FedSC: Provable Federated Self-supervised Learning with Spectral Contrastive Objective over Non-i.i.d. Data

Shusen Jing [1] **Anlan Yu** [2] **Shuai Zhang** [3] **Songyang Zhang** [4]

Abstract

Recent efforts have been made to integrate self-supervised learning (SSL) with the framework of federated learning (FL). One unique challenge of federated self-supervised learning (FedSSL) is that the global objective of FedSSL usually does not equal the weighted sum of local SSL objectives. Consequently, conventional approaches, such as federated averaging (FedAvg), fail to precisely minimize the FedSSL global objective, often resulting in suboptimal performance, especially when data is non-i.i.d.. To fill this gap, we propose a provable FedSSL algorithm, named FedSC, based on the spectral contrastive objective. In FedSC, clients share correlation matrices of data representations in addition to model weights periodically, which enables inter-client contrast of data samples in addition to intra-client contrast and contraction, resulting in improved quality of data representations. Differential privacy (DP) protection is deployed to control the additional privacy leakage on local datasets when correlation matrices are shared. We also provide theoretical analysis on the convergence and extra privacy leakage. The experimental results validate the effectiveness of our proposed algorithm.

1. Introduction

As a type of unsupervised learning, self-supervised learning (SSL) aims to learn a structured representation space, in which data similarity can be measured by simple metrics, such as cosine and Euclidean distances, with unlabeled data (Chen et al., 2020; Chen & He, 2021; Grill et al., 2020; He et al., 2020; Zbontar et al., 2021; Bardes et al., 2021; HaoChen et al., 2021). On top of the foundation model trained with SSL, a simple linear layer, also known as linear probe, is sufficient to perform well on a wide range of downstream tasks with minimal labeled data. Resulting from its high label efficiency, SSL has been adopted in a variety of applications, such as natural language processing (He et al., 2021; Brown et al., 2020) and computer vision (Ravi & Larochelle, 2016; Hu et al., 2021).

However, SSL algorithms are often executed on massive amounts of unlabeled data that may be dispersed across various locations. Moreover, the progressively tightening privacy-protection regulations frequently inhibit the centralization of data. Within this context, the federated learning (FL) framework is often favored, wherein a central server can learn from private data located on clients without the data being shared directly (McMahan et al., 2017; Stich, 2018; Li et al., 2019).

Despite the extensive study and theoretical guarantees (Stich, 2018; Li et al., 2019) associated with conventional FL, its generalization to incorporate with SSL is not straightforward. The *fundamental challenge* arises from the fact that, unlike FL within supervised learning, the global objective of FedSSL usually does not equal the weighted sum of local SSL objectives. Consequently, conventional FL approaches, e.g. federated averaging (FedAvg), can not minimize the exact global objective of FedSSL especially when data is non-independent and identically distributed (non-i.i.d.). From the perspective of contrastive learning, FedAvg only contrasts data samples within the same client (intra-client) rather than those across different clients (inter-client). Therefore, the learned representation might not be as effective at distinguishing inter-client data samples as it is with intra-client data samples.

Although recent works on FedSSL have shown great numerical success (Zhuang et al., 2021; 2022; Zhang et al., 2023; Han et al., 2022), the majority of them either overlook previously mentioned challenge or fail to offer a theoretical analysis. FedU (Zhuang et al., 2021) and FedEMA (Zhuang

[1]Department of Radiation Oncology, University of California, San Francisco, California, USA [2]Department of Electrical and Computer Engineering, Lehigh University, Bethlehem, Pennsylvania, USA [3]Department of Data Science, New Jersey Institute of Technology, Newark, New Jersey, USA [4]Department of Electrical and Computer Engineering, University of Louisiana at Lafayette, Lafayette, Louisiana, USA. Correspondence to: Shusen Jing <shusen.jing@ucsf.edu>, Songyang Zhang <songyang.zhang@louisiana.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Table 1. A comparison with SOTAs: FedSC (proposed) is the only one applying DP mechanism on components other than encoder. Moreover, FedSC is the only provable method among them.

	Info. shared besides encoder	Privacy Protection	Provable
FedU	predictor	×	×
FedEMA	predictor	×	×
FedX	N/A	×	×
FedCA	representations	×	×
FedSC	correlation matrices	√	√

et al., 2021) lack the formulation of global objective and thus fail to provide theoretical analysis. FedCA (Zhang et al., 2023) notices the unique challenge and proposes to share data representations, which, however, results in significant privacy leakage and communication overhead. Unlike FedU and FedEMA, which involve sharing predictors, and FedCA, which shares data representations, our proposed FedSC results in much lower communication costs, since sharing correlation matrices requires transmitting far fewer parameters than what is needed for predictors or data representations. FedX (Han et al., 2022) does not share additional information besides encoders, but still lacks theoretical analysis. Among all these works, only our proposed FedSC deploys differential privacy (DP) protection to mitigate the extra privacy leakage from components other than encoders. Moreover, FedSC is the only provable FedSSL method to the best knowledge of the authors. Table 1 summarizes the difference between this work and state of the arts (SOTAs).

Contribution. In this work, we propose a novel FedSSL formulation based on the spectral contrastive (SC) objective (HaoChen et al., 2021). The formulation clarifies all the necessary components in FedSSL encompassing intra-client contraction, intra-client contrast and inter-client contrast. Building upon this formulation, we propose the first provable FedSSL method, namely **FedSC**, with the convergence guarantee to the solutions of centralized SSL. Unlike FedAvg, clients in FedSC share correlation matrices of their local data representations in addition to the weights of local models. By leveraging the aggregated correlation matrix from the server, inter-client contrast of data samples, which is overlooked in FedAvg, can be performed in addition to local contrast and contraction. To better control and quantify the extra privacy leakage, we apply DP mechanism to correlation matrices when they are shared. We made theoretical analysis of FedSC, demonstrating the convergence of the global objective and efficacy of our method. Our contributions are summarized as follows:

• We propose a novel FedSSL formulation delineating all essential components of FedSSL, which encompasses intra-client contraction, intra-client contrast and inter-client contrast. This highlights the limitations of FedAvg due to its neglect of the inter-client contrast.

• We propose FedSC, in which clients are able to perform inter-client contrast of data samples by leveraging the correlation matrices of data representations shared from others, resulting in improved quality of data representations.

• DP protection is applied, which effectively constrains the privacy leakage resulting from sharing correlation matrices with only negligible utility degradation.

• Theoretical analysis of FedSC is made, providing extra privacy leakage and convergence guarantee for the global FedSSL objective. We prove that FedSC can achieve a $\mathcal{O}(1/\sqrt{T})$ convergence rate, while FedAvg will have a constant error floor.

• Through extensive experimentation involving 3 datasets across 4 SOTAs, we affirm that FedSC achieves superior or comparable performance compared with other methods.

2. Related Works

Self-supervised learning. SSL can be mainly categorized into contrastive and non-contrastive SSL. The mechanisms and explicit objective of non-contrastive SSL algorithms are still not fully understood despite a few recent attempts (Halvagal et al., 2023; Tian et al., 2021; Zhang et al., 2022). In contrast, contrastive SSL is more intuitive and explainable. Contrastive SSL explicitly penalizes the distance between positive pairs (two samples share the same semantic meaning), while encouraging distance between negative pairs (two samples share different semantic meanings). For example, SimCLR (Chen et al., 2020) objective accounts for the mutual information between positive pairs (Tschannen et al., 2019) preserved by representations. The SC objective (HaoChen et al., 2021) is equivalent to performing a spectral decomposition of the augmentation graph.

Federated Self-supervised Learning. In FedU (Zhuang et al., 2021), clients make decisions on whether the local model should be updated by the global based on the distances of two model weights when receiving global models from the server. As a follow up, FedEMA (Zhuang et al., 2022) is proposed, in which the hard decision in FedU is replaced with a weighted combination of local and global models. FedX (Han et al., 2022) designs local and global objectives using the idea of cross knowledge distillation to mitigate the effects of non-i.i.d. data. The authors of FedCA (Zhang et al., 2023) propose to share features of individual data samples in addition to local model weights for inter-client contrast, which however, results in significant privacy leakage and communication overhead.

Differential Privacy. Gaussian and Laplace mechanisms are most common DP approaches to protect a dataset from membership attack (Dwork, 2006). To better analyze DP, (Mironov, 2017) proposes Rényi differential privacy (RDP),

which characterizes the operations on mechanisms, such as composition, in a more elegant way, and proves the equivalence between DP and RDP. Currently, DP has been widely deployed in FL (Wei et al., 2020; Truex et al., 2020; Hu et al., 2020; Geyer et al., 2017; Noble et al., 2022).

3. Preliminaries: Spectral Contrastive (SC) Self-supervised Learning

Spectral contrastive (SC) SSL is proposed in (HaoChen et al., 2021) with the following objective:

$$\mathcal{L}^{SC}(\theta;\mathcal{D}) \triangleq \frac{1}{2}\mathsf{E}_{x,x^-\sim\mathcal{A}(\cdot|\mathcal{D})}\left[\left(z(x;\theta)^T z(x^-;\theta)\right)^2\right] - \mathsf{E}_{\bar{x}\sim\mathcal{D}}\mathsf{E}_{x,x^+\sim\mathcal{A}(\cdot|\bar{x})}\left[z(x;\theta)^T z(x^+;\theta)\right],$$

where $\mathcal{D}$ is the dataset; $z(\cdot;\theta):\mathbb{R}^d\to\mathbb{R}^H$ is the representation mapping parameterized by θ; E is referred to as the operator of expectation; $\mathcal{A}(\cdot|\bar{x})$ is referred to as the augmentation kernel, which is essentially a conditional distribution, and $\mathcal{A}(\cdot|\mathcal{D}) \triangleq \mathsf{E}_{\bar{x}\sim\mathcal{D}}\mathcal{A}(\cdot|\bar{x})$. We use (x,x^-) to denote negative pairs, where sample x and x^- have different semantic meanings, and (x,x^+) to denote positive pairs, where x and x^+ have same semantic meaning. Intuitively, minimizing the SC objective encourages the orthogonality of representations of a negative pair, and simultaneously promotes linear alignment of representations of a positive pair. It has been proved that solving this optimization problem is equivalent to doing spectral decomposition of a well-defined augmentation graph, whose nodes are augmented images, i.e, from $\mathcal{A}(\cdot|\mathcal{D})$, and edges describe the semantic similarity of two images determined by the kernel $\mathcal{A}(\cdot|\cdot)$, which results in high-quality and explainable data (node) representations (HaoChen et al., 2021).

After rearranging the original SC objective, we first propose the following equivalent form not reported in (HaoChen et al., 2021).

$$\mathcal{L}^{SC}(\theta;\mathcal{D}) = -\mathsf{E}_{\bar{x}\sim\mathcal{D}}Tr\{R^+(\bar{x};\theta)\} + \frac{1}{2}\left\|\mathsf{E}_{\bar{x}\sim\mathcal{D}}R(\bar{x};\theta)\right\|_F^2, \tag{1}$$

where $R^+(\bar{x};\theta)\in\mathbb{R}^{H\times H}$ and $R(\bar{x};\theta)\in\mathbb{R}^{H\times H}$ are defined respectively as

$$R^+(\bar{x};\theta) \triangleq \mathsf{E}_{x,x^+\sim\mathcal{A}(\cdot|\bar{x})}\left[z(x;\theta)z(x^+;\theta)^T\right],$$
$$R(\bar{x};\theta) \triangleq \mathsf{E}_{x\sim\mathcal{A}(\cdot|\bar{x})}\left[z(x;\theta)z(x;\theta)^T\right].$$

The detailed derivations are given in Appendix A.

4. Problem Formulation

In an FL system consists of a server and J clients, the j-th client owns a private local dataset $\mathcal{D}_j$ disjoint with others. The goal of FedSSL is to optimize the SSL model (SC model in this work) over the union of all local datasets, i.e,

$$\min_\theta \mathcal{L}^{SC}(\theta;\mathcal{D}), \tag{2}$$

where $\mathcal{D} = \cup_{j=1}^J \mathcal{D}_j$. Like the majority of SSL objectives, the global SC objective typically does not equal the weighted sum of local SC objectives, especially with non-i.i.d. data distribution. For the purpose of rigor, we make it an assumption instead of a claim in this work as follows.

$$\mathcal{L}^{SC}(\theta;\mathcal{D}) \neq \sum_{j=1}^J q_j \mathcal{L}^{SC}(\theta;\mathcal{D}_j), \tag{3}$$

where $\{q_j\}$ are weights depending on the amount of local data. As a result, FedAvg is not guaranteed to minimize the global objective $\mathcal{L}^{SC}(\theta;\mathcal{D})$ when data is non-i.i.d..

In addition, we adopt SC framework for the following reasons: First, SC has solid theoretical derivations and simultaneously achieves performance comparable to SOTA SSL methods (HaoChen et al., 2021). Second, the SC objective suggests that correlation matrices of data representations are sufficient for contrasting negative-pairs. Sharing correlation matrices only results in constant negligible extra communication overheads and quantifiable privacy leakage.

5. FedSC: A Provable FedSSL Method

For the simplification of notations, we denote $R_j^+(\theta) = \mathsf{E}_{\bar{x}\sim\mathcal{D}_j}R^+(\bar{x};\theta)$ and $R_j(\theta) = \mathsf{E}_{\bar{x}\sim\mathcal{D}_j}R(\bar{x};\theta)$ the positive correlation matrix and correlation matrix, respectively. We start with manipulating the global objective

$$\begin{aligned}\mathcal{L}^{SC}(\theta;\mathcal{D}) &= -\sum_{j=1}^J q_j Tr\{R_j^+(\theta)\} + \frac{1}{2}\left\|\sum_{j=1}^J q_j R_j(\theta)\right\|_F^2 \\ &= \sum_{j=1}^J q_j\Bigg(\underbrace{-Tr\{R_j^+(\theta)\}}_{\text{intra-client contraction}} + \underbrace{\frac{1}{2}q_j\left\|R_j(\theta)\right\|_F^2}_{\text{intra-client contrast}} \\ &\quad + \underbrace{\frac{1}{2}(1-q_j)Tr\{R_j(\theta)R_{-j}(\theta)\}}_{\text{inter-client contrast}}\Bigg),\end{aligned} \tag{4}$$

where $R_{-j}(\theta) \triangleq \frac{1}{1-q_j}\sum_{j'\neq j} q_{j'}R_{j'}(\theta)$. From Eq. (4), we notice that $\mathcal{L}^{SC}(\theta;\mathcal{D})$ can be decomposed into a weighted sum of J terms corresponding to J clients, where each term consists of three sub-terms accounting for intra-client contraction (of positive pairs), intra-client contrast (of negative pairs), and inter-clients contrast (of negative pairs), respectively. Inspired by this decomposition, we construct the

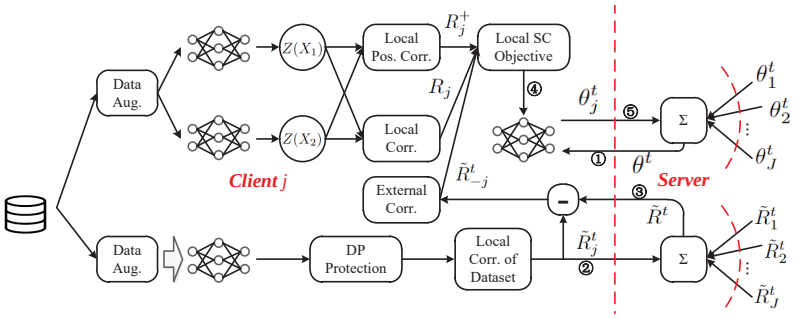


Figure 1. Diagram of the proposed FedSC. 1) The server synchronizes local models with the global model. 2) Clients compute their local correlation matrices of dataset and send them to the server. 3) The server distributes the aggregated global correlation matrices back to the clients. 4) The clients proceed to update their local models in accordance with the local objective specified in Eq. (5). 5) The server aggregates the local models and initiates the next iteration.

following local objective:

$$\mathcal{L}_j^{SC}(\theta;\bar{R}_{-j}) = -Tr\{R_j^+(\theta)\} + \frac{1}{2}q_j \left\|R_j(\theta)\right\|_F^2 + (1-q_j)Tr\{R_j(\theta)\bar{R}_{-j}\}, \tag{5}$$

where $\bar{R}_{-j} \in \mathbb{R}^{H\times H}$ is an estimate of $R_{-j}(\theta)$, whose updates relying on the communication with the server. Since $\bar{R}_{-j}$ is treated as a constant (stop gradient) in local objectives, we intentionally remove the coefficient $1/2$ before the third term for gradient alignment between local and global objectives. That is to say, when $\bar{R}_{-j} = R_{-j}(\theta)$, we have

$$\nabla\mathcal{L}^{SC}(\theta;\mathcal{D}) = \sum_{j=1}^{J} q_j \nabla\mathcal{L}_j^{SC}(\theta,\bar{R}_{-j}). \tag{6}$$

Note that directly applying FedAvg results in a misalignment of gradients, which is inherited from the fact that the global objective of FedSSL does not equal to the weighted sum of local objectives as suggested in Eq. (3).

The process of FedSC is similar to FedAvg, except sharing and aggregating local correlation matrices $\tilde{R}_j^t$ besides model weights. To begin with, the server synchronizes local models with the global model. Subsequently, clients compute their local correlation matrices and send them to the server. Following this, the server distributes the aggregated global correlation matrices back to the clients. The clients then proceed to update their local models in accordance with the local objective specified in Eq. (5). Finally, the server aggregates the local models and initiates the next iteration. The process is summarized in Fig. 1.

The detailed algorithm of FedSC is shown in Algorithm 1. Here, clients use Algorithm 2 `DP-CalR` to calculate local correlation matrices to be shared $\tilde{R}_j^t$ with differential privacy (DP) protection, which is detailed in Sec. 5.1. During local training, clients minimize $\mathcal{L}_j^{SC}$ through stochastic gradient descent (SGD), which is detailed in Sec. 5.2. It can be noticed that both clients and the server maintain the knowledge of global correlation matrix $\tilde{R}^t$.

Intuitively, since the averaged local gradients align with the global gradient as shown in Eq. (6), the drift and variance of local gradients contribute $\mathcal{O}(1/T)$ and $\mathcal{O}(1/\sqrt{T})$ to the convergence rate, respectively, which has been extensively studied by previous works on FedAvg. The difference is that the shared correlation matrix $\tilde{R}_j^t$ introduces additional perturbation due to its aging (compared with instant correlation matrix $R_{-j}(\theta)$) and DP noise. The perturbation caused by aging is proportional to the movements of weights, which is proportional to the squared learning rate η^2, thus contributing an additional $\mathcal{O}(1/T)$ factor to the convergence rate. This is what motivates the design of FedSC.

5.1. Correlation Matrices Sharing

DP protection is applied when correlation matrices are shared to mitigate additional privacy leakage on local dataset. A typical Gaussian mechanism is adopted, with parameters μ and σ^2 controlling sensitivity and noise scale, respectively. The process is summarized in Algorithm 2.

5.2. Local Training

The local training process follows mini-batch stochastic gradient descent (SGD). At each iteration, consider a batch of B samples $\bar{X} \sim \mathcal{D}_j$. Let $X_1, X_2, ..., X_{2V} \sim \mathcal{A}(\bar{X})$ be $2V$ views augmented from $\bar{X}$. The empirical correlation matrices are calculated as follows:

$$\hat{R}_j^+(\{X_v\}_v;\theta_j) = \frac{1}{2BV}\sum_{v=1}^{V}\Big[Z(X_v;\theta_j)Z(X_{v+V};\theta_j)^T + Z(X_{v+V};\theta_j)Z(X_v;\theta_j)^T\Big];$$

$$\hat{R}_j(\{X_v\}_v;\theta_j) = \frac{1}{2BV}\sum_{v=1}^{2V} Z(X_v;\theta_j)Z(X_v;\theta_j)^T.$$

The batch loss $\hat{\mathcal{L}}_j^{SC}(\theta_j;\{X_v\}_v,\bar{R}_{-j})$ can be obtained by substitute $R_j^+(\theta)$ and $R_j(\theta)$ in Eq. (5) with $\hat{R}_j^+(\{X_v\}_v,\theta_j)$

Algorithm 1 `FedSC`

Initialization: θ^0 and a set of clients $[J]$
for $t = 1...T$ **do**
  Server samples a subset of clients $\mathcal{J}^t \subset [J]$
  if $t = 1$ **then**
    for $j \in [J]$ **do**
      Server sends θ^{t-1} to client j.
      Client j uploads $\tilde{R}_j^t = \texttt{DP-CalR}(\theta^{t-1}, \mathcal{D}_j)$.
    end for
    Server sends $\tilde{R}^t = \sum_{j=1}^J q_j \tilde{R}_j^t$ to all clients.
  else
    for $j \in \mathcal{J}^t$ **do**
      Server sends θ^{t-1} to client j.
      Client j uploads $\tilde{R}_j^t = \texttt{DP-CalR}(\theta^{t-1}, \mathcal{D}_j)$.
      Server updates $\tilde{R}^t = \tilde{R}^{t-1} - q_j \tilde{R}_j^{t-1} + q_j \tilde{R}_j^t$.
    end for
    Server sends $\tilde{R}^t$ to all clients.
    Server sets $\tilde{R}_j^t = \tilde{R}_j^{t-1}$ for $j \notin \mathcal{J}^t$.
  end if
  for $j \in \mathcal{J}^t$ **do**
    Client j calculates $\tilde{R}_{-j}^t = \frac{1}{1-q_j}(\tilde{R}^t - q_j \tilde{R}_j^t)$.
    Client j trains local model with procedures in Sec. 5.2, and returns updated weights θ_j^{t-1}.
  end for
  Server aggregation: $\theta^t = \frac{1}{|\mathcal{J}^t|} \sum_{j \in \mathcal{J}^t} \theta_j^{t-1}$.
end for
return: θ^T

Algorithm 2 `DP-CalR`

Inputs: θ and local dataset $\mathcal{D}_j$
$\tilde{R}_j^t = 0$
for $\bar{x} \in \mathcal{D}_j$ **do**
  Sample $x_1, x_2, ..., x_V \sim \mathcal{A}(\cdot|\bar{x})$
  Calculate $\check{z}_v = \texttt{NormClip}(z(x_v; \theta), \sqrt{\mu})$, for $v = 1, 2, ..., V$.
  $\tilde{R}_j^t = \tilde{R}_j^t + \frac{1}{|\mathcal{D}_j| V} \sum_{v=1}^V \check{z}_v \check{z}_v^T$
end for
$\tilde{R}_j^t = \tilde{R}_j^t + \mathcal{N}(0, \sigma^2 \mathbf{I})$.
return: $\tilde{R}_j^t$

and $\hat{R}_j(\{X_v\}_v, \theta_j)$, respectively. The local training follows by back-propagating the batch loss and updating the model weights iteratively.

5.3. Comparison with existing FedSSL frameworks

In this subsection, we discuss the privacy leakage and communication overhead of FedSC in comparison with other FedSSL frameworks.

Sharing correlation matrices only results in negligible communication overhead: Although FedSC shares correlation matrices in addition, it still results in less communication overhead than SOTA non-contrastive FedSSL frameworks (Zhuang et al., 2021; 2022; He et al., 2020), due to the implementation of predictor in non-contrastive SSL methods. For example, the feature dimension is $H = 512$ in our experiments, thus the correlation matrices yield $H \times H \approx 260,000$ additional parameters to be communicated. In contrast, the structure of the predictor is often a three-layer multilayer perceptron (MLP), which contains parameters that are multiples of the correlation matrices. In our case, we choose a typical size of $(512 - 1024 - 512)$ resulting in $1,000,000$ parameters. The overhead of correlation matrices is negligible compared with the encoders. Therefore, even compared with contrastive SSL, which does not have a predictor, the communication overhead resulting from sharing correlation matrices is not a concern.

The extra privacy leakage is probably comparable to that caused by sharing predictors: The predictors in non-contrastive SSL also lead to potential privacy leakages. Although theoretical characterization has not been established, recent works shed lights on the operational meaning of the predictors (Halvagal et al., 2023; Tian et al., 2021), suggesting what information is probably leaked. Particularly, (Tian et al., 2021) reports that linear predictors in BYOL align with the correlation matrices $R(\theta)$ during training. This interesting finding suggests that predictors probably contain similar information as the correlation matrices.

6. Theoretical Analysis

In this section, we first analyze the additional privacy leakage and convergence of FedSC. Our findings are summarized as follows:

- We prove that sharing correlation matrices through `DP-CalR` results in $\left(\frac{T\mu^2}{2\sigma^2} + \sqrt{\frac{2T\mu^2 \log 1/\delta}{\sigma^2}}, \delta\right)$-DP.
- We provide the convergence analysis of FedSC. Specifically, with large batch size B, large number of views V and small scale of DP noise σ, we can achieve a convergence rate close to $\mathcal{O}(1/\sqrt{T})$.
- The analysis indicates superior performance of FedSC over FedAvg whose convergence is dominated by a $\mathcal{O}(1)$ constant meaning error floor.

6.1. Additional Privacy Leakage

In this subsection, we analyze the Gaussian mechanism in Algorithm 2 `DP-CalR`. We start with definitions of variations of differential privacy (DP).

Definition 6.1 ((ϵ, δ)-DP)**.** A mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ is (ϵ, δ)-DP, if for any neighboring $X, X' \in \mathcal{X}$ and $\mathcal{S} \subset \mathcal{Y}$,

the following inequality is satisfied.

$$Pr(\mathcal{M}(X) \in \mathcal{S}) \leq e^{\epsilon} Pr(\mathcal{M}(X') \in \mathcal{S}) + \delta.$$

DP protects the inputs of a mechanism from membership inference attacks. For a mechanism satisfying DP, we expect that one can hardly tell whether the input contains a certain entry by only looking at the output. In our case, we do not want the server to know whether a local dataset contains a particular data point.

Definition 6.2 ((α, ϵ)-RDP (Mironov, 2017))**.** A mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ has (α, ϵ)-Rényi differential privacy, if for any neighboring $X, X' \in \mathcal{X}$, $Y = \mathcal{M}(X)$ and $Y' = \mathcal{M}(X')$, the following inequality is satisfied:

$$D_\alpha(P_Y || P_{Y'}) \leq \epsilon,$$

where $D_\alpha(P_Y || P_{Y'})$ is Rényi-divergence of order $\alpha > 1$

$$D_\alpha(P_Y || P_{Y'}) \triangleq \frac{1}{\alpha - 1} \log \mathsf{E}_{Y'} \left(\frac{P_Y(Y')}{P_{Y'}(Y')} \right)^{\alpha}.$$

RDP is a variation of DP with many good properties, which are summarized in the following Lemmas.

Lemma 6.3 (Gaussian Mechanism of RDP (Mironov, 2017))**.** *Let $f : \mathcal{X} \to \mathbb{R}^n$ be a function with l_2 sensitivity W, then the Gaussian mechanism $G_f(\cdot) = f(\cdot) + \mathcal{N}(0, \mathbf{I}_n \sigma^2)$ is $(\alpha, \frac{\alpha W^2}{2\sigma^2})$-RDP.*

Lemma 6.4 (Composition of RDP (Mironov, 2017))**.** *Let $\mathcal{M}_1 : \mathcal{X} \to \mathcal{Y}$ be (α, ϵ_1)-RDP, and $\mathcal{M}_2 : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ be (α, ϵ_2)-RDP. Then the mechanism $\mathcal{M}_3 : \mathcal{X} \to \mathcal{Y} \times \mathcal{Z}, X \mapsto (\mathcal{M}_1(X), \mathcal{M}_2(X, \mathcal{M}_1(X)))$ is $(\alpha, \epsilon_1 + \epsilon_2)$-RDP to X.*

Lemma 6.5 ((Mironov, 2017))**.** *If a mechanism is (α, ϵ)-RDP, then it is $(\epsilon + \frac{\log 1/\delta}{\alpha - 1}, \delta)$-DP.*

With all these preparations, we use the following proposition to characterize the additional privacy leakage of FedSC.

Proposition 6.6 (Additional Privacy Leakage of FedSC)**.** *Sharing correlation matrices for T_j times through Algorithm 2 `DP-CalR` results in $\left(\frac{T_j \mu^2}{2\sigma^2} + \sqrt{\frac{2T_j \mu^2 \log 1/\delta}{\sigma^2}}, \delta \right)$-DP.*

Proof. We start with the sensitivity of `DP-CalR`.

$$\begin{aligned}
\left\| \tilde{R}_j^t \right\|_F &= \left\| \frac{1}{|\mathcal{D}_j|} \sum_{\bar{x} \in \mathcal{D}_j} \frac{1}{V} \sum_{v=1}^{V} \check{z}_v(\bar{x}) \check{z}_v(\bar{x})^T \right\|_F \\
&\leq \left\| \frac{1}{|\mathcal{D}_j|} \sum_{\bar{x}' \in \mathcal{D}_j / \bar{x}} \frac{1}{V} \sum_{v=1}^{V} \check{z}_v(\bar{x}') \check{z}_v(\bar{x}')^T \right\|_F \\
&\quad + \frac{1}{|\mathcal{D}_j|} \frac{1}{V} \sum_{v=1}^{V} \left\| \check{z}_v(\bar{x}) \check{z}_v(\bar{x})^T \right\|_F
\end{aligned}$$

where $\check{z}_v(\bar{x})$ is the representation of the v-th view of data $\bar{x}$. Notice that for any $\bar{x}$

$$\begin{aligned}
\left\| \check{z}_v(\bar{x}) \check{z}_v(\bar{x})^T \right\|_F &= \sqrt{Tr\left\{ \check{z}_v(\bar{x}) \check{z}_v(\bar{x})^T \check{z}_v(\bar{x}) \check{z}_v(\bar{x})^T \right\}} \\
&\leq \mu
\end{aligned}$$

The sensitivity is finally bounded by $\mu / |\mathcal{D}_j|$. With Lemma 6.3, we have `DP-CalR` is $\left(\alpha, \frac{\alpha \mu^2}{2\sigma^2 |\mathcal{D}_j|^2} \right)$-RDP. With Lemma 6.4, sharing correlation matrices for T_j times results in $\left(\alpha, \frac{T_j \alpha \mu^2}{2\sigma^2 |\mathcal{D}_j|^2} \right)$-RDP, which is $\left(\frac{T_j \mu^2}{2\sigma^2 |\mathcal{D}_j|^2} + \sqrt{\frac{2T_j \mu^2 \log 1/\delta}{\sigma^2 |\mathcal{D}_j|^2}}, \delta \right)$-DP using Lemma 6.5 with $\alpha = \sqrt{\frac{2\sigma^2 |\mathcal{D}_j|^2 \log 1/\delta}{T_j \mu^2}} + 1$. □

From the results, we can notice that for arbitrarily δ, T_j and σ, the parameter $\epsilon = \frac{T_j \mu^2}{2\sigma^2 |\mathcal{D}_j|^2} + \sqrt{\frac{2T_j \mu^2 \log 1/\delta}{\sigma^2 |\mathcal{D}_j|^2}}$ approaches to zero when the size of local dataset approaches to infinity, indicating no differential privacy leakage.

6.2. Convergence of FedSC

This subsection presents the convergence of FedSC. We begin with the following assumptions.

Assumption 6.7. For any θ and x, NN's output is bounded in norm $||z(x, \theta)||_2 \leq A_0$ for some constant A_0.

Assumption 6.8. For any θ and x, the Jacobian matrix of NN's output is bounded in norm $||\nabla_\theta z(x, \theta)||_F \leq A_1$ for some constant A_1.

Assumption 6.9. The function represented by NN has bounded second order derivatives, i.e, for any θ and x,

$$\sum_{m,p} \|\partial_m \partial_p z(x; \theta)\|_2^2 \leq A_2^2$$

for some constant A_2, where ∂_p refers to the partial derivation with respect to the p-th entry of θ.

Assumption 6.7 can be satisfied when the NN has a normalization layer at the end or uses bounded activation functions, such as sigmoid, at the output layer. Assumption 6.8 accounts for the Lipschitz continuity of $z(x, \theta)$, which is often the case when hidden layers of a NN uses activation functions, such as tanh, sigmoid and relu. Note that Assumption 6.8 is weaker than the bounded gradient norm assumption used in previous works (Li et al., 2019; Noble et al., 2022). However, in our case, it can lead to bounded gradient norm due to the structure of SC objectives, which is detailed in the appendices. Assumption 6.9 accounts for the strong smoothness of the NN, which is widely adopted in existing works (Karimireddy et al., 2020b; Li et al., 2019; Karimireddy et al., 2020a) . With these assumptions, we demonstrate the convergence of FedSC with the following theorem.

Table 2. Performance comparison between FedSC and SOTAs on benchmark tasks: FedSC outperforms most of the SOTAs under different settings. Here we use bold and underline to mark the highest and second highest accuracy, respectively.

	SVHN	CIFAR10	CIFAR100	SVHN	CIFAR10	CIFAR100
Participation	5/5	10/10	20/20	2/5	2/10	4/20
FedAvg + BYOL	87.85 ± 0.49	68.14 ± 0.51	43.54 ± 1.12	87.10 ± 0.79	65.28 ± 0.83	38.77 ± 1.04
FedAvg + SC	90.52 ± 0.42	77.82 ± 0.82	56.24 ± 0.19	89.89 ± 0.94	75.36 ± 0.36	42.95 ± 0.52
FedU	87.92 ± 0.31	68.39 ± 0.69	43.81 ± 0.98	87.40 ± 0.75	65.52 ± 0.51	39.11 ± 0.92
FedEMA	$\mathbf{91.87} \pm 0.30$	68.78 ± 0.25	44.18 ± 0.73	88.97 ± 0.82	65.93 ± 0.63	39.78 ± 1.20
FedX	74.60 ± 0.72	59.17 ± 0.93	39.70 ± 0.39	73.34 ± 0.88	57.42 ± 0.91	33.54 ± 0.67
FedCA	89.92 ± 0.14	78.22 ± 0.22	52.35 ± 0.09	89.28 ± 0.44	$\mathbf{77.22} \pm 0.65$	51.58 ± 0.18
FedSC (Proposed)	$\underline{91.78} \pm 0.49$	$\mathbf{80.06} \pm 0.35$	$\mathbf{58.35} \pm 0.15$	$\mathbf{91.03} \pm 0.58$	$\underline{77.12} \pm 0.44$	$\mathbf{56.64} \pm 0.65$
Centralized SC	93.17 ± 0.13	90.21 ± 0.08	64.32 ± 0.05	–	–	–

Theorem 6.10. *Let Assumption 6.7, 6.8 and 6.9 hold. Choose $\mu > A_0^2$, and the local learning rate $\eta = \mathcal{O}\left(\frac{1}{\sqrt{TE}}\right)$, where T and E are the number of communication rounds and local updates, respectively. Then FedSC achieves*

$$\begin{aligned}
&\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathsf{E}\left[\left\|\nabla\mathcal{L}^{SC}(\theta^{t,e};\mathcal{D})\right\|^2\right] \\
&\leq \mathcal{O}\Bigg(\frac{E^2(H^2\sigma^2+C_2)}{TE} + \frac{\sqrt{E}\sqrt{\frac{J/|\mathcal{J}|-1}{J-1}}}{\sqrt{T}} \\
&+ \frac{\left(\frac{1}{V}+\max_j\frac{|\mathcal{D}_j|/B-1}{|\mathcal{D}_j|-1}\right)\left(H^2\sigma^2+C_4\right)}{\sqrt{TE}} \\
&+ \frac{1}{V}+\max_j\frac{|\mathcal{D}_j|/B-1}{|\mathcal{D}_j|-1}+H^2\sigma^2\Bigg)
\end{aligned} \tag{7}$$

where $\theta^{t,e} \triangleq \frac{1}{|\mathcal{J}^t|}\sum_{j\in\mathcal{J}^t}\theta_j^{t,e}$ is the virtual averaged weights, and $\theta_j^{t,e}$ the local weights of client j at the e-th step in the t-th round; B and V are batch size and number of augmented views, respectively; C_2 and C_4 here are constants only depending on A_0, A_1 and A_2.

6.2.1. Superior performance of FedSC

The convergence rate of FedSC is dominated by the following term when T approaches infinity

$$\gamma = \mathcal{O}\left(\frac{1}{V}+\max_j\frac{|\mathcal{D}_j|/B-1}{|\mathcal{D}_j|-1}+H^2\sigma^2\right). \tag{8}$$

The bias of local batch gradients results in a rate of $\left(\frac{1}{V}+\max_j\frac{|\mathcal{D}_j|/B-1}{|\mathcal{D}_j|-1}\right)$, in which specifically, sampling the data set $\mathcal{D}_j$ (without replacement) leads to the rate of $\frac{|\mathcal{D}_j|/B-1}{|\mathcal{D}_j|-1}$, and sampling the augmentation kernel $\mathcal{A}(\cdot|\bar{x})$ results in $\frac{1}{V}$. Note that this bias also exists in centralized SSL training, and does not result from federation. Sampling variance, from the augmentation kernel, in the shared correlation matrix contributes $\frac{1}{V}$ to the convergence. The DP noise contributes $H^2\sigma^2$. If we set batch size $B = |\mathcal{D}_j|$, generate infinite number of views $V = \infty$ and not apply DP protection, i.e., $\sigma^2 = 0$, then the error floor will disappear, which results in convergence rate similar to FedAvg in supervised FL. In comparison, if we directly use the SC objective without modification and apply FedAvg, there will be a constant error floor independent with batch size B and number of views V, due to the misalignment between the averaged local objectives and the global objectives.

6.2.2. Sketch of Proof

We begin with the case of full clients participation. The convergence is determined by two terms: 1) The squared norm of the bias of the averaged local gradient and 2) The variance of the averaged local gradient. The norm of bias can be factorized into three components: 1.a) The "drift" of the local weights, leading to a factor of $\mathcal{O}(1/T)$. 1.b) The bias in the batch gradient of local objectives $\mathcal{L}_j^{SC}(\theta;\bar{R}_{-j})$, contributes a factor of $\mathcal{O}\left(\frac{1}{V}+\max_j\frac{|\mathcal{D}j|/B-1}{|\mathcal{D}j|-1}\right)$. The bias is due to the fact that the SSL objective can not be written as a sum of samples losses like in the supervised learning cases. 1.c) The impact of aging, sample variance, and DP noise in the shared correlation matrices $\tilde{R}_{-j}^t$. Note that $\tilde{R}_{-j}^t$ remains constant during local training. The aging (compared with $R_{-j}(\theta^{t,e})$) leads to a bias proportional to the drift of local weights, resulting in a factor of $\mathcal{O}(1/T)$. The sampling variance in $\tilde{R}_{-j}^t$ contributes a factor of $\mathcal{O}(1/V)$. DP noise contributes a factor of $\mathcal{O}(H^2\sigma^2)$, where H is the dimension of the representation. The variance in the averaged local gradient contributes a factor of $\mathcal{O}(1/\sqrt{T})$, considering 2.a) the variance in batch gradient sampling and 2.b) the DP noise in $\tilde{R}_{-j}^t$.For partial client participation, we need to consider the variance in aggregation and additional aging of $\tilde{R}_{-j}^t$. Given bounded gradient norm, the variance due to client sampling is $\mathcal{O}\left(\sqrt{E}\sqrt{\frac{J/|\mathcal{J}|-1}{T(J-1)}}\right)$. Additional aging is proportional to the extra drift, leading to a rate of $\mathcal{O}(1/T)$.

Table 3. FedSC with different levels of DP protections ($\delta = 10^{-2}$): deploying DP leads to negligible performance degradation.

	SVHN	CIFAR10	SVHN	CIFAR10
Participation	$5/5$	$10/10$	$2/5$	$2/10$
FedAvg+SC	90.52 ± 0.42	77.82 ± 0.82	89.89 ± 0.94	75.36 ± 0.36
$\epsilon = 3$	90.90 ± 0.52	79.21 ± 0.59	90.00 ± 0.72	76.97 ± 0.64
$\epsilon = 6$	91.32 ± 0.87	79.42 ± 0.63	90.12 ± 0.61	77.08 ± 0.46
$\epsilon = \infty$	91.78 ± 0.49	80.06 ± 0.35	91.03 ± 0.58	77.12 ± 0.44

7. Experiments

7.1. Experimental Setup

Datasets: Three datasets, SVHN, CIFAR10 and CIFAR100, are used for evaluation. SVHN is split into 5 disjoint local datasets, each of which contains 2 classes. CIFAR10 is split into 10 disjoint local datasets according to the 10 classes. CIFAR100 is split into 20 disjoint local datasets, each of which contains 5 classes. Therefore, the size of local datasets for SVHN, CIFAR10 and CIFAR100 tasks are $10,000$, $5,000$ and $2,500$, respectively.

Models: For SVHN and CIFAR10, we use a modified version of ResNet20 as backbones. For CIFAR100, the backbone is a modified version of ResNet50.

Hyper-parameters: For all three tasks, the number of communication round $T = 200$, and the number of local epochs is $E = 5$. For SVHN and CIFAR10, the batch size is $B = 512$. For CIFAR100, the batch size is $B = 256$. The number of views $V = 2$ for all experiments. For correlation matrices sharing, the number of views is set as $V = 5$.

Benchmarks: Besides FedAvg+BYOL and FedAvg+SC, we also compare with the following state of the arts: FedU (Zhuang et al., 2021), FedEMA (Zhuang et al., 2022), FedX (Han et al., 2022) and FedCA (Zhang et al., 2023).

7.2. Experimental Results

Comparison with SOTA approaches: Table 2 presents the performance comparisons of various algorithms under linear evaluation, where the centralized SC serves as an ideal upper bound. We conclude the following **three observations**: (1) Our proposed algorithm, FedSC, demonstrates better or comparable performance across different tasks compared to other 6 methods. (2) FedBYOL, FedU, and FedEMA show good results on SVHN but underperform on CIFAR10 and CIFAR100. We believe that this disparity is caused by the larger local dataset size in SVHN, leading to increased local updates. Since these methods incorporate momentum updates in the target encoder, a larger number of updates might be necessary to effectively initiate local training. (3) FedSC and FedCA exhibit less performance degradation when switched to the partial client participation case. We believe this is because clients in both FedSC and FedCA have extra global information about representations. Additionally, predictors in FedBYOL, FedU, and FedEMA are under the effect of client sampling, hindering their global information provision.

Table 4. FedSC with different levels of DP protections ($\delta = 10^{-2}$): deploying DP leads to negligible performance degradation.

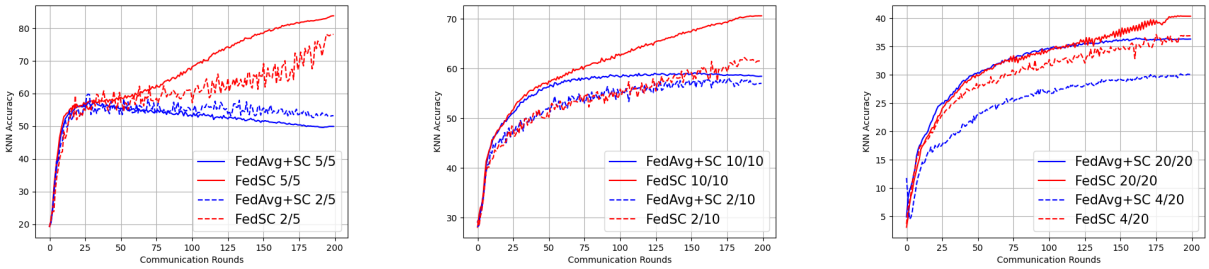
	CIFAR100	
Participation	$20/20$	$4/20$
FedAvg+SC	56.24 ± 0.19	42.95 ± 0.52
$\epsilon = 6$	57.10 ± 0.82	54.87 ± 0.62
$\epsilon = 12$	57.63 ± 0.57	55.76 ± 0.58
$\epsilon = \infty$	58.35 ± 0.15	56.64 ± 0.65

DP Impact: Table 3, 4 and 5 illustrate the impact of the DP mechanism on FedSC's performance. It is shown that with a reasonable degree of DP protection, there is only a modest decline in FedSC's performance, which remains better than that of FedAvg+SC. Given that our focus is on data level DP, the extra privacy leakage shown in the tables is typically insignificant when compared to the leakage resulting from the encoders. On the other hand, according to the analysis in Sec. 6.1, a smaller dataset necessitates a higher level of DP noise to maintain the same degree of privacy protection. The local dataset sizes for SVHN, CIFAR10, and CIFAR100 tasks are $10,000$, $5,000$, and $2,500$, respectively. As a result, for the CIFAR100 task, we choose a slightly higher privacy budget compared to the other two tasks.

Convergence: Fig. 2 compares the convergence of proposed FedSC and FedAvg+SC, in terms of communication rounds and KNN accuracy. The figures reveal that FedAvg+SC tends to experience either a high error rate or

Table 5. FedSC with different levels of DP protections ($\delta = 10^{-4}$): deploying DP leads to negligible performance degradation.

	SVHN	CIFAR10	CIFAR100
Participation	$2/5$	$2/10$	$4/20$
FedAvg+SC	89.89 ± 0.94	75.36 ± 0.36	42.95 ± 0.52
$\epsilon = 3$	89.95 ± 0.81	76.75 ± 0.62	54.22 ± 0.72
$\epsilon = 8$	90.12 ± 0.61	77.08 ± 0.46	54.87 ± 0.62
$\epsilon = \infty$	91.03 ± 0.58	77.12 ± 0.44	56.64 ± 0.65



(a) SVHN (b) CIFAR10 (c) CIFAR100

Figure 2. Convergence of FedSC and FedAvg+SC. 1) FedAvg+SC tends to experience either a high error floor or overfitting. 2) FedSC is able to consistently enhance KNN accuracy. This observation validates our theoretical analysis in Sec. 6.

overfitting as the number of communication rounds grows. In contrast, FedSC can consistently enhance KNN accuracy. This validates our theoretical analysis in Sec. 6.

8. Conclusion

In this paper, we proposed FedSC, a novel FedSSL framework based on spectral contrastive objectives. In FedSC, clients share correlation matrices besides local weights periodically. With shared correlation matrices, clients are able to contrast inter-client sample contrast in addition to intra-client contrast and contraction. To mitigate the extra privacy leakage on local dataset, we adopted DP mechanism on shared correlation matrices. We provided theoretical analysis on privacy leakage and convergence, demonstrating the efficacy of FedSC. To the best knowledge of the authors, this is the first provable FedSSL method.

9. Impact Statements

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Chen, X. and He, K. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15750–15758, 2021.

Dwork, C. Differential privacy. In *International colloquium on automata, languages, and programming*, pp. 1–12. Springer, 2006.

Geyer, R. C., Klein, T., and Nabi, M. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Halvagal, M. S., Laborieux, A., and Zenke, F. Implicit variance regularization in non-contrastive ssl. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Han, S., Park, S., Wu, F., Kim, S., Wu, C., Xie, X., and Cha, M. Fedx: Unsupervised federated learning with cross knowledge distillation. In *European Conference on Computer Vision*, pp. 691–707. Springer, 2022.

HaoChen, J. Z., Wei, C., Gaidon, A., and Ma, T. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

He, J., Zhou, C., Ma, X., Berg-Kirkpatrick, T., and Neubig, G. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*, 2021.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Hu, R., Guo, Y., Li, H., Pei, Q., and Gong, Y. Personalized federated learning with differential privacy. *IEEE Internet of Things Journal*, 7(10):9530–9539, 2020.

Karimireddy, S. P., Jaggi, M., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. Mime: Mimicking centralized stochastic algorithms in federated learning. *arXiv preprint arXiv:2008.03606*, 2020a.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning*, pp. 5132–5143. PMLR, 2020b.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Singh, A. and Zhu, J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017.

Mironov, I. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pp. 263–275. IEEE, 2017.

Noble, M., Bellet, A., and Dieuleveut, A. Differentially private federated learning on heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pp. 10110–10145. PMLR, 2022.

Ravi, S. and Larochelle, H. Optimization as a model for few-shot learning. In *International conference on learning representations*, 2016.

Stich, S. U. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

Tian, Y., Chen, X., and Ganguli, S. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pp. 10268–10278. PMLR, 2021.

Truex, S., Liu, L., Chow, K.-H., Gursoy, M. E., and Wei, W. Ldp-fed: Federated learning with local differential privacy. In *Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking*, pp. 61–66, 2020.

Tschannen, M., Djolonga, J., Rubenstein, P. K., Gelly, S., and Lucic, M. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.

Wei, K., Li, J., Ding, M., Ma, C., Yang, H. H., Farokhi, F., Jin, S., Quek, T. Q., and Poor, H. V. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

Zhang, C., Zhang, K., Zhang, C., Pham, T. X., Yoo, C. D., and Kweon, I. S. How does simsiam avoid collapse without negative samples? a unified understanding with self-supervised contrastive learning. *arXiv preprint arXiv:2203.16262*, 2022.

Zhang, F., Kuang, K., Chen, L., You, Z., Shen, T., Xiao, J., Zhang, Y., Wu, C., Wu, F., Zhuang, Y., et al. Federated unsupervised representation learning. *Frontiers of Information Technology & Electronic Engineering*, 24(8): 1181–1193, 2023.

Zhuang, W., Gan, X., Wen, Y., Zhang, S., and Yi, S. Collaborative unsupervised visual representation learning from decentralized data. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4912–4921, 2021.

Zhuang, W., Wen, Y., and Zhang, S. Divergence-aware federated self-supervised learning. *arXiv preprint arXiv:2204.04385*, 2022.

A. Derivation of SC objective

$$\begin{aligned}
\mathcal{L}^{SC}(\theta;\mathcal{D}) &\triangleq -\mathsf{E}_{\bar{x}\sim\mathcal{D}}\mathsf{E}_{x,x^+\sim\mathcal{A}(\cdot|\bar{x})}\left[z(x;\theta)^T z(x^+;\theta)\right] + \frac{1}{2}\mathsf{E}_{x,x^-\sim\mathcal{A}(\cdot|\mathcal{D})}\left[\left(z(x;\theta)^T z(x^-;\theta)\right)^2\right] \\
&= -\mathsf{E}_{\bar{x}\sim\mathcal{D}}\mathsf{E}_{x,x^+\sim\mathcal{A}(\cdot|\bar{x})}\left[Tr\left\{z(x^+;\theta)z(x;\theta)^T\right\}\right] \\
&+ \frac{1}{2}\mathsf{E}_{x,x^-\sim\mathcal{A}(\cdot|\mathcal{D})}\left[Tr\left\{z(x;\theta)z(x;\theta)^T z(x^-;\theta)z(x^-;\theta)^T\right\}\right] \\
&= -\mathsf{E}_{\bar{x}\sim\mathcal{D}}\left[Tr\left\{R^+(\theta)\right\}\right] + \frac{1}{2}Tr\left\{\mathsf{E}_{x\sim\mathcal{A}(\cdot|\mathcal{D})}z(x;\theta)z(x;\theta)^T\mathsf{E}_{x^-\sim\mathcal{A}(\cdot|\mathcal{D})}z(x^-;\theta)z(x^-;\theta)^T\right\} \\
&= -\mathsf{E}_{\bar{x}\sim\mathcal{D}}\left[Tr\left\{R^+(\theta)\right\}\right] + \frac{1}{2}\left\|\mathsf{E}_{x\sim\mathcal{A}(\cdot|\mathcal{D})}z(x;\theta)z(x;\theta)^T\right\|_F^2 \\
&= -\mathsf{E}_{\bar{x}\sim\mathcal{D}}Tr\{R^+(\bar{x};\theta)\} + \frac{1}{2}\left\|\mathsf{E}_{\bar{x}\sim\mathcal{D}}R(\bar{x};\theta)\right\|_F^2
\end{aligned} \tag{9}$$

B. Proof of Theorem 6.10

B.1. Additional Notations

Let $\theta_j^{t,e}$ and $v_j^{t,e}$ be the local weights and local SGD direction, respectively, at the e-th update in the t-th communication round. Denote $\theta^{t,e} \triangleq \sum_j q_j\theta_j^{t,e}$ and $v_j^{t,e} \triangleq \sum_j q_j v_j^{t,e}$ the virtual averaged weights and moving direction, respectively. Since the server aggregates periodically, we have $\theta^t = \theta^{t,0}$. For simplicity, we remove the up-script "SC" in $\mathcal{L}^{SC}(\theta)$ and $\mathcal{L}^{SC}(\theta, \tilde{R}_{-j}^t)$ without ambiguity.

B.2. Assumptions

Assumption B.1. For any θ and x, NN's output is bounded in norm $||z(x,\theta)||_2 < A_0$.

Assumption B.2. For any θ and x, the Jacobin of NN's output is bounded in norm $||\nabla z(x,\theta)||_F < A_1$.

Assumption B.3. The function represented by NN has bounded second order derivatives, i.e, for any θ and x

$$\sum_{m,p}\left\|\partial_m\partial_p z(x;\theta)\right\|_2^2 \le A_2^2 \tag{10}$$

B.3. Lemmas

Lemma B.4. *For any $\bar{x}$, $\{X_v\}_v$, θ and j, the following inequalities hold*

$$\left\|R(\bar{x},\theta)\right\|_F^2, \left\|R(\theta)\right\|_F^2, \left\|R_j(\theta)\right\|_F^2, \left\|\hat{R}(\{X_v\}_v,\theta)\right\|_F^2 \le A_0^4 \tag{11}$$

$$\sum_p\left\|\partial_p R(\bar{x},\theta)\right\|_F^2, \sum_p\left\|\partial_p R(\theta)\right\|_F^2, \sum_p\left\|\partial_p R_j(\theta)\right\|_F^2, \sum_p\left\|\partial_p\hat{R}(\{X_v\}_v,\theta)\right\|_F^2 \le 4A_1^2A_0^2. \tag{12}$$

Proof.

$$\begin{aligned}
\left\|R(\bar{x},\theta)\right\|_F^2 &= \left\|\mathsf{E}_{x\sim\mathcal{A}(\cdot|\bar{x})}z(x;\theta)z^T(x;\theta)\right\|_F^2 \\
&\le \mathsf{E}_{x\sim\mathcal{A}(\cdot|\bar{x})}\left\|z(x;\theta)\right\|^2\left\|z(x;\theta)\right\|^2 \\
&\le A_0^4
\end{aligned} \tag{13}$$

$$\begin{aligned}
\sum_p\left\|\partial_p R(\bar{x},\theta)\right\|_F^2 &\le \sum_p\mathsf{E}_{x\sim\mathcal{A}(\cdot|\bar{x})}\left\|\partial_p z(x;\theta)z^T(x;\theta) + z(x;\theta)\partial_p z^T(x;\theta)\right\|_F^2 \\
&\le 4\sum_p\mathsf{E}_{x\sim\mathcal{A}(\cdot|\bar{x})}\left\|\partial_p z(x;\theta)\right\|^2\left\|z(x;\theta)\right\|^2 \\
&\le 4A_1^2A_0^2
\end{aligned} \tag{14}$$

The remaining results directly follows Jansen's inequality. □

Lemma B.5. *The following function, whose p-th entry is defined as*

$$u_j^{t,e}(\theta)[p] = -\partial_p Tr\left\{R_j^+(\theta)\right\} + Tr\left\{\partial_p R_j(\theta)\sum_{j'} q_{j'} R_{j'}(\theta)\right\} \tag{15}$$

is β-Lipschitz continuous with

$$\beta^2 = 24(A_0^4+1)(A_1^4+A_2^2A_0^2) + 48A_1^4A_0^4. \tag{16}$$

Proof. We start with the derivative of $u_j^{t,e}(\theta)[p]$

$$\partial_m u_j^{t,e}(\theta)[p] = -\partial_m\partial_p Tr\left\{R_j^+(\theta)\right\} + Tr\left\{\partial_m\partial_p R_j(\theta)\sum_{j'} q_{j'} R_{j'}(\theta)\right\} + Tr\left\{\partial_p R_j(\theta)\sum_{j'} q_{j'}\partial_m R_{j'}(\theta)\right\} \tag{17}$$

Using AM-GM, we have

$$(\partial_m u_j^{t,e}(\theta)[p])^2 \le 3\left(\partial_m\partial_p Tr\left\{R_j^+(\theta)\right\}\right)^2 + 3Tr\left\{\partial_m\partial_p R_j(\theta)\sum_{j'} q_{j'} R_{j'}(\theta)\right\}^2 + 3Tr\left\{\partial_p R_j(\theta)\sum_{j'} q_{j'}\partial_m R_{j'}(\theta)\right\}^2 \tag{18}$$

For the first term, recall the definition of $R_j^+(\theta)$, we have

$$\begin{aligned}\partial_m\partial_p Tr\{R_j^+(\theta)\} &= 2\mathsf{E}_{\bar{x}\sim\mathcal{D}_j}\partial_p \mathsf{E}_{x\sim\mathcal{A}(\cdot|\bar{x})} z^T(x;\theta)\partial_m \mathsf{E}_{x\sim\mathcal{A}(\cdot|\bar{x})} z(x;\theta)\\ &\quad + 2\mathsf{E}_{\bar{x}\sim\mathcal{D}_j}\partial_m\partial_p \mathsf{E}_{x\sim\mathcal{A}(\cdot|\bar{x})} z^T(x;\theta)\mathsf{E}_{x\sim\mathcal{A}(\cdot|\bar{x})} z(x;\theta).\end{aligned} \tag{19}$$

Consequently, we have

$$\begin{aligned}&\sum_{m,p}(\partial_m\partial_p Tr\left\{R_j^+(\theta)\right\})^2\\ &\le 8\left(\mathsf{E}_{\bar{x}\sim\mathcal{D}_j}\partial_p \mathsf{E}_{x\sim\mathcal{A}(\cdot|\bar{x})} z^T(x;\theta)\partial_m \mathsf{E}_{x\sim\mathcal{A}(\cdot|\bar{x})} z(x;\theta)\right)^2 + 8\left(\mathsf{E}_{\bar{x}\sim\mathcal{D}_j}\partial_m\partial_p \mathsf{E}_{x\sim\mathcal{A}(\cdot|\bar{x})} z^T(x;\theta)\mathsf{E}_{x\sim\mathcal{A}(\cdot|\bar{x})} z(x;\theta)\right)^2\\ &\le 8\sum_{m,p}\mathsf{E}_{\bar{x}\sim\mathcal{D}_j}\mathsf{E}_{x_1,x_2\sim\mathcal{A}(\bar{x})}\left(\left(\partial_p z^T(x_1;\theta)\partial_m z(x_2;\theta)\right)^2 + \left(\partial_m\partial_p z^T(x_1;\theta)z(x_2;\theta)\right)^2\right)\\ &\le 8\sum_{m,p}\mathsf{E}_{\bar{x}\sim\mathcal{D}_j}\mathsf{E}_{x_1,x_2\sim\mathcal{A}(\bar{x})}\left(\|\partial_p z(x_1;\theta)\|_2^2\|\partial_m z(x_2;\theta)\|_2^2 + \|\partial_m\partial_p z(x_1;\theta)\|_2^2\|z(x_2;\theta)\|_2^2\right)\\ &\le 8(A_1^4+A_2^2A_0^2)\end{aligned} \tag{20}$$

where the first inequality uses AM-GM, and the second inequality uses Jensen's inequality.

For the second term, we have

$$Tr\left\{\partial_m\partial_p R_j(\theta)\sum_{j'} q_{j'} R_{j'}(\theta)\right\}^2 \le \|\partial_m\partial_p R_j(\theta)\|_F^2\left\|\sum_{j'} q_{j'} R_{j'}(\theta)\right\|_F^2. \tag{21}$$

Notice that

$$\begin{aligned}\partial_m\partial_p R_j(\theta) &= \mathsf{E}_{x\sim\mathcal{A}(\cdot|\mathcal{D}_j)}\partial_m\partial_p z(x;\theta)z^T(x;\theta) + z(x;\theta)\partial_m\partial_p z^T(x;\theta)\\ &\quad + \partial_m z(x;\theta)\partial_p z^T(x;\theta) + \partial_p z(x;\theta)\partial_m z^T(x;\theta)\end{aligned} \tag{22}$$

We have

$$\begin{aligned}\sum_{m,p}\|\partial_m\partial_p R_j(\theta)\|_F^2 &\le 4\left\|\mathsf{E}_{x\sim\mathcal{A}(\cdot|\mathcal{D}_j)}\partial_m\partial_p z(x;\theta)z^T(x;\theta)\right\|_F^2 + 4\left\|\mathsf{E}_{x\sim\mathcal{A}(\cdot|\mathcal{D}_j)}z(x;\theta)\partial_m\partial_p z^T(x;\theta)\right\|_F^2 \\ &+ 4\left\|\mathsf{E}_{x\sim\mathcal{A}(\cdot|\mathcal{D}_j)}\partial_m z(x;\theta)\partial_p z^T(x;\theta)\right\|_F^2 + 4\left\|\mathsf{E}_{x\sim\mathcal{A}(\cdot|\mathcal{D}_j)}\partial_p z(x;\theta)\partial_m z^T(x;\theta)\right\|_F^2 \\ &\le 4\mathsf{E}_{x\sim\mathcal{A}(\cdot|\mathcal{D}_j)}\Big[\left\|\partial_m\partial_p z(x;\theta)z^T(x;\theta)\right\|_F^2 + \left\|z(x;\theta)\partial_m\partial_p z^T(x;\theta)\right\|_F^2 \\ &+ \left\|\partial_m z(x;\theta)\partial_p z^T(x;\theta)\right\|_F^2 + \left\|\partial_p z(x;\theta)\partial_m z^T(x;\theta)\right\|_F^2\Big] \\ &\le 8\sum_{m,p}\mathsf{E}_{x\sim\mathcal{A}(\cdot|\mathcal{D}_j)}\left(\|\partial_m\partial_p z(x;\theta)\|_2^2\|z(x;\theta)\|_2^2 + \|\partial_m z(x;\theta)\|_2^2\|\partial_p z(x;\theta)\|_2^2\right) \\ &\le 8(A_1^4 + A_2^2A_0^2)\end{aligned} \tag{23}$$

Apply Lemma B.4, we have

$$\left\|\sum_{j'} q_{j'}R_{j'}(\theta)\right\|_F^2 \le \sum_{j'} q_{j'}\|R_{j'}(\theta)\|_F^2 \le A_0^4 \tag{24}$$

Substitute eq. (21) with eq. (23) and eq. (24), we have

$$Tr\left\{\partial_m\partial_p Rj(\theta)\sum_{j'} q_{j'}R_{j'}(\theta)\right\}^2 \le 8A_0^4(A_1^4 + A_2^2A_0^2) \tag{25}$$

For the third term, we have

$$\begin{aligned}\sum_{m,p}Tr\left\{\partial_p R_j(\theta)\sum_{j'} q_{j'}\partial_m R_{j'}(\theta)\right\}^2 &\le \sum_{m,p}\|\partial_p R_j(\theta)\|_F^2\left\|\sum_{j'} q_{j'}\partial_m R_{j'}(\theta)\right\|_F^2 \\ &\le \sum_{m,p}\|\partial_p R_j(\theta)\|_F^2\sum_{j'} q_{j'}\|\partial_m R_{j'}(\theta)\|_F^2 \\ &\le 16A_1^4A_0^4\end{aligned} \tag{26}$$

where the last inequality uses Lemma B.4.

Combine eq. (20), eq. (25) and eq. (26), we have

$$\left\|\nabla u_j^{t,e}(\theta)\right\|_F^2 = \sum_{m,p}(\partial_m u_j^{t,e}(\theta)[p])^2 \le 24(A_0^4+1)(A_1^4 + A_2^2A_0^2) + 48A_1^4A_0^4 \tag{27}$$

and thus $\beta^2 = 4(A_0^4+1)(A_1^4 + A_2^2A_0^2) + 48A_1^4A_0^4$. □

Corollary B.6. *The global loss $\mathcal{L}(\theta)$ is β-smooth.*

Proof. Notice that $\nabla\mathcal{L}(\theta) = \sum_j q_j u_j^{t,e}(\theta^{t,e})$. The result follows after applying Lemma B.5. □

Lemma B.7. *For any random matrix X, Y, we have*

$$\mathsf{Var}\left[Tr\left\{XY\right\}\right] \le 2\left\|\mathsf{E}[Y]\right\|_F^2\mathsf{Var}[X] + 2\left\|\mathsf{E}[X]\right\|_F^2\mathsf{Var}[Y] \tag{28}$$

Proof.

$$\begin{aligned}\mathsf{Var}\left[Tr\left\{XY\right\}\right] &\le 2\mathsf{Var}[Tr\{(X-\mathsf{E}[X])Y\}] + 2\mathsf{Var}[Tr\{\mathsf{E}[X]Y\}] \\ &\le 2\mathsf{E}[Tr\{(X-\mathsf{E}[X])Y\}^2] + 2\left\|\mathsf{E}[X]\right\|_F^2\mathsf{Var}[Y] \\ &\le 2\left\|\mathsf{E}[Y]\right\|_F^2\mathsf{Var}[X] + 2\left\|\mathsf{E}[X]\right\|_F^2\mathsf{Var}[Y]\end{aligned} \tag{29}$$

□

Lemma B.8. *The local stochastic gradient with p-th entry defined as*

$$\begin{aligned} v_j^{t,e}[p] = &-\partial_p Tr\left\{\hat{R}_j^+(\{X_v\}_v, \theta_j^{t,e})\right\} + q_j Tr\left\{\hat{R}_j(\{X_v\}_v, \theta_j^{t,e})\partial_p \hat{R}_j(\{X_v\}_v, \theta_j^{t,e})\right\} \\ &+ (1-q_j) Tr\left\{\partial_p \hat{R}_j(\{X_v\}_v, \theta_j^{t,e}))\tilde{R}_{-j}^t\right\}. \end{aligned} \tag{30}$$

has bounded norm

$$\mathsf{E}\left[\left\|v_j^{t,e}\right\|^2 |\mathcal{F}^{t,0}\right] \leq 12A_1^2A_0^2(H^2\sigma^2+1) + 12A_1^2A_0^6. \tag{31}$$

where $\mathcal{F}^{t,0}$ is the history before the t-th round; σ^2 is the variance of the DP noise and H is the dimension of the representation $z(x,\theta)$.

Proof.

$$\begin{aligned} \sum_p (v_j^{t,e}[p])^2 \leq &\, 2\sum_p Tr\left\{\partial_p \hat{R}_j^+(\{X_v\}_v, \theta_j^{t,e})\right\}^2 \\ &+ 2\sum_p Tr\left\{\partial_p \hat{R}_j(\{X_v\}_v, \theta_j^{t,e}))\left(q_j\hat{R}_j(\{X_v\}_v, \theta_j^{t,e})) + (1-q_j)\tilde{R}_{-j}^t\right)\right\}^2. \end{aligned} \tag{32}$$

For the first term,

$$Tr\left\{\hat{R}_j^+(\{X_v\}_v, \theta_j^{t,e})\right\} = \frac{1}{BV}\sum_{b=1}^B\sum_{v=1}^V z^T(x_{b,v}, \theta_j^{t,e}) z(x_{b,v+V}, \theta_j^{t,e}). \tag{33}$$

Then we have

$$Tr\left\{\partial_p \hat{R}_j^+(\{X_v\}_v, \theta_j^{t,e})\right\} = \frac{1}{BV}\sum_{b=1}^B\sum_{v=1}^V \partial_p z(x_{b,v}, \theta_j^{t,e}) z(x_{b,v+V}, \theta_j^{t,e})^T + z(x_{b,v}, \theta_j^{t,e})^T \partial_p z(x_{b,v+V}, \theta_j^{t,e}) \tag{34}$$

and

$$\begin{aligned} &\sum_p Tr\left\{\partial_p \hat{R}_j^+(\{X_v\}_v, \theta_j^{t,e})\right\}^2 \\ &\leq \sum_p \frac{2}{BV}\sum_{b=1}^B\sum_{v=1}^V \left\|\partial_p z(x_{b,v}, \theta_j^{t,e})\right\|^2 \left\|z(x_{b,v+V}, \theta_j^{t,e})\right\|^2 + \left\|z(x_{b,v}, \theta_j^{t,e})\right\|^2 \left\|\partial_p z(x_{b,v+V}, \theta_j^{t,e})\right\|^2 \\ &\leq 4A_1^2A_0^2 \end{aligned} \tag{35}$$

where the line uses Jensen's inequality and AM-GM. For the second term,

$$\begin{aligned} &\sum_p Tr\left\{\partial_p \hat{R}_j(\{X_v\}_v, \theta_j^{t,e}))\left(q_j\hat{R}_j(\{X_v\}_v, \theta_j^{t,e})) + (1-q_j)\tilde{R}_{-j}^t\right)\right\}^2 \\ &\leq \left\|q_j\hat{R}_j(\{X_v\}_v, \theta_j^{t,e})) + (1-q_j)\tilde{R}_{-j}^t\right\|_F^2 \sum_p \left\|\partial_p \hat{R}_j(\{X_v\}_v, \theta_j^{t,e})\right\|_F^2 \\ &\leq \left\|q_j\hat{R}_j(\{X_v\}_v, \theta_j^{t,e})) + (1-q_j)\tilde{R}_{-j}^t\right\|_F^2 4A_1^2A_0^2 \end{aligned} \tag{36}$$

where the last inequality uses Lemma B.4. Combine the above results we have

$$\begin{aligned} \mathsf{E}\left[\left\|v_j^{t,e}\right\|^2 |\mathcal{F}^{t,0}\right] &\leq 8A_1^2A_0^2 + 8A_1^2A_0^2\mathsf{E}\left[\left\|q_j\hat{R}_j(\{X_v\}_v, \theta_j^{t,e})) + (1-q_j)\tilde{R}_{-j}^t\right\|_F^2 |\mathcal{F}^{t,0}\right] \\ &= 8A_1^2A_0^2 + 8A_1^2A_0^2\left(A_0^4 + \sum_{j'\neq j} q_{j'}^2 H^2\sigma^2\right) \\ &\leq 8A_1^2A_0^2 + 8A_1^2A_0^2\left(A_0^4 + H^2\sigma^2\right) \end{aligned} \tag{37}$$

where we use the fact that $q_j\hat{R}_j(\{X_v\}_v, \theta_j^{t,e})) + (1-q_j)\tilde{R}_{-j}^t$ is essentially a correlation matrix plus DP noise with scale $\sum_{j'\neq j} q_{j'}^2 H^2\sigma^2$. □

B.4. Proof of the fully participation

From the β-smoothness of $\mathcal{L}(\theta)$ given by Corollary B.6, we have

$$\mathcal{L}(\theta^{t,e+1}) \leq \mathcal{L}(\theta^{t,e}) - \eta\langle\nabla\mathcal{L}(\theta^{t,e}), v^{t,e}\rangle + \frac{\beta\eta^2}{2}\left\|v^{t,e}\right\|^2. \tag{38}$$

Denote the history of the optimization process as $\mathcal{F}^{t,e}$, then we have

$$\mathsf{E}\left[\mathcal{L}(\theta^{t,e+1})|\mathcal{F}^{t,e}\right] \leq \mathcal{L}(\theta^{t,e}) - \eta\langle\nabla\mathcal{L}(\theta^{t,e}), \mathsf{E}\left[v^{t,e}|\mathcal{F}^{t,e}\right]\rangle + \frac{\beta\eta^2}{2}\mathsf{E}\left[\left\|v^{t,e}\right\|^2|\mathcal{F}^{t,e}\right]. \tag{39}$$

Let $\bar{v}^{t,e} = \mathsf{E}[v^{t,e}|\mathcal{F}^{t,e}]$ and $\bar{v}_j^{t,e} = \mathsf{E}\left[v_j^{t,e}|\mathcal{F}^{t,e}\right]$, we have

$$\begin{aligned}\mathsf{E}\left[\mathcal{L}(\theta^{t,e+1})|\mathcal{F}^{t,e}\right] \leq \mathcal{L}(\theta^{t,e}) &+ \frac{\beta\eta^2}{2}\mathsf{E}\left[\left\|v^{t,e}\right\|^2|\mathcal{F}^{t,e}\right] \\ &+ \frac{\eta}{2}\left[-\left\|\nabla\mathcal{L}(\theta^{t,e})\right\|^2 - \left\|\bar{v}^{t,e}\right\|^2 + \left\|\nabla\mathcal{L}(\theta^{t,e}) - \bar{v}^{t,e}\right\|^2\right].\end{aligned} \tag{40}$$

By the choice of $\eta \leq 1/\beta$, we have

$$\mathsf{E}\left[\mathcal{L}(\theta^{t,e+1})|\mathcal{F}^{t,e}\right] \leq \mathcal{L}(\theta^{t,e}) - \frac{\eta}{2}\left\|\nabla\mathcal{L}(\theta^{t,e})\right\|^2 + \frac{\eta}{2}\underbrace{\left\|\nabla\mathcal{L}(\theta^{t,e}) - \bar{v}^{t,e}\right\|^2}_{T_1} + \frac{\beta\eta^2}{2}\underbrace{\mathsf{Var}\left[v^{t,e}|\mathcal{F}^{t,e}\right]}_{T_2} \tag{41}$$

B.4.1. Bounding the term T_1

Recall the definition of local batch loss

$$\hat{\mathcal{L}}_j^{SC}(\theta_j^{t,e}) = -Tr\{\hat{R}_j^+(\{X_v\}_v, \theta_j^{t,e})\} + \frac{1}{2}q_j\left\|\hat{R}_j(\{X_v\}_v, \theta_j^{t,e})\right\|_F^2 + (1-q_j)Tr\{\hat{R}_j(\{X_v\}_v, \theta_j^{t,e})\tilde{R}_{-j}^t\} \tag{42}$$

The p-th entry of $v_j^{t,e} = \nabla\hat{\mathcal{L}}_j^{SC}(\theta_j^{t,e})$ is

$$\begin{aligned}v_j^{t,e}[p] = &-\partial_p Tr\left\{\hat{R}_j^+(\{X_v\}_v, \theta_j^{t,e})\right\} + q_j Tr\left\{\hat{R}_j(\{X_v\}_v, \theta_j^{t,e})\partial_p\hat{R}_j(\{X_v\}_v, \theta_j^{t,e})\right\} \\ &+ (1-q_j)Tr\left\{\partial_p\hat{R}_j(\{X_v\}_v, \theta_j^{t,e}))\tilde{R}_{-j}^t\right\}.\end{aligned} \tag{43}$$

Take expectation over $\{X_v\}_v$, we have

$$\begin{aligned}\bar{v}_j^{t,e}[p] = &-\partial_p Tr\left\{R_j^+(\theta_j^{t,e})\right\} + q_j\mathsf{E}_{\{X_v\}_v}\left[Tr\left\{\hat{R}_j(\{X_v\}_v, \theta_j^{t,e})\partial_p\hat{R}_j(\{X_v\}_v, \theta_j^{t,e})\right\}\right] \\ &+ (1-q_j)Tr\left\{\partial_p R_j(\theta_j^{t,e})\tilde{R}_{-j}^t\right\}.\end{aligned} \tag{44}$$

The p-th entry of the global loss gradient is

$$\begin{aligned}\partial_p\mathcal{L}(\theta^{t,e}) &= -\sum_j q_j\partial_p Tr\left\{R_j^+(\theta^{t,e})\right\} + \sum_j q_j Tr\left\{\partial_p R_j(\theta^{t,e})\sum_{j'}q_{j'}R_{j'}(\theta^{t,e})\right\} \\ &= \sum_j q_j\left(-\partial_p Tr\left\{R_j^+(\theta^{t,e})\right\} + q_j Tr\left\{\partial_p R_j(\theta^{t,e})R_j(\theta^{t,e})\right\} + Tr\left\{\partial_p R_j(\theta^{t,e})\sum_{j'\neq j}q_{j'}R_{j'}(\theta^{t,e})\right\}\right)\end{aligned} \tag{45}$$

Decompose $v_j^{t,e}[p] = u_j^{t,e}(\theta_j^{t,e})[p] + q_j b_j^{t,e}[p] + c_j^{t,e}[p]$, where the terms are defined as follows.

$$\begin{aligned}u_j^{t,e}(\theta)[p] &= -\partial_p Tr\left\{R_j^+(\theta)\right\} + Tr\left\{\partial_p R_j(\theta)\sum_{j'}q_{j'}R_{j'}(\theta)\right\} \\ b_j^{t,e}[p] &= \mathsf{E}_{\{X_v\}_v}\left[Tr\left\{\hat{R}_j(\{X_v\}_v, \theta_j^{t,e})\partial_p\hat{R}_j(\{X_v\}_v, \theta_j^{t,e})\right\}\right] - Tr\left\{R_j(\theta_j^{t,e})\partial_p R_j(\theta_j^{t,e})\right\} \\ c_j^{t,e}[p] &= (1-q_j)Tr\left\{\partial_p R_j(\theta_j^{t,e})\tilde{R}_{-j}^t\right\} - Tr\left\{\partial_p R_j(\theta_j^{t,e})\sum_{j'\neq j}q_{j'}R_{j'}(\theta_j^{t,e})\right\}\end{aligned} \tag{46}$$

Then we have

$$\bar{v}^{t,e}[p] - \partial_p \mathcal{L}(\theta^{t,e}) = \sum_j q_j \left(u_j^{t,e}(\theta_j^{t,e})[p] - u_j^{t,e}(\theta^{t,e})[p] + q_j b_j^{t,e}[p] + c_j^{t,e}[p] \right). \tag{47}$$

The term T_1 can be written as follows

$$\begin{aligned}
T_1 &= \sum_p \left(\bar{v}^{t,e}[p] - \partial_p \mathcal{L}(\theta^{t,e}) \right)^2 \\
&= \sum_p \left(\sum_j q_j \left(u_j^{t,e}(\theta_j^{t,e})[p] - u_j^{t,e}(\theta^{t,e})[p] + q_j b_j^{t,e}[p] + c_j^{t,e}[p] \right) \right)^2 \\
&\leq \sum_j q_j \sum_p \left(u_j^{t,e}(\theta_j^{t,e})[p] - u_j^{t,e}(\theta^{t,e})[p] + q_j b_j^{t,e}[p] + c_j^{t,e}[p] \right)^2 \\
&\leq 3 \sum_j q_j \left[\underbrace{\sum_p (u_j^{t,e}(\theta_j^{t,e})[p] - u_j^{t,e}(\theta^{t,e})[p])^2}_{T_3} + q_j^2 \underbrace{\sum_p (b_j^{t,e}[p])^2}_{T_4} + \underbrace{\sum_p (c_j^{t,e}[p])^2}_{T_5} \right]
\end{aligned} \tag{48}$$

where the third line uses Jensen's inequality and the last line uses AM-GM. By Lemma , we have

$$T_3 \leq \beta^2 \left\| \theta^{t,e} - \theta_j^{t,e} \right\|^2. \tag{49}$$

For the term T_4, we have

$$\begin{aligned}
T_4 &= \sum_p Tr \left\{ \mathsf{E}_{\{X_v\}_v} \left[\hat{R}_j(\{X_v\}_v, \theta_j^{t,e}) - R_j(\theta_j^{t,e}) \right] \left[\partial_p \hat{R}_j(\{X_v\}_v, \theta_j^{t,e}) - \partial_p R_j(\theta_j^{t,e}) \right] \right\}^2 \\
&\leq \sum_p \mathsf{E}_{\{X_v\}_v} \left[\left\| \hat{R}_j(\{X_v\}_v, \theta_j^{t,e}) - R_j(\theta_j^{t,e}) \right\|_F^2 \left\| \partial_p \hat{R}_j(\{X_v\}_v, \theta_j^{t,e}) - \partial_p R_j(\theta_j^{t,e}) \right\|_F^2 \right] \\
&\leq 2\mathsf{E}_{\{X_v\}_v} \left[\left(\left\| \hat{R}_j(\{X_v\}_v, \theta_j^{t,e}) \right\|_F^2 + \left\| R_j(\theta_j^{t,e}) \right\|_F^2 \right) \sum_p \left\| \partial_p \hat{R}_j(\{X_v\}_v, \theta_j^{t,e}) - \partial_p R_j(\theta_j^{t,e}) \right\|_F^2 \right] \\
&\leq 4A_0^4 \mathsf{E}_{\{X_v\}_v} \sum_p \left\| \partial_p \hat{R}_j(\{X_v\}_v, \theta_j^{t,e}) - \partial_p R_j(\theta_j^{t,e}) \right\|_F^2 \\
&= 4A_0^4 \mathsf{E}_{\bar{X} \sim \mathcal{D}_j} \mathsf{E}_{\{X_v\}_v | \bar{X}} \sum_p \left\| \partial_p \hat{R}_j(\{X_v\}_v, \theta_j^{t,e}) - \partial_p R(\bar{X}_j^{t,e}; \theta_j^{t,e}) + \partial_p R(\bar{X}_j^{t,e}; \theta_j^{t,e}) - \partial_p R_j(\theta_j^{t,e}) \right\|_F^2
\end{aligned} \tag{50}$$

where the third inequality uses Lemma B.4 $\bar{X}$ is a batch of samples drawn from $\mathcal{D}_j$, and $\{X_v\}_v$ are augmented views of $\bar{X}$. Notice that

$$\mathsf{E}_{\{X_v\}_v | \bar{X}} \langle \partial_p \hat{R}_j(\{X_v\}_v, \theta_j^{t,e}) - \partial_p R(\bar{X}_j^{t,e}; \theta_j^{t,e}), \partial_p R(\bar{X}_j^{t,e}; \theta_j^{t,e}) - \partial_p R_j(\theta_j^{t,e}) \rangle = 0 \tag{51}$$

we have

$$\begin{aligned}
T_4 &\leq 4A_0^4 \sum_p \mathsf{E}_{\bar{X} \sim \mathcal{D}_j} \mathsf{E}_{\{X_v\}_v | \bar{X}} \left(\left\| \partial_p \hat{R}_j(\{X_v\}_v, \theta_j^{t,e}) - \partial_p R(\bar{X}_j^{t,e}; \theta_j^{t,e}) \right\|_F^2 + \left\| \partial_p R(\bar{X}_j^{t,e}; \theta_j^{t,e}) - \partial_p R_j(\theta_j^{t,e}) \right\|_F^2 \right) \\
&= 4A_0^4 \sum_p \left(\mathsf{E}_{\bar{X} \sim \mathcal{D}_j} \mathsf{Var}_{\{X_v\}_v | \bar{X}} \left[\partial_p \hat{R}_j(\{X_v\}_v, \theta_j^{t,e}) \right] + \mathsf{Var}_{\bar{X} \sim \mathcal{D}_j} \left[\partial_p R(\bar{X}; \theta_j^{t,e}) \right] \right) \\
&\leq 16 A_1^2 A_0^6 \left(\frac{1}{2V} + \frac{|\mathcal{D}_j|/B - 1}{|\mathcal{D}_j| - 1} \right)
\end{aligned} \tag{52}$$

where the last inequality uses Lemma B.4, the fact $\mathsf{Var}[X] \leq \mathsf{E} \left\| X \right\|^2$ and sampling with and without replacement.

For the term T_5, we have

$$\begin{aligned}
T_5 &= Tr\left\{\partial_p R_j(\theta_j^{t,e})\left((1-q_j)\tilde{R}_{-j}^t - \sum_{j'\neq j} q_{j'} R_{j'}(\theta_j^{t,e})\right)\right\}^2 \\
&= \sum_p Tr\left\{\partial_p R_j(\theta_j^{t,e}) \sum_{j'\neq j} q_{j'}\left(\tilde{R}_{j'}^t - R_{j'}(\theta_j^{t,e})\right)\right\}^2 \\
&\leq \sum_p \left\|\partial_p R_j(\theta_j^{t,e})\right\|_F^2 \left\|\sum_{j'\neq j} q_{j'}\left(\tilde{R}_{j'}^t - R_{j'}(\theta_j^{t,e})\right)\right\|_F^2 \\
&\leq 4A_1^2A_0^2 \left\|\sum_{j'\neq j} q_{j'}\left(\tilde{R}_{j'}^t - R_{j'}(\theta_j^{t,e})\right)\right\|_F^2 \\
&\leq 4(1-q_j)A_1^2A_0^2 \sum_{j'\neq j} q_{j'} \left\|\tilde{R}_{j'}^t - R_{j'}(\theta_j^{t,e})\right\|_F^2 \\
&= 4(1-q_j)A_1^2A_0^2 \sum_{j'\neq j} q_{j'} \left\|\tilde{R}_{j'}^t - R_{j'}(\theta^t) + R_{j'}(\theta^t) - R_{j'}(\theta_j^{t,e})\right\|_F^2 \\
&\leq 8(1-q_j)A_1^2A_0^2 \sum_{j'\neq j} q_{j'} \left(\left\|\tilde{R}_{j'}^t - R_{j'}(\theta^t)\right\|_F^2 + \left\|R_{j'}(\theta^t) - R_{j'}(\theta_j^{t,e})\right\|_F^2\right)
\end{aligned} \tag{53}$$

where the second inequality uses Lemma B.4, and the third inequality uses Jensen's inequality.

Use Lemma B.4 and mean-value theorem, we have

$$\left\|R_{j'}(\theta^t) - R_{j'}(\theta_j^{t,e})\right\|_F^2 \leq 4A_1^2A_0^2\left\|\theta^t - \theta_j^{t,e}\right\|_2^2. \tag{54}$$

Denote $D_j^t = \frac{1}{1-q_j}\sum_{j'\neq j} q_{j'}\left\|\tilde{R}_{j'}^t - R_{j'}(\theta^t)\right\|_F^2$, we have

$$T_5 \leq 8(1-q_j)^2A_1^2A_0^2\left(4A_1^2A_0^2\left\|\theta^t - \theta_j^{t,e}\right\|^2 + D_j^t\right) \tag{55}$$

Notice that

$$\begin{aligned}
\sum_j q_j\left\|\theta^{t,e} - \theta_j^{t,e}\right\|^2 &= \sum_j q_j\left\|\theta^{t,e} - \theta^t + \theta^t - \theta_j^{t,e}\right\|^2 \\
&\leq \sum_j q_j\left\|\theta^t - \theta_j^{t,e}\right\|^2
\end{aligned} \tag{56}$$

then, substitute eq. (48) with eq. (49), eq. (52) and eq. (55), we have

$$\begin{aligned}
T_1 &\leq 3\Bigg(\beta^2\sum_j q_j\left\|\theta^{t,e} - \theta_j^{t,e}\right\|^2 + 16A_1^2A_0^6\sum_j q_j^3\left(\frac{1}{2V} + \frac{|\mathcal{D}_j|/B - 1}{|\mathcal{D}_j| - 1}\right) \\
&\quad + 8A_1^2A_0^2\sum_j q_j(1-q_j)^2\left(4A_1^2A_0^2\left\|\theta^t - \theta_j^{t,e}\right\|^2 + D_j^t\right)\Bigg) \\
&\leq 3\left((\beta^2 + 32A_1^6A_0^6)\sum_j q_j\left\|\theta^t - \theta_j^{t,e}\right\|^2 + 16A_1^2A_0^6\left(\frac{1}{2V} + \max_j\frac{|\mathcal{D}_j|/B - 1}{|\mathcal{D}_j| - 1}\right) + 8A_1^2A_0^2\sum_j q_j D_j^t\right)
\end{aligned} \tag{57}$$

B.4.2. BOUNDING THE TERM T_2

$$T_2 = \mathsf{Var}\left[v^{t,e}|\mathcal{F}^{t,e}\right] = \mathsf{Var}\left[\sum_j q_j v_j^{t,e}|\mathcal{F}^{t,e}\right] = \sum_j q_j^2 \mathsf{Var}\left[v_j^{t,e}|\mathcal{F}^{t,e}\right] \tag{58}$$

Compare eq. (43) and eq. (44), we have

$$\begin{aligned} v_j^{t,e}[p] - \bar{v}_j^{t,e}[p] = &-\partial_p Tr\left\{\hat{R}_j^+(\{X_v\}_v, \theta_j^{t,e})\right\} + \partial_p Tr\left\{R_j^+(\theta_j^{t,e})\right\} \\ &+ q_j Tr\left\{\hat{R}_j(\{X_v\}_v, \theta_j^{t,e})\partial_p \hat{R}_j(\{X_v\}_v, \theta_j^{t,e})\right\} - q_j \mathsf{E}_{\{X_v\}_v}\left[Tr\left\{\hat{R}_j(\{X_v\}_v, \theta_j^{t,e})\partial_p \hat{R}_j(\{X_v\}_v, \theta_j^{t,e})\right\}\right] \\ &+ (1-q_j) Tr\left\{\partial_p \hat{R}_j(\{X_v\}_v, \theta_j^{t,e}))\tilde{R}_{-j}^t\right\} - (1-q_j) Tr\left\{\partial_p R_j(\theta_j^{t,e})\tilde{R}_{-j}^t\right\} \end{aligned} \tag{59}$$

$$\begin{aligned} &\mathsf{Var}\left[v_j^{t,e}|\mathcal{F}^{t,e}\right] = \mathsf{E}_{\{X_v\}_v}\left[\left\|v_j^{t,e} - \bar{v}_j^{t,e}\right\|^2\right] \\ &\leq \underbrace{3\,\mathsf{E}_{\{X_v\}_v} \sum_p \left(\partial_p Tr\left\{\hat{R}_j^+(\{X_v\}_v, \theta_j^{t,e})\right\} - \partial_p Tr\left\{R_j^+(\theta_j^{t,e})\right\}\right)^2}_{T_6} \\ &+ \underbrace{3q_j^2\,\mathsf{E}_{\{X_v\}_v} \sum_p \left(Tr\left\{\hat{R}_j(\{X_v\}_v, \theta_j^{t,e})\partial_p \hat{R}_j(\{X_v\}_v, \theta_j^{t,e})\right\} - \mathsf{E}_{\{X_v\}_v}\left[Tr\left\{\hat{R}_j(\{X_v\}_v, \theta_j^{t,e})\partial_p \hat{R}_j(\{X_v\}_v, \theta_j^{t,e})\right\}\right]\right)^2}_{T_7} \\ &+ \underbrace{3(1-q_j)^2\,\mathsf{E}_{\{X_v\}_v} \sum_p \left(Tr\left\{\partial_p \hat{R}_j(\{X_v\}_v, \theta_j^{t,e}))\tilde{R}_{-j}^t\right\} - Tr\left\{\partial_p R_j(\theta_j^{t,e})\tilde{R}_{-j}^t\right\}\right)^2}_{T_8} \end{aligned} \tag{60}$$

For term T_6, we have

$$\begin{aligned} T_6 &\leq \mathsf{E}_{\bar{X}\sim\mathcal{D}_j}\mathsf{E}_{\{X_v\}_v|\bar{X}} \sum_p \Big(\partial_p Tr\left\{\hat{R}_j^+(\{X_v\}_v, \theta_j^{t,e})\right\} - \partial_p Tr\left\{R^+(\bar{X}, \theta_j^{t,e})\right\} \\ &\quad + \partial_p Tr\left\{R^+(\bar{X}, \theta_j^{t,e})\right\} - \partial_p Tr\left\{R_j^+(\theta_j^{t,e})\right\}\Big)^2 \\ &= \mathsf{E}_{\bar{X}\sim\mathcal{D}_j}\mathsf{E}_{\{X_v\}_v|\bar{X}} \sum_p \left(\partial_p Tr\left\{\hat{R}_j^+(\{X_v\}_v, \theta_j^{t,e})\right\} - \partial_p Tr\left\{R^+(\bar{X}, \theta_j^{t,e})\right\}\right)^2 \\ &\quad + \mathsf{E}_{\bar{X}\sim\mathcal{D}_j}\left(\partial_p Tr\left\{R^+(\bar{X}, \theta_j^{t,e})\right\} - \partial_p Tr\left\{R_j^+(\theta_j^{t,e})\right\}\right)^2 \\ &= \sum_p \left(\mathsf{E}_{\bar{X}\sim\mathcal{D}_j}\mathsf{Var}_{\{X_v\}_v|\bar{X}}\left[\partial_p Tr\left\{\hat{R}_j^+(\{X_v\}_v, \theta_j^{t,e})\right\}\right] + \mathsf{Var}_{\bar{X}\sim\mathcal{D}_j}\left[\partial_p Tr\left\{R^+(\bar{X}, \theta_j^{t,e})\right\}\right]\right) \\ &\leq 4A_1^2A_0^2\left(\frac{1}{V} + \frac{|\mathcal{D}_j|/B - 1}{|\mathcal{D}_j| - 1}\right) \end{aligned} \tag{61}$$

where the first equality uses $\mathsf{E}\left\|X - \mathsf{E}[X]\right\|^2 \leq \mathsf{E}\left\|X\right\|^2$; the last inequality uses the variance under sampling without replacement.

For the term T_7, we have

$$\begin{aligned} T_7 &= \sum_p \mathsf{Var}_{\{X_v\}_v}\left[Tr\left\{\hat{R}_j(\{X_v\}_v,\theta_j^{t,e})\partial_p\hat{R}_j(\{X_v\}_v,\theta_j^{t,e})\right\}\right] \\ &\le 2\left\|R_j(\theta_j^{t,e})\right\|_F^2\sum_p \mathsf{Var}_{\{X_v\}_v}\left[\partial_p\hat{R}_j(\{X_v\}_v,\theta_j^{t,e})\right] + 2\sum_p\left\|\partial_p R_j(\theta_j^{t,e})\right\|_F^2\mathsf{Var}_{\{X_v\}_v}\left[\hat{R}_j(\{X_v\}_v,\theta_j^{t,e})\right] \\ &\le 2A_0^4\sum_p \mathsf{Var}_{\{X_v\}_v}\left[\partial_p\hat{R}_j(\{X_v\}_v,\theta_j^{t,e})\right] + 8A_1^2A_0^2\mathsf{Var}_{\{X_v\}_v}\left[\hat{R}_j(\{X_v\}_v,\theta_j^{t,e})\right] \end{aligned} \tag{62}$$

Notice that

$$\begin{aligned} &\sum_p \mathsf{Var}_{\{X_v\}_v}\left[\partial_p\hat{R}_j(\{X_v\}_v,\theta_j^{t,e})\right] \\ &= \sum_p \mathsf{E}_{\bar{X}\sim\mathcal{D}_j}\mathsf{Var}_{\{X_v\}_v|\bar{X}}\left[\partial_p\hat{R}_j(\{X_v\}_v,\theta_j^{t,e})\right] + \sum_p \mathsf{Var}_{\bar{X}\sim\mathcal{D}_j}\left[\partial_p R(\bar{X};\theta_j^{t,e})\right] \\ &\le 4A_1^2A_0^2\left(\frac{1}{2V}+\frac{|\mathcal{D}_j|/B-1}{|\mathcal{D}_j|-1}\right) \end{aligned} \tag{63}$$

where the second line uses $\mathsf{Var}[Y] = \mathsf{E}_X\mathsf{Var}[Y|X] + \mathsf{Var}[\mathsf{E}[Y|X]]$; the third line uses $\mathsf{Var}[X] \le \mathsf{E}\left\|X\right\|_F^2$, Lemma B.4, and sampling with and without replacement. Similarly, we have

$$\begin{aligned} &\mathsf{Var}_{\{X_v\}_v}\left[\hat{R}_j(\{X_v\}_v,\theta_j^{t,e})\right] \\ &= \mathsf{E}_{\bar{X}\sim\mathcal{D}_j}\mathsf{Var}_{\{X_v\}_v|\bar{X}}\left[\hat{R}_j(\{X_v\}_v,\theta_j^{t,e})\right] + \mathsf{Var}_{\bar{X}\sim\mathcal{D}_j}\left[R(\bar{X};\theta_j^{t,e})\right] \\ &\le A_0^4\left(\frac{1}{2V}+\frac{|\mathcal{D}_j|/B-1}{|\mathcal{D}_j|-1}\right) \end{aligned} \tag{64}$$

Plug eq. (63) and eq. (64) into eq. (62), we have

$$T_7 \le 8A_1^2A_0^6\left(\frac{1}{2V}+\frac{|\mathcal{D}_j|/B-1}{|\mathcal{D}_j|-1}\right) \tag{65}$$

For the term T_8, using eq. (63) we have

$$\begin{aligned} T_8 &\le \sum_p \mathsf{Var}_{\{X_v\}_v}\left[\partial_p\hat{R}_j(\{X_v\}_v,\theta_j^{t,e})\right]\left\|\tilde{R}_{-j}^t\right\|_F^2 \\ &\le 4A_1^2A_0^2\left(\frac{1}{2V}+\frac{|\mathcal{D}_j|/B-1}{|\mathcal{D}_j|-1}\right)\left\|\tilde{R}_{-j}^t\right\|_F^2 \end{aligned} \tag{66}$$

Substitute eq. (58) with eq. (60), eq. (61), eq. (65) and eq. (66), we have

$$\begin{aligned} T_2 &\le 12A_1^2A_0^2\sum_j q_j^2\left(\frac{1}{V}+\frac{|\mathcal{D}_j|/B-1}{|\mathcal{D}_j|-1}\right) + 24A_1^2A_0^6\sum_j q_j^4\left(\frac{1}{V}+\frac{|\mathcal{D}_j|/B-1}{|\mathcal{D}_j|-1}\right) \\ &\quad + 12A_1^2A_0^2\sum_j q_j^2(1-q_j)^2\left(\frac{1}{V}+\frac{|\mathcal{D}_j|/B-1}{|\mathcal{D}_j|-1}\right)\left\|\tilde{R}_{-j}^t\right\|_F^2 \\ &\le \left(\frac{1}{V}+\max_j\frac{|\mathcal{D}_j|/B-1}{|\mathcal{D}_j|-1}\right)\left(12A_1^2A_0^2 + 24A_1^2A_0^6 + 12A_1^2A_0^2\sum_j q_j^2(1-q_j)^2\left\|\tilde{R}_{-j}^t\right\|_F^2\right) \end{aligned} \tag{67}$$

B.4.3. Combine the results

Take expectation on both sides of eq. (41) conditioned on $\mathcal{F}^{t,0}$, we have

$$\mathsf{E}\left[\mathcal{L}(\theta^{t,e+1})|\mathcal{F}^{t,0}\right] \le \mathsf{E}\left[\mathcal{L}(\theta^{t,e})|\mathcal{F}^{t,0}\right] - \frac{\eta}{2}\mathsf{E}\left[\left\|\nabla\mathcal{L}(\theta^{t,e})\right\|^2|\mathcal{F}^{t,0}\right] + \frac{\eta}{2}\mathsf{E}\left[T_1|\mathcal{F}^{t,0}\right] + \mathsf{E}\left[T_2|\mathcal{F}^{t,0}\right]. \tag{68}$$

Notice that

$$\begin{aligned}\mathsf{E}\Big[\big\|\theta^t-\theta_j^{t,e}\big\|^2|\mathcal{F}^{t,0}\Big]&=\mathsf{E}\left[\left\|\sum_{e'=0}^{e-1}\eta v_j^{t,e'}\right\|_2^2|\mathcal{F}^{t,0}\right]\\&\le E\sum_{e'=0}^{e-1}\eta^2\mathsf{E}\Big[\big\|v_j^{t,e}\big\|_2^2|\mathcal{F}^{t,0}\Big]\\&\le E^2\eta^2(8A_1^2A_0^2+8A_1^2A_0^2(A_0^4+H^2\sigma^2))\end{aligned}\tag{69}$$

where the last line uses Lemma B.8. Also notice that

$$\begin{aligned}\mathsf{E}\big[D_j^t|\mathcal{F}^{t,0}\big]&=\frac{1}{1-q_j}\sum_{j'\neq j}q_{j'}\mathsf{E}\left[\left\|\tilde{R}_{j'}^t-R_{j'}(\theta^t)\right\|_F^2|\mathcal{F}^{t,0}\right]\\&=\frac{1}{1-q_j}\sum_{j'\neq j}q_{j'}\left[\mathsf{E}\left[\left\|\tilde{R}_{j'}^t-\check{R}_{j'}(\theta^t)\right\|_F^2|\mathcal{F}^{t,0}\right]+\mathsf{E}\Big[\big\|\check{R}_{j'}-R_{j'}(\theta^t)\big\|_F^2|\mathcal{F}^{t,0}\Big]\right]\\&=H^2\sigma^2+\frac{1}{2V}A_0^4\end{aligned}\tag{70}$$

where $\check{R}_{j'}$ is the empirical correlation matrix before applying DP noise. Then we have

$$\begin{aligned}\mathsf{E}\big[T_1|\mathcal{F}^{t,0}\big]\le 3\Bigg(&\left(\beta^2+32A_1^6A_0^6\right)E^2\eta^2\left(8A_1^2A_0^2+8A_1^2A_0^2\left(A_0^4+H^2\sigma^2\right)\right)\\&+16A_1^2A_0^6\left(\frac{1}{V}+\max_j\frac{|\mathcal{D}_j|/B-1}{|\mathcal{D}_j|-1}\right)+8A_1^2A_0^2\left(H^2\sigma^2+\frac{1}{2V}A_0^4\right)\Bigg).\end{aligned}\tag{71}$$

For the term $\mathsf{E}\big[T_2|\mathcal{F}^{t,0}\big]$, we have

$$\begin{aligned}\mathsf{E}\left[\left\|\tilde{R}_{-j}^t\right\|_F^2|\mathcal{F}^{t,0}\right]&=\mathsf{E}\Big[\big\|\bar{R}_{-j}^t\big\|_F^2|\mathcal{F}^{t,0}\Big]+\mathsf{E}\Big[\big\|n_{-j}^t\big\|_F^2|\mathcal{F}^{t,0}\Big]\\&=A_0^4+\frac{1}{(1-q_j)^2}\sum_{j'\neq j}q_j'^2H^2\sigma^2\end{aligned}\tag{72}$$

$$\mathsf{E}\big[T_2|\mathcal{F}^{t,0}\big]\le\left(\frac{1}{V}+\max_j\frac{|\mathcal{D}_j|/B-1}{|\mathcal{D}_j|-1}\right)\left(12A_1^2A_0^2+26A_1^2A_0^6+12A_1^2A_0^2H^2\sigma^2\right)\tag{73}$$

where we use fact $q_j(1-q_j)^2\le\frac{4}{27}$. Combine eq. (68), eq. (71) and eq. (73), we have

$$\begin{aligned}&\mathsf{E}\big[\mathcal{L}(\theta^{t,e+1})|\mathcal{F}^{t,0}\big]-\mathsf{E}\big[\mathcal{L}(\theta^{t,e})|\mathcal{F}^{t,0}\big]\\&\le-\frac{\eta}{2}\mathsf{E}\Big[\big\|\nabla\mathcal{L}(\theta^{t,e})\big\|^2|\mathcal{F}^{t,0}\Big]+\frac{\eta^3}{2}C_1E^2(H^2\sigma^2+C_2)+\frac{\eta^2}{2}C_3\left(\frac{1}{V}+\max_j\frac{|\mathcal{D}_j|/B-1}{|\mathcal{D}_j|-1}\right)\left(H^2\sigma^2+C_4\right)\\&+\frac{\eta}{2}C_5\left(\frac{1}{V}+\max_j\frac{|\mathcal{D}_j|/B-1}{|\mathcal{D}_j|-1}\right)+\frac{\eta}{2}C_6\left(H^2\sigma^2+\frac{1}{V}C_7\right)\end{aligned}\tag{74}$$

where $C_1, C_2, ..., C_7$ are constant depending A_0, A_1 and A_2. Telescope eq. (74) and take expectation, we have

$$\begin{aligned}&\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathsf{E}\Big[\big\|\nabla\mathcal{L}(\theta^{t,e})\big\|^2\Big]\\&\le\frac{2}{\eta TE}\left(\mathsf{E}\big[\mathcal{L}(\theta^{t,0})\big]-\mathsf{E}\big[\mathcal{L}(\theta^{t,E})\big]\right)+\eta^2C_1E^2(H^2\sigma^2+C_2)+\eta C_3\left(\frac{1}{V}+\max_j\frac{|\mathcal{D}_j|/B-1}{|\mathcal{D}_j|-1}\right)\left(H^2\sigma^2+C_4\right)\\&+C_5\left(\frac{1}{V}+\max_j\frac{|\mathcal{D}_j|/B-1}{|\mathcal{D}_j|-1}\right)+C_6\left(H^2\sigma^2+\frac{1}{V}C_7\right).\end{aligned}\tag{75}$$

Recall the choice $\eta = \mathcal{O}\left(\frac{1}{\sqrt{TE}}\right)$, we have

$$\begin{aligned}&\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathsf{E}\left[\left\|\nabla\mathcal{L}(\theta^{t,e})\right\|^2\right]\\&\leq\mathcal{O}\left(\frac{E^2(H^2\sigma^2+C_2)}{TE}+\frac{\mathcal{L}(\theta^0)+\left(\frac{1}{V}+\max_j\frac{|\mathcal{D}_j|/B-1}{|\mathcal{D}_j|-1}\right)\left(H^2\sigma^2+C_4\right)}{\sqrt{TE}}+\frac{1}{V}+\max_j\frac{|\mathcal{D}_j|/B-1}{|\mathcal{D}_j|-1}+H^2\sigma^2\right)\end{aligned}\tag{76}$$

B.5. Partial Participation Case

Partial participation results in perturbation in aggregation.

$$\begin{aligned}&\mathsf{E}\left[\left\|\nabla\mathcal{L}(\theta^{t,E}_{\mathcal{J}^t})\right\|^2-\left\|\nabla\mathcal{L}(\theta^{t,E})\right\|^2|\mathcal{J}_t\right]\\&=2\mathsf{E}_{\mathcal{J}_t}\left[\nabla\mathcal{L}^T(\theta^{t,E}+\lambda h)\nabla^2\mathcal{L}(\theta^{t,E}+\lambda h,\mathcal{D})h\right]\\&\leq 2W\beta\mathsf{E}_{\mathcal{J}_t}\left[\left\|\theta^{t,E}-\theta^{t,E}_{\mathcal{J}^t}\right\|\right]\\&\leq 2W\beta\sqrt{\mathsf{E}_{\mathcal{J}^t}\left[\left\|\theta^{t,E}-\theta^{t,E}_{\mathcal{J}^t}\right\|_2^2\right]}\\&\leq 2\eta\beta EW^2\sqrt{\frac{J/|\mathcal{J}^t|-1}{J-1}}\end{aligned}\tag{77}$$

where one can easily verify that $W=8A_1^2A_0^2+8A_1^2A_0^6$ from Lemma B.8 servers a bound for gradient norm; the second line uses mean-value theorem. Another aspect is that $\tilde{R}_j$ is less frequently updated on sever. Therefore the term $\mathsf{E}\left[D_j^t|\mathcal{F}^{t,0}\right]$ should involve an additional term accounting to aging of correlation matrix,

$$\mathsf{E}\left[D_j^t|\mathcal{F}^{t,0}\right]\leq H^2\sigma^2+\frac{1}{2V}A_0^4+C_8\eta^2.\tag{78}$$

The reason is that the difference between the current and old correlation matrix is proportional to the distance between the current and old variables (shown in eq. (54)), which is proportional to $E^2\eta^2$ (shown in eq. (69)). Thus we finally have

$$\begin{aligned}\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{e=0}^{E-1}\mathsf{E}\left[\left\|\nabla\mathcal{L}(\theta^{t,e})\right\|^2\right]\leq\mathcal{O}\Bigg(&\frac{E^2(H^2\sigma^2+C_2)}{TE}\\&+\frac{\mathcal{L}(\theta^0)+\left(\frac{1}{V}+\max_j\frac{|\mathcal{D}_j|/B-1}{|\mathcal{D}_j|-1}\right)\left(H^2\sigma^2+C_4\right)+E\sqrt{\frac{J/|\mathcal{J}^t|-1}{J-1}}}{\sqrt{TE}}\\&+\frac{1}{V}+\max_j\frac{|\mathcal{D}_j|/B-1}{|\mathcal{D}_j|-1}+H^2\sigma^2\Bigg)\end{aligned}\tag{79}$$

C. Detailed Implementation of FedSC

Recall the local objective

$$\mathcal{L}_j^{SC}(\theta;\bar{R}_{-j})=-Tr\{R_j^+(\theta)\}+\frac{1}{2}\alpha_j\left\|R_j(\theta)\right\|_F^2+(1-\alpha_j)Tr\{R_j(\theta)\bar{R}_{-j}\}\tag{80}$$

here we replace q_j with a general coefficient α_j, and decay it linearly from 1 to 0.2 along with communication round indices. The behind motivation is as follows. At the beginning of the training, moving direction from the global objective and the average local objective tend to align closely. Moreover, the correlation matrices of clients are not yet stable at this stage,

Table 6. Implementation details of FedSC with DP protection: full client participation

	μ	σ	round indices	local dataset size
SVHN ($\epsilon = 3, \delta = 10^{-2}$)	2	0.0034	$t > 100$	$10,000$
SVHN ($\epsilon = 6, \delta = 10^{-2}$)	2	0.0018	$t > 100$	$10,000$
SVHN ($\epsilon = 3, \delta = 10^{-4}$)	2	0.0048	$t > 100$	$10,000$
SVHN ($\epsilon = 8, \delta = 10^{-4}$)	2	0.0018	$t > 100$	$10,000$
CIFAR10 ($\epsilon = 3, \delta = 10^{-2}$)	4	0.01	$t > 150$	$5,000$
CIFAR10 ($\epsilon = 6, \delta = 10^{-2}$)	4	0.0052	$t > 100, t\%2 = 0$	$5,000$
CIFAR10 ($\epsilon = 3, \delta = 10^{-4}$)	4	0.012	$t > 150$	$5,000$
CIFAR10 ($\epsilon = 8, \delta = 10^{-4}$)	4	0.0051	$t > 100, t\%2 = 0$	$5,000$
CIFAR100 ($\epsilon = 6, \delta = 10^{-2}$)	5	0.013	$t > 100, t\%2 = 0$	$2,500$
CIFAR100($\epsilon = 12, \delta = 10^{-2}$)	5	0.0075	$t > 100, t\%2 = 0$	$2,500$
CIFAR100 ($\epsilon = 3, \delta = 10^{-4}$)	5	0.04	$t > 100, t\%2 = 0$	$2,500$
CIFAR100($\epsilon = 8, \delta = 10^{-4}$)	5	0.013	$t > 100, t\%2 = 0$	$2,500$

making it less critical to at early stages. Therefore, we choose large α_j for quicker start. Conversely, correlation matrices converges and becomes stable at the end of training, thus we give the inter-client contrast larger weights, i.e., smaller α_j.

We also make modifications when DP protection is applied. Based on the above analysis, we start sharing at the middle or late stages of the training to save privacy budgets. Following are the detailed implementation details. For partial client participation, we only change σ according to the ratio of participation.