N}_d$.]

We now present the one-step contraction lemma of `CENTAUR` in the LRL setting. The proof can be found in Appendix D.1.

Lemma C.5 (One-step contraction). *Consider the instance of `CENTAUR` for the LRL setting with its* CLIENT *and* INITIALIZATION *procedures defined in Algorithms 4 and 3 respectively. Suppose that the matrix $\boldsymbol{B}^0$ returned by* INITIALIZATION *satisfies $\mathrm{dist}(\boldsymbol{B}^0,\boldsymbol{B}^*) \le \epsilon_0 = 0.2$ and suppose that the mini-batch size parameter satisfies*

$$\bar{m} \ge c_m \max\{\kappa^2 k^2 \log n, \frac{k^2 d + k^3\log n}{n}\},$$

for some universal constant c_m. Set the clipping threshold of the Gaussian mechanism ζ_g according to Lemma C.1 and set the global step size $\eta_g = 1/4s_1^2$. Suppose that the level of manually injected noise is sufficiently small: For some universal constant c_σ, it satisfies

$$\sigma_g \le \frac{c_\sigma n}{\mu^2(\sqrt{d}+\sqrt{k\log n})}\min\left(\frac{1}{k^2\log n}, \frac{\kappa^4}{\sqrt{dk\log n}}\right). \tag{19}$$

We have the following one-step contraction from a single iteration of `CENTAUR`

$$\operatorname{dist}(\boldsymbol{B}^*, \boldsymbol{B}^{t+1}) \le \operatorname{dist}(\boldsymbol{B}^*, \boldsymbol{B}^{t})\sqrt{1-E_0/8\kappa^2} + 3C_{\mathcal{N}}\frac{\eta_g\sigma_g\zeta_g}{n}\sqrt{d}, \tag{20}$$

holds with probability at least $1 - c_p\bar{m}n^{-k}$, where c_p is some universal constant. Moreover, with the same probability, we also have

$$\operatorname{dist}(\boldsymbol{B}^*, \boldsymbol{B}^{t}) \le \operatorname{dist}(\boldsymbol{B}^*, \boldsymbol{B}^{0}). \tag{21}$$

Remark C.1. *The lower bound of the mini-batch size parameter $\bar{m}$ is derived to satisfy the following inequalities:*

$$\max\left\{\frac{\delta^{(2)}\sqrt{k}}{1-\delta^{(1)}}, \frac{(\delta^{(2)})^2k}{(1-\delta^{(1)})^2}, \delta^{(3)}k\right\} \le \frac{s_k^2(1-\epsilon_0^2)}{36s_1^2},$$

which is required to establish the above one-step contraction lemma. The upper bound of the noise multiplier σ_g is derived to satisfy the following inequalities:

$$\frac{\eta_g\sigma_g\zeta_g}{n} \le \min\left(\frac{\sqrt{1-\epsilon_0^2}}{4C_{\mathcal{N}}\kappa^2\sqrt{k\log n}}, \frac{8\kappa^2\epsilon_0}{3C_{\mathcal{N}}\sqrt{d}E_0}\right). \tag{22}$$

Proof of Theorem 5.1. Denote $E_0 = \sqrt{1-\epsilon_0^2}$. Using the recursion (20), we have

$$\begin{aligned}\operatorname{dist}(\boldsymbol{B}^*, \boldsymbol{B}^{t}) &\le (1-E_0/8\kappa^2)^{t/2}\operatorname{dist}(\boldsymbol{B}^*, \boldsymbol{B}^{0}) + 3C_{\mathcal{N}}\frac{\eta_g\sigma_g\zeta_g}{n}\sqrt{d}/(1-\sqrt{1-E_0/8\kappa^2}) \\ &\le (1-E_0/8\kappa^2)^{t/2}\operatorname{dist}(\boldsymbol{B}^*, \boldsymbol{B}^{0}) + 48C_{\mathcal{N}}\kappa^2\frac{\eta_g\sigma_g\zeta_g}{n}\sqrt{d}/E_0. \end{aligned}\tag{23}$$

With the choice of T specified in the theorem, we have that

$$\operatorname{dist}(\boldsymbol{B}^*, \boldsymbol{B}^{t}) \le c_d\kappa^2\frac{\eta_g\sigma_g\zeta_g}{n}\sqrt{d}, \tag{24}$$

for some universal constant c_d. By plugging the choices of η_g, ζ_g, we obtain the simplified bound

$$\operatorname{dist}(\boldsymbol{B}^*, \boldsymbol{B}^{t}) \le \tilde{c}_d\sigma_g\mu^2k^{1.5}d/n,$$

where $\tilde{c}_d$ hides the constant and $\log$ terms. ☐

D Proof of the One-step Contraction Lemma

D.1 Proof of Lemma C.5

Proof. The following discussion will be conditioned on the event that all the clipping operations are inactive, whose probability is proved to be at least $1 - 5n^{-k}$ in Lemma C.1. Using union bound to combine the following result and Lemma C.1 leads to the result.

Recall that the average gradient $\boldsymbol{G}^t = \frac{1}{n}\sum_{i=1}^n \boldsymbol{G}_i^t$ and denote $\boldsymbol{Q}^t = \boldsymbol{B}^t{\boldsymbol{V}^{t+1}}^\top - \boldsymbol{B}^*{\boldsymbol{V}^*}^\top$. Denote

$$\bar{\boldsymbol{B}}^{t+1} = \boldsymbol{B}^t + \eta_g\cdot\boldsymbol{G}^t. \tag{25}$$

Recall that $\mathbb{C}^t = \{1,\ldots,n\}$ given $p_g = 1$. Since the clipping operations are inactive, we have

$$\bar{\boldsymbol{B}}^{t+1} = \boldsymbol{B}^t - \eta_g\boldsymbol{Q}^t\boldsymbol{V}^{t+1} + \eta_g\left(\boldsymbol{Q}^t\boldsymbol{V}^{t+1} - \boldsymbol{G}^t\right) + \frac{\eta_g\sigma_g\zeta_g}{n}\boldsymbol{W}^t, \tag{26}$$

where $\boldsymbol{W}^t$ denotes the noise added by the Gaussian mechanism in the t^{th} global round. Note that ${\boldsymbol{B}_{\perp}^{*}}^{\top}\boldsymbol{Q}^t = {\boldsymbol{B}_{\perp}^{*}}^{\top}\boldsymbol{B}^t{\boldsymbol{V}^{t+1}}^{\top}$, and denote the QR decomposition $\bar{\boldsymbol{B}}^{t+1} = \boldsymbol{B}^{t+1}\boldsymbol{R}^{t+1}$. We have

$$\begin{aligned}
&{\boldsymbol{B}_{\perp}^{*}}^{\top}\boldsymbol{B}^{t+1} \\
&= {\boldsymbol{B}_{\perp}^{*}}^{\top}\left(\boldsymbol{B}^t - \eta_g\boldsymbol{Q}^t\boldsymbol{V}^{t+1} + \eta_g\left(\boldsymbol{Q}^t\boldsymbol{V}^{t+1} - \boldsymbol{G}^t\right) + \eta_g\sigma_g\zeta_g\boldsymbol{W}^t\right)\left(\boldsymbol{R}^{t+1}\right)^{-1} \\
&= {\boldsymbol{B}_{\perp}^{*}}^{\top}\boldsymbol{B}^t\left(\boldsymbol{I}_k - \eta_g{\boldsymbol{V}^{t+1}}^{\top}\boldsymbol{V}^{t+1}\right)\left(\boldsymbol{R}^{t+1}\right)^{-1} \\
&\quad + \eta_g{\boldsymbol{B}_{\perp}^{*}}^{\top}\left(\boldsymbol{Q}^t\boldsymbol{V}^{t+1} - \boldsymbol{G}^t\right)\left(\boldsymbol{R}^{t+1}\right)^{-1} + \frac{\eta_g\sigma_g\zeta_g}{n}{\boldsymbol{B}_{\perp}^{*}}^{\top}\boldsymbol{W}^t(\boldsymbol{R}^{t+1})^{-1}.
\end{aligned}$$

Recall the definition $\mathrm{dist}(\boldsymbol{B}^*, \boldsymbol{B}^t) = \|{\boldsymbol{B}_{\perp}^{*}}^{\top}\boldsymbol{B}^t\|$. We bound

$$\begin{aligned}
\mathrm{dist}(\boldsymbol{B}^*, \boldsymbol{B}^{t+1}) \leq{}& \mathrm{dist}(\boldsymbol{B}^*, \boldsymbol{B}^t)\|\boldsymbol{I}_k - \eta_g{\boldsymbol{V}^{t+1}}^{\top}\boldsymbol{V}^{t+1}\|\|\left(\boldsymbol{R}^{t+1}\right)^{-1}\| \\
&+ \eta\|\boldsymbol{Q}^t\boldsymbol{V}^{t+1} - \boldsymbol{G}^t\|\|\left(\boldsymbol{R}^{t+1}\right)^{-1}\| + \frac{\eta_g\sigma_g\zeta_g}{n}\|{\boldsymbol{B}_{\perp}^{*}}^{\top}\boldsymbol{W}^t\|\|\left(\boldsymbol{R}^{t+1}\right)^{-1}\|. \qquad (27)
\end{aligned}$$

In the following, we show that the factor $\|\boldsymbol{I}_k - \eta_g{\boldsymbol{V}^{t+1}}^{\top}\boldsymbol{V}^{t+1}\|\|\left(\boldsymbol{R}^{t+1}\right)^{-1}\| < 1$ which leads to a contraction in the principal angle distance and treat the rest two terms as controllable noise for sufficiently small constants $(\delta^{(1)}, \delta^{(2)}, \delta^{(3)})$ (see Lemma C.1) and a sufficiently smaller noise multiplier σ_g (see equation 19).

Bound $\|\boldsymbol{I}_k - \eta_g{\boldsymbol{V}^{t+1}}^{\top}\boldsymbol{V}^{t+1}\|$. Recall that $\eta_g = 1/4s_1^2$. Using Lemma D.5, we have

$$\|\boldsymbol{I}_k - \eta_g{\boldsymbol{V}^{t+1}}^{\top}\boldsymbol{V}^{t+1}\| \leq 1 - \eta_g(E_0 s_k^2 - \frac{2\delta^{(2)}s_1^2\sqrt{k}}{1-\delta^{(1)}}\mathrm{dist}(\boldsymbol{B}^*, \boldsymbol{B}^t)) \leq 1 - \eta_g E_0 s_k^2 \cdot 0.75, \quad (28)$$

where we use the following inequality in Remark C.1

$$\frac{2\delta^{(2)}s_1^2\sqrt{k}}{1-\delta^{(1)}}\mathrm{dist}(\boldsymbol{B}^*, \boldsymbol{B}^t) \leq \frac{2\delta^{(2)}s_1^2\sqrt{k}}{1-\delta^{(1)}} \leq E_0 s_k^2/4. \quad (29)$$

Bound $\|\left(\boldsymbol{R}^{t+1}\right)^{-1}\|$. With the choice of $\bar{m}$ stated in the lemma, the tuple $(\delta^{(1)}, \delta^{(2)}, \delta^{(3)})$ satisfies the requirements in Remark C.1. Using Lemma D.7, we have with probability at least $1 - 4n^{-k}$

$$\|(\boldsymbol{R}^{t+1})^{-1}\| \leq 1/\sqrt{1 - \eta_g s_k^2 E_0/2}. \quad (30)$$

Combining equation 28 and equation 30, we have the contraction

$$\|\boldsymbol{I}_k - \eta_g{\boldsymbol{V}^{t+1}}^{\top}\boldsymbol{V}^{t+1}\|\|(\boldsymbol{R}^{t+1})^{-1}\| \leq \frac{1 - \eta_g s_k^2 E_0 \cdot 0.75}{\sqrt{1 - \eta_g s_k^2 E_0 \cdot 0.5}} < 1.$$

We now bound the last two terms of equation 27.

Bound $\|\boldsymbol{Q}^t\boldsymbol{V}^{t+1} - \boldsymbol{G}^t\|$. Using Lemma D.6, we have

$$\|\boldsymbol{Q}^t\boldsymbol{V}^{t+1} - \boldsymbol{G}^t\| \leq \delta^{(3)}s_1^2 k\,\mathrm{dist}(\boldsymbol{B}^*, \boldsymbol{B}^t) \leq E_0 s_k^2\,\mathrm{dist}(\boldsymbol{B}^*, \boldsymbol{B}^t) \cdot 0.25, \quad (31)$$

where we use condition in Remark C.1 $\delta^{(3)}s_1^2 k \leq E_0 s_k^2/4$.

Bound ${\boldsymbol{U}_{\perp}^{*}}^{\top}\boldsymbol{W}^t$. Due to the rotational invariance of independent Gaussian random variables, every entry in ${\boldsymbol{B}_{\perp}^{*}}^{\top}\boldsymbol{W}^t \in \mathbb{R}^{(n-k)\times k}$ is distributed as $\mathcal{N}(0,1)$. Using Lemma I.1, we have with probability at least $1 - n^{-k}$

$$\|{\boldsymbol{B}_{\perp}^{*}}^{\top}\boldsymbol{W}^t\| \leq C_{\mathcal{N}}(\sqrt{d-k} + \sqrt{k} + \sqrt{\ln(k\log n)}) \leq 3C_{\mathcal{N}}\sqrt{d}, \quad (32)$$

where we assume $d = \Omega(\log\log n)$ to simplify the above bound.

Final Result. Combining the above bounds, we conclude that the following one-step contraction holds under the assumptions stated in the theorem.

$$\begin{aligned}\operatorname{dist}(\boldsymbol{B}^*, \boldsymbol{B}^{t+1}) &\leq \operatorname{dist}(\boldsymbol{B}^*, \boldsymbol{B}^t)\frac{1-\eta_g E_0 s_k^2 \cdot (0.75-0.25)}{\sqrt{1-\eta_g E_0 s_k^2 \cdot 0.5}} + 3C_{\mathcal{N}}\frac{\eta_g \sigma_g \zeta_g}{n}\sqrt{d} \\ &\leq \operatorname{dist}(\boldsymbol{B}^*, \boldsymbol{B}^t)\sqrt{1-\eta_g E_0 s_k^2 \cdot 0.5} + 3C_{\mathcal{N}}\frac{\eta_g \sigma_g \zeta_g}{n}\sqrt{d}\end{aligned}$$

□

D.2 LEMMAS FOR THE UTILITY ANALYSIS

Lemma D.1. *Use* $vec(\cdot)$ *and* $\otimes$ *to denote the standard vectorization operation and Kronecker product respectively. Recall that* $\boldsymbol{V}^{t+1} := \arg\min_{\boldsymbol{V}\in\mathbb{R}^{n\times k}} \mathcal{F}(\boldsymbol{B}^t, \boldsymbol{V}; \mathcal{A}^{t,1})$*. We have*

$$vec(\boldsymbol{V}^{t+1}) = vec((\boldsymbol{B}^*\boldsymbol{V}^{*\top})^\top \boldsymbol{B}^t) - \boldsymbol{f}^t \tag{33}$$

where

$$\boldsymbol{f}^t = \left(\boldsymbol{B}^{t\top}\boldsymbol{B}^* \otimes \boldsymbol{I}_d - \boldsymbol{H}^{tt^{-1}}\boldsymbol{H}^{t*}\right) vec(\boldsymbol{V}^*) = \boldsymbol{H}^{tt^{-1}}\left(\boldsymbol{H}^{tt}(\boldsymbol{B}^{t\top}\boldsymbol{B}^* \otimes \boldsymbol{I}_d) - \boldsymbol{H}^{t*}\right) vec(\boldsymbol{V}^*). \tag{34}$$

Here we denote

$$\boldsymbol{H}^{t*} = \frac{1}{n}\sum_{i=1}^n \frac{1}{\bar{m}}\sum_{j\in\mathbb{S}_i^{t,1}} vec(\boldsymbol{A}_{ij}^\top \boldsymbol{B}^t)\, vec(\boldsymbol{A}_{ij}^\top \boldsymbol{B}^*)^\top \in \mathbb{R}^{nk\times nk}, \tag{35}$$

$$\boldsymbol{H}^{tt} = \frac{1}{n}\sum_{i=1}^n \frac{1}{\bar{m}}\sum_{j\in\mathbb{S}_i^{t,1}} vec(\boldsymbol{A}_{ij}^\top \boldsymbol{B}^t)\, vec(\boldsymbol{A}_{ij}^\top \boldsymbol{B}^t)^\top \in \mathbb{R}^{nk\times nk}. \tag{36}$$

We use $\boldsymbol{F}^t \in \mathbb{R}^{n\times k}$ *to denote the matrix satisfying* $\boldsymbol{f}^t = vec(\boldsymbol{F}^t)$*.*

Proof. Recall that $\bar{N} = \bar{m}n$. For the simplicity of notations, we use the collection $\{\boldsymbol{A}_l\}$, $l = 1, \ldots, \bar{N}$ to denote $\{\boldsymbol{A}_{ij}\}$, $i \in [n]$, $j \in \mathcal{S}_i^{t,1}$ (there exists a one-to-one mapping between the indices l and (i, j)). Compute that for any $\boldsymbol{B} \in \mathbb{R}^{d\times k}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times k}$

$$\begin{aligned}&vec(\partial_{\boldsymbol{V}}\mathcal{F}(\boldsymbol{B}, \boldsymbol{V})) \\ &= \frac{1}{\bar{N}}\sum_{l=1}^{\bar{N}}(vec(\boldsymbol{A}_l^\top \boldsymbol{B})^\top vec(\boldsymbol{V}) - vec(\boldsymbol{A}_l^\top \boldsymbol{B}^*)^\top vec(\boldsymbol{V}^*))\, vec(\boldsymbol{A}_l^\top \boldsymbol{B}) \\ &= \left(\frac{1}{\bar{N}}\sum_{l=1}^{\bar{N}} vec(\boldsymbol{A}_l^\top \boldsymbol{B})\, vec(\boldsymbol{A}_l^\top \boldsymbol{B})^\top\right) vec(\boldsymbol{V}) - \left(\frac{1}{\bar{N}}\sum_{l=1}^{\bar{N}} vec(\boldsymbol{A}_l^\top \boldsymbol{B})\, vec(\boldsymbol{A}_l^\top \boldsymbol{B}^*)^\top\right) vec(\boldsymbol{V}^*).\end{aligned}$$

Since $\boldsymbol{V}^{t+1} = \arg\min_{\boldsymbol{V}\in\mathbb{R}^{n\times k}} \mathcal{F}(\boldsymbol{B}^t, \boldsymbol{V}; \mathcal{A}^{t,1})$, we have $\partial_{\boldsymbol{V}}\mathcal{F}(\boldsymbol{B}^t, \boldsymbol{V}^{t+1}; \mathcal{A}^{t,1}) = 0$ and hence

$$\begin{aligned}vec(\boldsymbol{V}^{t+1}) &= \left(\frac{1}{\bar{N}}\sum_{l=1}^{\bar{N}} vec(\boldsymbol{A}_l^\top \boldsymbol{B}^t)\, vec(\boldsymbol{A}_l^\top \boldsymbol{B}^t)^\top\right)^{-1}\left(\frac{1}{\bar{N}}\sum_{l=1}^{\bar{N}} vec(\boldsymbol{A}_l^\top \boldsymbol{B}^t)\, vec(\boldsymbol{A}_l^\top \boldsymbol{B}^*)^\top\right) vec(\boldsymbol{V}^*) \\ &= \boldsymbol{H}^{tt^{-1}}\boldsymbol{H}^{t*}\, vec(\boldsymbol{V}^*),\end{aligned}$$

where we denote

$$\boldsymbol{H}^{t*} = \frac{1}{\bar{N}}\sum_{l=1}^{\bar{N}} vec(\boldsymbol{A}_l^\top \boldsymbol{B}^t)\, vec(\boldsymbol{A}_l^\top \boldsymbol{B}^*)^\top \in \mathbb{R}^{nk\times nk},$$

$$\boldsymbol{H}^{tt} = \frac{1}{\bar{N}}\sum_{l=1}^{\bar{N}} vec(\boldsymbol{A}_l^\top \boldsymbol{B}^t)\, vec(\boldsymbol{A}_l^\top \boldsymbol{B}^t)^\top \in \mathbb{R}^{nk\times nk}.$$

Use $\otimes$ to denote the Kronecker product of matrices. Recall that

$$(\boldsymbol{B}^\top \otimes \boldsymbol{A})\, vec(\boldsymbol{X}) = vec(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B}). \tag{37}$$

We have

$$vec\left((\boldsymbol{B}^*{\boldsymbol{V}^*}^\top)^\top \boldsymbol{B}^t\right) = vec\left(\boldsymbol{V}^*{\boldsymbol{B}^*}^\top \boldsymbol{B}^t\right) = ({\boldsymbol{B}^t}^\top \boldsymbol{B}^* \otimes I_d)\, vec(\boldsymbol{V}^*). \tag{38}$$

$\square$

Lemma D.2. *Recall the definition of $\boldsymbol{H}^{tt}$ in equation 36 and that $s_{\min}$ denotes the minimum singular value of a matrix. Suppose that the matrix sensing operator $\mathcal{A}^{t,1}$ satisfies Condition C.2 with constant $\delta^{(1)}$. We can bound*

$$s_{\min}(\boldsymbol{H}^{tt}) \geq 1 - \delta^{(1)}.$$

Proof. From the definition of $s_{\min}$, we have

$$s_{\min}(\boldsymbol{H}^{tt}) = \min_{\substack{\boldsymbol{P}\in\mathbb{R}^{n\times k}\\ \|\boldsymbol{P}\|_F=1}} vec(\boldsymbol{P})^\top \boldsymbol{H}^{tt}\, vec(\boldsymbol{P}) = \frac{1}{\bar{N}}\langle \mathcal{A}^{t,1}(\boldsymbol{B}^t\boldsymbol{P}^\top), \mathcal{A}^{t,1}(\boldsymbol{B}^t\boldsymbol{P}^\top)\rangle \geq 1-\delta^{(1)}.$$

$\square$

Lemma D.3. *Recall the definitions of $\boldsymbol{H}^{tt}$ and $\boldsymbol{H}^{t*}$ in equation 36 and equation 35 and recall that $\mathrm{dist}(\boldsymbol{B}^t, \boldsymbol{B}^*)$ is the principal angle distance between the current variable $\boldsymbol{B}^t$ and the ground truth $\boldsymbol{B}^*$. Suppose that the matrix sensing operator $\mathcal{A}^{t,1}$ satisfies Condition C.3 with constant $\delta^{(2)}$. We can bound*

$$\|\boldsymbol{H}^{tt}({\boldsymbol{B}^t}^\top \boldsymbol{B}^* \otimes I_d) - \boldsymbol{H}^{t*}\|_2 \leq \delta^{(2)}\mathrm{dist}(\boldsymbol{B}^t, \boldsymbol{B}^*). \tag{39}$$

Proof. Recall that $\bar{N} = \bar{m}n$. For the simplicity of notations, we use the collection $\{\boldsymbol{A}_l\}$, $l = 1, \ldots, \bar{N}$ to denote $\{\boldsymbol{A}_{ij}\}$, $i \in [n]$, $j \in \mathcal{S}_i^{t,1}$ (there can be a one-to-one mapping between the indices l and (i,j)). For arbitrary $\boldsymbol{W} \in \mathbb{R}^{n\times k}, \boldsymbol{P} \in \mathbb{R}^{n\times k}$ with $\|\boldsymbol{W}\|_F = \|\boldsymbol{P}\|_F = 1$, we have

$$\begin{aligned}
vec(\boldsymbol{W})^\top \boldsymbol{H}^{tt}({\boldsymbol{B}^t}^\top \boldsymbol{B}^* \otimes \boldsymbol{I}_d)\, vec(\boldsymbol{P}) &= \frac{1}{\bar{N}}\sum_{i=1}^{\bar{N}} vec(\boldsymbol{W})^\top\, vec(\boldsymbol{A}_i^\top \boldsymbol{B}^t)\, vec(\boldsymbol{A}_i^\top \boldsymbol{B}^t)^\top\, vec(\boldsymbol{P}{\boldsymbol{B}^*}^\top \boldsymbol{B}^t)\\
&= \frac{1}{\bar{N}}\sum_{i=1}^{\bar{N}} \langle \boldsymbol{A}_i, \boldsymbol{B}^t\boldsymbol{W}^\top\rangle\langle \boldsymbol{A}_i, \boldsymbol{B}^t{\boldsymbol{B}^t}^\top \boldsymbol{B}^*\boldsymbol{P}^\top\rangle\\
&= \frac{1}{\bar{N}}\langle \mathcal{A}^{t,1}(\boldsymbol{B}^t\boldsymbol{W}^\top), \mathcal{A}^{t,1}(\boldsymbol{B}^t{\boldsymbol{B}^t}^\top\boldsymbol{B}^*\boldsymbol{P}^\top)\rangle,
\end{aligned}$$

where we use $(\boldsymbol{B}^\top \otimes \boldsymbol{A})\, vec(\boldsymbol{X}) = vec(\boldsymbol{A}\boldsymbol{X}\boldsymbol{B})$ in the first equality. Similarly, we can compute that

$$vec(\boldsymbol{W})^\top \boldsymbol{H}^{t*}\, vec(\boldsymbol{P}) = \frac{1}{\bar{N}}\langle \mathcal{A}^{t,1}(\boldsymbol{B}^t\boldsymbol{W}^\top), \mathcal{A}^{t,1}(\boldsymbol{B}^*\boldsymbol{P}^\top)\rangle, \tag{40}$$

and hence

$$vec(\boldsymbol{W})^\top \left(\boldsymbol{H}^{tt}({\boldsymbol{B}^t}^\top\boldsymbol{B}^* \otimes \boldsymbol{I}_d) - \boldsymbol{H}^{t*}\right) vec(\boldsymbol{P}) = \frac{1}{\bar{N}}\langle \mathcal{A}^{t,1}(\boldsymbol{B}^t W^\top), \mathcal{A}^{t,1}((\boldsymbol{B}^t{\boldsymbol{B}^t}^\top - \boldsymbol{I}_k)\boldsymbol{B}^*\boldsymbol{P}^\top)\rangle. \tag{41}$$

Using Condition C.3 and the definition of $s_{\max}$, we have the result. $\square$

Lemma D.4. *Recall the definitions of $\boldsymbol{F}^t$ and f^t in equation 34 and recall that $\mathrm{dist}(\boldsymbol{B}^t, \boldsymbol{B}^*)$ is the principal angle distance between the current variable $\boldsymbol{B}^t$ and the ground truth $\boldsymbol{B}^*$. Suppose that the matrix sensing operator $\mathcal{A}$ satisfies Conditions C.2 and C.3 with constants $\delta^{(1)}$ and $\delta^{(2)}$ respectively. We can bound*

$$\|\boldsymbol{F}_t\|_F = \|\boldsymbol{f}^t\|_2 \leq \frac{\delta^{(2)}s_1\sqrt{k}}{1-\delta^{(1)}}\mathrm{dist}(\boldsymbol{B}^*, \boldsymbol{B}^t), \tag{42}$$

and

$$\|\boldsymbol{V}^{t+1}\|_F \leq \sqrt{k}s_1 + \frac{\delta^{(2)}s_1\sqrt{k}}{1-\delta^{(1)}}\mathrm{dist}(\boldsymbol{B}^*, \boldsymbol{B}^t) \leq \frac{s_1\sqrt{k}(1-\delta^{(1)}+\delta^{(2)})}{1-\delta^{(1)}}. \tag{43}$$

Lemma D.5. *Suppose that the matrix sensing operator $\mathcal{A}^{t,1}$ satisfies Conditions C.2 and C.3 with constants $\delta^{(1)}$ and $\delta^{(2)}$ respectively. Further, suppose that* $\max\{\frac{\delta^{(2)}\sqrt{k}}{1-\delta^{(1)}}, \frac{(\delta^{(2)})^2 k}{(1-\delta^{(1)})^2}\} \leq \frac{s_k^2 E_0}{36 s_1^2}$*. For a sufficiently small step size $\eta_g \leq 1/(4s_1^2)$, we have $\boldsymbol{I}_k - \eta_g \boldsymbol{V}^{t+1\top}\boldsymbol{V}^{t+1} \succcurlyeq 0$. Moreover, we can bound*

$$\|\boldsymbol{I}_k - \eta_g \boldsymbol{V}^{t+1\top}\boldsymbol{V}^{t+1}\|_2 \leq 1 - \eta_g(E_0 s_k^2 - \frac{2\delta^{(2)} s_1^2 \sqrt{k}}{1-\delta^{(1)}}\mathrm{dist}(\boldsymbol{B}^*, \boldsymbol{B}^t)). \tag{44}$$

Proof. We first show that the matrix $\boldsymbol{I}_k - \eta_g \boldsymbol{V}^{t+1\top}\boldsymbol{V}^{t+1}$ is positive semi-definite for a sufficiently small step-size η_g. Note that following the idea of the seminal work Jain et al. (2013), the update of $\boldsymbol{V}^{t+1} = V(\boldsymbol{B}^t; \mathcal{A}^{t,1})$ can be regarded as a noisy power iteration, as detailed in Lemma D.1. This allows us to compute

$$\begin{aligned}\|\boldsymbol{V}^{t+1\top}\boldsymbol{V}^{t+1}\|_2 &= \|\boldsymbol{V}^*\boldsymbol{B}^{*\top}\boldsymbol{B}^t - \boldsymbol{F}^t\|_2^2 \leq 2\|\boldsymbol{V}^*\boldsymbol{B}^{*\top}\boldsymbol{B}^t\|_2^2 + 2\|\boldsymbol{F}^t\|_2^2 \\ &\leq 2s_1^2 + 2\left(\frac{\delta^{(2)} s_1\sqrt{k}}{1-\delta^{(1)}}\mathrm{dist}(\boldsymbol{B}^*, \boldsymbol{B}^t)\right)^2 \leq 4s_1^2,\end{aligned}$$

where we use Lemma D.4 in the first inequality and use $\frac{\delta^{(2)}\sqrt{k}}{1-\delta^{(1)}} \leq 1$ in the last. Consequently, for $\eta_g \leq \frac{1}{4s_1^2}$, $\boldsymbol{I}_k - \eta_g \boldsymbol{V}^{t+1\top}\boldsymbol{V}^{t+1} \succcurlyeq 0$.

Given the positive semi-definiteness of the matrix $\boldsymbol{I}_k - \eta_g \boldsymbol{V}^{t+1\top}\boldsymbol{V}^{t+1}$, we can bound

$$\|\boldsymbol{I}_k - \eta_g \boldsymbol{V}^{t+1\top}\boldsymbol{V}^{t+1}\|_2 \leq 1 - \eta_g s_{\min}(\boldsymbol{V}^{t+1\top}\boldsymbol{V}^{t+1}).$$

By using Lemma D.1 again. We have

$$\begin{aligned}\boldsymbol{V}^{t+1\top}\boldsymbol{V}^{t+1} &= \left(\boldsymbol{V}^*\boldsymbol{B}^{*\top}\boldsymbol{B}^t - \boldsymbol{F}^t\right)^\top \left(\boldsymbol{V}^*\boldsymbol{B}^{*\top}\boldsymbol{B}^t - \boldsymbol{F}^t\right) \\ &= \boldsymbol{B}^{t\top}\boldsymbol{B}^*\boldsymbol{V}^{*\top}\boldsymbol{V}^*\boldsymbol{B}^{*\top}\boldsymbol{B}^t - \boldsymbol{F}^{t\top}\boldsymbol{V}^*\boldsymbol{B}^{*\top}\boldsymbol{B}^t - \boldsymbol{B}^{t\top}\boldsymbol{B}^*\boldsymbol{V}^{*\top}\boldsymbol{F}^t + \boldsymbol{F}^{t\top}\boldsymbol{F}^t.\end{aligned}$$

Note that $\boldsymbol{F}^{t\top}\boldsymbol{F}^t$ is PSD which makes nonnegative contribution to $s_{\min}(\boldsymbol{V}^{t+1\top}\boldsymbol{V}^{t+1})$ and hence

$$\begin{aligned}s_{\min}(\boldsymbol{V}^{t+1\top}\boldsymbol{V}^{t+1}) &\geq s_{\min}(\boldsymbol{B}^{t\top}\boldsymbol{B}^*\boldsymbol{V}^{*\top}\boldsymbol{V}^*\boldsymbol{B}^{*\top}\boldsymbol{B}^t - \boldsymbol{F}^{t\top}\boldsymbol{V}^*\boldsymbol{B}^{*\top}\boldsymbol{B}^t - \boldsymbol{B}^{t\top}\boldsymbol{B}^*\boldsymbol{V}^{*\top}\boldsymbol{F}^t) \\ &\geq s_{\min}(\boldsymbol{B}^{t\top}\boldsymbol{B}^*\boldsymbol{V}^{*\top}\boldsymbol{V}^*\boldsymbol{B}^{*\top}\boldsymbol{B}^t) - 2s_{\max}(\boldsymbol{F}^{t\top}\boldsymbol{V}^*\boldsymbol{B}^{*\top}\boldsymbol{B}^t) \\ &\geq s_{\min}^2(\boldsymbol{B}^{t\top}\boldsymbol{B}^*)s_{\min}^2(\boldsymbol{V}^*) - 2\|\boldsymbol{F}^t\|s_{\max}(\boldsymbol{V}^*)\end{aligned} \tag{45}$$

To bound the first term of Eq. equation 45, recall that $s_{\min}^2(\boldsymbol{B}^{t\top}\boldsymbol{B}^*) = 1 - (\mathrm{dist}(\boldsymbol{B}^*, \boldsymbol{B}^t))^2$ from Eq. equation 6 and $\mathrm{dist}(\boldsymbol{B}^*, \boldsymbol{B}^t) \leq \mathrm{dist}(\boldsymbol{B}^*, \boldsymbol{B}^0)$ from the induction, we have $s_{\min}^2(\boldsymbol{B}^{t\top}\boldsymbol{B}^*) \geq E_0$. To bound the last term of Eq. equation 45, we use Lemma D.4 to obtain $\|\boldsymbol{F}^t\|_2 \leq \|\boldsymbol{F}^t\|_F \leq \frac{\delta^{(2)} s_1\sqrt{k}}{1-\delta^{(1)}}\mathrm{dist}(\boldsymbol{B}^*, \boldsymbol{B}^t)$. Combining the above results, we have

$$\|\boldsymbol{I}_k - \eta_g \boldsymbol{V}^{t+1\top}\boldsymbol{V}^{t+1}\|_2 \leq 1 - \eta_g(E_0 s_k^2 - \frac{2\delta^{(2)} s_1^2 \sqrt{k}}{1-\delta^{(1)}}\mathrm{dist}(\boldsymbol{B}^*, \boldsymbol{B}^t)). \tag{46}$$

□

Lemma D.6. *Recall that $\boldsymbol{V}^{t+1} = V(\boldsymbol{B}^t; \mathcal{A}^{t,1})$ (see the definition of $V(\cdot; \mathcal{A})$ in Eq. equation 16), $\boldsymbol{Q}^t = \boldsymbol{B}^t\boldsymbol{V}^{t+1\top} - \boldsymbol{B}^*\boldsymbol{V}^{*\top}$, $\boldsymbol{G}^t = \frac{1}{n}\sum_{i=1}^n \boldsymbol{G}_i^t$ is the global average of local gradient, and recall that $\mathrm{dist}(\boldsymbol{B}^t, \boldsymbol{B}^*)$ is the principal angle distance between the current variable $\boldsymbol{B}^t$ and the ground truth $\boldsymbol{B}^*$. Suppose that the matrix sensing operator $\mathcal{A}^{t,2}$ satisfies Condition C.4 with a constant $\delta^{(3)}$. We have*

$$\|\boldsymbol{G}^t - \boldsymbol{Q}^t\boldsymbol{V}^{t+1}\|_2 \leq \delta^{(3)} s_1^2 k \mathrm{dist}(\boldsymbol{B}^*, \boldsymbol{B}^t). \tag{47}$$

Proof. Recall that $\bar{N} = \bar{m}n$. For the simplicity of notations, we use the collection $\{\boldsymbol{A}_l\}$, $l = 1, \ldots, \bar{N}$ to denote $\{\boldsymbol{A}_{ij}\}$, $i \in [n]$, $j \in \mathcal{S}_i^{t,2}$ (there can be a one-to-one mapping between the indices l and (i,j)). With this notation, we can compactly write $\boldsymbol{G}^t$ as

$$\boldsymbol{G}^t = \frac{1}{\bar{N}}\sum_{l=1}^{\bar{N}} \langle \boldsymbol{A}_l, \boldsymbol{Q}^t\rangle \boldsymbol{A}_l \boldsymbol{V}^{t+1}. \tag{48}$$

From the definition of $s_{\max}$, for any $\boldsymbol{a} \in \mathbb{R}^d$ with $\|\boldsymbol{a}\|_2 = 1$ and any $\boldsymbol{b} \in \mathbb{R}^k$ with $\|\boldsymbol{b}\|_2 = 1$, we have

$$\|\frac{1}{\bar{N}}\sum_{l=1}^{\bar{N}}\langle \boldsymbol{A}_l, \boldsymbol{Q}^t\rangle \boldsymbol{A}_l \boldsymbol{V}^{t+1} - \boldsymbol{Q}^t\boldsymbol{V}^{t+1}\|_2 = \max_{\|\boldsymbol{a}\|_2=\|\boldsymbol{b}\|_2=1} \boldsymbol{a}^\top \left(\frac{1}{N}\sum_{l=1}^{\bar{N}}\langle \boldsymbol{A}_i, \boldsymbol{Q}^t\rangle \boldsymbol{A}_l \boldsymbol{V}^{t+1} - \boldsymbol{Q}^t\boldsymbol{V}^{t+1}\right)\boldsymbol{b}.$$

We obtain the result from Condition C.4

$$|\frac{1}{\bar{N}}\langle \mathcal{A}^{t,2}(\boldsymbol{B}^t{\boldsymbol{V}^{t+1}}^\top - \boldsymbol{B}^*{\boldsymbol{V}^*}^\top), \mathcal{A}^{t,2}(\boldsymbol{a}\boldsymbol{b}^\top{\boldsymbol{V}^{t+1}}^\top)\rangle| \leq \delta^{(3)}s_1^2k\mathrm{dist}(\boldsymbol{B}^t, \boldsymbol{B}^*). \tag{49}$$

□

Lemma D.7. *Recall that $\bar{\boldsymbol{B}}^{t+1} = \boldsymbol{B}^{t+1}\boldsymbol{R}^{t+1}$ is the QR decomposition of $\bar{\boldsymbol{B}}^{t+1}$. Denote $E_0 = 1-\epsilon_0^2$ and $\sigma = \zeta_g\sigma_g\eta_g/n$. Suppose that Conditions C.2 to C.4 are satisfied with constants $\delta^{(1)}$, $\delta^{(2)}$, and $\delta^{(3)}$. Further, suppose that $\max\{\frac{\delta^{(2)}\sqrt{k}}{1-\delta^{(1)}}, \frac{(\delta^{(2)})^2k}{(1-\delta^{(1)})^2}, \delta^{(3)}k\} \leq \frac{s_k^2E_0}{36s_1^2}$ and the level of manually injected noise is sufficiently small $\sigma \leq \frac{E_0}{4C_{\mathcal{N}}\kappa^2\sqrt{k\log n}}$. We have with probability at least $1-4\exp(-k\log n)$*

$$\|(\boldsymbol{R}^{t+1})^{-1}\|_2 \leq \frac{1}{\sqrt{1-\eta_g s_k^2E_0/2}}. \tag{50}$$

Proof. We now focus on bounding $s_{\min}(\boldsymbol{R}^{t+1})$. Recall that $\boldsymbol{G}^t = \partial_{\boldsymbol{B}}\mathcal{F}(\boldsymbol{B}^t, \boldsymbol{V}^{t+1}; \mathcal{A}^{t,2})$ with $\boldsymbol{V}^{t+1} = V(\boldsymbol{B}^t; \mathcal{A}^{t,1})$ (see the definition of $V(\cdot;\mathcal{A})$ in equation 16) and recall that $\bar{\boldsymbol{B}}^{t+1} := \boldsymbol{B}^t - \eta_g\boldsymbol{G}^t + \sigma\boldsymbol{W}^t$. Compute

$$\begin{aligned}
&{\boldsymbol{R}^{t+1}}^\top\boldsymbol{R}^{t+1} = {\bar{\boldsymbol{B}}^{t+1}}{}^\top\bar{\boldsymbol{B}}^{t+1}\\
&= \boldsymbol{I}_k + \eta_g^2{\boldsymbol{G}^t}^\top\boldsymbol{G}^t + \sigma^2{\boldsymbol{W}^t}^\top\boldsymbol{W}^t - \eta_g{\boldsymbol{B}^t}^\top\boldsymbol{G}^t - \eta_g{\boldsymbol{G}^t}^\top\boldsymbol{B}^t + \sigma{\boldsymbol{B}^t}^\top\boldsymbol{W}^t + \sigma{\boldsymbol{W}^t}^\top\boldsymbol{B}^t - \eta_g\sigma{\boldsymbol{G}^t}^\top\boldsymbol{W}^t - \eta_g\sigma{\boldsymbol{W}^t}^\top\boldsymbol{G}^t.
\end{aligned}$$

Therefore, we have

$$s_{\min}({\boldsymbol{R}^{t+1}}^\top\boldsymbol{R}^{t+1}) \geq 1 - 2\eta_g s_{\max}\left({\boldsymbol{B}^t}^\top\boldsymbol{G}^t\right) - 2\sigma s_{\max}({\boldsymbol{B}^t}^\top\boldsymbol{W}^t) - 2\eta_g\sigma s_{\max}\left({\boldsymbol{G}^t}^\top\boldsymbol{W}^t\right). \tag{51}$$

We now bound the last three terms of the R.H.S. of the above inequality.

1. $s_{\max}\left({\boldsymbol{B}^t}^\top\boldsymbol{G}^t\right)$: To bound $s_{\max}\left({\boldsymbol{B}^t}^\top\boldsymbol{G}^t\right)$, compute that

$${\boldsymbol{B}^t}^\top\boldsymbol{G}^t = {\boldsymbol{B}^t}^\top\boldsymbol{Q}^t\boldsymbol{V}^{t+1} + {\boldsymbol{B}^t}^\top\left(\boldsymbol{G}^t - \boldsymbol{Q}^t\boldsymbol{V}^{t+1}\right), \tag{52}$$

where we recall the definition of $\boldsymbol{Q}^t$ as $\boldsymbol{Q}^t = \boldsymbol{B}^t{\boldsymbol{V}^{t+1}}^\top - \boldsymbol{B}^*{\boldsymbol{V}^*}^\top$. Note that the spectral norm of second term in Eq. equation 52 is bounded by Lemma D.6 and we hence focus on the spectral norm of the first term ${\boldsymbol{B}^t}^\top\boldsymbol{Q}^t\boldsymbol{V}^{t+1}$. Recall the noisy power interpretation of $\boldsymbol{V}^{t+1}$ in Lemma D.1. We can write

$${\boldsymbol{B}^t}^\top\boldsymbol{Q}^t\boldsymbol{V}^{t+1} = {\boldsymbol{B}^t}^\top\left(\boldsymbol{B}^t({\boldsymbol{B}^t}^\top\boldsymbol{B}^*{\boldsymbol{V}^*}^\top - {\boldsymbol{F}^t}^\top) - \boldsymbol{B}^*{\boldsymbol{V}^*}^\top\right)\boldsymbol{V}^{t+1} = {\boldsymbol{F}^t}^\top\boldsymbol{V}^{t+1} = {\boldsymbol{F}^t}^\top\left(\boldsymbol{V}^*{\boldsymbol{B}^*}^\top\boldsymbol{B}^t - \boldsymbol{F}^t\right), \tag{53}$$

where we use ${\boldsymbol{B}^t}^\top\left(\boldsymbol{B}^t{\boldsymbol{B}^t}^\top - \boldsymbol{I}_d\right)\boldsymbol{B}^*{\boldsymbol{V}^*}^\top = 0$. Consequently, we can bound $s_{\max}\left({\boldsymbol{B}^t}^\top\boldsymbol{G}^t\right)$ as follows

$$\begin{aligned}
s_{\max}\left({\boldsymbol{B}^t}^\top\boldsymbol{G}^t\right) &\leq s_1\|\boldsymbol{F}^t\|_2 + \|\boldsymbol{F}^t\|_2^2 + \|\boldsymbol{G}^t - \boldsymbol{Q}^t\boldsymbol{V}^{t+1}\|_2\\
&\leq \frac{\delta^{(2)}s_1^2\sqrt{k}}{1-\delta^{(1)}} + \frac{(\delta^{(2)})^2s_1^2k}{(1-\delta^{(1)})^2} + \delta^{(3)}s_1^2k \leq \frac{s_k^2E_0}{12},
\end{aligned}$$

where we use the assumptions that $\max\{\frac{\delta^{(2)}\sqrt{k}}{1-\delta^{(1)}}, \frac{(\delta^{(2)})^2k}{(1-\delta^{(1)})^2}, \delta^{(3)}k\} \leq \frac{s_k^2E_0}{36s_1^2}$ and $\mathrm{dist}(\boldsymbol{B}^*, \boldsymbol{B}^t) \leq 1$ for the last inequality.

2. $s_{\max}\left(\boldsymbol{B}^{t^\top}\boldsymbol{W}^t\right)$**:** Due to the rotational invariance of independent Gaussian random variables, every entry in $\boldsymbol{B}^{t^\top}\boldsymbol{W}^t \in \mathbb{R}^{k\times k}$ is distributed as $\mathcal{N}(0,1)$. According to Theorem 4.4.5 in Vershynin (2018), with probability at least $1-2\exp(-k\log n)$, we have the bound $\|\boldsymbol{B}^{t^\top}\boldsymbol{W}^t\| \le \frac{C_{\mathcal{N}}\sqrt{k\log n}}{48\sqrt{2}}$ for some universal constant $C_{\mathcal{N}}$.

3. $s_{\max}\left(\boldsymbol{G}^{t^\top}\boldsymbol{W}^t\right)$**:** Let $\boldsymbol{G}^t = U_{\boldsymbol{G}}\boldsymbol{S}_{\boldsymbol{G}}V_{\boldsymbol{G}}^\top$ be the compact singular value decomposition of $\boldsymbol{G}^t$ such that $U_{\boldsymbol{G}} \in \mathbb{R}^{d\times k}$ and $U_{\boldsymbol{G}}^\top U_{\boldsymbol{G}} = \boldsymbol{I}_k$. We can bound

$$s_{\max}\left(\boldsymbol{G}^{t^\top}\boldsymbol{W}^t\right) \le \|\boldsymbol{S}_{\boldsymbol{G}}\|\|U_{\boldsymbol{G}}^\top\boldsymbol{W}^t\|. \tag{54}$$

Due to the rotational invariance of independent Gaussian random variables, every entry in ${\boldsymbol{U}_{\boldsymbol{G}}^t}^\top\boldsymbol{W}^t \in \mathbb{R}^{k\times k}$ is distributed as $\mathcal{N}(0,1)$ and hence with probability at least $1-2\exp(-k\log n)$, we have the bound $\|\boldsymbol{U}_{\boldsymbol{G}}^\top\boldsymbol{W}^t\| \le \frac{C_{\mathcal{N}}\sqrt{k\log n}}{48\sqrt{2}}$ for some universal constant $C_{\mathcal{N}}$. We now focus on bounding $\|\boldsymbol{S}_{\boldsymbol{G}}\|_2 = \|\boldsymbol{G}^t\|_2$. Note that

$$\boldsymbol{G}^t = \boldsymbol{Q}^t\boldsymbol{V}^{t+1} + \left(\boldsymbol{G}^t - \boldsymbol{Q}^t\boldsymbol{V}^{t+1}\right), \tag{55}$$

where the spectral norm of second term in Eq. equation 55 is bounded by Lemma D.6. Recall the noisy power interpretation of $\boldsymbol{V}^{t+1}$ in Lemma D.1. We can write

$$\begin{aligned}\boldsymbol{Q}^t\boldsymbol{V}^{t+1} &= \left(\boldsymbol{B}^t(\boldsymbol{B}^{t^\top}\boldsymbol{B}^*\boldsymbol{V}^{*^\top} - \boldsymbol{F}^{t^\top}) - \boldsymbol{B}^*\boldsymbol{V}^{*^\top}\right)\boldsymbol{V}^{t+1} \\ &= \left((\boldsymbol{B}^t\boldsymbol{B}^{t^\top} - \boldsymbol{I}_d)\boldsymbol{B}^*\boldsymbol{V}^{*^\top} - \boldsymbol{B}^t\boldsymbol{F}^{t^\top}\right)\left(\boldsymbol{V}^*\boldsymbol{B}^{*^\top}\boldsymbol{B}^t - \boldsymbol{F}^t\right),\end{aligned}$$

and therefore we can bound

$$\|\boldsymbol{Q}^t\boldsymbol{V}^{t+1}\|_2 \le (s_1 + \|\boldsymbol{F}^t\|_2)^2 \le \left(s_1 + \frac{\delta^{(2)}s_1\sqrt{k}}{1-\delta^{(1)}}\mathrm{dist}(\boldsymbol{B}^*,\boldsymbol{B}^t)\right)^2 \le 4s_1^2, \tag{56}$$

since we assume that $\frac{\delta^{(2)}\sqrt{k}}{1-\delta^{(1)}} \le 1$.

Combining the above three points, we have with probability at least $1-4\exp(-k\log n)$

$$s_{\min}(\boldsymbol{R}^{t+1^\top}\boldsymbol{R}^{t+1}) \ge 1 - \eta_g s_k^2 E_0/6 - \sigma\cdot C_{\mathcal{N}}\sqrt{k\log n}/6 - \eta_g\sigma\cdot C_{\mathcal{N}}\sqrt{k\log n}\cdot s_1^2/6. \tag{57}$$

Hence, if we choose σ such that (recall that $\eta_g \le 1/4s_1^2$)

$$\sigma\cdot C_{\mathcal{N}}\sqrt{k\log n} \le \eta_g s_k^2 E_0 \text{ and } \eta_g\sigma\cdot C_{\mathcal{N}}\sqrt{k\log n}\cdot s_1^2 \le \eta_g s_k^2 E_0 \Rightarrow \sigma \le \frac{E_0}{4C_{\mathcal{N}}\kappa^2\sqrt{k\log n}}, \tag{58}$$

we have

$$s_{\min}(\boldsymbol{R}^{t+1^\top}\boldsymbol{R}^{t+1}) \ge 1 - \eta_g s_k^2 E_0/2 \Rightarrow s_{\max}((\boldsymbol{R}^{t+1})^{-1}) \le \frac{1}{\sqrt{1-\eta_g s_k^2 E_0/2}}. \tag{59}$$

□

E ESTABLISH LEMMAS C.2 TO C.4 FOR THE LRL CASE

E.1 PROOF OF LEMMA C.2

For the simplicity of the notations, we omit the superscript of $\mathcal{A}_i^{t,1}$ and $\mathcal{A}^{t,1}$. Moreover, recall that $\boldsymbol{v}_i = \boldsymbol{V}_{i,:} \in \mathbb{R}^k$ denotes the ith row of the matrix $\boldsymbol{V}$.

While we can directly use an ϵ-net argument to establish the desired property on the set of matrices $\boldsymbol{V} \in \mathbb{R}^{n\times k}$, $\|\boldsymbol{V}\|_F = 1$, it leads to a suboptimal bound since the size of the ϵ-net is $\mathcal{O}((\frac{2}{\epsilon}+1)^{nk})$. In the following, we show that by exploiting the special structure of the operator $\mathcal{A}$, i.e. $\boldsymbol{V}$ is row-wise separable in $\mathcal{A}(\boldsymbol{B}^t\boldsymbol{V}^\top)$, we can reduce the size of the ϵ-net to $\mathcal{O}((\frac{2}{\epsilon}+1)^k)$: Compute that

$$\langle\mathcal{A}(\boldsymbol{B}^t\boldsymbol{V}^\top), \mathcal{A}(\boldsymbol{B}^t\boldsymbol{V}^\top)\rangle = \frac{1}{n}\sum_{i=1}^n \frac{1}{\bar{m}}\langle\mathcal{A}_i(\boldsymbol{B}^t\boldsymbol{v}_i\boldsymbol{e}^{(i)^\top}), \mathcal{A}_i(\boldsymbol{B}^t\boldsymbol{v}_i\boldsymbol{e}^{(i)^\top})\rangle.$$

If for any $\boldsymbol{v} \in \mathbb{R}^k, \|\boldsymbol{v}\|_2 = 1$, we have $\frac{1}{\bar{m}}\langle \mathcal{A}_i(\boldsymbol{B}^t\boldsymbol{v}\boldsymbol{e}^{(i)^\top}), \mathcal{A}_i(\boldsymbol{B}^t\boldsymbol{v}\boldsymbol{e}^{(i)^\top})\rangle = n\left(1 \pm \mathcal{O}(\delta^{(1)})\right)$, we can show

$$\begin{aligned}\langle \mathcal{A}(\boldsymbol{B}^t\boldsymbol{V}^\top), \mathcal{A}(\boldsymbol{B}^t\boldsymbol{V}^\top)\rangle &= \frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{v}_i\|_2^2\frac{1}{\bar{m}}\langle \mathcal{A}_i(\boldsymbol{B}^t\frac{\boldsymbol{v}_i\boldsymbol{e}^{(i)^\top}}{\|\boldsymbol{v}_i\|_2}), \mathcal{A}_i(\boldsymbol{B}^t\frac{\boldsymbol{v}_i\boldsymbol{e}^{(i)^\top}}{\|\boldsymbol{v}_i\|_2})\rangle \\ &= \sum_{i=1}^{n}\|\boldsymbol{v}_i\|_2^2\left(1 \pm \mathcal{O}(\delta^{(1)})\right) = 1 \pm \mathcal{O}(\delta^{(1)}), \forall \boldsymbol{V} \in \mathbb{R}^{n\times k}, \|\boldsymbol{V}\|_F = 1,\end{aligned}$$

which is the desired result (note that $\langle \boldsymbol{B}^t\boldsymbol{V}^\top, \boldsymbol{B}^t\boldsymbol{V}^\top\rangle = \|\boldsymbol{B}^t\boldsymbol{V}^\top\|_F = 1$). We now establish

$$\frac{1}{\bar{m}}\langle \mathcal{A}_i(\boldsymbol{B}^t\boldsymbol{v}\boldsymbol{e}^{(i)^\top}), \mathcal{A}_i(\boldsymbol{B}^t\boldsymbol{v}\boldsymbol{e}^{(i)^\top})\rangle = \frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}} n((\boldsymbol{x}_{ij}^\top\boldsymbol{B}^t)^\top\boldsymbol{v})^2 = n\left(1 \pm \mathcal{O}(\delta^{(1)})\right)$$

holds for any $\boldsymbol{v} \in \mathbb{R}^k, \|\boldsymbol{v}\|_2 = 1$. Let $\mathcal{S}^{k-1}$ be the sphere in the k-dimensional Euclidean space and let $\mathcal{N}_k$ be the $1/4$-net of cardinality 9^k (see Corollary 4.2.13 in Vershynin (2018)). Note that $\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}} n((\boldsymbol{x}_{ij}^\top\boldsymbol{B}^t)^\top\boldsymbol{v})^2 - n = n\boldsymbol{v}^\top\left(\boldsymbol{B}^{t^\top}(\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top)\boldsymbol{B}^t - \boldsymbol{I}_k\right)\boldsymbol{v}$ and we have

$$\sup_{\boldsymbol{v}\in\mathcal{S}^{k-1}}\boldsymbol{v}^\top\left(\boldsymbol{B}^{t^\top}(\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top)\boldsymbol{B}^t - \boldsymbol{I}_k\right)\boldsymbol{v} \le 2\sup_{v\in\mathcal{N}_k}\boldsymbol{v}^\top\left(\boldsymbol{B}^{t^\top}(\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top)\boldsymbol{B}^t - \boldsymbol{I}_k\right)\boldsymbol{v}, \tag{60}$$

where we use Lemma 4.4.1 in Vershynin (2018). In the following, we prove

$$\sup_{\boldsymbol{v}\in\mathcal{N}_k}\boldsymbol{v}^\top\left(\boldsymbol{B}^{t^\top}(\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top)\boldsymbol{B}^t - \boldsymbol{I}_k\right)\boldsymbol{v} \le \delta^{(1)}/2$$

so that we have $\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}} n((\boldsymbol{x}_{ij}^\top\boldsymbol{B}^t)^\top\boldsymbol{v})^2 = n\left(1 + \mathcal{O}(\delta^{(1)})\right)$.

For any fixed index i, denote $Z_{ij} = (\boldsymbol{x}_{ij}^\top\boldsymbol{B}^t\boldsymbol{v})^2 = n(\boldsymbol{x}_{ij}^\top\boldsymbol{B}^t\boldsymbol{v})^2$ and note that $\mathbb{E}_{\boldsymbol{x}_{ij}}[Z_{ij}] = n$. Since $\boldsymbol{x}_{ij}$'s are independent subgaussian variables, Z_{ij}'s are independent subexponential variables. Recall that $\|\cdot\|_{\psi_2}$ and $\|\cdot\|_{\psi_1}$ denote the subgaussian norm and subexponential norm respectively and we have $\|XY\|_{\psi_1} \le \|X\|_{\psi_2}\|Y\|_{\psi_2}$ for subgaussian random variables X and Y (see Lemma 2.2.7 in Vershynin (2018)). Therefore, we can bound $\|Z_{ij}\|_{\psi_1} \le n\|\boldsymbol{B}^t\boldsymbol{v}\|_2^2 = n$. Using the centering property (see Exercise 2.7.10 of Vershynin (2018)) and Bernstein's inequality (see Theorem 2.8.1 of Vershynin (2018)) of the zero mean subexponential variables, we can bound, for every $\tau \ge 0$

$$\begin{aligned}\Pr\{|\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}} Z_{ij} - n| \ge n\tau\} &\le 2\exp\left(-c\min\{\frac{\bar{m}^2n^2\tau^2}{\sum_{j=1}^{\bar{m}}\|Z_{ij}\|_{\psi_1}^2}, \frac{\bar{m}n\tau}{\max_{j\in[\bar{m}]}\|Z_{ij}\|_{\psi_1}}\}\right) \\ &= 2\exp\left(-c\min\{\bar{m}\tau^2, \bar{m}\tau\}\right) \overset{\tau\le 1}{=} 2\exp\left(-c\bar{m}\tau^2\right)\end{aligned}$$

where $c > 0$ is an absolute constant. Using the union bound over $\mathcal{N}_k$ and set $\tau = \sqrt{k\log n}/\sqrt{\bar{m}}$. We obtain $\delta^{(1)} \le \sqrt{k\log n}/\sqrt{\bar{m}}$ with probability at least $1 - \exp(-c'k\log n)$ for some constant c'. Similarly, we can show that $\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}} n((\boldsymbol{x}_{ij}^\top\boldsymbol{B}^t)^\top\boldsymbol{v})^2 = n\left(1 - \mathcal{O}(\delta^{(1)})\right)$. We therefore have the result.

E.2 PROOF OF LEMMA C.3

Recall that $\boldsymbol{v}_i \in \mathbb{R}^k$ denotes the ith row of the matrix $\boldsymbol{V}$ and denote $\boldsymbol{W} = (\boldsymbol{B}^t\boldsymbol{B}^{t^\top} - \boldsymbol{I}_k)\boldsymbol{B}^*$.

Following a similar argument as Lemma C.2, we simply need to focus on showing for any $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^k, \|\boldsymbol{v}_1\| = \|\boldsymbol{v}_2\| = 1$, we have $\frac{1}{\bar{m}}\langle \mathcal{A}_i(\boldsymbol{W}\boldsymbol{v}_1\boldsymbol{e}^{(i)^\top}), \mathcal{A}_i(\boldsymbol{B}^t\boldsymbol{v}_2\boldsymbol{e}^{(i)^\top})\rangle = \pm\mathcal{O}(n\delta^{(2)})$ (Note that $\langle \boldsymbol{W}\boldsymbol{v}_1\boldsymbol{e}^{(i)^\top}, \boldsymbol{B}^t\boldsymbol{v}_2\boldsymbol{e}^{(i)^\top}\rangle = 0$). Let $\mathcal{S}^{k-1}$ be the sphere in the k-dimensional Euclidean space and let $\mathcal{N}_k$ be the $1/4$-net of cardinality 9^k (see Corollary 4.2.13 in Vershynin (2018)). Note that

$$\frac{1}{\bar{m}}\langle \mathcal{A}_i(\boldsymbol{W}\boldsymbol{B}^*\boldsymbol{v}_1\boldsymbol{e}^{(i)^\top}), \mathcal{A}_i(\boldsymbol{B}^t\boldsymbol{v}_2\boldsymbol{e}^{(i)^\top})\rangle = n\boldsymbol{v}_1^\top\left(\boldsymbol{W}^\top(\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top)\boldsymbol{B}^t\right)\boldsymbol{v}_2. \tag{61}$$

Moreover, according to Exercise 4.4.3 in Vershynin (2018), we have

$$\sup_{v_1,v_2\in\mathcal{S}^{k-1}} \boldsymbol{v}_1^\top \left(\boldsymbol{W}^\top(\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top)\boldsymbol{B}^t\right)\boldsymbol{v}_2 \le 2\sup_{v_1,v_2\in\mathcal{N}_k} \boldsymbol{v}_1^\top \left(\boldsymbol{W}^\top(\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}}\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top)\boldsymbol{B}^t\right)\boldsymbol{v}_2.$$

In the following, we prove that the quantity on the RHS is bounded by $\delta^{(2)}/2$ so that we have the desired result.

For any fixed index i, denote $Z_{ij} = n(\boldsymbol{x}_{ij}^\top\boldsymbol{W}\boldsymbol{v}_1)(\boldsymbol{x}_{ij}^\top\boldsymbol{B}^t\boldsymbol{v}_2)$ and note that $\mathbb{E}_{\boldsymbol{x}_{ij}}[Z_{ij}] = 0$. Since $\boldsymbol{x}_{ij}$'s are independent subgaussian variables, Z_{ij}'s are independent subexponential variables. We can bound $\|Z_{ij}\|_{\psi_1} \le n\|\boldsymbol{W}\boldsymbol{v}_1\|_2\|\boldsymbol{B}^t\boldsymbol{v}_2\|_2 = n\mathrm{dist}(\boldsymbol{B}^t,\boldsymbol{B}^*)$. Using the centering property (see Exercise 2.7.10 of Vershynin (2018)) and Bernstein's inequality (see Theorem 2.8.1 of Vershynin (2018)) of the zero mean subexponential variables, we can bound, for every $t \ge 0$

$$\begin{aligned}\Pr\{|\frac{1}{\bar{m}}\sum_{j=1}^{\bar{m}} Z_{ij}| \ge n\tau\} &\le 2\exp\left(-c\min\{\frac{\bar{m}^2n^2\tau^2}{\sum_{j=1}^{\bar{m}}\|Z_{ij}\|_{\psi_1}^2}, \frac{\bar{m}n\tau}{\max_{j\in[\bar{m}]}\|Z_{ij}\|_{\psi_1}}\}\right)\\ &= 2\exp\left(-c\min\{\frac{\bar{m}\tau^2}{(\mathrm{dist}(\boldsymbol{B}^t,\boldsymbol{B}^*))^2}, \frac{\bar{m}\tau}{\mathrm{dist}(\boldsymbol{B}^t,\boldsymbol{B}^*)}\}\right)\end{aligned}$$

where $c > 0$ is an absolute constant. Use the union bound over $\mathcal{N}_k^2$, set $\delta^{(2)} \le \sqrt{k\log n}/\sqrt{\bar{m}}$ and set $\tau = \mathrm{dist}(\boldsymbol{B}^t,\boldsymbol{B}^*)\delta^{(2)}$. We show that Condition C.3 is satisfied with parameter $\delta^{(2)}$ with probability at least $1 - \exp(-c'k\log n)$ for some constant c'.

E.3 PROOF OF LEMMA C.4

Denote $\boldsymbol{Q}^t = \boldsymbol{B}^t{\boldsymbol{V}^{t+1}}^\top - \boldsymbol{B}^*{\boldsymbol{V}^*}^\top$ and recall that $\boldsymbol{V}_{i,:} \in \mathbb{R}^k$ denotes the ith row of the matrix $\boldsymbol{V}$. For simplicity, denote $\boldsymbol{W} = \frac{1}{N}\sum_{i=1,j=1}^{n,\bar{m}}\langle\boldsymbol{A}_{ij},\boldsymbol{Q}^t\rangle\boldsymbol{A}_{ij}\boldsymbol{V}^{t+1}$ and note that $\mathbb{E}_{\boldsymbol{x}_{ij}}[\boldsymbol{W}] = \boldsymbol{Q}^t\boldsymbol{V}^{t+1}$.

We use an ϵ-net argument to establish the desired property on the set of vectors $\boldsymbol{a} \in \mathbb{R}^d, \boldsymbol{b} \in \mathbb{R}^k, \|\boldsymbol{a}\| = \|\boldsymbol{b}\| = 1$. Let $\mathcal{S}^{k-1}$ and $\mathcal{S}^{d-1}$ be the spheres in the k-dimensional and d-dimensional Euclidean spaces and let $\mathcal{N}_k$ and $\mathcal{N}_d$ be the $1/4$-net of cardinality 9^k and 9^d respectively (see Corollary 4.2.13 in Vershynin (2018)). Note that

$$\langle\mathcal{A}(\boldsymbol{Q}^t), \mathcal{A}(\boldsymbol{a}\boldsymbol{b}^\top{\boldsymbol{V}^{t+1}}^\top)\rangle = \boldsymbol{a}^\top\boldsymbol{W}\boldsymbol{b}. \tag{62}$$

Moreover, according to Exercise 4.4.3 in Vershynin (2018), we have

$$\sup_{\boldsymbol{a}\in\mathcal{S}^{d-1},\boldsymbol{b}\in\mathcal{S}^{k-1}} \boldsymbol{a}^\top(\boldsymbol{W} - \boldsymbol{Q}^t\boldsymbol{V}^{t+1})\boldsymbol{b} \le 2\sup_{\boldsymbol{a}\in\mathcal{N}_d,\boldsymbol{b}\in\mathcal{N}_k} \boldsymbol{a}^\top(\boldsymbol{W} - \boldsymbol{Q}^t\boldsymbol{V}^{t+1})\boldsymbol{b}.$$

In the following, we prove that the quantity on the RHS is bounded by $\delta^{(3)}s_1^2k\mathrm{dist}(\boldsymbol{B}^t,\boldsymbol{B}^*)/2$ so that we have the desired result. We first bound $\|\boldsymbol{Q}^t\boldsymbol{e}^{(i)}\|_2$ and $\|\boldsymbol{V}_{i,:}^{t+1}\|_2$ as they will be used in our concentration argument.

Bound $\|\boldsymbol{V}_{i,:}^{t+1}\|_2$. Using Lemma D.1, we can write $\boldsymbol{V}^{t+1} = \boldsymbol{V}^*{\boldsymbol{B}^*}^\top\boldsymbol{B}^t - \boldsymbol{F}^t$ and therefore $\boldsymbol{V}_{i,:}^{t+1} = {\boldsymbol{B}^t}^\top\boldsymbol{B}^*\boldsymbol{V}_{i,:}^* - \boldsymbol{F}_{i,:}^t$. Using Assumption 5.2, we have $\|\boldsymbol{V}_{i,:}^*\|_2 \le \mu\sqrt{k}s_k/\sqrt{n}$. Further, we can compute that

$$\boldsymbol{F}_{i,:}^t = \left(\frac{1}{\bar{m}}{\boldsymbol{B}^t}^\top\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top\boldsymbol{B}^t\right)^{-1}\left(\frac{1}{\bar{m}}{\boldsymbol{B}^t}^\top\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top(\boldsymbol{B}^t{\boldsymbol{B}^t}^\top - \boldsymbol{I}_d)\boldsymbol{B}^*\right)\boldsymbol{V}_{i,:}^*. \tag{63}$$

Using the variational formulation of the spectral norm and Conditions C.2 and C.3 (note that $\delta^{(1)} = \delta^{(2)} = \sqrt{k\log n}/\sqrt{\bar{m}}$), we have $\|\boldsymbol{F}_{i,:}^t\|_2 \le \frac{\delta^{(2)}\mathrm{dist}(\boldsymbol{B}^t,\boldsymbol{B}^*)}{1-\delta^{(1)}}\cdot\frac{\mu\sqrt{k}s_k}{\sqrt{n}}$. Therefore we obtain

$$\|\boldsymbol{V}_{i,:}^{t+1}\|_2 \le 2\mu\sqrt{k}s_k/\sqrt{n}. \tag{64}$$

The high probability upper bound of $\boldsymbol{w}_i^t$ in Lemma C.1 can be derived from the above inequality by noting that $\boldsymbol{w}_i^t = \sqrt{n}\boldsymbol{V}_{i,:}^t$.

Bound $\|\boldsymbol{Q}^t\boldsymbol{e}^{(i)}\|_2$. Recall the definition of $\boldsymbol{Q}^t = \boldsymbol{B}^t\boldsymbol{V}^{t+1^\top} - \boldsymbol{B}^*\boldsymbol{V}^{*\top}$. Using Lemma D.1, we obtain

$$\boldsymbol{Q}^t = \boldsymbol{B}^t(\boldsymbol{V}^*\boldsymbol{B}^{*\top}\boldsymbol{B}^t - \boldsymbol{F}^t)^\top - \boldsymbol{B}^*\boldsymbol{V}^{*\top} = (\boldsymbol{B}^t\boldsymbol{B}^{t^\top} - \boldsymbol{I}_d)\boldsymbol{B}^*\boldsymbol{V}^* - \boldsymbol{B}^t\boldsymbol{F}^{t^\top}. \tag{65}$$

We can bound $\|\boldsymbol{Q}^t\boldsymbol{e}^{(i)}\|_2 \leq \|(\boldsymbol{B}^t\boldsymbol{B}^{t^\top} - \boldsymbol{I}_d)\boldsymbol{B}^*\|_2\|\boldsymbol{v}_i^*\|_2 + \|\boldsymbol{F}_{i,:}^t\|_2 \leq 2\mathrm{dist}(\boldsymbol{B}^t, \boldsymbol{B}^*)\mu\sqrt{k}s_k/\sqrt{n}$.

Denote $Z_{ij} = n(\boldsymbol{x}_{ij}^\top\boldsymbol{Q}^t\boldsymbol{e}^{(i)})(\boldsymbol{x}_{ij}^\top\boldsymbol{a}\boldsymbol{b}^\top\boldsymbol{V}_{i,:}^{t+1})$ and note that $\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\boldsymbol{x}_{ij}}[Z_{ij}] = n\boldsymbol{b}^\top\boldsymbol{V}^{t+1^\top}\boldsymbol{Q}^t\boldsymbol{a}$. Since $\boldsymbol{x}_{ij}$'s are independent subgaussian variables, Z_{ij}'s are independent subexponential variables. We can bound $\|Z_{ij}\|_{\psi_1} \leq n\|\boldsymbol{Q}^t\boldsymbol{e}^{(i)}\|_2\|\boldsymbol{a}\boldsymbol{b}^\top\boldsymbol{V}_{i,:}^{t+1}\|_2 \leq 4k\mu^2 s_k^2\mathrm{dist}(\boldsymbol{B}^t, \boldsymbol{B}^*)$. Using the centering property (see Exercise 2.7.10 of Vershynin (2018)) and Bernstein's inequality (see Theorem 2.8.1 of Vershynin (2018)) of the zero mean subexponential variables, we can bound, for every $t \geq 0$

$$\begin{aligned}
&\Pr\{|\frac{1}{n\bar{m}}\sum_{i=1,j=1}^{n,\bar{m}} Z_{ij} - \boldsymbol{a}^\top(\boldsymbol{Q}^t\boldsymbol{V}^{t+1})\boldsymbol{b}| \geq \tau\} \\
\leq\ & 2\exp\left(-c\min\{\frac{\bar{m}^2n^2\tau^2}{\sum_{i=1,j=1}^{n,\bar{m}}\|Z_{ij}\|_{\psi_1}^2}, \frac{\bar{m}n\tau}{\max_{i\in[n],j\in[\bar{m}]}\|Z_{ij}\|_{\psi_1}}\}\right) \\
=\ & 2\exp\left(-c\min\{\frac{\bar{m}n\tau^2}{(4k\mu^2s_k^2\mathrm{dist}(\boldsymbol{B}^t, \boldsymbol{B}^*))^2}, \frac{\bar{m}n\tau}{4k\mu^2s_k^2\mathrm{dist}(\boldsymbol{B}^t, \boldsymbol{B}^*)}\}\right)
\end{aligned}$$

Set $\tau = \delta^{(3)}s_1^2k\mathrm{dist}(\boldsymbol{B}^t, \boldsymbol{B}^*)$. We have that when $\delta^{(3)} = \frac{4(\sqrt{d}+\sqrt{k\log n})}{\sqrt{\bar{m}n}\kappa^2}$,

$$\Pr\{|\frac{1}{n\bar{m}}\sum_{i=1,j=1}^{n,\bar{m}} Z_{ij} - \boldsymbol{a}^\top(\boldsymbol{Q}^t\boldsymbol{V}^{t+1})\boldsymbol{b}| \geq \tau\} \leq \exp\left(-c(d + k\log n)\right). \tag{66}$$

Use the union bound over $\mathcal{N}_k \times \mathcal{N}_d$, we have that with probability at least $1 - \exp(-c''k\log n)$ Condition C.4 holds with $\delta^{(3)} = \frac{4(\sqrt{d}+\sqrt{k\log n})}{\sqrt{\bar{m}n}\kappa^2}$.

F ANALYSIS OF THE INITIALIZATION PROCEDURE

In this section, we present the privacy and utility guarantees of the INITIALIZATION procedure (Algorithm 3), when the local data points $\boldsymbol{x}_{ij}$ are Gaussian variables. Recall that to establish the utility guarantee for `CENTAUR` in the LRL setting, a column orthonormal initialization $\boldsymbol{B}^0 \in \mathbb{R}^{d\times k}$ such that the initial error $\mathrm{dist}(\boldsymbol{B}^0, \boldsymbol{B}^*) \leq \epsilon_0 = 0.2$ is required. While such a requirement can be ensured by the private power method (`PPM`, presented in Algorithm 5), the utility guarantee only holds *with a constant probability*, e.g. 0.99.

A key contribution of our work is to propose a *cross-validation* type scheme to boost the success probability of `PPM` to $1 - O(n^{-k})$ with a small extra cost of $O(k\log n)$. Note that boosting the success probability has a small cost of utility and hence we need a higher target accuracy $\epsilon_i = 0.01$[2] for `PPM` compared to $\epsilon_0 = 0.2$. The most important novelty of our selection scheme is that it only takes as input the results of $O(k\log n)$ independent `PPM` runs, which hence can be treated as *post-processing* and is free of privacy leakage.

F.1 UTILITY AND PRIVACY GUARANTEES OF THE `PPM` PROCEDURE

In this section, we present the guarantees for the PPM procedure, under the additional assumption that $\boldsymbol{x}_{ij}$ are Gaussian variables. One can prove that the choice of ζ_0 described in Lemma F.1 is a high probability upper bound of $\|\boldsymbol{Y}_i^l\|_F$ (see Lemma G.1 in the appendix). Therefore, with the same high probability, the clipping operation in the Gaussian mechanism will not take effect.

Conditioned on the above event (clipping takes no effect), to establish the utility guarantee of Algorithm 5, we view `PPM` as a specific instance of the perturbed power method presented in Algorithm 6 Hardt & Price (2014). Suppose that the level of perturbation is sufficiently small, we can exploit the analysis from (Hardt & Price, 2014) to prove the following lemma.

---

[2] The choice of ϵ_i should satisfy equation 70. $\epsilon_i = 0.01$ is one valid example.

Lemma F.1 (Utility Guarantee of a Single `PPM` Trial)**.** *Consider the LRL setting. Suppose that Assumptions 5.2 and 5.1 hold and $\boldsymbol{x}_{ij}$ are Gaussian variables. Let $\epsilon_i = 0.01$ be the target accuracy. For `PPM` presented in Algorithm 5, set $\zeta_0 = c_0'\mu^2 k^{1.5} s_k^2 d^{0.5} \log n \cdot (\sqrt{\log n} + \sqrt{k})$, $L = c_L'(s_k^2 + \sum_{j=1}^k s_j^2)/s_k^2 \cdot \log(kd/\epsilon_i)$. Suppose that n is sufficiently large so that there exists $\bar{m}_0$ such that*

$$\frac{\bar{m}n}{\log^6 \bar{m}n} \geq c_1' \frac{d \cdot \log^6 d \cdot k^3 \cdot \mu^2 \cdot \sum_{i=1}^k s_i^2}{s_k^2}. \tag{67}$$

Choose a noise multiplier $\sigma_0 = ns_k^2/c_2'\zeta_0 k\sqrt{d \log L}$. We have with probability at least 0.99, we have $\operatorname{dist}(\boldsymbol{B}^0, \boldsymbol{B}^*) \leq \epsilon_i$.

Remark F.1 (Remark the d dependence)**.**

The following RDP guarantee of `PPM` can be established using the Gaussian mechanism of RDP and the RDP composition lemma.

Lemma F.2 (Privacy Guarantee)**.** *Consider `PPM` presented in Algorithm 5. Set inputs L, the number of communication rounds, σ_0, the noise multiplier according to Lemma F.1. We have that `PPM` is an $(\alpha, \epsilon_{init}')$-RDP mechanism with*

$$\epsilon_{init}' = \frac{\alpha L \cdot (c_2'\zeta_0 k\sqrt{d \log L})^2}{n^2 s_k^4} = \tilde{O}(\frac{\alpha\kappa^2 k^7 \mu^4 d^2}{n^2}), \tag{68}$$

where $\tilde{O}$ hides the constants and the log *terms and we treat $(s_k^2 + \sum_{j=1}^k s_j^2)/s_k^2 = O(k)$.*

F.2 BOOST THE SUCCESS PROBABILITY WITH CROSS-VALIDATION

The following lemma shows that if the output of `PPM` has utility at least ϵ_i, e.g. $\epsilon_i = 0.01$, with probability p, e.g. $p = 0.99$, then any candidate that passes the test (69) has utility no less than ϵ_0, e.g. $\epsilon_0 = 0.2$, with a high probability $(1 - \delta)$. The proof is provided in Appendix H.

Lemma F.3. *Use ϵ_0 to denote the accuracy required by `CENTAUR` in the LRL setting and use ϵ_i to denote the accuracy of a single `PPM` trial. Recall that $p = 0.99$ is the probability of success of `PPM`. Use $\boldsymbol{B}_{0,c}$ to denote the output of `PPM` in the c^{th} trial and set $T_0 = 8p \log 1/\delta$ in the procedure* INITIALIZATION *(Algorithm 3). We have with probability at least $1 - \delta$ that there exists one element $\hat{c}$ in $\{1, \ldots, T_0\}$ such that for at least half of $c \in \{1, \ldots, T_0\}$,*

$$s_i(\boldsymbol{B}_{0,c}^\top \boldsymbol{B}_{0,\hat{c}}) \geq 1 - 2\epsilon_i^2, \forall i \in [k]. \tag{69}$$

Moreover, $\boldsymbol{B}_{0,\hat{c}}$ must satisfy $\operatorname{dist}(\boldsymbol{B}_{0,\hat{c}}, \boldsymbol{B}^*) \leq \epsilon_0$ with a sufficiently small $\epsilon_i \in [0, 1]$ such that

$$\sqrt{1 - \epsilon_0^2} + 1 - \sqrt{1 - \epsilon_i^2} + \epsilon_i < 1 - 2\epsilon_i^2. \tag{70}$$

One valid example is $\epsilon_i = 0.01$ and $\epsilon_0 = 0.2$, which is the one chosen in our previous discussions.

Corollary F.1. *Consider the* INITIALIZATION *procedure presented in Algorithm 3. By setting $T_0 = c_T' k \log n$, $\epsilon_i = 0.01$, and setting L_0, σ_0 and ζ_0 according to Lemma F.1, we have with probability at least $1 - n^{-k}$, the output $\boldsymbol{B}_{\hat{c}}$ satisfies $\operatorname{dist}(\boldsymbol{B}_{\hat{c}}, \boldsymbol{B}^*) \leq \epsilon_0 = 0.2$. Moreover, the* INITIALIZATION *procedure is an $(\alpha, \epsilon_{init})$-RDP mechanism with*

$$\epsilon_{init} = \epsilon_{init}' \cdot T_0 = \tilde{O}(\frac{\alpha\kappa^2 k^8 \mu^4 d^2}{n^2}). \tag{71}$$

The idea of boosting the success probability is inspired by Algorithm 5 of (Liang et al., 2014). The major improvement of our approach is that given the outputs of $O(k \log n)$ `PPM` trials, it no long requires access to the dataset and hence can be treated as postprocessing, while (Liang et al., 2014) requires an extra data-dependent SVD operation which violates the purpose of DP protection in the first place.

G GUARANTEES FOR THE PPM PROCEDURE

In this section, we present the analysis to establish the guarantees for the PPM procedure. We first show that the choice of ζ_0 described in Lemma F.1 is a high probability upper bound of $\|\boldsymbol{Y}_i^l\|_F$. Therefore, with the same high probability, the clipping operation in the Gaussian mechanism will not take effect.

Algorithm 5 `PPM`: Private Power Method (Adapted from Hardt & Price (2014))

```
1: procedure PPM($L$, $\sigma_0$, $\zeta_0$, $\bar{m}_0$)
2:     Choose $\boldsymbol{X}^0 \in \mathbb{R}^{d\times k}$ to be a random column orthonormal matrix;
3:     for $l = 1$ to $L$ do
4:         for $i = 1$ to $n$ do
5:             Sample without replacement a subset $\mathbb{S}_i^{l,0}$ from $\mathbb{S}_i$ with cardinality $\bar{m}_0$.
6:             Denote $\boldsymbol{M}_i^l = \frac{1}{\bar{m}} \sum_{j\in\mathbb{S}_i^{l,0}} y_{ij}^2 \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top$.
7:             Compute $\boldsymbol{Y}_i^l := \boldsymbol{M}_i^l \boldsymbol{X}^{l-1}$.
8:         Let $\boldsymbol{X}^l = \mathcal{QR}(\boldsymbol{Y}^l)$, where $\boldsymbol{Y}^l = \mathrm{GM}_{\zeta_0,\sigma_0}(\{\boldsymbol{Y}_i^l\}_{i=1}^n)$.
9:         // $\mathcal{QR}(\cdot)$ denotes the QR decomposition and only returns the $Q$ matrix.
```

Algorithm 6 `NPM`: Noisy Power Method (Adapted from Hardt & Price (2014))

```
procedure NPM($\boldsymbol{A}$, $L$)
    // $\boldsymbol{A}$ is the target matrix
    Choose $\boldsymbol{X}^0 \in \mathbb{R}^{d\times k}$ to be a random column orthonormal matrix;
    for $l = 1$ to $L$ do
        Let $\boldsymbol{X}^l = \mathcal{QR}(\boldsymbol{Y}^l)$, where $\boldsymbol{Y}^l = \boldsymbol{A}\boldsymbol{X}^{l-1} + \boldsymbol{G}^l$.
        // $\boldsymbol{G}^l$ is some perturbation matrix.
        // $\mathcal{QR}(\cdot)$ denotes the QR decomposition and only returns the $Q$ matrix.
```

Lemma G.1. *Consider the LRL setting. Under the assumptions 5.2 and 5.1, we have with a probability at least* $1 - 3\bar{m}n^{-100}$,

$$\|\boldsymbol{Y}_i^l\|_F \le \zeta_0 \doteq c_0\mu^2 k^{1.5} s_k^2 (\log n)(\sqrt{\log n} + \sqrt{d})(\sqrt{\log n} + \sqrt{k})$$

where $\boldsymbol{Y}_i^l$ *is computed in line 7 of Algorithm 5 and* c_0 *is some universal constant.*

Proof. The detailed expression of $\boldsymbol{Y}_i^l$ in line 7 of Algorithm 5 can be calculated as follows:

$$\boldsymbol{Y}_i^l = \boldsymbol{M}_i^l \boldsymbol{X}^{l-1} = \frac{1}{\bar{m}} \sum_{j\in\mathbb{S}_i^{l,0}} y_{ij}^2 \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top \boldsymbol{X}^{l-1}. \tag{72}$$

Using the triangle inequality of the matrix norm, ζ is a high probability upper bound of $\|\boldsymbol{Y}_i^l\|_F$ if the following inequality holds with a high probability,

$$\|y_{ij}^2 \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top \boldsymbol{X}^{l-1}\|_F = |y_{ij}|^2 \|\boldsymbol{x}_{ij}\| \|\boldsymbol{x}_{ij}^\top \boldsymbol{X}^{l-1}\| \le \zeta_0. \tag{73}$$

To bound $|y_{ij}|^2$, recall that in Assumption 5.1, we assume that $\boldsymbol{x}_{ij}$ is a sub-Gaussian random vector with $\|\boldsymbol{x}_{ij}\|_{\psi_2} = 1$. Using the definition of a sub-Gaussian random vector, we have

$$\mathbb{P}\{|y_{ij}| \ge \tau\} \le 2\exp(-c_s\tau^2/\|\boldsymbol{w}_i^*\|^2) \le \exp(-100\log n), \tag{74}$$

with the choice $\tau = \mu\sqrt{k}s_k \cdot \sqrt{(100\log n + \log 2)/c_s}$ since $\|\boldsymbol{w}_i^*\|_2 \le \mu\sqrt{k}s_k$. Here c_s is some constant and we recall that s_k is a shorthand for $s_k(\boldsymbol{W}^*/\sqrt{n})$.

To bound $\|\boldsymbol{x}_{ij}\|$, recall that $\boldsymbol{x}_{ij}$ is a sub-Gaussian random vector with $\|\boldsymbol{x}_{ij}\|_{\psi_2} = 1$ and therefore with probability at least $1 - \delta$,

$$\|\boldsymbol{x}_{ij}\| \le 4\sqrt{d} + 2\sqrt{\log\frac{1}{\delta}}. \tag{75}$$

Therefore by taking $\delta = \exp(-100\log n)$, we have that $4\sqrt{d} + 2\sqrt{\log\frac{1}{\delta}} = 4\sqrt{d} + 2\sqrt{100\log n} = \zeta_x$.

To bound $\|\boldsymbol{x}_{ij}^\top \boldsymbol{X}^{l-1}\|$, note that due to the rotational invariance of the Gaussian random vector $\boldsymbol{x}_{ij}$ (recall that $\boldsymbol{X}^{l-1}$ is an column orthonormal matrix), the ℓ_2 norm $\|\boldsymbol{x}_{ij}^\top \boldsymbol{X}^{l-1}\|$ is distributed like the ℓ_2 norm of a Gaussian random vector drawn from $\mathcal{N}(0, \boldsymbol{I}_k)$. Therefore, w.p. at least $1 - n^{-100}$, $\|\boldsymbol{x}_{ij}^\top \boldsymbol{X}^{l-1}\| \le c(\sqrt{k} + \sqrt{\log n}) \doteq \zeta$.

Using the union bound and the fact that $\zeta_x\zeta_y^2\zeta \le \zeta_0$ leads to the conclusion. □

Conditioned on the above event (clipping takes no effect), to establish the utility guarantee of Algorithm 5, we view `PPM` as a specific instance of the noisy power method (`NPM`) presented in Algorithm 6 Hardt & Price (2014), where the target matrix is $\boldsymbol{A} = (2\Gamma + \mathrm{trace}(\Gamma)\boldsymbol{I}_d)$ with $\Gamma = \boldsymbol{B}^*\boldsymbol{V}^{*\top}\boldsymbol{V}^*\boldsymbol{B}^{*\top}$ and the perturbation matrix $\boldsymbol{G}^l = \boldsymbol{P}_1^l + \boldsymbol{P}_2^l$ is the sum of the noise matrix added by the Gaussian mechanism, $\boldsymbol{P}_2^l = \frac{\sigma_0\zeta_0}{n}\boldsymbol{W}^l$, and the error matrix $\boldsymbol{P}_1^l = (\boldsymbol{M}^l - \boldsymbol{A})\boldsymbol{X}^{l-1}$. One can easily check that with these choices, we recover line 8 of Algorithm 5

$$\boldsymbol{Y}^l = \boldsymbol{A}\boldsymbol{X}^{l-1} + \boldsymbol{G}^l = \boldsymbol{M}^l\boldsymbol{X}^{l-1} + \frac{\sigma_0\zeta_0}{n}\boldsymbol{W}^l = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{M}_i^l\boldsymbol{X}^{l-1} + \frac{\sigma_0\zeta_0}{n}\boldsymbol{W}^l. \tag{76}$$

Suppose that the level of perturbation is sufficiently small, we can exploit the following analysis of `NPM` from (Hardt & Price, 2014).

Theorem G.1 (Adapted from Corollary 1.1 of Hardt & Price (2014))**.** *Consider the noisy power method (`NPM`) presented in Algorithm 6. Let $\boldsymbol{U} \in \mathbb{R}^{d\times k}$ represent the top-k eigenvectors of the input matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$. Suppose that the perturbation matrix $\boldsymbol{G}^l$ satisfies for all $l \in \{1, \ldots, L\}$*

$$5\|\boldsymbol{G}^l\| \le \epsilon(s_k(\boldsymbol{A}) - s_{k+1}(\boldsymbol{A})), 5\|\boldsymbol{U}^\top\boldsymbol{G}^l\| \le (s_k(\boldsymbol{A}) - s_{k+1}(\boldsymbol{A}))\frac{C}{\tau\sqrt{kd}} \tag{77}$$

for some fixed parameter τ and $\epsilon < 1/2$. Then with all but $1/\tau + e^{-\Omega(d)}$ probability, there exists an $L = \mathcal{O}(\frac{s_k(\boldsymbol{A})}{s_k(\boldsymbol{A})-s_{k+1}(\boldsymbol{A})}\log(kd\tau/\epsilon))$ so that after L steps we have that $\|(\boldsymbol{I} - \boldsymbol{X}_L\boldsymbol{X}_L^\top)\boldsymbol{U}\| \le \epsilon$. Here $C > 0$ is a constant defined in Lemma I.3.

To prove that the perturbation matrix $\boldsymbol{G}^l$ satisfies the conditions required by the above theorem, we bound $\boldsymbol{M}^l - \boldsymbol{A}$ and $\frac{\sigma_0\zeta_0}{n}\boldsymbol{W}^l$ individually, both with high probabilities.

Proof of Lemma F.1. Recall that $\boldsymbol{A} = 2\Gamma + \mathrm{trace}(\Gamma)\boldsymbol{I}_d$ and note that $rank(\Gamma) = k$ and that its singular values are $s_1^2, \ldots, s_k^2$. Therefore $s_i(\boldsymbol{A}) = 2s_i^2 + \sum_{j=1}^k s_j^2$ for $i \le k$ and $s_i(\boldsymbol{A}) = \sum_{j=1}^k s_j^2$ for $i > k$.

In this proof, we will show that for both matrices $\boldsymbol{P}_1^l = (\boldsymbol{M}^l - \boldsymbol{A})\boldsymbol{X}^{l-1}$ and $\boldsymbol{P}_2^l = \frac{\sigma_0\zeta_0}{n}\boldsymbol{W}^l$ the following inequalities hold for all $l \in \{1, \ldots, L\}$, which is a sufficient condition for Lemma G.1 to hold

$$10\|\boldsymbol{P}^l\| \le \epsilon(s_k(\boldsymbol{A}) - s_{k+1}(\boldsymbol{A})), 10\|\boldsymbol{U}^\top\boldsymbol{P}^l\| \le (s_k(\boldsymbol{A}) - s_{k+1}(\boldsymbol{A}))\frac{C}{\tau\sqrt{kd}}. \tag{78}$$

Control terms related to $\boldsymbol{P}_1^l$. 1. We first bound $\|\boldsymbol{P}_1^l\|$. By the independence between M^l and X^{l-1}, we have that

$$\mathbb{E}\left[(\boldsymbol{M}^l - \boldsymbol{A})\boldsymbol{X}^{l-1}\right] = \mathbb{E}\left[\boldsymbol{M}^l - \boldsymbol{A}\right]\cdot\mathbb{E}\left[\boldsymbol{X}^{l-1}\right] = 0.$$

To bound the norm term $\mathbb{E}\left[\|(\boldsymbol{M}^l - \boldsymbol{A})\boldsymbol{X}^{l-1}\|\right]$, we begin by controlling the norms of $\boldsymbol{Z}_{ij}^l \cdot \boldsymbol{X}^{l-1}$, where $\boldsymbol{Z}_{ij}^l = y_{ij}^2\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top$. By the proof for Lemma G.1, with a probability at least $1-\delta$, we have that $\|\boldsymbol{Z}_{ij}^l \cdot \boldsymbol{X}^{l-1}\| \le O\left((\sqrt{d} + \log 1/\delta)\cdot(\sqrt{k} + \log 1/\delta)\cdot\|\boldsymbol{w}_i^*\|^2\right) \le O\left((\sqrt{d} + \log 1/\delta)\cdot(\sqrt{k} + \log 1/\delta)\mu^2 k s_k^2\right)$ since $\|\boldsymbol{w}_i^*\| \le \mu\sqrt{k}s_k$. We then compute an upper bound on the matrix variance term

$$\begin{aligned}\|\mathbb{E}\left[(\boldsymbol{Z}_{ij}^l\boldsymbol{X}^{l-1})^\top\boldsymbol{Z}_{ij}^l\boldsymbol{X}^{l-1}\right]\| &\le \|\mathbb{E}\left[y_{ij}^4\|\boldsymbol{x}_{ij}\|^2(\boldsymbol{x}_{ij}^\top\boldsymbol{X}^{l-1})^\top\boldsymbol{x}_{ij}^\top\boldsymbol{X}^{l-1}\right]\| && (79)\\ &= \|(\boldsymbol{X}^{l-1})^\top\mathbb{E}\left[y_{ij}^4\|\boldsymbol{x}_{ij}\|^2\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top\right]\boldsymbol{X}^{l-1}\| && (80)\end{aligned}$$

Due to isotropy of the Gaussian, by (Tripuraneni et al., 2021, Lemma 5), we have that

$$\mathbb{E}\left[y_{ij}^4\|\boldsymbol{x}_{ij}\|^2\boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top\right] = \|\boldsymbol{w}_i^*\|^4\cdot\left((2d+75)\boldsymbol{e}_1\boldsymbol{e}_1^\top + (3d+15)\boldsymbol{I}_d\right) \tag{81}$$

By plugging equation 81 into equation 80, we prove the following inequality.

$$\begin{aligned}\|\mathbb{E}\left[(\boldsymbol{Z}_{ij}^l\boldsymbol{X}^{l-1})^\top\boldsymbol{Z}_{ij}^l\boldsymbol{X}^{l-1}\right]\| &\le \|\boldsymbol{w}_i^*\|^4\cdot\left((2d+75)(\boldsymbol{X}^{l-1})^\top\boldsymbol{e}_1\boldsymbol{e}_1^\top\cdot\boldsymbol{X}^{l-1} + (3d+15)\boldsymbol{I}_k\right) && (82)\\ &\le O(d\|\boldsymbol{w}_i^*\|^4) \le O(d\mu^2 k s_k^2\sum_{i=1}^k s_i^2) && (83)\end{aligned}$$

By combining both the norm bound and the matrix variance bound and using the modified Bernstein matrix inequality (Tripuraneni et al., 2021, Lemma 31), we have

$$\|(\boldsymbol{M}^l - \boldsymbol{A})\boldsymbol{X}^{l-1}\| \leq \log^3 \bar{m}n \cdot \log^3 d \cdot \mathcal{O}\left(\sqrt{\frac{d\mu^2 k s_k^2 \sum_{i=1}^k s_i^2}{\bar{m}n}} + \frac{\sqrt{kd}\mu^2 k s_k^2}{\bar{m}n}\right). \tag{84}$$

2. We then proceed to bound the term $\|\boldsymbol{U}^\top \boldsymbol{P}_1^l\| = \|\boldsymbol{U}^\top(\boldsymbol{M}^l - \boldsymbol{A})\boldsymbol{X}^{l-1}\|$. By using proof of Lemma G.1, we first bound the norms of $\|\boldsymbol{U}^\top \boldsymbol{Z}_{ij}^l \boldsymbol{X}^{l-1}\| \leq O((\sqrt{k} + \log 1/\delta)(\sqrt{k} + \log 1/\delta)\mu^2 k s_k^2)$, where $\boldsymbol{Z}_{ij}^l = y_{ij}^2 \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top$. We then compute an upper bound on the matrix variance term.

$$\|\mathbb{E}[(\boldsymbol{U}^\top \boldsymbol{Z}_{ij}^l \boldsymbol{X}^{l-1})^\top \boldsymbol{U}^\top \boldsymbol{Z}_{ij}^l \boldsymbol{X}^{l-1}]\| \leq \|\mathbb{E}\left[y_{ij}^4 (\boldsymbol{U}^\top \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top \boldsymbol{X}^{l-1})^\top \boldsymbol{U}^\top \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top \boldsymbol{X}^{l-1}\right]\| \tag{85}$$

$$= \|(\boldsymbol{X}^{l-1})^\top \mathbb{E}\left[y_{ij}^4 \|\boldsymbol{U}^\top \boldsymbol{x}_{ij}\|^2 \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top\right] \boldsymbol{X}^{l-1}\| \tag{86}$$

$$= \|(\boldsymbol{X}^{l-1})^\top \mathbb{E}\left[(\boldsymbol{w}_i^{*\top} \boldsymbol{B}^{*\top} \boldsymbol{x}_{ij})^4 \|\boldsymbol{U}^\top \boldsymbol{x}_{ij}\|^2 \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top\right] \boldsymbol{X}^{l-1}\| \tag{87}$$

$$= \|(\boldsymbol{X}^{l-1})^\top \mathbb{E}\left[(\boldsymbol{w}_i^{*\top} \boldsymbol{U}^\top \boldsymbol{x}_{ij})^4 \|\boldsymbol{U}^\top \boldsymbol{x}_{ij}\|^2 \boldsymbol{x}_{ij}\boldsymbol{x}_{ij}^\top\right] \boldsymbol{X}^{l-1}\|, \tag{88}$$

where the last equality is because by definition, we have $\boldsymbol{B}^* = \boldsymbol{U}$. We now perform variable transformation and denote $\boldsymbol{v}_{ij} = \boldsymbol{V}^\top \boldsymbol{x}_{ij}$ where $\boldsymbol{V} = [\boldsymbol{U}, \boldsymbol{U}']$ is an orthogonal matrix that is extended based on $\boldsymbol{U}$. Therefore, we equivalently write the above equation 88 to the following inequality.

$$\|\mathbb{E}\left[(\boldsymbol{U}^\top \boldsymbol{Z}_{ij}^l \boldsymbol{X}^{l-1})^\top \boldsymbol{U}^\top \boldsymbol{Z}_{ij}^l \boldsymbol{X}^{l-1}\right]\| \tag{89}$$

$$\leq \|(\boldsymbol{V}^\top \boldsymbol{X}^{l-1})^\top \mathbb{E}\left[[\boldsymbol{w}_i^{*\top}(\boldsymbol{v}_{ij}[1], \cdots, \boldsymbol{v}_{ij}[k])^\top]^4 (\sum_{a=1}^k \boldsymbol{v}_{ij}[a]^2) \boldsymbol{v}_{ij}\boldsymbol{v}_{ij}^\top\right] \boldsymbol{V}^\top \boldsymbol{X}^{l-1}\|, \tag{90}$$

where $\boldsymbol{v}_{ij}[a]$ means the a-th coordinate of the k-dimensional vector $\boldsymbol{v}_{ij}$, which is also distributed as Gaussian. Due to the isotropy of the Gaussian it suffices to compute the expectation assuming $\boldsymbol{w}_i^* \propto \|\boldsymbol{w}_i\|\boldsymbol{e}_1$ where $\boldsymbol{e}_1 = (1, 0, \cdots, 0)^\top$. Then following the proof for (Tripuraneni et al., 2021, Lemma 5), by combinatorics, we have the following equation.

$$\mathbb{E}[(\boldsymbol{w}_i^{*\top}(\boldsymbol{v}_{ij}[1], \cdots, \boldsymbol{v}_{ij}[k])^\top)^4 (\sum_{a=1}^k \boldsymbol{v}_{ij}[a])^2 \boldsymbol{v}_{ij}\boldsymbol{v}_{ij}^\top] \tag{91}$$

$$= \|\boldsymbol{w}_i^*\|^4 \mathbb{E}\left[(\boldsymbol{v}_{ij}[1])^4 (\sum_{a=1}^k \boldsymbol{v}_{ij}[a]^2) \boldsymbol{v}_{ij}\boldsymbol{v}_{ij}^\top\right] = O(\|\boldsymbol{w}_i^*\|^4 k) \tag{92}$$

Therefore, by plugging the above equation into equation 90, we prove the following inequality.

$$\|\mathbb{E}\left[(\boldsymbol{U}^\top \boldsymbol{Z}_{ij}^l \boldsymbol{X}^{l-1})^\top \boldsymbol{U}^\top \boldsymbol{Z}_{ij}^l \boldsymbol{X}^{l-1}\right]\| \leq O(k\|\boldsymbol{w}_i^*\|^4) \leq O(k\mu^2 k s_k^2 \sum_{i=1}^k s_i^2) \tag{93}$$

By combining both the norm bound and the above matrix variance bound and using the matrix Bernstein inequality (Tripuraneni et al., 2021, Lemma 31), we have

$$\|\boldsymbol{U}^\top(\boldsymbol{M}^l - \boldsymbol{A})\boldsymbol{X}^{l-1}\| \leq \log^3 \bar{m}n \cdot \log^3 d \cdot \mathcal{O}\left(\sqrt{\frac{k^2\mu^2 s_k^2 \sum_{i=1}^k s_i^2}{\bar{m}n}} + \frac{\mu^2 k^2 s_k^2}{\bar{m}n}\right). \tag{94}$$

Therefore, to ensure that equation 78 holds, it suffices to use $\bar{m}n$ sufficiently large such that

$$\log^3 \bar{m}n \cdot \log^3 d \cdot \mathcal{O}\left(\sqrt{\frac{d\mu^2 k s_k^2 \sum_{i=1}^k s_i^2}{\bar{m}n}} + \frac{\sqrt{kd}\mu^2 k s_k^2}{\bar{m}n}\right) \leq \epsilon(s_k(\boldsymbol{A}) - s_{k+1}(\boldsymbol{A})) \tag{95}$$

$$\log^3 \bar{m}n \cdot \log^3 d \cdot \mathcal{O}\left(\sqrt{\frac{k^2\mu^2 s_k^2 \sum_{i=1}^k s_i^2}{\bar{m}n}} + \frac{\mu^2 k^2 s_k^2}{\bar{m}n}\right) \leq (s_k(\boldsymbol{A}) - s_{k+1}(\boldsymbol{A}))\frac{C}{\tau\sqrt{kd}} \tag{96}$$

For simplicity, assume that $\sqrt{\frac{d\mu^2 k s_k^2 \sum_{i=1}^k s_i^2}{\bar{m}n}} \geq \frac{\sqrt{kd}\mu^2 k s_k^2}{\bar{m}n}$ and $\epsilon_a \geq 1/\tau\sqrt{kd}$. The above inequalities can be simplified as follows, with c_1 being some constant.

$$\frac{\bar{m}n}{\log^6 \bar{m}n} \geq c_1 \frac{d \cdot \log^6 d \cdot k^3 \cdot \mu^2 \cdot \sum_{i=1}^k s_i^2}{s_k^2}. \tag{97}$$

Control terms related to $\boldsymbol{P}_2^l$. Recall that $\boldsymbol{P}_2^l \sim \mathcal{N}(0, \sigma^2)^{d\times k}$, with $\sigma = \sigma_0\zeta_0/n$. Using Lemma I.2, we have with probability $1 - 2e^{-x}$:

- $\max_{l\in[L]} \|\boldsymbol{P}_2^l\| \leq C_{\mathcal{N}}\sigma(\sqrt{d} + \sqrt{p} + \sqrt{\ln(2L + x)})$;
- $\max_{l\in[L]} \|\boldsymbol{U}^\top \boldsymbol{P}_2^l\| \leq C_{\mathcal{N}}\sigma(2\sqrt{p} + \sqrt{\ln(2L + x)})$.

Consequently, we can obtain bound the terms related to $\boldsymbol{P}_2^l$ by setting σ_0 sufficiently small such that the following inequalities hold with $x = 100 \log n$

$$10C_{\mathcal{N}}\sigma(\sqrt{d} + \sqrt{k} + \sqrt{\ln(2L + x)}) \leq \epsilon_a s_k^2, 10C_{\mathcal{N}}\sigma(2\sqrt{k} + \sqrt{\ln(2L + x)}) \leq s_k^2/(\tau\sqrt{dk}).$$

To simplify the above inequalities, suppose that ϵ_a is a constant and neglect the $\log\log$ terms. We obtain

$$\frac{c_2\sigma_0\zeta_0}{n} \leq \frac{s_k^2}{\tau k\sqrt{d}\sqrt{\log L}}. \tag{98}$$

for some constant c_2.

Having established equation 78 for both $\boldsymbol{P}_1^l$ and $\boldsymbol{P}_2^l$, we can then use Lemma G.1 to obtain the target result.

□

H PROOF OF LEMMA F.3

Proof. In the following, we first show that 1. the index $\hat{c}$ exists with a high probability and we then show that 2. any candidate $\boldsymbol{B}_{\hat{c}}$ that passes the test equation 69 has utility no less than b, i.e. $\mathrm{dist}(\boldsymbol{B}_{\hat{c}}, \boldsymbol{B}^*) \leq b$.

Existence of $\hat{c}$ Suppose that both $\boldsymbol{B}_{c_1}$ and $\boldsymbol{B}_{c_2}$ are successful, i.e. $\mathrm{dist}(\boldsymbol{B}_{c_i}, \boldsymbol{B}^*) \leq a$ for $i = 1, 2$ or equivalently $s_{\min}(\boldsymbol{B}_{c_i}^\top \boldsymbol{B}^*) \geq \sqrt{1 - a^2}$ for $i = 1, 2$. Recall that $\boldsymbol{B}^*{\boldsymbol{B}^*}^\top + \boldsymbol{B}_\perp^*{\boldsymbol{B}_\perp^*}^\top = \boldsymbol{I}_d$. Compute that

$$\begin{aligned}
& s_{\min}(\boldsymbol{B}_{c_1}^\top \boldsymbol{B}_{c_2}) \\
= & s_{\min}\left(\boldsymbol{B}_{c_1}^\top(\boldsymbol{B}^*{\boldsymbol{B}^*}^\top + \boldsymbol{B}_\perp^*{\boldsymbol{B}_\perp^*}^\top)\boldsymbol{B}_{c_2}\right) \\
\geq & s_{\min}(\boldsymbol{B}_{c_1}^\top \boldsymbol{B}^*{\boldsymbol{B}^*}^\top \boldsymbol{B}_{c_2}) - s_{\max}(\boldsymbol{B}_{c_1}^\top \boldsymbol{B}_\perp^*{\boldsymbol{B}_\perp^*}^\top \boldsymbol{B}_{c_2}) \\
\geq & s_{\min}(\boldsymbol{B}_{c_1}^\top \boldsymbol{B}^*)s_{\min}({\boldsymbol{B}^*}^\top \boldsymbol{B}_{c_2}) - s_{\max}(\boldsymbol{B}_{c_1}^\top \boldsymbol{B}_\perp^*)s_{\max}({\boldsymbol{B}_\perp^*}^\top \boldsymbol{B}_{c_2}) \geq 1 - 2a^2.
\end{aligned} \tag{99}$$

Define the binomial random variable $X_c = \mathbf{1}_{\boldsymbol{B}_c \text{ is successful}}$. We have that $\mathbb{E}[X_c] = \mathbb{P}\{\boldsymbol{B}_c \text{ is successful}\} \geq p$.

Using the concentration of the binomial random variable, we have

$$\mathbb{P}\{\sum_{c=1}^{T_0} X_c \leq T_0 \cdot \mathbb{E}[X_c] - t\} \leq \exp(-t^2/(2T_0p)). \tag{100}$$

Therefore, with the choice $t = T_0/2$ and $T_0 \geq 8p \log 1/\delta$, we have with probability at least $1 - \delta$, at least half of the outputs of T_0 independent `NPM` runs are successful. Consequently, there exists at least $T_0/2$ pairs of $\boldsymbol{B}_{c_1}$ and $\boldsymbol{B}_{c_2}$ such that Eq. equation 99 holds, which shows the existence of $\hat{c}$.

Utility of $\boldsymbol{B}_{\hat{c}}$ We now show that the candidate $\boldsymbol{B}_{\hat{c}}$ that passes the test Eq. equation 69 must satisfy $\operatorname{dist}(\boldsymbol{B}_{\hat{c}}, \boldsymbol{B}^*) \leq b$. We prove via contradiction. Suppose that there exists a candidate $\boldsymbol{B}$ that passes the test, but with $\operatorname{dist}(\boldsymbol{B}, \boldsymbol{B}^*) > b$. This means that there exists $\hat{x} \in \mathbb{R}^k$ and $\hat{y} \in \mathbb{R}^k$ with $\|\hat{x}\|_2 = \|\hat{y}\|_2 = 1$ achieving the minimum singular value of $\boldsymbol{B}^\top \boldsymbol{B}^*$, such that $\langle \boldsymbol{B}\hat{x}, \boldsymbol{B}^*\hat{y}\rangle \leq \sqrt{1-b^2}$.

Let $\boldsymbol{B}_i$ be a successful candidate, i.e. $\operatorname{dist}(\boldsymbol{B}_i, \boldsymbol{B}^*) \leq a$ (note that with a high probability they are in the majority according to the discussion above). We have

$$s_{\min}(\boldsymbol{B}^\top \boldsymbol{B}_i) = s_{\min}(\boldsymbol{B}^\top \left(\boldsymbol{B}^*\boldsymbol{B}^{*\top} + \boldsymbol{B}^*_\perp \boldsymbol{B}^{*\top}_\perp\right)\boldsymbol{B}_i) \leq s_{\min}(\boldsymbol{B}^\top \boldsymbol{B}^*\boldsymbol{B}^{*\top}\boldsymbol{B}_i) + s_{\max}(\boldsymbol{B}^\top \boldsymbol{B}^*_\perp \boldsymbol{B}^{*\top}_\perp \boldsymbol{B}_i).$$

For the second term, we have $s_{\max}(\boldsymbol{B}^\top \boldsymbol{B}^*_\perp \boldsymbol{B}^{*\top}_\perp \boldsymbol{B}_i) \leq s_{\max}(\boldsymbol{B}^\top \boldsymbol{B}^*_\perp) \cdot s_{\max}(\boldsymbol{B}^{*\top}_\perp \boldsymbol{B}_i) \leq 1 \cdot a \leq a$. To bound the first term, recall the variational formulation of the minimum singular value $s_{\min}(A) = \min_{\|x\|=\|y\|=1} x^\top A y$ and hence

$$s_{\min}(\boldsymbol{B}^\top \boldsymbol{B}^*\boldsymbol{B}^{*\top}\boldsymbol{B}_i) \leq \hat{x}^\top \boldsymbol{B}^\top \boldsymbol{B}^* \boldsymbol{B}^{*\top}\boldsymbol{B}_i \hat{y} = \hat{x}^\top \boldsymbol{B}^\top \boldsymbol{B}^*\hat{y} + \hat{x}^\top \boldsymbol{B}^\top \boldsymbol{B}^*(\hat{y} - \boldsymbol{B}^{*\top}\boldsymbol{B}_i\hat{y})$$
$$\leq \sqrt{1-b^2} + \hat{x}^\top \boldsymbol{B}^\top \boldsymbol{B}^*(\hat{y} - \boldsymbol{B}^{*\top}\boldsymbol{B}_i\hat{y}) \leq \sqrt{1-b^2} + \|\boldsymbol{I}_k - \boldsymbol{B}^{*\top}\boldsymbol{B}_i\|_2 \leq \sqrt{1-b^2} + 1 - \sqrt{1-a^2}.$$

where we recall the definitions of $\hat{x}$ and $\hat{y}$ in the above paragraph. Combining the above bounds, we obtain $s_{\min}(\boldsymbol{B}^\top \boldsymbol{B}_i) \leq \sqrt{1-b^2} + 1 - \sqrt{1-a^2} + a$ which is strictly smaller than $1 - 2a^2$ for a sufficiently small a and a sufficiently large b, e.g. $a = 0.01$ and $b = 0.2$. To put it in other words, we obtain that $\boldsymbol{B}$ will fail test (69). This leads to a contradiction and hence we have proved that any candidate that passes the test (69) must satisfy $\operatorname{dist}(\boldsymbol{B}_{\hat{c}}, \boldsymbol{B}^*) \leq b$. $\square$

I PRELIMINARY ON MATRIX CONCENTRATION INEQUALITIES

Lemma I.1 (Theorem 4.4.5 in Vershynin (2018))**.** *Let $G_1, \ldots, G_L \sim \mathcal{N}(0, \sigma^2)^{d \times n}$. There exists a constant $C_{\mathcal{N}}$ such that with probability at least $1 - e^{-x}$,*

$$\max_{l \in [L]} \|G_l\| \leq C_{\mathcal{N}}\sigma(\sqrt{n} + \sqrt{d} + \sqrt{\ln(2L + x)}). \tag{101}$$

Lemma I.2 (Lemma A.2 of Hardt & Price (2014))**.** *Let $U \in \mathbb{R}^{d \times p}$ be a matrix with orthonormal columns. Let $G_1, \ldots, G_L \sim \mathcal{N}(0, \sigma^2)^{d \times p}$ with $0 \leq p \leq d$. There exists a constant $C_{\mathcal{N}}$ such that with probability at least $1 - e^{-x}$,*

$$\max_{l \in [L]} \|U^\top G_l\| \leq C_{\mathcal{N}}\sigma(2\sqrt{p} + \sqrt{\ln(2L + x)}). \tag{102}$$

Lemma I.3 (Minimum Singular Value of a Square Gaussian Matrix (Theorem 1.2 of Rudelson & Vershynin (2008)))**.** *Let $A \in \mathbb{R}^{k \times k}$ be a Gaussian random matrix, i.e. $\boldsymbol{A}_{ij} \sim \mathcal{N}(0, 1)$. Then, for every $\epsilon > 0$, we have*

$$\Pr\{s_{\min}(A) \leq \epsilon k^{-1/2}\} \leq C\epsilon + c^k, \tag{103}$$

where $C > 0$ and $c \in (0, 1)$ are absolute constants.