LSSCA: A UNIFIED PRIVACY-PRESERVING FEDERATED LEARNING FRAMEWORK FOR FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Differentially private federated fine-tuning enables collaborative adaptation of language models while protecting sensitive training records, but its utility is challenged by client heterogeneity and privacy-induced noise. Control variates mitigate client drift, yet noisy updates also perturb local trajectories and subsequent control states. We propose LSSCA, a federated fine-tuning framework that combines stochastic controlled averaging with fixed random subspace filtering. Constructed independently of private data without requiring public reference data, the filter preserves reference-subspace components and attenuates orthogonal components of the complete control-corrected noisy update. An exact signal–noise decomposition characterizes the trade-off between noise reduction and signal distortion, without assuming that the random subspace aligns with useful gradients. Under the stated assumptions, including inactive clipping, we establish an $\mathcal{O}(T^{-1/2})$ non-convex stationarity rate at a fixed noise scale without an additional gradient-dissimilarity assumption. The analysis explicitly accounts for noise propagation through local trajectories and control states, retaining the costs of filtering in the convergence constants. Experiments across NLP benchmarks, privacy budgets, and client heterogeneity levels demonstrate improved privacy–utility trade-offs over the evaluated DP-FL baselines.

INTRODUCTION

Federated learning (FL) enables collaborative training without centralizing clients' raw data McMahan et al. (2017). Selected clients optimize a shared model locally and return updates for server aggregation. Keeping data local, however, neither eliminates information leakage nor resolves statistical heterogeneity Feng et al. (2018). Different client distributions can produce diverging update directions, known as *client drift* Karimireddy et al. (2020). Control variates mitigate drift by correcting client-specific updates, but these corrections evolve through noisy training. Shared gradients and models remain vulnerable to model inversion and membership inference Fredrikson et al. (2014); Yeom et al. (2018). Differential privacy (DP) bounds individual records' influence on released outputs through clipping and calibrated noise injection Dwork et al. (2006). DP-FedAvg McMahan et al. (2018) and DP-SCAFFOLD Noble et al. (2022a) incorporate this protection into federated optimization. However, noise does more than increase update variance: it alters local trajectories, subsequent gradients, and future control states. Privacy noise thus propagates through both model updates and the mechanism used to correct client drift.

Motivation: Correcting Direction Is Not Enough. Drift correction and noise attenuation address different properties of an update. Control variates adjust its direction to compensate for client-specific optimization, but do not by themselves selectively attenuate freshly injected isotropic noise. Conversely, uniformly shrinking an update reduces noise while shrinking the useful signal by the same factor. The central challenge is therefore to attenuate perturbations selectively without discarding the directions needed for learning across heterogeneous clients. Figure 1 provides a simulated illustration of the joint difficulty, motivating the question:

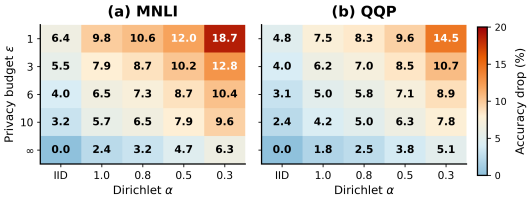

Figure 1: Accuracy degradation under joint DP noise and data heterogeneity.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

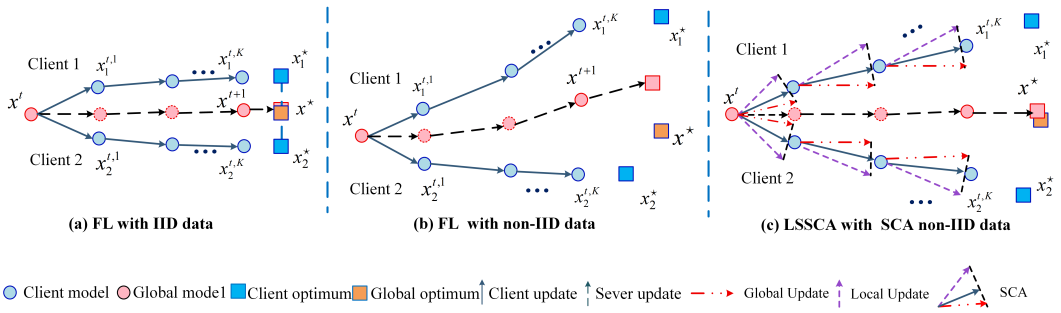

Figure 2: Illustration of DP-induced perturbation and client drift in DP-FedAvg, and the proposed LSSCA framework and SCA.

Can we jointly correct client drift and selectively attenuate DP noise, while explicitly accounting for the useful signal lost through filtering?

Our Idea: Subspace-Preserving Controlled Updates. We combine stochastic controlled averaging with a shared *soft subspace filter*. Let U contain orthonormal reference directions, constructed independently of the private training datasets and fixed throughout training. For a control-corrected noisy direction v, we apply $P_q v = UU^\top v + q(I - UU^\top)v, q \in (0, 1]$. The filter preserves updates within the reference subspace and attenuates their orthogonal components. Unlike hard projection, a positive q retains the ability to learn outside the reference subspace, which is important when it does not capture every client's useful directions.

The key insight is a measurable trade-off between *signal alignment* and *noise dimensionality*. Isotropic Gaussian noise occupies all d parameter directions, whereas useful updates may concentrate in a smaller reference subspace. For a fixed signal s and independent noise $\xi \sim \mathcal{N}(0, \tau^2 I_d)$,

$$\mathbb{E}_\xi \|P_q(s+\xi) - s\|_2^2 = (1-q)^2 \|(I - UU^\top)s\|_2^2 + \tau^2 \big[r + q^2(d-r)\big], \tag{1}$$

where r is the subspace dimension. This identity makes the benefit and its cost explicit: filtering removes noise energy from the orthogonal complement while distorting any signal in that complement. Its effectiveness therefore depends on subspace alignment, rather than on variance reduction alone.

This perspective leads to `LSSCA`, which filters the complete control-corrected direction and retains the resulting filter-dependent control recursion. Our analysis accounts for noise in both the local trajectories and the control states, and characterizes convergence through the filter's trace terms and minimum eigenvalue. The resulting framework connects drift correction with directional noise attenuation without assuming that privacy noise and heterogeneity occupy separate frequency bands.

Our Contributions. We summarize our main contributions as follows.

• **Subspace filtering for controlled DP-FL.** We propose `LSSCA`, which combines stochastic controlled averaging with a shared soft subspace filter. The method corrects client-specific updates while preserving reference-subspace components and attenuating orthogonal components. An exact signal–noise decomposition characterizes when noise reduction outweighs signal distortion.

• **Convergence analysis with explicit filter dependence.** Under the stated assumptions, including inactive clipping, we establish an $\mathcal{O}(T^{-1/2})$ non-convex stationarity bound at a fixed noise scale. The analysis retains the filter-dependent control recursion and accounts for noise propagation through local trajectories and control states. It explicitly captures the trade-off between noise attenuation, governed by $\mathrm{tr}(P_q^2)$, and directional progress, governed by $\lambda_{\min}(P_q)$.

•**Evaluation on language tasks.** We evaluate `LSSCA` with GPT-2 XL and RoBERTa-Base across privacy budgets and Non-iid data, comparing its privacy–utility trade-off with DP-FL baselines.

Related Work

Although FL keeps data decentralized, privacy leakage can still occur through shared updates or the final global model, as shown by gradient inversion and membership inference attacks Geiping

et al. (2020); Shokri et al. (2017). Differential privacy (DP) has therefore become a standard protection mechanism in FL. Classical methods such as DP-FedSGD Geyer et al. (2017) and DP-FedAvg McMahan et al. (2018) provide sample-level privacy by adding Gaussian noise to client updates, but suffer severe utility loss under small client populations or strong data heterogeneity.

Existing DP-FL methods address privacy-induced variance and heterogeneity-induced drift with different emphases. DP-SCAFFOLD Noble et al. (2022a) uses control variates to correct client drift, but drift correction alone does not explicitly distinguish useful update directions from noise-dominated ones. DP-FedAvg-LS Liang et al. (2024) reduces noise variance through Laplacian smoothing, without explicitly correcting heterogeneity-induced client drift. DP-FedSAM Shi et al. (2023) improves robustness through sharpness-aware updates, at the cost of additional local computation. Under an honest-but-curious server, local differential privacy (LDP) Duchi et al. (2018) protects client updates before they reach the server, making the privacy–utility trade-off particularly challenging. These challenges also motivate examining how LDP and personalized DP-FL methods Zhao et al. (2021); Girgis et al. (2021); Hu et al. (2020) balance noise suppression with adaptation to heterogeneous data.

Summary and Gap. The key challenge is not simply to combine privacy protection with drift correction, as methods such as DP-SCAFFOLD already do, but to coordinate noise attenuation with drift correction. Indiscriminate variance reduction may suppress useful signal, while modifying noisy updates can also affect the evolution of control variates and their subsequent drift corrections. This motivates a unified DP-FL approach that selectively attenuates privacy-induced noise while accounting for its interaction with client drift and evolving control states.

A Unified Differentially Private Federated Learning Framework

We introduce `LSSCA`, which combines subspace filtering with stochastic controlled averaging for sample-level DP-FL. Control variates correct client-specific updates, while a shared filter preserves reference-subspace components and attenuates orthogonal components, including DP noise.

For N clients with data distributions $\{\mathcal{D}_i\}_{i=1}^N$, we minimize

$$f(\boldsymbol{x}) := \frac{1}{N}\sum_{i=1}^{N} f_i(\boldsymbol{x}), \qquad f_i(\boldsymbol{x}) := \mathbb{E}_{z\sim\mathcal{D}_i}[\ell(\boldsymbol{x}; z)], \tag{2}$$

where $\boldsymbol{x} \in \mathbb{R}^d$ is the shared model and ℓ is the per-example loss. The distributions $\mathcal{D}_i$ may differ across clients; uniform sampling over a finite dataset D_i recovers the corresponding empirical risk.

Subspace Filtering

Isotropic DP noise perturbs every parameter direction equally, although useful optimization signals need not be distributed uniformly across the parameter space. This motivates a selective filtering principle: *preserve updates in a shared reference subspace while attenuating their orthogonal components*. The potential benefit depends on how well this subspace captures useful updates, rather than on noise reduction alone.

Shared reference subspace. Before private training, we sample a Gaussian matrix $G \in \mathbb{R}^{d\times r}$ with independent $G_{ij} \sim \mathcal{N}(0,1)$ and obtain $U \in \mathbb{R}^{d\times r}$ through a thin QR decomposition, so that $U^\top U = I_r$ and $1 \le r < d$. The resulting random orthonormal basis is independent of the private training data, shared by all clients, and fixed throughout optimization, consistently with our analysis. This construction requires no public reference data. It defines a reference subspace for filtering, but does not guarantee alignment with task-relevant directions.

Soft subspace filtering. For an attenuation parameter $q \in (0, 1]$, define

$$P_q := UU^\top + q(I - UU^\top). \tag{3}$$

For any direction v, let $v_\parallel = UU^\top v$ and $v_\perp = (I - UU^\top)v$. Then $P_q v = v_\parallel + q v_\perp$. The filter preserves the reference-subspace component and reduces the orthogonal component without eliminating it. In particular, $q > 0$ permits adaptation outside the reference subspace, whereas $q = 1$ recovers the unfiltered update. Since P_q is positive definite, it does not introduce additional stationary points for the deterministic direction: $P_q \nabla f(x) = 0$ if and only if $\nabla f(x) = 0$.

In LSSCA, filtering acts on the entire control-corrected noisy direction:

$$v_i^{t,k} := \widehat{g}_i^{t,k} - \boldsymbol{c}_i^t + \boldsymbol{c}^t, \omega_i^{t,k+1} = \omega_i^{t,k} - \eta_l P_q v_i^{t,k}, \tag{4}$$

where $\widehat{g}_i^{t,k}$ denotes the clipped minibatch gradient perturbed by independent Gaussian noise. Control correction adjusts client-specific directions, while filtering determines how strongly their components are propagated. Using a common filter ensures that all clients apply the same directional scaling.

Signal–noise trade-off. The filter's effect can be characterized exactly. Let s be a fixed noise-free direction and $\xi \sim \mathcal{N}(0, \tau^2 I_d)$. Writing $s_\perp = (I - UU^\top)s$, we obtain

$$\mathbb{E}_\xi \|P_q(s+\xi) - s\|_2^2 = (1-q)^2 \|s_\perp\|_2^2 + \tau^2 \big[r + q^2(d-r)\big]. \tag{5}$$

The first term measures signal distortion outside the reference subspace; the second measures the remaining noise energy. For $0 < q < 1$, the filtered direction has smaller mean-squared error relative to s than the unfiltered direction $s + \xi$ precisely when $\|s_\perp\|_2^2 < \frac{1+q}{1-q}\tau^2(d-r)$. Thus, filtering is beneficial in this estimation sense when the discarded signal energy is sufficiently small relative to the noise energy in the orthogonal complement. This identifies subspace alignment as a central design criterion rather than treating smaller update variance as sufficient evidence of better optimization.

Equation 5 describes a fixed direction with fresh independent noise. It does not assume that later local gradients are independent of earlier perturbations; these dependencies are retained in our convergence analysis.

Spectral characterization. The filter has eigenvalue 1 on $\mathrm{span}(U)$ and q on its orthogonal complement. Consequently, $\Lambda_{\min} = q, d_q := \mathrm{tr}(P_q) = r + q(d-r), \widetilde{d}_q := \mathrm{tr}(P_q^2) = r + q^2(d-r)$. In particular, $\mathbb{E}\|P_q\xi\|_2^2 = \tau^2 \widetilde{d}_q$, compared with $\tau^2 d$ without filtering. Averaging KS independent perturbations yields $\tau^2 \widetilde{d}_q/(KS)$, the explicit noise contribution used in our analysis. A smaller q strengthens noise attenuation but also slows updates outside the reference subspace. Our convergence bounds retain this trade-off through $\Lambda_{\min} = q$ and the control and local-drift terms.

Efficient implementation. The filtered direction is computed as $P_q v = qv + (1-q)U(U^\top v)$, without constructing a dense $d \times d$ matrix. An explicitly stored rank-r basis requires $\mathcal{O}(dr)$ storage and arithmetic per application. For large models, we apply the filter blockwise using fixed bases $U_b \in \mathbb{R}^{d_b \times r_b}$ and a common q, giving a block-diagonal operator and an application cost of $\mathcal{O}(\sum_b d_b r_b)$ for positive block ranks. The spectral formulas above remain valid with $d = \sum_b d_b$ and $r = \sum_b r_b$.

Our LSSCA Algorithm

We introduce LSSCA (Algorithm 1), which combines shared subspace filtering with stochastic controlled averaging for sample-level differentially private federated learning. Each selected client filters the *entire control-corrected noisy direction*, coupling directional attenuation with client-drift correction. The server samples S clients uniformly without replacement, broadcasts the global model and control variate, and aggregates the resulting updates. Crucially, each client uploads only its model increment: the server reconstructs the corresponding control increment exactly, halving the uplink vector payload relative to transmitting both increments separately.

Client-side DP updates with subspace filtering. Let $P_q = UU^\top + q(I - UU^\top), \quad U^\top U = I_r, \quad q \in (0,1]$, where the reference basis U and attenuation parameter q are shared by all clients and fixed independently of private data before training, as described in Section . Upon receiving $(\boldsymbol{x}^t, \boldsymbol{c}^t)$, each selected client $i \in S_t$ initializes $\boldsymbol{x}_i^{t,0} \leftarrow \boldsymbol{x}^t$ and performs K local steps. At step k, client i samples a fresh minibatch $B_i^{t,k} \subset D_i$ uniformly without replacement, with $|B_i^{t,k}| = b$. For each sample $z \in B_i^{t,k}$, it computes $g_i^{t,k}(z) = \nabla \ell(\boldsymbol{x}_i^{t,k}; z)$ and applies ℓ_2 clipping: $\bar{g}_i^{t,k}(z) = g_i^{t,k}(z)/\max\{1, \|g_i^{t,k}(z)\|_2/C\}$. The clipped minibatch gradient is $\bar{g}_i^{t,k} = b^{-1}\sum_{z \in B_i^{t,k}} \bar{g}_i^{t,k}(z)$. The client then adds an independent Gaussian perturbation:

$$\widehat{g}_i^{t,k} = \bar{g}_i^{t,k} + \xi_i^{t,k}, \qquad \xi_i^{t,k} \sim \mathcal{N}\left(0, \frac{\sigma^2 C^2}{b^2} I_d\right). \tag{6}$$

Here σ is the noise multiplier, calibrated for the specified neighboring-dataset relation, sampling scheme, and composition across training. The Gaussian perturbation is conditionally zero-mean; clipping may nevertheless bias the gradient estimator.

Algorithm 1: LSSCA with Subspace Filtering and Single-Increment Uploads

Input: Noise multiplier $\sigma > 0$, clipping threshold $C > 0$, shared fixed basis $U \in \mathbb{R}^{d\times r}$ with $U^\top U = I_r$ and $1 \le r < d$, attenuation parameter $q \in (0, 1]$, global stepsize $\eta_g > 0$, local stepsize $\eta_l > 0$, local steps K, total rounds T, participating clients $S \in \{1, \dots, N\}$, batch size $b \le \min_i |D_i|$.
Initialize: Data-independent global model $\boldsymbol{x}^0 \in \mathbb{R}^d$; local controls $\boldsymbol{c}_i^0 \leftarrow \mathbf{0}$ for all $i \in [N]$; global control $\boldsymbol{c}^0 \leftarrow \mathbf{0}$. Fix U and q independently of the private training datasets and share them with all clients. Define $P_q = UU^\top + q(I_d - UU^\top)$.
for $t = 0, 1, \dots, T-1$ **do**
  Sample $S_t \subset [N]$ uniformly without replacement, with $|S_t| = S$.
  Server broadcasts $(\boldsymbol{x}^t, \boldsymbol{c}^t)$ to all $i \in S_t$.
  $\boldsymbol{c}_i^{t+1} \leftarrow \boldsymbol{c}_i^t$ for all $i \notin S_t$.
  for clients $i \in S_t$ **in parallel do**
    $\boldsymbol{x}_i^{t,0} \leftarrow \boldsymbol{x}^t$.
    for $k = 0, 1, \dots, K-1$ **do**
      Sample a fresh minibatch $B_i^{t,k} \subset D_i$ uniformly without replacement, with $|B_i^{t,k}| = b$.
      for $z \in B_i^{t,k}$ **do**
        $g_i^{t,k}(z) \leftarrow \nabla \ell(\boldsymbol{x}_i^{t,k}; z)$; $\bar{g}_i^{t,k}(z) \leftarrow \dfrac{g_i^{t,k}(z)}{\max\{1, \|g_i^{t,k}(z)\|_2 / C\}}$.
      end for
      $\bar{g}_i^{t,k} \leftarrow \dfrac{1}{b}\sum_{z\in B_i^{t,k}} \bar{g}_i^{t,k}(z)$; $\xi_i^{t,k} \sim \mathcal{N}\left(0, \dfrac{\sigma^2 C^2}{b^2} I_d\right)$; $\widehat{g}_i^{t,k} \leftarrow \bar{g}_i^{t,k} + \xi_i^{t,k}$.
      $v_i^{t,k} \leftarrow \widehat{g}_i^{t,k} - \boldsymbol{c}_i^t + \boldsymbol{c}^t$; $\boldsymbol{x}_i^{t,k+1} \leftarrow \boldsymbol{x}_i^{t,k} - \eta_l \left[q v_i^{t,k} + (1-q) U (U^\top v_i^{t,k})\right]$.
    end for
    $\Delta \boldsymbol{x}_i^t \leftarrow \boldsymbol{x}_i^{t,K} - \boldsymbol{x}^t$; $\boldsymbol{c}_i^{t+1} \leftarrow \boldsymbol{c}_i^t - \boldsymbol{c}^t - \dfrac{\Delta \boldsymbol{x}_i^t}{K\eta_l}$.
    Client sends only $\Delta \boldsymbol{x}_i^t$ to the server.
  end for
  $\Delta \boldsymbol{x}^t \leftarrow \dfrac{1}{S}\sum_{i\in S_t} \Delta \boldsymbol{x}_i^t$; $\boldsymbol{x}^{t+1} \leftarrow \boldsymbol{x}^t + \eta_g \Delta \boldsymbol{x}^t$.
  $\Delta \boldsymbol{c}^t \leftarrow -\boldsymbol{c}^t - \dfrac{\Delta \boldsymbol{x}^t}{K\eta_l}$; $\boldsymbol{c}^{t+1} \leftarrow \boldsymbol{c}^t + \dfrac{S}{N}\Delta \boldsymbol{c}^t$. ▷ Exact reconstruction
end for
Output: $\boldsymbol{x}^T$ for final-model evaluation; for the non-convex guarantee, return $\boldsymbol{x}^R$.

The local model is updated by filtering the entire control-corrected direction:

$$\boldsymbol{x}_i^{t,k+1} = \boldsymbol{x}_i^{t,k} - \eta_l P_q \left(\widehat{g}_i^{t,k} - \boldsymbol{c}_i^t + \boldsymbol{c}^t\right), \tag{7}$$

where η_l is the local stepsize. The filter preserves components in $\mathrm{span}(U)$ and scales orthogonal components by q. Thus, it attenuates orthogonal noise while retaining updates in every direction when $q > 0$, with the accompanying signal attenuation explicitly accounted for in our analysis.

Control update and communication. After K local steps, client i updates its control variate: $\boldsymbol{c}_i^{t+1} = \boldsymbol{c}_i^t - \boldsymbol{c}^t + \frac{\boldsymbol{x}^t - \boldsymbol{x}_i^{t,K}}{K\eta_l}$. This update reuses the local model displacement and requires no additional gradient evaluation. Under subspace filtering, it satisfies

$$\boldsymbol{c}_i^{t+1} = (I - P_q)\left(\boldsymbol{c}_i^t - \boldsymbol{c}^t\right) + \frac{P_q}{K}\sum_{k=0}^{K-1} \widehat{g}_i^{t,k}. \tag{8}$$

The residual control term shows that filtering changes the control dynamics as well as the model updates; our analysis explicitly retains this interaction. Define $\Delta \boldsymbol{x}_i^t = \boldsymbol{x}_i^{t,K} - \boldsymbol{x}^t$ and $\Delta \boldsymbol{c}_i^t = \boldsymbol{c}_i^{t+1} - \boldsymbol{c}_i^t$. The control update gives the exact identity $\Delta \boldsymbol{c}_i^t = -\boldsymbol{c}^t - \frac{\Delta \boldsymbol{x}_i^t}{K\eta_l}$. Therefore, client i uploads only $\Delta \boldsymbol{x}_i^t$. Knowing $\boldsymbol{c}^t$, K, and η_l, the server reconstructs $\Delta \boldsymbol{c}_i^t$ without changing the algorithmic model or control updates. Relative to separately transmitting two d-dimensional increments at the same precision, this reduces each participating client's uplink vector payload from $2d$ to d scalars, a 50% reduction. Nonparticipating clients retain their control variates.

Server aggregation. The server averages the received model increments: $\Delta \boldsymbol{x}^t = \frac{1}{S}\sum_{i\in S_t} \Delta \boldsymbol{x}_i^t$, $\boldsymbol{x}^{t+1} = \boldsymbol{x}^t + \Delta \boldsymbol{x}^t$. where η_g is the global stepsize. Using Equation equation , the average control increment is $\Delta \boldsymbol{c}^t = -\boldsymbol{c}^t - \Delta \boldsymbol{x}^t/(K\eta_l)$. Hence the server updates its

control directly as

$$\boldsymbol{c}^{t+1} = \boldsymbol{c}^t + \frac{S}{N}\Delta\boldsymbol{c}^t = \left(1 - \frac{S}{N}\right)\boldsymbol{c}^t - \frac{S}{NK\eta_l}\Delta\boldsymbol{x}^t. \tag{9}$$

With initialization $\boldsymbol{c}^0 = N^{-1}\sum_{i=1}^N \boldsymbol{c}_i^0$, this update preserves $\boldsymbol{c}^t = N^{-1}\sum_{i=1}^N \boldsymbol{c}_i^t$ at every round. The reconstructed controls are deterministic functions of the received noisy model increments and public algorithmic state, so this communication reduction introduces no additional privacy cost.

Differential Privacy Bounds

DP Dwork et al. (2014) provides a formal mathematical framework for quantifying disclosure risk. For privacy accounting, we adopt Rényi differential privacy (RDP) Mironov (2017), which supports adaptive composition. Following Noble et al. (2022b), we report final guarantees in the (ϵ, δ)-DP format via the standard RDP-to-DP conversion Mironov (2017).

Definition 1 (Sample-level DP)**.** *Fix a public roster of N clients and a public lower bound $n_{\min} \geq b$ on local dataset sizes. Two admissible dataset collections $\mathcal{D} = (D_1, \ldots, D_N)$ and $\mathcal{D}'$ are neighboring if they differ by the addition or removal of one record at one client, with both collections satisfying $|D_i| \geq n_{\min}$. A randomized mechanism $\mathcal{M}$ satisfies (ϵ, δ)-DP if, for every ordered neighboring pair and every measurable output set $O \subseteq \mathcal{Y}$, $\Pr[\mathcal{M}(\mathcal{D}) \in O] \leq e^\epsilon \Pr[\mathcal{M}(\mathcal{D}') \in O] + \delta$, where $\epsilon \geq 0$ and $\delta \in (0, 1)$. This definition protects individual records rather than entire clients.*

Definition 2 (Rényi DP Mironov (2017))**.** *For any order $\lambda > 1$, $\mathcal{M}$ satisfies $(\lambda, \epsilon_{\mathrm{R}})$-RDP if, for every ordered neighboring pair $\mathcal{D}, \mathcal{D}'$, $D_\lambda(\mathcal{M}(\mathcal{D}) \,\|\, \mathcal{M}(\mathcal{D}')) := \frac{1}{\lambda-1}\log \mathbb{E}_{Y\sim\mathcal{M}(\mathcal{D}')}\left[\left(\frac{p_{\mathcal{M}(\mathcal{D})}(Y)}{p_{\mathcal{M}(\mathcal{D}')}(Y)}\right)^\lambda\right] \leq \epsilon_{\mathrm{R}}$. The densities are taken with respect to a common dominating measure; the divergence is infinite if the required absolute continuity does not hold.*

Lemma 1 (RDP-to-DP Mironov (2017))**.** *If $\mathcal{M}$ is $(\lambda, \epsilon_{\mathrm{R}})$-RDP, then, for any $\delta \in (0, 1)$, it is $\left(\epsilon_{\mathrm{R}} + \frac{\log(1/\delta)}{\lambda-1}, \delta\right)$-DP.*

At each round, S clients are sampled uniformly without replacement from the fixed roster, independently of private data, with $l = S/N$. Each selected client performs K local steps, using a fresh uniform size-b minibatch at each step. Per-example gradients are clipped to norm at most C, and each minibatch average is perturbed by independent noise $\xi_i^{t,k} \sim \mathcal{N}\left(0, \frac{\sigma^2 C^2}{b^2} I_d\right)$. Here $\sigma > 0$ is the noise multiplier relative to C/b. The initialization, fixed filter P_q, training hyperparameters, and $n_{\min}$ are independent of private data, and controls are initialized to zero. All model and control states are computed from the noisy queries. Exact local dataset sizes, sampled record identities, and additional nonprivate statistics are not released.

Theorem 1 (Privacy guarantee (sample-level))**.** *Under the protocol above, fix $\delta \in (0, 1)$ and define, for every $\lambda > 1$, $\bar{s} := \frac{b}{n_{\min}+1}$, $\rho_{\mathrm{step}}(\lambda) := \frac{1}{\lambda-1}\log\left[1 - \bar{s} + \bar{s}\exp\left(\frac{2\lambda(\lambda-1)}{\sigma^2}\right)\right]$, $\rho_{\mathrm{round}}(\lambda) := \frac{1}{\lambda-1}\log[1 - l + l\exp((\lambda-1)K\rho_{\mathrm{step}}(\lambda))]$. The joint transcript of sampled client sets, uploaded model and control increments, and global model and control states is $(\lambda, T\rho_{\mathrm{round}}(\lambda))$-RDP. Consequently, the global model sequence $\{x^t\}_{t=0}^T$ produced by* `LSSCA` *satisfies (ϵ, δ)-DP towards a third party observing the released models, where $\epsilon = \inf_{\lambda>1}\left\{T\rho_{\mathrm{round}}(\lambda) + \frac{\log(1/\delta)}{\lambda-1}\right\}$.*

The proof is provided in Appendix. This conservative bound permits active clipping and accounts for adaptive local updates and persistent control states. The fixed filter and control updates are post-processing of the accounted noisy queries and incur no additional privacy cost. The bound does not claim an additional privacy amplification gain from aggregating client updates.

Convergence Analysis

We provide the convergence guarantees for the proposed `LSSCA` algorithm. First, we give some common assumptions for our theoretical analysis.

Assumption 1 (Smoothness)**.** *There exists a constant $\beta > 0$ such that, for every client $i \in \{1, \ldots, N\}$ and all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, $\|\nabla f_i(\boldsymbol{x}) - \nabla f_i(\boldsymbol{y})\|_2 \leq \beta\|\boldsymbol{x} - \boldsymbol{y}\|_2$.*

Table 1: Performance comparison under different data heterogeneity levels and privacy budgets on NLP benchmarks with RoBERTa-Base.

Dataset	Method	Data Heterogeneity ($\epsilon = 1$)					Privacy Budget (Dir-0.5)				
		Dir-0.3	Dir-0.5	Dir-0.8	Dir-1.0	IID	$\epsilon = 1$	$\epsilon = 3$	$\epsilon = 6$	$\epsilon = 10$	$\epsilon = \infty$
MNLI	DP-FedAvg	72.65±0.86	73.86±0.74	75.02±0.63	75.71±0.58	78.44±0.41	73.86±0.74	75.21±0.68	76.48±0.51	77.15±0.46	82.03±0.58
	DP-SCAFFOLD	73.28±0.74	74.32±0.62	75.55±0.54	76.20±0.49	79.16±0.38	74.32±0.62	75.74±0.55	76.93±0.47	77.62±0.39	82.50±0.52
	DP-FedSAM	71.39±0.91	72.64±0.83	73.92±0.69	74.63±0.61	77.52±0.46	72.64±0.83	74.18±0.71	75.36±0.56	76.24±0.48	81.35±0.64
	DP-FedAvg-LS	74.28±0.61	75.38±0.51	76.62±0.43	77.31±0.39	80.08±0.34	75.38±0.51	77.06±0.42	78.35±0.39	79.12±0.34	83.72±0.47
	LSSCA	**76.31**±0.52	**77.96**±0.45	**79.42**±0.38	**80.21**±0.34	**82.64**±0.30	**77.96**±0.45	**79.84**±0.36	**81.27**±0.31	**82.06**±0.29	**85.31**±0.42
SST-2	DP-FedAvg	92.55±0.39	92.84±0.35	93.04±0.31	93.18±0.28	93.62±0.23	92.84±0.35	93.12±0.31	93.38±0.27	93.56±0.24	94.05±0.21
	DP-SCAFFOLD	92.64±0.36	92.91±0.32	93.15±0.28	93.27±0.25	93.72±0.21	92.91±0.32	93.24±0.28	93.47±0.25	93.63±0.22	94.12±0.20
	DP-FedSAM	92.31±0.46	92.66±0.41	92.88±0.36	93.02±0.31	93.46±0.26	92.66±0.41	92.97±0.34	93.21±0.31	93.42±0.26	93.88±0.29
	DP-FedAvg-LS	92.89±0.31	93.18±0.27	93.42±0.24	93.55±0.21	93.91±0.18	93.18±0.27	93.55±0.23	93.81±0.20	93.96±0.18	94.12±0.17
	LSSCA	**93.18**±0.28	**93.44**±0.24	**93.72**±0.20	**93.86**±0.18	**94.18**±0.16	**93.44**±0.24	**93.82**±0.18	**94.09**±0.16	**94.24**±0.15	**94.48**±0.14
QQP	DP-FedAvg	86.32±0.48	86.92±0.42	87.38±0.36	87.71±0.32	88.36±0.26	86.92±0.42	87.58±0.36	88.13±0.31	88.44±0.28	89.02±0.25
	DP-SCAFFOLD	86.46±0.43	87.05±0.38	87.56±0.33	87.88±0.29	88.52±0.24	87.05±0.38	87.74±0.33	88.25±0.30	88.57±0.25	89.18±0.22
	DP-FedSAM	85.91±0.53	86.51±0.47	86.97±0.41	87.28±0.36	87.94±0.30	86.51±0.47	87.16±0.41	87.69±0.35	88.04±0.31	88.62±0.29
	DP-FedAvg-LS	87.09±0.36	87.74±0.31	88.24±0.27	88.57±0.23	89.08±0.20	87.74±0.31	88.41±0.27	88.96±0.23	89.25±0.20	89.56±0.18
	LSSCA	**87.62**±0.34	**88.36**±0.29	**88.91**±0.25	**89.23**±0.22	**89.76**±0.18	**88.36**±0.29	**89.03**±0.24	**89.47**±0.21	**89.73**±0.19	**90.06**±0.17
QNLI	DP-FedAvg	83.14±0.53	83.92±0.46	84.68±0.40	85.12±0.35	86.28±0.29	83.92±0.46	85.14±0.39	85.91±0.34	86.36±0.30	87.85±0.26
	DP-SCAFFOLD	83.42±0.49	84.18±0.42	84.91±0.37	85.37±0.32	86.55±0.27	84.18±0.42	85.39±0.35	86.13±0.31	86.57±0.28	88.04±0.24
	DP-FedSAM	82.63±0.58	83.41±0.51	84.12±0.45	84.58±0.39	85.86±0.33	83.41±0.51	84.72±0.43	85.36±0.37	85.82±0.33	87.21±0.30
	DP-FedAvg-LS	84.21±0.42	85.06±0.36	85.84±0.31	86.29±0.27	87.41±0.23	85.06±0.36	86.31±0.31	86.94±0.27	87.38±0.24	88.92±0.21
	LSSCA	**85.16**±0.38	**86.14**±0.32	**87.02**±0.27	**87.56**±0.24	**88.62**±0.20	**86.14**±0.32	**87.36**±0.28	**88.12**±0.24	**88.63**±0.21	**90.18**±0.19

Assumption 2 (Unbiased stochastic gradients and bounded variance)**.** *Let $\mathcal{F}$ denote the history before sampling at client i, and let $\boldsymbol{x}$ be an $\mathcal{F}$-measurable evaluation point. For every client i, a fresh sample z satisfies $\mathbb{E}[\nabla \ell_i(\boldsymbol{x}; z) \mid \mathcal{F}] = \nabla f_i(\boldsymbol{x})$ and $\mathbb{E}[\|\nabla \ell_i(\boldsymbol{x}; z) - \nabla f_i(\boldsymbol{x})\|_2^2 \mid \mathcal{F}] \le \varsigma^2$ almost surely, where $\varsigma^2 < \infty$ is uniform across clients and evaluation points. Each minibatch $\mathcal{B}_i$ consists of b independent draws or is a fresh uniformly sampled size-b subset of the local dataset, with fresh sampling at every local step. Consequently, the raw minibatch gradient $g_i(\boldsymbol{x}; \mathcal{B}_i) = b^{-1} \sum_{z \in \mathcal{B}_i} \nabla \ell_i(\boldsymbol{x}; z)$ satisfies $\mathbb{E}[g_i(\boldsymbol{x}; \mathcal{B}_i) \mid \mathcal{F}] = \nabla f_i(\boldsymbol{x})$ and $\mathbb{E}[\|g_i(\boldsymbol{x}; \mathcal{B}_i) - \nabla f_i(\boldsymbol{x})\|_2^2 \mid \mathcal{F}] \le \sigma_l^2 := \varsigma^2/b$. These conditions concern raw gradients before clipping and Gaussian noise injection.*

Theorem 2 (Non-convex convergence of LSSCA)**.** *Under Assumptions 1–2 and the sampling, noise, and initialization conditions of Theorem 3, let $P_q = UU^\top + q(I - UU^\top)$, where $U^\top U = I_r$, $1 \le r < d$, and $q \in (0, 1]$, with U and q fixed independently of private training data. Let $F_0 = f(\boldsymbol{x}^0) - f_{\inf}$, $L_Q = \beta^2 Q$, and $W_{20} = 8\rho(C^2 + \overline{\mathcal{E}}_{\mathrm{con}}) + 4(\sigma_l^2 + \tau^2 \widetilde{d}_q)/(KS)$, where $\widetilde{d}_q = r + q^2(d - r)$, $\tau = \sigma C/b$, and $Q, \rho, \overline{\mathcal{E}}_{\mathrm{con}}, \sigma_l^2$ are defined in the appendix. For $\eta_g \ge 1$, $0 < h \le 1/(8\beta)$, and $\eta_l = h/(K\eta_g\sqrt{T})$, the output $\boldsymbol{x}^R$, with R independent and uniform on $\{0, \dots, T-1\}$, satisfies*

$$\mathbb{E}\|\nabla f(\boldsymbol{x}^R)\|_{P_q}^2 \le (2F_0/h + \beta h W_{20})/\sqrt{T} + 2L_Q h^2/T + 4\beta L_Q h^3/T^{3/2}. \tag{10}$$

Consequently, $\mathbb{E}\|\nabla f(\boldsymbol{x}^R)\|_2^2 \le C_{\mathrm{nc}}/(q\sqrt{T})$, where $C_{\mathrm{nc}} = 2F_0/h + \beta h W_{20} + 2L_Q h^2 + 4\beta L_Q h^3$. Thus, $T \ge \max\{1, C_{\mathrm{nc}}^2/(q^2\epsilon^2)\}$ suffices for ϵ-stationarity in expected squared gradient norm, with all problem parameters, the filter, and the noise scale fixed.

Effect of subspace filtering and stochastic control. Theorem 2 reveals two pathways through which privacy noise affects convergence. Its direct contribution through W_{20} is $4\beta h \tau^2 \widetilde{d}_q/(KS\sqrt{T})$, where $\widetilde{d}_q = r + q^2(d - r)$ replaces the ambient dimension d. Subspace filtering therefore preserves noise within the reference subspace while attenuating its orthogonal components by q^2 in variance. Noise also propagates indirectly through the control states and local trajectories, as captured by $\overline{\mathcal{E}}_{\mathrm{con}}$ and Q. In particular, the control recursion retains the residual $(I - P_q)(\boldsymbol{c}_i^t - \boldsymbol{c}^t)$, showing that filtering and drift correction interact rather than contribute independent improvements. The resulting trade-off is explicit: decreasing q reduces the direct noise term but weakens directional progress through $\lambda_{\min}(P_q) = q$ and can enlarge the control bound. Thus, the theorem supports selective noise attenuation together with an $\mathcal{O}(T^{-1/2})$ fixed-noise stationarity rate, while retaining the costs of filtering in the convergence constants.

EXPERIMENTS

EXPERIMENTAL SETTINGS

Datasets and data dartition. We implement all experiments in Ray Moritz et al. (2017) based on Python. We verify the effectiveness of LSSCA by training the model on the NLP dataset, such as SST-

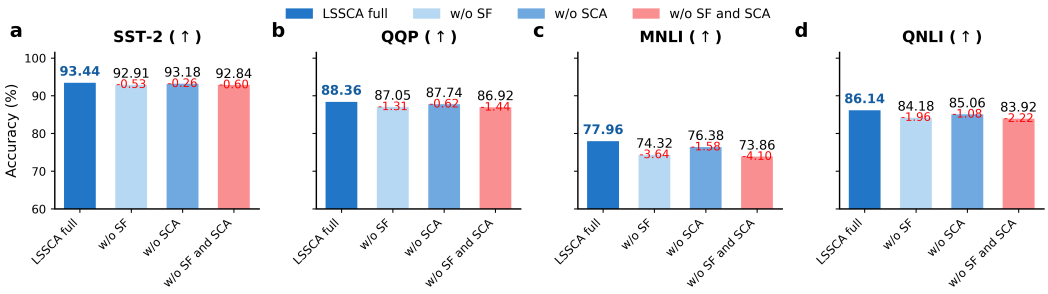

Figure 3: Ablation study of different components in LSSCA on NLP benchmarks.

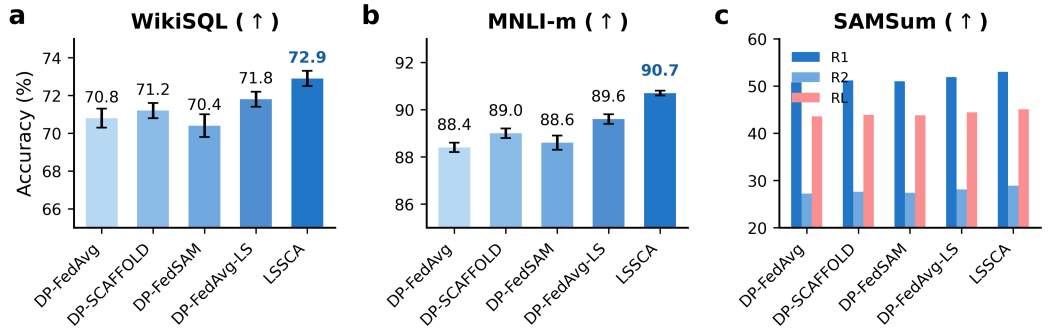

Figure 4: Comparison of LSSCA with DP FL methods on GPT-2 XL 1.5B across generation tasks.

2 Socher et al. (2013), QQP, MNLI, and QNLI Rajpurkar et al. (2018). The heterogeneous datasets we used in the experiment are generated by Dirichlet distribution Hsu et al. (2019): The subset of the dataset for each client obeys a Dirichlet distribution, which is a categorical distribution, *i.e.* $\mathcal{D}_i \sim Dir(\rho q)$ where $\rho > 0$ is the concentration parameter to determine the degree of heterogeneity and q is a uniform prior probability.

Model. We use many architectures: RoBERTa-Base Liu et al. (2019) and GPT-2 XL 1.5B.

Baseline algorithms. We compare the proposed `LSSCA` algorithm with several recently proposed algorithms of federated learning, including DP-FedAvg McMahan et al. (2018), DP-SCAFFOLD Noble et al. (2022a), DP-FedAvg-LS Liang et al. (2024) and DP-FedSAM Shi et al. (2023), and evaluate them on several NLP. In the experimental setup, we used following hyperparameters: total number of clients $N = 100$, mini-batch size $B = 50$. The clipping threshold is selected from $C \in \{2, 5, 10\}$. and client sampling rate $\gamma = \frac{S_t}{N} = 0.1$. And we set the local learning rate by $\eta_l \in [0.01, 0.025, 0.1, 0.25]$, best is 0.01 and the global learning rate by $\eta_g = 1.0$. The default noise multiplier σ is set to 10, and the subspace attenuation parameter q and rank r are set to 0.5 and 32, respectively, with the privacy failure probability $\delta = 1/(NR)$. R is the number of local samples per client. The number of steps for local training on clients, denoted as $K = 50$. All results are averaged over 5 runs with std reported with seeds 42, 43, 44, 45, 46. All experiments were performed on 8 NVIDIA RTX 4090 GPUs.

Effect of data heterogeneity Table 1 evaluates the effect of data heterogeneity under the strong privacy setting $\epsilon = 1$. As the Dirichlet parameter increases from Dir-0.3 to IID, all methods achieve higher accuracy, showing that non-IID client distributions substantially affect private federated fine-tuning. Across all datasets and heterogeneity levels, `LSSCA` consistently performs best. The gains are more pronounced on MNLI and QNLI, where client drift is stronger; for example, under Dir-0.5, `LSSCA` improves over DP-FedAvg-LS by 2.58% on MNLI and 1.08% on QNLI. These results demonstrate that combining Low-pass Smoothing with stochastic controlled averaging effectively mitigates both DP-induced noise and heterogeneity-induced drift, while introducing only marginal extra client-side computation compared with DP-FedAvg and DP-FedAvg-LS.

Effect of privacy budgets Table 1 compares different DP-FL methods on RoBERTa-Base under the Dir-0.5 non-IID setting. As expected, larger ϵ improves accuracy due to reduced DP noise. Across all privacy budgets and datasets, `LSSCA` consistently achieves the best performance, indicating a

superior privacy–utility trade-off. The gains are particularly evident on challenging tasks affected by privacy noise and heterogeneity. For example, on MNLI, LSSCA surpasses DP-FedAvg-LS by 1.59%–2.94% across different privacy budgets, while achieving over 1.0% improvement on QNLI. The smaller gain on SST-2 is attributed to its simpler binary classification nature, where all methods already reach high accuracy. Moreover, LSSCA introduces negligible computational overhead. On MNLI, it requires only 5.63s per round, comparable to DP-FedAvg-LS (5.49s) and substantially lower than DP-FedSAM (9.35s). These results demonstrate that LSSCA effectively alleviates DP perturbation and client drift with limited additional cost.

Scaling to GPT-2 XL. To further evaluate the scalability of LSSCA to large language models, we conduct GPT-2 XL experiments under the sample-level DP budget $\epsilon = 3$, following the LoRA evaluation protocol on WikiSQL, MNLI-matched, and SAMSum. As shown in Table 2, privacy noise causes clear degradation, but LSSCA-LoRA consistently outperforms all DP baselines under the same LoRA trainable scope and parameter budget. Compared with DP-FedAvg-LS-LoRA, LSSCA-LoRA improves WikiSQL accuracy from 71.8% to 72.9%, MNLI-m accuracy from 89.6% to 90.7%, and SAMSum ROUGE-L from 44.4 to 45.1. These results suggest that soft subspace filtering and stochastic controlled averaging remain effective for parameter-efficient private federated fine-tuning at GPT-2 XL scale, improving the privacy–utility trade-off without increasing the trainable parameter budget.

Table 2: GPT-2 XL results under sample-level DP ($\epsilon = 3$). We report logical-form validation accuracy on WikiSQL, matched validation accuracy on MNLI, and ROUGE-1/2/L on SAMSum.

Method	WikiSQL	MNLI-m	SAMSum
	Acc. (%)	Acc. (%)	R1/R2/RL
DP-FedAvg	$70.8_{\pm 0.5}$	$88.4_{\pm 0.2}$	$50.8/27.2/43.6$
DP-SCAFFOLD	$71.2_{\pm 0.4}$	$89.0_{\pm 0.2}$	$51.2/27.6/43.9$
DP-FedSAM	$70.4_{\pm 0.6}$	$88.6_{\pm 0.3}$	$51.0/27.4/43.8$
DP-FedAvg-LS	$71.8_{\pm 0.4}$	$89.6_{\pm 0.2}$	$51.9/28.1/44.4$
LSSCA	$\mathbf{72.9}_{\pm 0.4}$	$\mathbf{90.7}_{\pm 0.1}$	$\mathbf{53.0/28.9/45.1}$

Ablation Study

Ablation study of SF and SCA. Table 3 evaluates subspace filtering (SF) and stochastic controlled averaging (SCA) under $\epsilon = 1$ and Dir-0.5. Removing SF recovers DP-SCAFFOLD under the same experimental configuration, and its results match the corresponding entries in the main comparison. Full LSSCA achieves the highest mean accuracy on all four benchmarks. Removing SF reduces accuracy by 0.53–3.64 percentage points, with the largest drops on MNLI (3.64 points) and QNLI (1.96 points). Removing SCA also lowers accuracy across all benchmarks, by 0.26–1.58 points. Removing both components yields the largest degradation on every benchmark, including 4.10 points on MNLI and 2.22 points on QNLI. These results support combining subspace filtering with control-based drift correction for private federated fine-tuning under heterogeneous data.

Table 3: Component ablation on RoBERTa-Base ($\epsilon = 1$, Dir-0.5). Accuracy is reported in %; Δ is the change from full LSSCA in percentage points.

Variant	SST-2		QQP		MNLI		QNLI	
	Acc.	Δ	Acc.	Δ	Acc.	Δ	Acc.	Δ
LSSCA full	**93.44**	–	**88.36**	–	**77.96**	–	**86.14**	–
w/o SF	92.91	-0.53	87.05	-1.31	74.32	-3.64	84.18	-1.96
w/o SCA	93.18	-0.26	87.74	-0.62	76.38	-1.58	85.06	-1.08
w/o SF & SCA	92.84	-0.60	86.92	-1.44	73.86	-4.10	83.92	-2.22

Conclusion

We introduced LSSCA, which combines shared soft subspace filtering with stochastic controlled averaging for differentially private federated learning. The method preserves reference-subspace updates while attenuating orthogonal components and correcting client-specific directions. Our analysis establishes non-convex convergence under the stated assumptions and quantifies how filtering affects noise, local trajectories, and control states. It exposes a central trade-off: stronger attenuation reduces noise energy but can weaken directional progress. Experiments across six federated datasets with Dirichlet-partitioned clients evaluate this trade-off against DP-SCAFFOLD, DP-FedAvg-LS, and DP-FedSAM. These findings motivate jointly designing update geometry and drift correction, with subspace alignment and computational overhead remaining important considerations.

REFERENCES

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology-EUROCRYPT 2006: 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503, 2006.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Wei Feng, Zheng Yan, Hengrun Zhang, Kai Zeng, Yu Xiao, and Y. Thomas Hou. A survey on security, privacy, and trust in mobile crowdsourcing. *IEEE Internet of Things Journal*, 5(4):2971–2992, 2018.

Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An End-to-End case study of personalized warfarin dosing. In *23rd USENIX Security Symposium (USENIX Security 14)*, pp. 17–32, 2014.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients - how easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems*, volume 33, pp. 16937–16947, 2020.

Robin C. Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *CoRR*, abs/1712.07557, 2017.

Antonious M. Girgis, Deepesh Data, Suhas Diggavi, Peter Kairouz, and Ananda Theertha Suresh. Shuffled model of federated learning: Privacy, accuracy and communication trade-offs. *IEEE Journal on Selected Areas in Information Theory*, 2(1):464–478, 2021.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the Effects of Non-Identical Data Distribution for Federated Visual Classification. *arXiv e-prints*, art. arXiv:1909.06335, September 2019.

Rui Hu, Yuanxiong Guo, Hongning Li, Qingqi Pei, and Yanmin Gong. Personalized federated learning with differential privacy. *IEEE Internet of Things Journal*, 7(10):9530–9539, 2020.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning, 2020.

Zhicong Liang, Bao Wang, Quanquan Gu, Stanley Osher, and Yuan Yao. Differentially private federated learning with laplacian smoothing. *Applied and Computational Harmonic Analysis*, pp. 101660, 2024.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1273–1282, 2017.

H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations*, 2018.

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

Philipp Moritz, Robert Nishihara, Stephanie Wang, Alexey Tumanov, Richard Liaw, Eric Liang, Melih Elibol, Zongheng Yang, William Paul, Michael I. Jordan, and Ion Stoica. Ray: A distributed framework for emerging ai applications, 2017.

Maxence Noble, Aurélien Bellet, and Aymeric Dieuleveut. Differentially private federated learning on heterogeneous data. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pp. 10110–10145, 2022a.

Maxence Noble, Aurélien Bellet, and Aymeric Dieuleveut. Differentially private federated learning on heterogeneous data. In *International conference on artificial intelligence and statistics*, pp. 10110–10145. PMLR, 2022b.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.

Yifan Shi, Yingqi Liu, Kang Wei, Li Shen, Xueqian Wang, and Dacheng Tao. Make landscape flatter in differentially private federated learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24552–24562, 2023.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, 2017.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *2018 IEEE 31st Computer Security Foundations Symposium (CSF)*, pp. 268–282, 2018.

Yang Zhao, Jun Zhao, Mengmeng Yang, Teng Wang, Ning Wang, Lingjuan Lyu, Dusit Niyato, and Kwok-Yan Lam. Local differential privacy-based federated learning for internet of things. *IEEE Internet of Things Journal*, 8(11):8836–8853, 2021.

ETHICS STATEMENT

This work studies differentially private federated fine-tuning using existing NLP benchmarks: SST-2, QQP, MNLI, QNLI, WikiSQL, and SAMSum. Our experiments simulate federated clients using benchmark datasets rather than collecting new data from users. LSSCA combines control-based drift correction with data-independent random subspace filtering to improve the privacy–utility trade-off. The privacy analysis provides record-level differential privacy under the stated protocol and accounting assumptions; it does not provide client-level protection or eliminate all security risks. Models may also inherit biases and harmful content from pretrained models and benchmark datasets, and differential privacy does not ensure fairness or safe model behavior. Deployment therefore requires validated privacy accounting, appropriate security safeguards, and application-specific safety evaluation.

REPRODUCIBILITY STATEMENT

The paper describes the LSSCA algorithm, including local gradient clipping and noise injection, control-variate updates, and fixed random subspace filtering. The main text and appendices present the theoretical assumptions, convergence analysis, and record-level privacy analysis with detailed proofs. The convergence guarantees require the stated assumptions, including inactive clipping, whereas the privacy analysis permits active clipping. The experimental section describes the benchmarks, client heterogeneity settings, model architectures, baseline methods, optimization and filtering hyperparameters, and computational setup. Experiments cover RoBERTa-Base and GPT-2 XL, with comparisons across privacy budgets and heterogeneity levels, component ablations, and runtime measurements. Results are averaged over five runs using seeds 42, 43, 44, 45, and 46. We will publicly release the implementation and experiment scripts upon acceptance.

AI USE STATEMENT

Large language models were used to assist with language editing, manuscript organization, and drafting supplementary material. They also assisted in identifying potential inconsistencies, examining mathematical formulations, and suggesting additional experiments and evaluation protocols. AI-generated content does not constitute independent evidence for theoretical or empirical claims. The authors are responsible for verifying all incorporated material and for the accuracy and integrity of the final manuscript, including its proofs, experimental results, and citations.

NON-CONVEX CONVERGENCE OF LSSCA

We first introduce the assumptions and auxiliary lemmas, and then establish finite-time guarantees for LSSCA with fixed subspace filtering. We explicitly assume that clipping is inactive at the analyzed gradient evaluation points. Both minibatch randomness and Gaussian perturbations are included in the noisy stochastic oracle. The control recursion below retains the subspace filter and the noise.

ASSUMPTIONS AND NOTATION

For a positive semidefinite matrix M, write $\|u\|_M^2 = u^\top M u$ and $\langle u, v\rangle_M = u^\top M v$. All sums over local-step indices run from 0 to $K-1$ unless stated otherwise.

Let $f(x) = N^{-1}\sum_{i=1}^N f_i(x)$, where $N \geq 1$, $1 \leq S \leq N$, $K \geq 1$, and $b \geq 1$. The objective f_i uses the same sample distribution as the local stochastic gradient. In particular, uniform sampling from a finite client dataset defines its empirical risk.

Assumption 3 (Smoothness and lower boundedness)**.** *There exists $\beta > 0$ such that each $f_i : \mathbb{R}^d \to \mathbb{R}$ is differentiable and, for all $x, y \in \mathbb{R}^d$,*

$$\|\nabla f_i(x) - \nabla f_i(y)\|_2 \leq \beta\|x - y\|_2.$$

The average objective satisfies $f(x) \geq f_{\inf} > -\infty$ for all x. No convexity or existence of a minimizer is assumed.

Assumption 4 (Conditionally unbiased stochastic gradients)**.** *For a fresh sample z from client i and each evaluation point ω,*

$$\mathbb{E}_z[\nabla \ell_i(\omega; z)] = \nabla f_i(\omega), \qquad \mathbb{E}_z\|\nabla \ell_i(\omega; z) - \nabla f_i(\omega)\|_2^2 \leq \varsigma^2.$$

The same statements hold conditional on the history before each fresh minibatch. Each minibatch consists of b independent draws, or is a fresh uniform size-b subset of the client dataset, whose size is at least b. Fresh sampling across local steps is required; arbitrary random reshuffling is not silently identified with this conditional model. Thus the raw minibatch variance is at most $\sigma_l^2 := \varsigma^2/b$.

Assumption 5 (Inactive clipping)**.** *There is a fixed $C > 0$, independent of the client, sample, and local step, such that at all local evaluation points in the analyzed run,*

$$\|\nabla \ell_i(\omega_i^{t,k}; z)\|_2 \leq C$$

almost surely under the fresh sample distribution. The condition includes $k = 0$, where $\omega_i^{t,0} = x^t$. It holds almost surely over training randomness for all clients and steps, including the virtual trajectories defined below.

Remark 1 (Meaning and limitation of inactive clipping)**.** *This is a genuine restriction on the raw gradients, not a consequence of applying the clipping operator. Under this condition $\mathrm{clip}_C(\nabla \ell_i) = \nabla \ell_i$, so the minibatch estimator stays unbiased. If clipping is active in an implementation, the theorem below does not automatically apply to that implementation. A clipping-bias term must then be restored. Also, a uniform bound independent of the horizon is needed to interpret asymptotic rates with fixed constants.*

Subspace filtering, sampling, and noise. Assume $\beta > 0$ and a deterministic initial model x^0. Fix $d \geq 2$, $1 \leq r < d$, an orthonormal basis $U \in \mathbb{R}^{d\times r}$ with $U^\top U = I_r$, and $q \in (0, 1]$. The basis and attenuation parameter are shared by all clients and fixed throughout training, independently of the private training datasets. Define

$$P_q := UU^\top + q(I - UU^\top). \tag{11}$$

For any update vector v, write $v_\parallel = UU^\top v$ and $v_\perp = (I - UU^\top)v$. Subspace filtering computes

$$P_q v = v_\parallel + q v_\perp = qv + (1 - q)U(U^\top v).$$

Thus reference-subspace components are preserved and orthogonal components are attenuated. Before private training, we sample $G \in \mathbb{R}^{d\times r}$ with independent standard Gaussian entries and obtain an orthonormal basis U through a thin QR decomposition. This basis is sampled independently of the private training data and subsequent training randomness, shared by all clients, and fixed throughout

optimization. No public reference data are required. Our convergence analysis conditions on the realized basis U; the resulting bounds also hold after averaging over its randomness. No alignment of private gradients with the random subspace is assumed, and this construction alone does not imply improvement over the unfiltered method. Updating the basis using private training gradients is outside the present analysis.

The two orthogonal projectors in equation 11 imply

$$\begin{aligned} &0 \prec P_q \preceq I, \qquad \Lambda_{\min} = \lambda_{\min}(P_q) = q, \\ &d_q := \operatorname{tr}(P_q) = r + q(d-r), \\ &\widetilde{d}_q := \operatorname{tr}(P_q^2) = r + q^2(d-r). \end{aligned} \tag{12}$$

Indeed, P_q has eigenvalue 1 on $\operatorname{span}(U)$ and q on its orthogonal complement. Hence $P_q^2 \preceq P_q \preceq I$. These are precisely the spectral properties used below. The restriction $q > 0$ is necessary for the conversion to Euclidean stationarity and the uniform control bound used here; hard projection with $q = 0$ is not covered. The matrix-vector product requires $\mathcal{O}(dr)$ operations and storage for an explicit basis, without forming a dense $d \times d$ matrix.

Blockwise implementation. Partition the trainable parameters into B_{blk} fixed blocks of sizes d_j. Choose data-independent fixed bases $U_j \in \mathbb{R}^{d_j \times r_j}$ satisfying $U_j^\top U_j = I_{r_j}$, with $1 \le r_j \le d_j$, a common $q \in (0, 1]$, and $\sum_j r_j < \sum_j d_j$. With $U = \operatorname{diag}(U_1, \dots, U_{B_{\mathrm{blk}}})$, $d = \sum_j d_j$, and $r = \sum_j r_j$, the same spectral formulas apply. The application cost is $\mathcal{O}(\sum_j d_j r_j)$. Here r is the total rank, not the rank of a single block. The proof requires only the stated spectral properties, not a globally uniform random subspace; it therefore also covers this blockwise construction.

At each round the server samples a uniform size-S subset $\boldsymbol{S}_t$, independently of the fresh local randomness conditional on the past $\mathcal{F}_t$, which contains the fixed basis, all model and control states, and all randomness revealed before the current selection. Stepsizes are deterministic and positive. At every client and step use an independent $\xi_i^{t,k} \sim \mathcal{N}(0, \tau^2 I)$, also independent of the same-step minibatch. For a noise multiplier $\sigma > 0$, the LSSCA noise convention is

$$\tau = \sigma C / b. \tag{13}$$

The parameter τ is the actual coordinate standard deviation. Any privacy calibration must use the same noise normalization, adjacency relation, sampling scheme, and released transcript as the implementation. The result below is an optimization guarantee at this specified noise scale.

Let $g_i^{t,k}$ be the raw minibatch gradient. Under Assumption 5, it equals the clipped minibatch gradient. Write $\widehat{g}_i^{t,k} = g_i^{t,k} + \xi_i^{t,k}$. The updates are

$$\omega_i^{t,0} = x^t, \qquad \omega_i^{t,k+1} = \omega_i^{t,k} - \eta_{l,t} P_q (\widehat{g}_i^{t,k} - \boldsymbol{c}_i^t + \boldsymbol{c}^t), \tag{14}$$

$$v^t = \frac{1}{KS} \sum_{i \in \boldsymbol{S}_t} \sum_{k=0}^{K-1} (\widehat{g}_i^{t,k} - \boldsymbol{c}_i^t + \boldsymbol{c}^t), \qquad x^{t+1} = x^t - \bar{\eta}_t P_q v^t, \tag{15}$$

$$\bar{\eta}_t = K \eta_{l,t} \eta_g, \qquad \eta_g \ge 1,$$

$$\boldsymbol{c}_i^{t+1} = \boldsymbol{c}_i^t - \boldsymbol{c}^t + \frac{x^t - \omega_i^{t,K}}{K \eta_{l,t}} \quad (i \in \boldsymbol{S}_t), \tag{16}$$

$$\boldsymbol{c}_i^{t+1} = \boldsymbol{c}_i^t \quad (i \notin \boldsymbol{S}_t), \qquad \boldsymbol{c}^{t+1} = \boldsymbol{c}^t + \frac{1}{N} \sum_{i \in \boldsymbol{S}_t} (\boldsymbol{c}_i^{t+1} - \boldsymbol{c}_i^t). \tag{17}$$

Initialize $\boldsymbol{c}_i^0 = 0$ for every client and $\boldsymbol{c}^0 = 0$, as in the algorithm. All expectations below are over client selection, fresh minibatches, and Gaussian noise, conditional on the fixed data and basis. For each finite horizon, the gradient bound, finite Gaussian moments, and the linear update rules give finite second moments of the iterates and controls. Smoothness then ensures integrability of $f(x^t)$, justifying the expectations used below.

Virtual trajectories. Conditional on $\mathcal{F}_t$, generate independent local trajectories for all clients, and independently select $\boldsymbol{S}_t$. Keeping only the selected trajectories gives the same distribution as the real algorithm. Unselected trajectories are analytical devices only: no additional updates or communications are performed.

Gradient heterogeneity. No additional (G, B) dissimilarity assumption is imposed. The raw-gradient bound and unbiasedness imply $\|\nabla f_i(x^t)\|_2 \le C$ for every client. This implication is used in the client-sampling variance bound below. In particular, this already implies

$$\frac{1}{N}\sum_i \|\nabla f_i(x^t) - \nabla f(x^t)\|_2^2 = \frac{1}{N}\sum_i \|\nabla f_i(x^t)\|_2^2 - \|\nabla f(x^t)\|_2^2 \le C^2.$$

Thus no separate dissimilarity parameter is introduced, but the analysis does use the heterogeneity bound implied by inactive clipping.

AUXILIARY LEMMAS

Lemma 2 (Noise attenuation by subspace filtering)**.** *Let $\boldsymbol{n} \sim \mathcal{N}(0, \mathcal{L}^2\sigma^2 I_d)$, where $\mathcal{L} = C/b$, so that $\tau = \mathcal{L}\sigma$. Then*

$$\begin{aligned} \mathbb{E}\|\boldsymbol{n}\|_{P_q}^2 &= \mathcal{L}^2\sigma^2 d_q, \\ \mathbb{E}\|P_q\boldsymbol{n}\|_2^2 &= \mathcal{L}^2\sigma^2 \widetilde{d}_q. \end{aligned} \tag{18}$$

Proof. For any fixed symmetric M,

$$\mathbb{E}[\boldsymbol{n}^\top M \boldsymbol{n}] = \operatorname{tr}(M\mathbb{E}[\boldsymbol{n}\boldsymbol{n}^\top]) = \mathcal{L}^2\sigma^2 \operatorname{tr}(M).$$

Use $M = P_q$ for the first identity and $M = P_q^2$ for the second. Furthermore, orthogonality of the projectors gives

$$P_q^2 = UU^\top + q^2(I - UU^\top), \qquad \operatorname{Cov}(P_q\boldsymbol{n}) = \tau^2 P_q^2.$$

Thus the noise variance per unit direction is τ^2 inside the reference subspace and $q^2\tau^2$ in its orthogonal complement. For the explicit averaged noise $\overline{\xi}^t = (KS)^{-1}\sum_{i\in\boldsymbol{S}_t,k}\xi_i^{t,k}$, independence of the fresh perturbations yields

$$\mathbb{E}\|P_q\overline{\xi}^t\|_2^2 = \frac{\tau^2}{KS}\operatorname{tr}(P_q^2) = \frac{\tau^2[r + q^2(d-r)]}{KS}. \tag{19}$$

This is smaller than the unfiltered variance $\tau^2 d/(KS)$ when $q < 1$. It describes the explicit noise term, not the full update variance: later local gradients also depend on earlier perturbations.

□

Lemma 3 (Conditional oracle moments)**.** *Let $e_i^{t,k} = \widehat{g}_i^{t,k} - \nabla f_i(\omega_i^{t,k})$ and let $\mathcal{F}_{t,i,k}$ denote the history immediately before the fresh minibatch and noise at that step. Define*

$$V_2 = \sigma_l^2 + \tau^2\widetilde{d}_q. \tag{20}$$

Then

$$\mathbb{E}[e_i^{t,k} \mid \mathcal{F}_{t,i,k}] = 0, \qquad \mathbb{E}[\|P_q e_i^{t,k}\|_2^2 \mid \mathcal{F}_{t,i,k}] \le V_2.$$

For $M_t = (KS)^{-1}\sum_{i\in\boldsymbol{S}_t,k} e_i^{t,k}$,

$$\mathbb{E}[M_t \mid \mathcal{F}_t] = 0, \qquad \mathbb{E}\|P_q M_t\|_2^2 \le \frac{V_2}{KS}. \tag{21}$$

Proof. Write $\epsilon_i^{t,k} = g_i^{t,k} - \nabla f_i(\omega_i^{t,k})$. Inactive clipping gives $e_i^{t,k} = \epsilon_i^{t,k} + \xi_i^{t,k}$. Conditional unbiasedness and fresh zero-mean Gaussian noise imply $\mathbb{E}[e_i^{t,k} \mid \mathcal{F}_{t,i,k}] = 0$. Moreover,

$$\begin{aligned} \mathbb{E}[\|P_q e_i^{t,k}\|_2^2 \mid \mathcal{F}_{t,i,k}] &= \mathbb{E}[\|P_q\epsilon_i^{t,k}\|_2^2 \mid \mathcal{F}_{t,i,k}] + \mathbb{E}[\|P_q\xi_i^{t,k}\|_2^2 \mid \mathcal{F}_{t,i,k}] \\ &\le \sigma_l^2 + \tau^2\operatorname{tr}(P_q^2) = V_2. \end{aligned} \tag{22}$$

The cross term vanishes because the same-step Gaussian noise is independent of the minibatch and has conditional mean zero; $P_q^2 \preceq I$ bounds the minibatch contribution.

Conditional on $\mathcal{F}_t$ and $\boldsymbol{S}_t$, order the KS oracle calls so that each client's local order is preserved. Let e_j be the jth error and $\mathcal{G}_{j-1}$ the history before that call. Independence of fresh randomness across clients and local steps gives $\mathbb{E}[e_j \mid \mathcal{G}_{j-1}] = 0$. For $a < j$,

$$\mathbb{E}\langle P_q e_a, P_q e_j\rangle = \mathbb{E}\langle P_q e_a, P_q\mathbb{E}[e_j \mid \mathcal{G}_{j-1}]\rangle = 0.$$

Consequently,

$$\mathbb{E}\|P_q M_t\|_2^2 = \frac{1}{K^2 S^2} \sum_{j=1}^{KS} \mathbb{E}\|P_q e_j\|_2^2 \le \frac{V_2}{KS}.$$

The tower property gives the stated conditional mean of M_t. This argument does not assume that later gradients are independent of previous Gaussian perturbations. □

Lemma 4 (Uniform client-sampling variance)**.** *Set $\rho = (N-S)/(S(N-1))$ if $N > 1$, and $\rho = 0$ if $N = 1$. For fixed vectors u_i, their mean $\bar{u}$, and a positive semidefinite matrix M,*

$$\mathbb{E}\left\| \frac{1}{S} \sum_{i \in \boldsymbol{S}_t} u_i - \bar{u} \right\|_M^2 = \frac{\rho}{N} \sum_i \|u_i - \bar{u}\|_M^2. \tag{23}$$

Proof. For $N > 1$, set $z_i = u_i - \bar{u}$, so that $\sum_i z_i = 0$, and let $I_i = \mathbf{1}\{i \in \boldsymbol{S}_t\}$. Uniform sampling without replacement gives

$$\mathbb{E} I_i = \frac{S}{N}, \qquad \mathbb{E}[I_i I_j] = \frac{S(S-1)}{N(N-1)} \quad (i \neq j).$$

Expanding the second moment and using the zero sum of the z_i gives

$$\begin{aligned} \mathbb{E}\left\| \frac{1}{S} \sum_i I_i z_i \right\|_M^2 &= \frac{1}{S^2} \left[\frac{S}{N} \sum_i \|z_i\|_M^2 + \frac{S(S-1)}{N(N-1)} \sum_{i \neq j} \langle z_i, z_j \rangle_M \right] \\ &= \frac{1}{S^2} \left[\frac{S}{N} - \frac{S(S-1)}{N(N-1)} \right] \sum_i \|z_i\|_M^2 \\ &= \frac{N-S}{SN(N-1)} \sum_i \|z_i\|_M^2. \end{aligned} \tag{24}$$

Here $\sum_{i \neq j} \langle z_i, z_j \rangle_M = -\sum_i \|z_i\|_M^2$. For $N = 1$, the sampled mean equals the full mean and both sides are zero. □

Control and drift quantities. We use $\mathcal{E}_{\mathrm{con}}^t$ for the second moment of the random controls, as defined below. It is not a squared error of the unconditional mean control, nor a gradient-tracking error. The quantity $\mathcal{E}_{\mathrm{cli}}^t$ measures Euclidean displacement at the gradient evaluation points $k = 0, \ldots, K-1$. These definitions are chosen to match the actual noisy recursion.

Lemma 5 (Exact control identity and stability)**.** *Define*

$$p = S/N, \qquad \mathcal{E}_{\mathrm{con}}^t = \frac{1}{N} \sum_i \mathbb{E}\|\boldsymbol{c}_i^t\|_2^2,$$

$$V_c = 2C^2 + \frac{2au^2 d_q}{K}, \qquad \overline{\mathcal{E}}_{\mathrm{con}} = \max\{\mathcal{E}_{\mathrm{con}}^0, V_c/\Lambda_{\min}\}.$$

Then $\boldsymbol{c}^t = N^{-1} \sum_i \boldsymbol{c}_i^t$, and for each selected client,

$$\boldsymbol{c}_i^{t+1} = (I - P_q)(\boldsymbol{c}_i^t - \boldsymbol{c}^t) + \frac{P_q}{K} \sum_k \widehat{g}_i^{t,k}. \tag{25}$$

Furthermore,

$$\begin{aligned} &\mathcal{E}_{\mathrm{con}}^{t+1} \le (1 - p\Lambda_{\min}) \mathcal{E}_{\mathrm{con}}^t + pV_c, \\ &\mathcal{E}_{\mathrm{con}}^t \le (1 - p\Lambda_{\min})^t \mathcal{E}_{\mathrm{con}}^0 + \left[1 - (1 - p\Lambda_{\min})^t\right] \frac{V_c}{\Lambda_{\min}} \le \overline{\mathcal{E}}_{\mathrm{con}}. \end{aligned} \tag{26}$$

Proof. The initialization and server update imply

$$\frac{1}{N}\sum_i \boldsymbol{c}_i^{t+1} = \frac{1}{N}\sum_i \boldsymbol{c}_i^t + \frac{1}{N}\sum_{i\in\boldsymbol{S}_t}(\boldsymbol{c}_i^{t+1} - \boldsymbol{c}_i^t) = \boldsymbol{c}^{t+1}. \tag{27}$$

Thus $\boldsymbol{c}^t = N^{-1}\sum_i \boldsymbol{c}_i^t$ holds by induction. Summing the local steps,

$$\begin{aligned}\frac{x^t - \omega_i^{t,K}}{K\eta_{l,t}} &= \frac{1}{K}\sum_{k=0}^{K-1} P_q(\widehat{g}_i^{t,k} - \boldsymbol{c}_i^t + \boldsymbol{c}^t) \\ &= P_q\left(\frac{1}{K}\sum_k \widehat{g}_i^{t,k} - \boldsymbol{c}_i^t + \boldsymbol{c}^t\right).\end{aligned} \tag{28}$$

Substituting this identity in the original control update gives

$$\begin{aligned}\boldsymbol{c}_i^{t+1} &= \boldsymbol{c}_i^t - \boldsymbol{c}^t + P_q\left(\frac{1}{K}\sum_k \widehat{g}_i^{t,k} - \boldsymbol{c}_i^t + \boldsymbol{c}^t\right) \\ &= (I - P_q)(\boldsymbol{c}_i^t - \boldsymbol{c}^t) + \frac{P_q}{K}\sum_k \widehat{g}_i^{t,k}.\end{aligned} \tag{29}$$

For any vectors a, v, diagonalizing P_q and applying convexity of the scalar square in each eigendirection yields

$$\|(I - P_q)a + P_q v\|_2^2 \le \|a\|_{I-P_q}^2 + \|v\|_{P_q}^2 \le (1 - \Lambda_{\min})\|a\|_2^2 + \|v\|_{P_q}^2.$$

Since $\|g_i^{t,k}\|_2 \le C$ pathwise,

$$\begin{aligned}\mathbb{E}\left[\left\|\frac{1}{K}\sum_k \widehat{g}_i^{t,k}\right\|_{P_q}^2 \middle| \mathcal{F}_t\right] &\le 2\mathbb{E}\left[\left\|\frac{1}{K}\sum_k g_i^{t,k}\right\|_{P_q}^2 \middle| \mathcal{F}_t\right] + 2\mathbb{E}\left\|\frac{1}{K}\sum_k \xi_i^{t,k}\right\|_{P_q}^2 \\ &\le 2C^2 + 2\tau^2 d_q/K.\end{aligned}$$

The first inequality, rather than a false independence assertion, handles the dependence of later gradients on earlier noise. Average over clients, use uniform selection, and note $N^{-1}\sum_i \|\boldsymbol{c}_i^t - \boldsymbol{c}^t\|_2^2 = N^{-1}\sum_i \|\boldsymbol{c}_i^t\|_2^2 - \|\boldsymbol{c}^t\|_2^2$. More explicitly, averaging selected and unchanged controls yields

$$\begin{aligned}\mathcal{E}_{\mathrm{con}}^{t+1} &\le (1-p)\mathcal{E}_{\mathrm{con}}^t + p(1-\Lambda_{\min})\frac{1}{N}\sum_i \mathbb{E}\|\boldsymbol{c}_i^t - \boldsymbol{c}^t\|_2^2 + pV_c \\ &\le (1 - p\Lambda_{\min})\mathcal{E}_{\mathrm{con}}^t + pV_c.\end{aligned} \tag{30}$$

Since $0 < p\Lambda_{\min} \le 1$, iteration of this inequality gives

$$\begin{aligned}\mathcal{E}_{\mathrm{con}}^t &\le (1 - p\Lambda_{\min})^t \mathcal{E}_{\mathrm{con}}^0 + pV_c\sum_{j=0}^{t-1}(1 - p\Lambda_{\min})^j \\ &= (1 - p\Lambda_{\min})^t \mathcal{E}_{\mathrm{con}}^0 + [1 - (1 - p\Lambda_{\min})^t]\frac{V_c}{\Lambda_{\min}} \le \overline{\mathcal{E}}_{\mathrm{con}}.\end{aligned} \tag{31}$$

For $t = 0$ the sum is empty and the claim follows from the definition of $\overline{\mathcal{E}}_{\mathrm{con}}$. □

Remark 2 (Why the reference tracking identity is unavailable)**.** *DP-SCAFFOLD has $P_q = I$, so equation 25 becomes the average of noisy gradients. For LSSCA, $P_q \neq I$ generally leaves a residual $(I - P_q)(\boldsymbol{c}_i^t - \boldsymbol{c}^t)$. Changing the update to $\boldsymbol{c}_i^t - \boldsymbol{c}^t + P_q^{-1}(x^t - \omega_i^{t,K})/(K\eta_{l,t})$ would remove this residual, but would define a different algorithm. The present proof makes no such change.*

Lemma 6 (Local drift)**.** *Let*

$$\begin{aligned}\mathcal{E}_{\mathrm{cli}}^t &= \frac{1}{NK}\sum_{i=1}^N\sum_{k=0}^{K-1}\mathbb{E}\|\omega_i^{t,k} - x^t\|_2^2, \\ a_K &= (K-1)(2K-1)/6, \qquad b_K = (K-1)/2, \\ Q &= \frac{4a_K(C^2 + \overline{\mathcal{E}}_{\mathrm{con}}) + 2b_K\tau^2\widetilde{d}_q}{K^2\eta_g^2}.\end{aligned} \tag{32}$$

Then $\mathcal{E}_{\mathrm{cli}}^t \le Q\bar{\eta}_t^2$, including $\mathcal{E}_{\mathrm{cli}}^t = 0$ when $K = 1$.

Proof. For each $0 \le k < K$,

$$\omega_i^{t,k} - x^t = -\eta_{l,t} P_q \left[\sum_{r<k} g_i^{t,r} + k(\boldsymbol{c}^t - \boldsymbol{c}_i^t) + \sum_{r<k} \xi_i^{t,r} \right].$$

Separate the explicit noise sum using $\|u+v\|^2 \le 2\|u\|^2 + 2\|v\|^2$, and use the same inequality on the first two terms. As $\|P_q\|_2 \le 1$,

$$\mathbb{E}\|\omega_i^{t,k} - x^t\|_2^2 \le 4\eta_{l,t}^2 k^2 (C^2 + \mathbb{E}\|\boldsymbol{c}_i^t - \boldsymbol{c}^t\|_2^2) + 2\eta_{l,t}^2 k \tau^2 \widetilde{d}_q.$$

Average over i and use $\mathcal{E}_{\mathrm{con}}^t \le \overline{\mathcal{E}}_{\mathrm{con}}$. Average over k using $K^{-1}\sum_{k=0}^{K-1} k^2 = a_K$ and $K^{-1}\sum_{k=0}^{K-1} k = b_K$. Finally substitute $\eta_{l,t} = \bar{\eta}_t/(K\eta_g)$. □

Lemma 7 (Server-direction decomposition)**.** *Let $g^t = \nabla f(x^t)$. Define*

$$J_t = \frac{1}{S}\sum_{i\in\boldsymbol{S}_t} (\nabla f_i(x^t) - \boldsymbol{c}_i^t) - \frac{1}{N}\sum_i (\nabla f_i(x^t) - \boldsymbol{c}_i^t),$$

$$r_t = \frac{1}{KS}\sum_{i\in\boldsymbol{S}_t,k} [\nabla f_i(\omega_i^{t,k}) - \nabla f_i(x^t)].$$

Then $v^t = g^t + J_t + r_t + M_t$, with $\mathbb{E}[J_t \mid \mathcal{F}_t] = \mathbb{E}[M_t \mid \mathcal{F}_t] = 0$. Set

$$E_a = \beta^2 Q a^2, \qquad W_2(a) = 8\rho(C^2 + \overline{\mathcal{E}}_{\mathrm{con}}) + 4E_a + \frac{4V_2}{KS}, \quad a \ge 0. \tag{33}$$

For any round,

$$\mathbb{E}\|r_t\|_2^2 \le E_{\bar{\eta}_t}, \tag{34}$$

$$\mathbb{E}\|P_q v^t\|_2^2 \le 4\mathbb{E}\|g^t\|_{P_q}^2 + W_2(\bar{\eta}_t). \tag{35}$$

Proof. The decomposition follows from the control invariant. Uniform client selection gives the zero mean of J_t, and Lemma 3 gives the zero mean of M_t. Jensen's inequality and smoothness give

$$\mathbb{E}\|r_t\|_2^2 \le \frac{\beta^2}{NK}\sum_{i,k}\mathbb{E}\|\omega_i^{t,k} - x^t\|_2^2 \le \beta^2 Q \bar{\eta}_t^2.$$

Uniform selection is valid here through the virtual-trajectory construction. At $k=0$, Assumptions 4–5 imply $\|\nabla f_i(x^t)\|_2 \le \mathbb{E}_z\|\nabla \ell_i(x^t; z)\|_2 \le C$. For $M = P_q^2$, Lemma 4 therefore yields

$$\mathbb{E}\|J_t\|_M^2 \le \frac{\rho}{N}\sum_i \mathbb{E}\|\nabla f_i(x^t) - \boldsymbol{c}_i^t\|_M^2 \le 2\rho(C^2 + \overline{\mathcal{E}}_{\mathrm{con}}).$$

Apply $\|a+b+c+d\|_M^2 \le 4(\|a\|_M^2 + \|b\|_M^2 + \|c\|_M^2 + \|d\|_M^2)$ to the exact decomposition and use $P_q^2 \preceq P_q \preceq I$, Lemma 3, and the residual bound. No independence between r_t and M_t is used. □

NON-CONVEX STATIONARITY GUARANTEE

Theorem 3 (Non-convex convergence under inactive clipping)**.** *Under Assumptions 3–5 and the initialization, fixed-filter, and sampling conditions above, fix $\eta_g \ge 1$ and $0 < \bar{\eta} \le 1/(8\beta)$. Set $\eta_l = \bar{\eta}/(K\eta_g)$ and run $T \ge 1$ rounds. Let $F_0 = f(x^0) - f_{\inf}$ and let R be uniform on $\{0, \dots, T-1\}$, independently of the complete training trajectory. With Q, V_2, and W_2 defined in equation 32, equation 20, and equation 33, respectively,*

$$\begin{aligned}\mathbb{E}\|\nabla f(x^R)\|_{P_q}^2 &= \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(x^t)\|_{P_q}^2 \\ &\le \frac{2F_0}{\bar{\eta}T} + 2\beta^2 Q\bar{\eta}^2 + \beta\bar{\eta}W_2(\bar{\eta}).\end{aligned} \tag{36}$$

Consequently,

$$\mathbb{E}\|\nabla f(x^R)\|_2^2 \le \frac{1}{q}\left[\frac{2F_0}{\bar{\eta}T} + 2\beta^2 Q\bar{\eta}^2 + \beta\bar{\eta}W_2(\bar{\eta})\right]. \tag{37}$$

PROOF OF THEOREM 3

Proof. Write $g^t = \nabla f(x^t)$ and $\Delta^t = x^{t+1} - x^t = -\bar{\eta} P_q v^t$. Smoothness gives

$$\begin{aligned} \mathbb{E}f(x^{t+1}) - \mathbb{E}f(x^t) &\le \mathbb{E}\langle g^t, \Delta^t \rangle + \frac{\beta}{2}\mathbb{E}\|\Delta^t\|_2^2 \\ &\le -B_1^t + \frac{\beta\bar{\eta}^2}{2} B_2^t + B_3^t, \end{aligned} \tag{38}$$

where

$$\begin{aligned} B_1^t &= \bar{\eta}\mathbb{E}\|g^t\|_{P_q}^2, \\ B_2^t &= \mathbb{E}\|P_q v^t\|_2^2, \\ B_3^t &= -\bar{\eta}\mathbb{E}\langle g^t, P_q r_t \rangle. \end{aligned} \tag{39}$$

This follows because the conditional expectations of J_t and M_t vanish, while g^t is measurable with respect to $\mathcal{F}_t$.

Bounding B_2^t. Apply Lemma 7 to obtain

$$\begin{aligned} B_2^t &\le 4\mathbb{E}\|g^t\|_{P_q}^2 + 8\rho(C^2 + \overline{\mathcal{E}}_{\mathrm{con}}) + 4E_{\bar{\eta}} + \frac{4V_2}{KS} \\ &= 4\mathbb{E}\|g^t\|_{P_q}^2 + W_2(\bar{\eta}). \end{aligned} \tag{40}$$

In particular, the explicit DP contribution to V_2 is $\tau^2 \widetilde{d}_q$, and its average is divided by KS.

Bounding B_3^t. Weighted Cauchy–Schwarz and Young's inequality yield

$$\begin{aligned} B_3^t &\le \bar{\eta}\mathbb{E}[\|g^t\|_{P_q}\|r_t\|_{P_q}] \\ &\le \frac{\bar{\eta}}{4}\mathbb{E}\|g^t\|_{P_q}^2 + \bar{\eta}\mathbb{E}\|r_t\|_{P_q}^2 \\ &\le \frac{\bar{\eta}}{4}\mathbb{E}\|g^t\|_{P_q}^2 + \bar{\eta}E_{\bar{\eta}}. \end{aligned} \tag{41}$$

The residual is a difference of gradients of the same client at two points. No bound on $\|\nabla f_i(x) - \nabla f(x)\|$ by local parameter movement is asserted.

One-round descent. Substitute equation 40–equation 41 into equation 38:

$$\begin{aligned} \mathbb{E}f(x^{t+1}) &\le \mathbb{E}f(x^t) - \bar{\eta}\left(\frac{3}{4} - 2\beta\bar{\eta}\right)\mathbb{E}\|g^t\|_{P_q}^2 + \bar{\eta}E_{\bar{\eta}} + \frac{\beta\bar{\eta}^2}{2}W_2(\bar{\eta}) \\ &\le \mathbb{E}f(x^t) - \frac{\bar{\eta}}{2}\mathbb{E}\|g^t\|_{P_q}^2 + \bar{\eta}E_{\bar{\eta}} + \frac{\beta\bar{\eta}^2}{2}W_2(\bar{\eta}). \end{aligned} \tag{42}$$

The second line follows from $2\beta\bar{\eta} \le 1/4$.

Summation and output. Summing equation 42 over $t = 0, \ldots, T-1$ gives

$$\begin{aligned} \frac{\bar{\eta}}{2}\sum_{t=0}^{T-1}\mathbb{E}\|g^t\|_{P_q}^2 &\le f(x^0) - \mathbb{E}f(x^T) + T\bar{\eta}E_{\bar{\eta}} + \frac{\beta T\bar{\eta}^2}{2}W_2(\bar{\eta}) \\ &\le F_0 + T\bar{\eta}E_{\bar{\eta}} + \frac{\beta T\bar{\eta}^2}{2}W_2(\bar{\eta}). \end{aligned} \tag{43}$$

Divide by $\bar{\eta}T/2$ and substitute $E_{\bar{\eta}} = \beta^2 Q\bar{\eta}^2$ to obtain equation 36. For an independent uniform index R,

$$\mathbb{E}\|\nabla f(x^R)\|_{P_q}^2 = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla f(x^t)\|_{P_q}^2.$$

Finally, $\|u\|_{P_q}^2 \ge \Lambda_{\min}\|u\|_2^2$ converts this to the Euclidean stationarity statement. This completes the proof. □

RATE AND COMMUNICATION COMPLEXITY

Corollary 1 (Fixed-noise non-convex rate)**.** *Under the conditions of Theorem 3, fix* $0 < h \le 1/(8\beta)$ *and use*

$$\bar{\eta} = \frac{h}{\sqrt{T}}, \qquad \eta_l = \frac{h}{K\eta_g\sqrt{T}}$$

throughout the T*-round run. Define*

$$L_Q = \beta^2 Q, \qquad W_{20} = 8\rho(C^2 + \overline{\mathcal{E}}_{\mathrm{con}}) + \frac{4(\sigma_l^2 + \tau^2\widetilde{d}_q)}{KS}.$$

Then

$$\mathbb{E}\|\nabla f(x^R)\|_{P_q}^2 \le \frac{2F_0/h + \beta h W_{20}}{\sqrt{T}} + \frac{2L_Q h^2}{T} + \frac{4\beta L_Q h^3}{T^{3/2}}. \tag{44}$$

Let

$$C_{\mathrm{nc}} = 2F_0/h + \beta h W_{20} + 2L_Q h^2 + 4\beta L_Q h^3.$$

For $T \ge 1$*,*

$$\mathbb{E}\|\nabla f(x^R)\|_2^2 \le \frac{C_{\mathrm{nc}}}{q\sqrt{T}}. \tag{45}$$

In particular, for a tolerance $\varepsilon_{\mathrm{opt}} > 0$*,*

$$T \ge \max\left\{1, \left\lceil \frac{C_{\mathrm{nc}}^2}{q^2\varepsilon_{\mathrm{opt}}^2} \right\rceil\right\} \tag{46}$$

suffices for $\mathbb{E}\|\nabla f(x^R)\|_2^2 \le \varepsilon_{\mathrm{opt}}$*. The* $\mathcal{O}(T^{-1/2})$ *rate holds with the noise scale, filter, problem parameters, and* h *fixed, and the assumptions uniform over the considered horizons. The tolerance here is for the expected squared gradient norm, not the gradient norm.*

Proof. Since $E_{\bar{\eta}} = L_Q h^2/T$ and $W_2(\bar{\eta}) = W_{20} + 4L_Q h^2/T$, substitution in equation 36 yields equation 44. For $T \ge 1$, both T^{-1} and $T^{-3/2}$ are at most $T^{-1/2}$. Using $\|u\|_{P_q}^2 \ge q\|u\|_2^2$ gives equation 45; solving this inequality for T proves equation 46. □

SCOPE AND INTERPRETATION

Output covered by the guarantee. The theorem controls a uniformly sampled iterate x^R and the average squared gradient norm over $x^0, \dots, x^{T-1}$. It does not establish the same guarantee for the final iterate x^T. If the implementation returns x^T, this distinction must be maintained. For the analyzed output, one may draw R before training independently of all training randomness and retain x^R, without any additional gradient evaluation.

Inactive clipping. The assumption concerns raw per-example gradients. It is not a consequence of applying clipping. If clipping is active, the conditional bias

$$b_i(x) = \mathbb{E}_z[\mathrm{clip}_C(\nabla\ell_i(x; z))] - \nabla f_i(x)$$

is generally nonzero, and the oracle and descent arguments require additional bias terms. The present theorem does not cover such runs without further analysis.

Noise attenuation and directional progress. The explicit averaged Gaussian variance is $\tau^2\widetilde{d}_q/(KS)$, while noise also enters the control bound through d_q and the local-drift constant Q. Decreasing q attenuates the explicit noise but also reduces the minimum eigenvalue and may increase $\overline{\mathcal{E}}_{\mathrm{con}}$. For any fixed direction s,

$$\|P_q s - s\|_2^2 = (1-q)^2\|(I - UU^\top)s\|_2^2.$$

Thus the theorem retains filtering costs and does not assert automatic improvement over the unfiltered method or automatic alignment of random subspaces with useful gradients.

Fixed noise versus fixed total privacy. These are optimization guarantees at the specified noise scale $\tau = \sigma C/b$. At a fixed total privacy budget, an accountant valid for the actual sampling and observation protocol must first determine the required noise. Any dependence $\tau = \tau(T)$ must be substituted into V_2, V_c, $\overline{\mathcal{E}}_{\mathrm{con}}$, Q, and W_{20}. Consequently, the fixed-noise rate is not by itself a fixed-total-privacy rate. This convergence appendix makes no separate privacy claim.

Differential Privacy

We protect individual records using differential privacy (DP). We use Rényi differential privacy (RDP) for adaptive composition and convert the resulting bound to (ϵ, δ)-DP Mironov (2017). The neighboring relation, sampling scheme, noise normalization, and released transcript are specified explicitly below.

Definition 3 (Record-level add/remove adjacency and DP)**.** *Fix a public roster of N clients and a public integer $n_{\min} \geq b$. The data domain consists of collections $\mathcal{D} = (D_1, \ldots, D_N)$ with $|D_i| \geq n_{\min}$. Two admissible collections $\mathcal{D} \simeq \mathcal{D}'$ are neighbors if one is obtained from the other by adding or removing one record at one client; the client roster and all other records remain unchanged. A randomized mechanism $\mathcal{M}$ is (ϵ, δ)-DP if, for every ordered neighboring pair and every measurable output event O,*

$$\Pr[\mathcal{M}(\mathcal{D}) \in O] \leq e^{\epsilon} \Pr[\mathcal{M}(\mathcal{D}') \in O] + \delta. \tag{47}$$

Here $\epsilon \geq 0$ and $\delta \in (0, 1)$ are fixed public parameters.

Definition 4 (Rényi differential privacy)**.** *For $\lambda > 1$, a mechanism is (λ, ρ)-RDP if*

$$D_\lambda(\mathcal{M}(\mathcal{D}) \| \mathcal{M}(\mathcal{D}')) := \frac{1}{\lambda - 1} \log \int p(y)^\lambda p'(y)^{1-\lambda} \, \mathrm{d}y \leq \rho \tag{48}$$

for every ordered neighboring pair, where p, p' are densities with respect to a common dominating measure. The divergence is infinite when the required absolute continuity does not hold.

Lemma 8 (RDP-to-DP conversion)**.** *If $\mathcal{M}$ is (λ, ρ)-RDP, then it is $(\rho + \log(1/\delta)/(\lambda - 1), \delta)$-DP for every $\delta \in (0, 1)$ Mironov (2017).*

Privacy Accounting

At each round, S clients are sampled uniformly without replacement from the fixed roster, independently of the private data and fresh training randomness. Write $l = S/N$. Each participating client performs K local steps, drawing a fresh uniform size-b subset of its local dataset at each step. The only direct accesses to private records are the clipped gradient queries

$$\widehat{g}_i^{t,k} = \frac{1}{b} \sum_{z \in B_i^{t,k}} \mathrm{clip}_C(\nabla \ell(x_i^{t,k}; z)) + \xi_i^{t,k}, \qquad \xi_i^{t,k} \sim \mathcal{N}(0, \tau^2 I_d), \quad \tau = \frac{\sigma C}{b}, \tag{49}$$

where $C > 0$, $\sigma > 0$, and all perturbations are independent. Thus σ is the multiplier relative to C/b, rather than the coordinate standard deviation or a multiplier relative to $2C/b$.

The initialization, U, $q \in (0, 1]$, stepsizes, C, σ, b, K, S, and T are fixed independently of the private records. Controls are initialized to zero. All subsequent model and control states are computed from the noisy queries and the algorithmic state. Exact local dataset sizes, sampled record identities, raw gradients, and additional nonprivate statistics are not released.

Theorem 4 (A conservative record-level privacy guarantee)**.** *Under the protocol above, define*

$$\bar{s} := \frac{b}{n_{\min} + 1}, \qquad a_\lambda := \frac{2\lambda}{\sigma^2}, \tag{50}$$

$$\rho_{\mathrm{step}}(\lambda) := \frac{1}{\lambda - 1} \log\left(1 - \bar{s} + \bar{s} \exp((\lambda - 1) a_\lambda)\right), \tag{51}$$

$$\rho_{\mathrm{round}}(\lambda) := \frac{1}{\lambda - 1} \log\left(1 - l + l \exp((\lambda - 1) K \rho_{\mathrm{step}}(\lambda))\right). \tag{52}$$

The joint transcript consisting of the sampled client sets, all uploaded model and control increments, and all global model and control states is $(\lambda, T\rho_{\mathrm{round}}(\lambda))$-RDP for every $\lambda > 1$. Consequently, for

any fixed $\delta \in (0,1)$, it is (ϵ, δ)-DP with

$$\epsilon = \inf_{\lambda>1} \left\{ T\rho_{\mathrm{round}}(\lambda) + \frac{\log(1/\delta)}{\lambda - 1} \right\}. \tag{53}$$

In particular, the released global model sequence $(x^0, \ldots, x^T)$ satisfies the same guarantee.

The theorem permits active clipping and does not require the inactive-clipping assumption used in the utility analysis. It accounts for client selection even when selected client identities are visible, but it does not claim an additional amplification gain from hiding identities or aggregating client updates. Equation equation 51 is a conservative coupling bound, not a tight subsampled-Gaussian accountant. A sharper, independently certified one-step RDP bound for exactly the same add/remove relation and fixed-size sampling scheme can replace ρ_{step} in Equations equation 52–equation 53. The fixed filter and the control updates incur no additional privacy cost beyond the adaptively composed noisy queries.

Proof of the Privacy Guarantee

A common-mixture inequality. Let π be a common distribution on an auxiliary variable Z, and let P_z, Q_z be two families of conditional distributions. For the joint laws $P_{Z,Y} = \pi(\mathrm{d}z)P_z(\mathrm{d}y)$ and $Q_{Z,Y} = \pi(\mathrm{d}z)Q_z(\mathrm{d}y)$, direct integration gives

$$\exp\left((\lambda-1)D_\lambda(P_{Z,Y}\|Q_{Z,Y})\right) = \mathbb{E}_{Z\sim\pi}\exp\left((\lambda-1)D_\lambda(P_Z\|Q_Z)\right). \tag{54}$$

Discarding Z cannot increase Rényi divergence. Therefore the right-hand side also upper-bounds the exponentiated divergence between the marginal distributions of Y. The same statement holds in the reverse direction.

Lemma 9 (One local query under fixed-size sampling)**.** *Fix an evaluation point x and a function $h_x(z) = \mathrm{clip}_C(\nabla\ell(x;z))$ with $\|h_x(z)\|_2 \le C$. Let*

$$\mathcal{Q}_x(D) = \frac{1}{b}\sum_{z\in B} h_x(z) + \mathcal{N}(0, \tau^2 I_d), \qquad B \sim \mathrm{Unif}\{A \subseteq D : |A| = b\}. \tag{55}$$

For every add/remove neighboring pair with both sizes at least $n_{\min}$, both ordered Rényi divergences of this query are at most $\rho_{\mathrm{step}}(\lambda)$.

Proof. Treat records as indexed entries, so repeated values cause no ambiguity. Write $D^+ = D \cup \{z_*\}$ and $|D| = n \ge n_{\min} \ge b$. Construct a coupling of the two uniform batches as follows: draw a uniform size-b subset B from D; independently draw $J \sim \mathrm{Bernoulli}(p)$ with $p = b/(n+1)$; and choose an entry V uniformly from B. Set

$$B^+ = \begin{cases} B, & J = 0, \\ (B \setminus \{V\}) \cup \{z_*\}, & J = 1. \end{cases} \tag{56}$$

The set $B \setminus \{V\}$ is uniform among the size-$(b-1)$ subsets of D. A uniform size-b subset of D^+ contains z_* with probability $b/(n+1)$, and conditional on either inclusion event the construction has the required uniform law. Thus B^+ has exactly the desired batch distribution.

For the common auxiliary variable $Z = (B, J, V)$, compare Gaussian conditional outputs with means $\mu_Z = b^{-1}\sum_{z\in B} h_x(z)$ and $\mu_Z^+ = b^{-1}\sum_{z\in B^+} h_x(z)$. Their marginal laws are $\mathcal{Q}_x(D)$ and $\mathcal{Q}_x(D^+)$. If $J = 0$, the means agree. If $J = 1$, then

$$\|\mu_Z - \mu_Z^+\|_2 = \frac{1}{b}\|h_x(V) - h_x(z_*)\|_2 \le \frac{2C}{b}. \tag{57}$$

For equal-covariance Gaussians, in either direction,

$$D_\lambda(\mathcal{N}(\mu, \tau^2 I_d)\|\mathcal{N}(\mu', \tau^2 I_d)) = \frac{\lambda\|\mu - \mu'\|_2^2}{2\tau^2}. \tag{58}$$

Consequently this conditional divergence is zero when $J = 0$ and at most $2\lambda/\sigma^2 = a_\lambda$ when $J = 1$. Equation equation 54 and marginalization give

$$D_\lambda(\mathcal{Q}_x(D)\|\mathcal{Q}_x(D^+)) \le \frac{\log(1 - p + pe^{(\lambda-1)a_\lambda})}{\lambda - 1}. \tag{59}$$

The same proof applies with the two Gaussian families interchanged. Since $p \le b/(n_{\min}+1) = \bar{s}$ and the bound is nondecreasing in p, the claimed uniform bound follows. The auxiliary variable is only a proof device; sampled record identities are not released by the protocol. □

Proof of Theorem 4. Fix an arbitrary ordered neighboring pair differing at client i_*. For analysis, consider the augmented transcript that reveals each selected client set and every noisy query $\widehat{g}_i^{t,k}$, but not the minibatch identities or the Gaussian random seeds. Public data-independent randomness, including the random reference basis, may also be included.

Step 1: conditioning on the augmented history. For a fixed common value of the preceding augmented transcript, all local model states and control states are determined by the same update rules and the same noisy query values under the two datasets. Hence each next evaluation point is identical in the two conditional mechanisms. This does not assert that unconditional training trajectories are independent of private data or earlier noise. It is precisely the common-history conditioning used in adaptive RDP composition. Fresh batch sampling and fresh Gaussian noise ensure that Lemma 9 applies at every query of i_*.

Step 2: local composition conditional on selection. Condition additionally on the selected client set A at round t. If $i_* \notin A$, the conditional round distributions coincide: all queried datasets are unchanged, and the states have already been fixed by the common history. The round divergence is zero. If $i_* \in A$, serialize the conditionally independent client computations in any fixed order. Only the K queries at i_* incur a nonzero conditional RDP cost; queries at the other clients have identical conditional laws. Adaptive composition therefore bounds the round divergence by $K\rho_{\mathrm{step}}(\lambda)$.

Step 3: accounting for visible client selection. Conditional on the past, the selected set has the same uniform distribution under both datasets, and $\Pr(i_* \in A) = l$. Because the selection is included in the augmented transcript, Equation equation 54 yields

$$\exp((\lambda-1)D_\lambda(\mathrm{round}_{\mathcal{D}} \| \mathrm{round}_{\mathcal{D}'})) \le 1 - l + l\exp((\lambda-1)K\rho_{\mathrm{step}}(\lambda)). \tag{60}$$

This is a bound for the joint random selection and round output, not a bound obtained by conditioning on a favorable realized participation count. It holds uniformly over preceding histories.

Step 4: composition over rounds. Adaptive RDP composition over T rounds gives $T\rho_{\mathrm{round}}(\lambda)$ for the augmented transcript. The initialization has no privacy cost because it is independent of the private records.

Step 5: post-processing of models and controls. The local updates

$$x_i^{t,k+1} = x_i^{t,k} - \eta_l P_q(\widehat{g}_i^{t,k} - c_i^t + c^t) \tag{61}$$

and control updates

$$c_i^{t+1} = c_i^t - c^t + \frac{x^t - x_i^{t,K}}{K\eta_l} \tag{62}$$

are deterministic functions of the augmented transcript and public parameters, as are all uploaded increments and server updates. Their joint release is therefore post-processing of that transcript. In particular, releasing both Δx_i^t and Δc_i^t does not require charging the noisy queries twice. Post-processing proves the claimed RDP guarantee for the actual transcript. Lemma 8, followed by optimization over λ, proves Equation equation 53. Releasing only the global models is further post-processing. □

Interpretation and implementation requirements. The following qualifications are part of the result.

- The roster is public and fixed; exact dataset sizes are not released. Publishing exact sizes is incompatible with this add/remove guarantee. The public lower bound $n_{\min}$ specifies the admissible data domain and must not be selected by nonprivate inspection of the datasets. The theorem does not cover preprocessing that repartitions many records when a single record is added or removed.

- The lower bound on dataset sizes is needed to make size-b sampling well-defined under both neighboring datasets. Actual local sizes may differ; the bound uses a common public worst-case value.
- Active clipping is allowed. The bound $2C/b$ arises from the coupled size-b batches differing by one replacement, even though the original datasets are add/remove neighbors. A C/b sensitivity claim for a different sampling convention cannot be substituted here.
- The selection accounting is valid for the joint transcript, including selected client identities. It does not give the same bound conditional on every fixed realized selection schedule.
- A fixed filter introduces no extra privacy cost, but need not improve the privacy parameter. Since P_q is invertible for $q > 0$, applying P_q alone to a complete noisy query preserves its Rényi divergences exactly. Filtering can still affect utility.
- Local trajectories depend on previous Gaussian perturbations. This proof uses adaptive composition and never treats an entire multi-step client update as a deterministic query plus an independent Gaussian perturbation. No $1/\sqrt{S}$ aggregation gain is claimed.
- Any private initialization, private hyperparameter selection, or extra private-data-dependent release needs separate accounting. If a preliminary mechanism has RDP cost $\rho_0(\lambda)$ and the subsequent conditional bounds remain uniform, add $\rho_0(\lambda)$ to $T\rho_{\mathrm{round}}(\lambda)$ before the DP conversion.
- For numerical accounting, minimizing over a finite set of orders $\mathcal{A} \subset (1, \infty)$ gives a valid upper bound. Use stable log-sum-exp evaluations. Report $N, S, b, K, T, C, \sigma, n_{\min}, \delta$, the order grid, and any additional privacy cost. Previously reported target budgets must be checked with an accountant valid for their actual protocol; this theorem alone does not validate those numerical labels.