PRIVACY-AMPLIFYING SKETCHING FOR FEDERATED LEARNING: SECURE AND DECOUPLED DESIGNS

Anonymous authors
Paper under double-blind review

ABSTRACT

Communication efficiency and client-level privacy are two central considerations in federated learning (FL). Privacy-amplifying sketching, as instantiated by the Federated Sketched Gaussian Mechanism (Fed-SGM) of Li et al. (2025), uses random sketching both to compress client updates and to strengthen privacy when the sketch randomness remains hidden from the server. This creates a deployment tension: communication-efficient aggregation requires participating clients to share a common sketch, whereas privacy amplification requires its randomness to remain hidden from the honest-but-curious server.

We develop two complementary designs for resolving this tension. First, we securely realize a shared, server-hidden sketching matrix for Fed-SGM through a threshold-resilient seed-sharing protocol that protects against server–client collusion under an honest-majority assumption. Second, as our main result, we propose Fed-DSSGM, which eliminates the shared-hidden-sketch requirement of Fed-SGM by decoupling privacy amplification from communication compression: each client uses a private sketch for the privacy mechanism, while a separate shared global sketch is used solely for communication efficiency. We establish privacy, security, and convergence guarantees for these designs. Experiments on vision and language benchmarks show that Fed-DSSGM closely matches the utility of Fed-SGM across privacy budgets, while being considerably more robust due to the decoupled design.

1 INTRODUCTION

Federated learning (FL) (McMahan et al., 2017) is a distributed training paradigm in which a central server coordinates a collection of clients that keep their data local. In each round, clients perform optimization steps on their private datasets and transmit only model updates (e.g., gradients or parameter differences) to the server, which aggregates these updates to improve a shared global model. By avoiding raw data transfer, FL can provide stronger privacy protection than fully centralized learning. However, this decentralization does not by itself yield a rigorous privacy guarantee such as differential privacy (DP) (Dwork et al., 2006), and FL has been shown vulnerable to a range of inference and poisoning-style attacks that can leak client information during training (Nasr et al., 2019; Pustozerova and Mayer, 2020; Xie et al., 2019; Zhao et al., 2020; Zhu et al., 2019). In addition, FL often incurs substantial communication cost due to frequent client–server exchanges (Li et al., 2014a). These two bottlenecks—privacy leakage and communication overhead—have driven much of the recent progress in FL.

To reduce communication, a central line of work focuses on lowering the per-round payload by compressing client updates into a lower-dimensional representation (Kairouz et al., 2021; Li et al., 2014b). Popular strategies include sparsification (Lin et al., 2017; Wang et al., 2018b; Barnes et al., 2020), quantization (Alistarh et al., 2017; Liu et al., 2023a; Reisizadeh et al., 2020), and sketching (Song et al., 2023; Rothchild et al., 2020; Ivkin et al., 2019). Among them, sketching is particularly appealing for its simplicity and ease of integration into existing FL pipelines. As an unbiased compressor, it avoids the bias-correction machinery (e.g., error feedback) that is typically required for sparsification, which can also introduce extra memory overhead (Seide et al., 2014; Stich et al., 2018). Moreover, because sketching is linear, it approximately preserves geometric structure under compression (Dasgupta and Gupta, 2003), whereas quantization can distort inner products and potentially slow optimization (Alistarh et al., 2017).

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.

On the privacy side, two DP notions are most commonly studied in FL: sample-level privacy and client-level privacy. Client-level privacy is the stronger requirement in that it seeks indistinguishability when an entire client's dataset changes, rather than only protecting the inclusion or exclusion of individual examples (Zhang et al., 2022). Many DP-FL algorithms (Geyer et al., 2017; Wang et al., 2020; Triastcyn and Faltings, 2019; Truex et al., 2020) target client-level guarantees by adapting canonical centralized DP training techniques—most notably clipping updates and injecting appropriately scaled Gaussian noise (Abadi et al., 2016; Fang et al., 2023)—to the federated setting.

Although privacy and communication efficiency in FL have mostly been studied separately, a recent line of work has begun to address them jointly. In particular, (Li et al., 2025) proposed the *Sketched Gaussian Mechanism* (SGM), which combines Gaussian sketching (Song et al., 2023) with the classical Gaussian mechanism, and its federated instantiation *Fed-SGM*. Using Rényi differential privacy (RDP) (Mironov, 2017) together with standard subsampling, post-processing, and composition tools (Dwork et al., 2014), they establish client-level privacy guarantees for Fed-SGM. Their analysis shows that, with other hyperparameters fixed, the privacy level $\epsilon = \mathcal{O}(1/\sqrt{b\sigma_g^2})$, where b is the sketching dimension and σ_g^2 is the Gaussian noise variance. The result shows that, for sufficiently large b, SGM can yield stronger privacy guarantees than the standard Gaussian mechanism, pointing to an inherent privacy benefit of sketching.

Despite this appealing privacy–communication tradeoff, the original Fed-SGM design face practical deployment challenges. In Fed-SGM (Li et al., 2025), the per-round sketching matrix R_t plays two roles simultaneously: (i) as part of the privacy mechanism, whose randomness should remain hidden from an honest-but-curious server to enable privacy amplification; and (ii) as a shared compression operator, which must be common across clients so that the server can aggregate compressed updates and reconstruct an ambient-space update (e.g., via transpose-based recovery). These two requirements are in tension. Privacy calls for server-hidden sketch randomness, whereas communication-efficient aggregation requires all participating clients to use the same sketch. Realizing a common yet server-hidden sketch matrix in a federated system is challenging (Section 2).

In this paper, we propose two complementary approaches to address this tension. First, we develop cryptographic seed-sharing protocols (Section 3) that allow clients to jointly generate the per round sketch while keeping it hidden from the server. We introduce an honest-majority threshold-resilient protocol that is robust to collusion: it remains secure even if the server colludes with less than half of the participating clients. The core idea is to keep the round seed (and hence the induced sketching matrix) secret-shared throughout the round, and to generate/apply the sketch using verifiable secret sharing (VSS), distributed pseudorandom generation, and honest-majority Multi-Party Computation (MPC), so that only the final aggregated noisy sketch is revealed to the server. Under the honest-majority assumption, the colluding coalition cannot reconstruct the secret-shared seed and thus cannot recover the sketch matrix R_t, while the MPC transcript leaks no additional information beyond the final secure aggregate. Note that equipping Fed-SGM with cryptographic protocols preserves its empirical performance, with mild additional communication.

Second, as our main result, we propose a statistical approach that removes the source of the conflict by *decoupling* the two roles of sketching (Section 4). The key observation is that privacy amplification and communication compression need not rely on the same sketch. We introduce *Fed-DSSGM*, which uses two types of sketches: client-specific, server-hidden sketches for the privacy mechanism; and a separate global sketch (shared across clients and potentially known to the server) solely for communication reduction. Such a decoupled design cleanly separates privacy and compression functionalities, avoiding the implementation bottleneck of a single dual-purpose sketch as in Fed-SGM Li et al. (2025). With an analysis paralleling that of Fed-SGM, we show that Fed-DSSGM attains the same level of client-level differential privacy, and in the same practically relevant operating regime as Fed-SGM, achieves the same-order optimization guarantees. Through experiments on vision and language modeling benchmarks, we demonstrate that the performance of Fed-DSSGM is comparable to that of Fed-SGM, while being considerably more robust due to the decoupled design.

2 Fed-SGM Privacy Degradation

An overview of Fed-SGM (Li et al., 2025) is given below, and the full algorithm is in Algorithm 1. At round t, each participating client $c \in \mathcal{C}_t$, with local dataset D_c, clips its local update to $\hat{\Delta}_{c,t}$ with

threshold τ and applies $\mathcal{SG}\Big(\hat{\Delta}_{c,t}; R_t, \mathbf{z}_{c,t}\Big) = R_t\hat{\Delta}_{c,t} + \mathbf{z}_{c,t}$, where $R_t \in \mathbb{R}^{b\times d}$ has i.i.d. entries from $\mathcal{N}(0, 1/b)$ and $\mathbf{z}_{c,t} \sim \mathcal{N}(0, \sigma_g^2 I_b)$. Li et al. (2025) shows that hiding both R_t and $\mathbf{z}_{c,t}$ from the server yields privacy amplification from sketch randomness, with effective privacy of $\mathcal{O}(1/\sqrt{b})$.

Client Update	**Server Update**
1. Initialize the local model from θ_t and perform local training to obtain the update $\Delta_{c,t}$.	1. Sample a subset of clients $\mathcal{C}_t$ of size N.
2. Clip $\Delta_{c,t}$ with threshold τ to obtain $\hat{\Delta}_{c,t}$.	2. Aggregate the compressed updates as $\bar{\tilde{\Delta}}_t = \frac{1}{N}\sum_{c\in\mathcal{C}_t}\tilde{\Delta}_{c,t}$.
3. Send $\tilde{\Delta}_{c,t} \propto \mathcal{SG}\Big(\hat{\Delta}_{c,t}; R_t, \mathbf{z}_{c,t}\Big) = R_t\hat{\Delta}_{c,t} + \mathbf{z}_{c,t}$ to the server.	3. Broadcast $\bar{\tilde{\Delta}}_t$ to the clients.

Global Step: Each client recovers the aggregated update via $R_t^\top \bar{\tilde{\Delta}}_t$ and updates

$$\theta_{t+1} \leftarrow \texttt{GLOBAL_OPT}\Big(\theta_t, R_t^\top \bar{\tilde{\Delta}}_t, \eta_{\text{global}}\Big).$$

However, if the server learns R_t, then the sketch randomness is no longer hidden and therefore cannot contribute to the client-level privacy analysis. The only remaining source of privacy randomness is the additive Gaussian noise $\mathbf{z}_{c,t}$, so the mechanism reduces to a Gaussian mechanism (GM) up to the fixed linear map R_t. To formalize this degradation in the round-level privacy analysis, we view round t as acting on the collection of sampled client datasets $D_t := \{D_c\}_{c\in\mathcal{C}_t}$ where $|\mathcal{C}_t| = N$, with neighboring round-level dataset D_t' obtained by replacing exactly one client dataset in D_t. We further define $\gamma_t(D_t) := \sum_{c\in\mathcal{C}_t}\hat{\Delta}_{c,t}$, $\delta_t := \gamma_t(D_t) - \gamma_t(D_t')$, and $\mathbf{z}_t := \sum_{c\in\mathcal{C}_t}\mathbf{z}_{c,t}$. Since the client-level noises are independent and satisfy $\mathbf{z}_{c,t} \sim \mathcal{N}(0, \sigma_g^2 I_b)$, we have $\mathbf{z}_t \sim \mathcal{N}(0, N\sigma_g^2 I_b)$. Conditioned on R_t being known, we obtain the following Rényi-divergence bound.

Lemma 2.1 (Conditional Rényi Divergence with Known Sketch)**.** *Conditioned on a fixed and known sketching matrix R_t, the Rényi divergence of order $\alpha > 1$ satisfies*

$$D_\alpha(R_t\gamma_t(D_t) + \mathbf{z}_t \,\|\, R_t\gamma_t(D_t') + \mathbf{z}_t) = \frac{\alpha}{2N\sigma_g^2}\|R_t\delta_t\|_2^2.$$

In practice, the server may learn only *partial* information about R_t, for example, a subset of its rows. Fix a round t and decompose the sketching matrix as $R_t = \begin{bmatrix} R_t^{\mathsf{kn}} \\ R_t^{\mathsf{un}} \end{bmatrix}$, where $R_t^{\mathsf{kn}} \in \mathbb{R}^{b_{\mathsf{kn}}\times d}$ denotes the portion revealed to the server and $R_t^{\mathsf{un}} \in \mathbb{R}^{b_{\mathsf{un}}\times d}$ remains hidden, with $b_{\mathsf{kn}} + b_{\mathsf{un}} = b$. Under the round-level notation introduced above, the aggregate mechanism can be written as

$$R_t\gamma_t(D_t) + \mathbf{z}_t = \begin{bmatrix} R_t^{\mathsf{kn}}\gamma_t(D_t) \\ R_t^{\mathsf{un}}\gamma_t(D_t) \end{bmatrix} + \begin{bmatrix} \mathbf{z}_t^{\mathsf{kn}} \\ \mathbf{z}_t^{\mathsf{un}} \end{bmatrix},$$

where $\mathbf{z}_t^{\mathsf{kn}} \sim \mathcal{N}(0, N\sigma_g^2 I_{b_{\mathsf{kn}}}), \mathbf{z}_t^{\mathsf{un}} \sim \mathcal{N}(0, N\sigma_g^2 I_{b_{\mathsf{un}}}), \mathbf{z}_t^{\mathsf{kn}} \perp \mathbf{z}_t^{\mathsf{un}}$. In this partial-knowledge setting, the Rényi divergence between the mechanism outputs on neighboring round-level datasets admits the following upper bound.

Theorem 2.2 (Hybrid RDP Bound: Known Rows + Hidden Rows)**.** *Let D_t and D_t' be neighboring round-level datasets, and let $\delta_t := \gamma_t(D_t) - \gamma_t(D_t')$. Assume the server knows $R_t^{\mathsf{kn}} \in \mathbb{R}^{b_{\mathsf{kn}}\times d}$ but not $R_t^{\mathsf{un}} \in \mathbb{R}^{b_{\mathsf{un}}\times d}$. Then, for any $\alpha > 1$, the round-t mechanism satisfies the conditional Rényi divergence bound*

$$D_\alpha\big(R_t\gamma_t(D_t) + \mathbf{z}_t \,\big\|\, R_t\gamma_t(D_t') + \mathbf{z}_t\big) \le \underbrace{\frac{\alpha}{2N\sigma_g^2}\left\|R_t^{\mathsf{kn}}\delta_t\right\|_2^2}_{\text{mean-shift term (known sketch)}} + \underbrace{\frac{\alpha^2 b_{\mathsf{un}}\tau^4}{(\alpha-1)b^2\,\sigma_g^4}}_{\text{variance-ratio term (hidden sketch)}}.$$

By directly applying the standard relationship between Rényi divergence and differential privacy (Mironov, 2017), we obtain the following corollary.

Corollary 2.3 ((ε, δ)-DP from the Hybrid RDP Bound)**.** *Under the partial-knowledge scenario, the round-t aggregate mechanism is (ϵ_p, δ_p)-client-level DP for any $\delta \in (0, 1)$ with*

$$\epsilon_p \le \inf_{\alpha>1}\left\{\frac{\alpha}{2N\sigma_g^2}\left\|R_t^{\mathsf{kn}}\delta_t\right\|_2^2 + \frac{\alpha^2 b_{\mathsf{un}}\tau^4}{(\alpha-1)b^2\,\sigma_g^4} + \frac{\log(1/\delta_p)}{\alpha-1}\right\}.$$

Remark 2.4. Theorem 2.2 and Corollary 2.3 interpolate smoothly between two extremes, where b_{kn} denotes the number of sketch rows known to the server and b_{un} the number of rows remaining hidden:

- If $b_{\mathsf{kn}} = 0$ (the server learns no rows), the setting reduces to the fully hidden-sketch regime of Li et al. (2025). In this case, the "known-rows" term in Theorem 2.2 vanishes, and the remaining term yields the sketch-amplified $O(1/\sqrt{b})$ privacy scaling after optimizing the (ε, δ) conversion as in (Li et al., 2025).
- If $b_{\mathsf{un}} = 0$ (the server learns all rows), then the entire sketching matrix is known to the server and the mechanism reduces to the standard Gaussian mechanism. Correspondingly, the "unknown-rows" term disappears, and we recover the classical Gaussian mean-shift RDP bound with sensitivity $\|R_t \delta_t\|_2$ as stated in Lemma 2.1.

Thus, revealing additional rows increases privacy loss, but the degradation is smooth: the bound transitions continuously between the sketch-amplified regime and the classical GM as b_{un} decreases.

3 Cryptographic Method: Secure Seed Sharing

In this section, we introduce a cryptographic approach that addresses the challenges in Fed-SGM. In the original Fed-SGM algorithm (Li et al., 2025) restated as Algorithm 1, the per-round sketching matrix R_t serves two distinct purposes. First, it is used within SGM and its randomness is leveraged to provide client-level privacy against an honest-but-curious server. Second, it reduces the dimensionality of each client update, thereby lowering the communication cost between clients and the server.

These two roles impose conflicting requirements on R_t. For privacy amplification, since client-level privacy is defined with respect to the server, the sketch randomness must remain hidden from the server. In contrast, for communication reduction, clients must employ a common sketch so that the server can aggregate sketches and clients can map the aggregated sketch back to the ambient space (e.g., via a transpose-based reconstruction) for the global model update.

Consequently, R_t must be shared across clients while remaining unknown to the server. However, under the standard federated learning communication model—where clients do not communicate directly and any information sharing must be relayed by the server—distributing a shared R_t in the clear through the server would reveal it to the server, thereby undermining the server-side privacy amplification that relies on the sketch randomness.

To realize such a shared yet server-hidden sketch under server-mediated communication, while remaining robust to server–client collusion, we use an honest-majority, threshold-resilient protocol. The protocol remains secure against an adversary controlling the server and up to κ participating clients, for any $\kappa < |\mathcal{C}_t|/2$. Its key idea is to keep the round seed, and hence the induced sketching matrix, in secret-shared form throughout the round, and to generate and apply the sketch using verifiable secret sharing and honest-majority secure computation so that only the final aggregate is revealed.

Protocol 3.1. *Assumption:* An honest majority of participating clients, $\kappa < |\mathcal{C}_t|/2$.

1. **Secret-shared round seed.** Each client $i \in \mathcal{C}_t$ samples $s_{i,t} \leftarrow \{0,1\}^\lambda$, where λ is the computational security parameter, and distributes it using $(\kappa{+}1)$-out-of-$|\mathcal{C}_t|$ VSS. Each client j locally computes $\langle \mathsf{seed}_t \rangle_j = \sum_{i \in \mathcal{C}_t} \langle s_{i,t} \rangle_j$, without reconstructing seed_t.
2. **Secret-shared sketch generation.** Clients jointly evaluate dPRG on $\{\langle \mathsf{seed}_t \rangle_j\}$ to obtain $\{\langle w_{t,\ell} \rangle_j\}_{\ell=1}^b$ and define $(R_t)_{\ell,:} = \mathsf{Gauss}(w_{t,\ell})$, with R_t remaining in secret-shared form.
3. **Secure local sketching and noise addition.** For each client j, honest-majority MPC computes $\langle \hat{\Delta}_{j,t}^{\mathrm{sk}} \rangle = \langle R_t \rangle \hat{\Delta}_{j,t}$ from its private clipped update $\hat{\Delta}_{j,t}$. Gaussian noise is added locally or in shares to obtain $\langle \tilde{\Delta}_{j,t} \rangle$.
4. **Secure aggregation.** Clients execute SecAgg so that the server learns only $\bar{\tilde{\Delta}}_t = \frac{1}{|\mathcal{C}_t|} \sum_{j \in \mathcal{C}_t} \tilde{\Delta}_{j,t}$.

The resulting protocol is summarized in Protocol 3.1. The construction combines several standard cryptographic primitives, each serving a distinct role. A $(\kappa{+}1)$-out-of-$|\mathcal{C}_t|$ perfectly hiding verifiable secret sharing (VSS) scheme based on threshold secret sharing (Shamir, 1979; Pedersen, 1991) allows the clients to jointly hold the aggregate round seed without reconstructing it in the clear. A distributed PRG primitive dPRG (Naor et al., 1999) expands this shared seed into b secret-shared pseudorandom states, where $w_{t,\ell}$ determines the ℓ-th sketch row through $(R_t)_{\ell,:} = \mathsf{Gauss}(w_{t,\ell}) \in \mathbb{R}^d$. Honest-majority MPC (Ben-Or et al., 2019) then enables the shared matrix R_t to be applied to each client's private clipped update without revealing either the sketch or the honest clients' updates to the

adversary. Finally, secure aggregation SecAgg (Bonawitz et al., 2017b) reveals only the averaged noisy sketch $\bar{\tilde{\Delta}}_t$ to the server.

Remark 3.1. Protocol 3.1 replaces the standard client-to-server transmission of b-dimensional sketches with an interactive secure-computation procedure that generates and applies the shared sketch in secret-shared form and reveals only the aggregate $\bar{\tilde{\Delta}}_t$. In particular, the additional interaction arises from four components: (i) VSS-based seed sharing, (ii) distributed PRG expansion, (iii) honest-majority MPC to form $\langle \hat{\Delta}_{j,t}^{\mathrm{sk}} \rangle = \langle R_t \rangle \hat{\Delta}_{j,t}$, and (iv) secure aggregation. We denote the total interaction/communication cost of this pipeline by $\mathrm{Comm}_{\mathrm{SMPC}}(N, b)$, where $N := |\mathcal{C}_t|$, suppressing dependence on fixed security parameters. Typically, $\mathrm{Comm}_{\mathrm{SMPC}}(N, b)$ is non-negligible: its dominant b-dependent component scales with the sketch dimension b (since the protocol must generate and apply the b sketch rows in shared form) and usually requires multiple rounds of interaction among the N participants for share-consistent PRG evaluation and MPC multiplications, in addition to the messaging required by secure aggregation. As a result, Protocol 3.1 typically incurs higher within-round communication than Fed-SGM, while retaining the same dominant $\Theta(Nb)$ scaling in the sketch dimension: Fed-SGM requires each of the N selected clients to send a single b-dimensional sketch to the server, whereas the secure protocol introduces additional communication for distributed PRG evaluation, MPC, and secure aggregation. The increased communication is the price paid for stronger robustness to collusion, tolerating up to $\kappa < N/2$ corrupted participants.

Next, we establish the correctness and threshold-privacy guarantees of Protocol 3.1 under the honest-majority semi-honest threat model. The formal statements and proofs are deferred to Appendix D.3.

Theorem 3.2 (Correctness and Threshold Privacy, Informal version of Theorem D.12 and D.13)**.** *Under standard correctness and security assumptions on the underlying distributed primitives, stated formally in Appendix D.3, the protocol guarantees the following:*

1. *An honest execution deterministically produces the exact noisy aggregate sketched update $\bar{\tilde{\Delta}}_t$ at the server.*
2. *The protocol securely computes this aggregate. In particular, the adversary's entire view during the execution can be computationally simulated using only the final aggregate and the inputs of the corrupted clients.*

Remark 3.3. Intuitively, these guarantees follow from the structural properties of the underlying primitives. The correctness guarantee follows from the linearity of the secret shares together with the correctness of the dPRG, honest-majority MPC, and SecAgg primitives: the intermediate values remain valid shares throughout the computation, and the final SecAgg reconstructs the exact sum. The privacy guarantee holds because the round seed seed_t is protected by a $(\kappa+1)$-out-of-$|\mathcal{C}_t|$ VSS scheme. An adversary controlling the server and at most κ clients does not meet the reconstruction threshold and therefore cannot reconstruct the round seed. Furthermore, the security of the distributed PRG, honest-majority MPC, and SecAgg ensures that the adversary's view of the remaining execution is computationally simulatable from the final aggregate and the inputs of the corrupted clients, and hence reveals no additional information about the hidden sketch R_t or the honest clients' seeds and updates.

4 Statistical Method: Decoupled Sketching

In this section, we introduce a new statistical approach that avoids the challenges in Fed-SGM without using any cryptographic protocols. As discussed in Section 3, requiring R_t to be shared across clients while remaining hidden from the server may necessitate additional machinery, such as cryptographic protocols; moreover, even with such protocols, the system may remain vulnerable to information leakage under adversarial behaviors, e.g., collusion. At a conceptual level, this requirement arises from coupling two roles—privacy and compression—within a single sketching matrix.

A natural remedy is to decouple the two functionalities by introducing two separate sketching matrices. We use a client-specific, server-hidden sketching matrix for the privacy mechanism and a global sketching matrix, shared across clients and potentially known to the server, solely for communication compression. With this separation, the privacy-amplification and communication-reduction components no longer interfere. This design yields Fed-DSSGM, a variant of Fed-SGM that employs *Decoupled Sketching* (DS).

Fed-DSSGM's high-level client-server procedure is summarized below, and the full algorithm is given in Algorithm 2. Relative to Fed-SGM, the key modification is to decouple the privacy and

communication roles of sketching: each client first applies SGM using a client-specific sketching matrix $R_{c,t} \in \mathbb{R}^{b_{\text{in}} \times d}$, lifts the resulting private sketched update back via $R_{c,t}^\top$, and then applies a global sketching matrix $S_t \in \mathbb{R}^{b_{\text{out}} \times d}$ solely for communication compression. The server aggregates these compressed client updates and broadcasts the aggregated compressed update, which is mapped back via $S_t^\top$ for the global model update.

Client Update

1. Initialize the local model from θ_t and perform local training to obtain the update $\Delta_{c,t} \in \mathbb{R}^d$.
2. Clip the scaled update with threshold τ to obtain $\hat{\Delta}_{c,t}$.
3. Compute the private sketched update $\breve{\Delta}_{c,t} = \mathcal{SG}\left(\hat{\Delta}_{c,t}; R_{c,t}, \mathbf{z}_{c,t}\right) = R_{c,t}\hat{\Delta}_{c,t} + \mathbf{z}_{c,t}$, using the client-specific sketching matrix $R_{c,t}$.
4. Lift the private sketched update back to the ambient dimension as $\breve{\Delta}_{c,t} = R_{c,t}^\top \breve{\Delta}_{c,t}$.
5. Apply the global sketching matrix S_t to form the compressed update $\tilde{\Delta}_{c,t}$ and send $\tilde{\Delta}_{c,t}$ to the server.

Server Update

1. Sample a subset of clients $\mathcal{C}_t$ of size N.
2. Aggregate the compressed updates as $\bar{\tilde{\Delta}}_t = \frac{1}{N}\sum_{c \in \mathcal{C}_t} \tilde{\Delta}_{c,t}$.
3. Broadcast $\bar{\tilde{\Delta}}_t$ to the clients.

Global Step: Each client recovers the aggregated update via $S_t^\top \bar{\tilde{\Delta}}_t$ and updates

$$\theta_{t+1} \leftarrow \texttt{GLOBAL_OPT}\left(\theta_t, S_t^\top \bar{\tilde{\Delta}}_t, \eta_{\text{global}}\right).$$

Remark 4.1 (Collusion robustness of Fed-DSSGM). In Fed-DSSGM, the randomness that underlies the privacy mechanism in SGM—namely, the sketching matrix $R_{c,t}$ and the Gaussian noise $\mathbf{z}_{c,t}$—is client-specific rather than shared across clients. As a result, these quantities need not be jointly established or revealed among clients. Therefore, even if the server colludes with a subset of clients, the server still cannot recover the private sketching randomness of any non-colluding client. Consequently, the privacy of those non-colluding clients is not affected by such collusion.

4.1 Privacy Guarantee

Adopting a similar privacy analysis as in Li et al. (2025), we establish the following client-level privacy guarantee for Algorithm 2.

Theorem 4.2. *There exist constants c_3 and c_4 so that given the sampling probability $q = \frac{N}{C}$ and the number of communication rounds T, for any $\epsilon_p \le c_3 q\sqrt{T}$, Algorithm 2 is (ϵ_p, δ_p)-client-level private for any $\delta_p > 0$ if we choose*

$$\sigma_g^2 \ge \frac{c_4 q \tau^2 \sqrt{T}\log(2qT/\delta_p)}{\sqrt{b_{in}}\epsilon_p}.$$

Remark 4.3. Although our procedure differs from the original Fed-SGM in that each client uses a client-specific sketching matrix for privacy and we additionally apply a global sketching step, the resulting client-level privacy guarantee is unchanged. This follows from the parallel composition and post-processing properties of differential privacy: the client-specific mechanisms compose in parallel across clients, and the additional global sketching is a post-processing of differentially private outputs, which does not affect privacy.

4.2 Convergence Analysis

In this section, we establish an optimization guarantee for Algorithm 2 with AMSGrad as the global optimizer (corresponding to `GLOBAL_OPT` in the algorithm). While the privacy analysis remains largely unchanged under the decoupled-sketching modification, the optimization analysis must be revisited. The key difference from Fed-SGM is that each local update in Fed-DSSGM undergoes two successive sketching–de-sketching transformations: $R_{c,t}^\top R_{c,t}$ followed by $S_t^\top S_t$ before contributing to the global update, rather than only one: $R_t^\top R_t$ for Fed-SGM. The structural change necessitates a new sketching analysis for the resulting two-step transformation (Appendix F).

Compared to concentration bounds for a single sketching operation (e.g., Lemma D.24 in (Song et al., 2023)), the two-stage sketching–de-sketching structure introduces an additional dependence on the ambient dimension d. We further show in Appendix F that this extra $\sqrt{d}$-order dependence is unavoidable up to logarithmic factors. Building on this guarantee, we obtain Theorem 4.4, which gives the convergence result for Algorithm 2 with AMSGrad. We defer the technical assumptions to Appendix G and the full formal statement and proof to Appendix I.

Theorem 4.4 (Informal version of Theorem I.1)**.** *Suppose $\{\theta_t\}_{t=0}^{T}$ is generated by Algorithm 2 with AMSGrad as* GLOBAL_OPT*. Denote $\mathcal{L}^*$ the minimum of the average empirical loss. Under standard assumptions stated in Appendix G, with learning rate $\eta = \eta_{global}\eta_{local}$, we have that with probability at least $1 - \Theta(\delta)$,*

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla \mathcal{L}(\theta_t)\|_2^2 \leq E_s^{AMSGrad} + E_c^{AMSGrad} + E_g^{AMSGrad}$$

in which $E_s^{AMSGrad}$, $E_c^{AMSGrad}$ and $E_g^{AMSGrad}$ denote the terms arising from the sketching-based FedAvg procedure, clipping, and Gaussian noise, respectively. Specifically, with $\tau_{K,G} = \min\{\tau, KG\}$,

$$E_s^{AMSGrad} = \mathcal{O}\left(\frac{\mathcal{L}(\theta_0) - \mathcal{L}^*}{\eta TK}\right) + \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{NT}}\right) + \tilde{\mathcal{O}}\left(\eta_{local} K\right) + \tilde{\mathcal{O}}\left(\frac{\sqrt{d}\,\tau_{K,G}}{\sqrt{b_{in} b_{out}}\sqrt{TK}}\right) + \tilde{\mathcal{O}}\left(\frac{(\eta_{local} + \eta\mathcal{I})\,\alpha_1^2 \tau_{K,G}^2}{K}\right) + \tilde{\mathcal{O}}\left(\frac{1}{\eta TK}\right),$$

$$E_c^{AMSGrad} = \max\left\{0, \frac{\left(\epsilon + \tilde{\mathcal{O}}\left(\eta_{local}\right)\right) G(KG - \tau)}{K\epsilon}\right\},$$

$$E_g^{AMSGrad} = \tilde{\mathcal{O}}\left(\frac{(\eta_{local} + \eta\mathcal{I})\,\alpha_2^2 \sigma_g^2}{NK}\right).$$

where $\alpha_1 = 1 + \tilde{\mathcal{O}}\left(\frac{\sqrt{d}}{\sqrt{b_{in} b_{out}}}\right)$, $\alpha_2 = 1 + \tilde{\mathcal{O}}\left(\sqrt{\frac{d}{b_{out}}}\right)$, and $\mathcal{I}$ denotes the absolute intrinsic dimension parameter of loss Hessian defined in Assumption G.4.

Remark 4.5. Relative to the optimization guarantee for the original Fed-SGM, our bound introduces an additional factor α_1 in E_s^{AMSGrad} and α_2 in E_g^{AMSGrad}. Both factors incur an extra dependence on the ambient dimension d through the new sketching guarantee proved in Appendix F. However, in the practically relevant regime where both sketching dimensions b_{in} and b_{out} are constant fractions of d, we have $\alpha_2 = \tilde{\mathcal{O}}(1)$ and $\alpha_1 = 1 + \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{d}}\right) \leq \mathcal{O}(1)$. Therefore, these additional factors do not change the order of the convergence guarantee, and the optimization result for Fed-DSSGM matches that of Fed-SGM up to logarithmic factors in this regime.

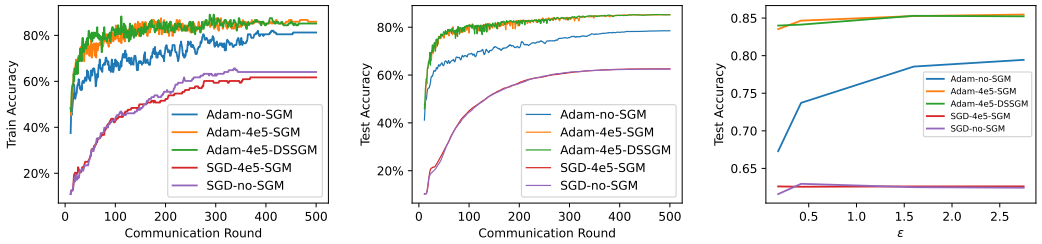
Remark 4.6. Allowing b_{in} and b_{out} to differ provides an additional degree of freedom beyond the balanced choice above. For a fixed product $b_{\text{in}}b_{\text{out}}$, the leading sketching distortion controlled by α_1 remains unchanged. Shifting more of the sketch dimension to b_{in} and less to b_{out} then strengthens privacy and reduces communication, since the required noise variance decreases with b_{in} according to Theorem 4.2 while the transmitted dimension is b_{out}, at the cost of a larger α_2 and hence larger Gaussian-noise-related optimization error. Moreover, for dense sketches, the client-side sketching and desketching cost scales with $b_{\text{in}} + b_{\text{out}}$, which is minimized by the balanced choice $b_{\text{in}} = b_{\text{out}}$ for a fixed product. Thus, the two sketch dimensions allow Fed-DSSGM to trade privacy and communication against computation and optimization error rather than tying all of them to a single parameter.

5 Experimental Results

We conduct empirical evaluations of Fed-DSSGM on model training for both vision and language benchmarks, assessing its predictive performance and communication efficiency, while also discussing its computational overhead. Across various privacy budgets ϵ_p, we compare Fed-DSSGM with Fed-SGM as well as with the corresponding unsketched baselines—DP-FedAvg (Zhang et al., 2022) for the GD variant and a matched unsketched Adam implementation for the Adam variant. Since the results in (Li et al., 2025) show that Adam-based variants consistently outperform their GD-based counterparts, we focus our Fed-DSSGM experiments on the Adam setting. We first describe the experimental setup, and then evaluate predictive performance and communication cost, followed by a brief discussion of computational overhead.

5.1 Experimental Setup

Datasets and Model Architectures. We consider two experimental settings covering both vision and language tasks. For vision, we use the full EMNIST ByClass dataset, which contains 814K training samples and 140K test samples across 62 classes corresponding to handwritten characters. We train a ResNet101 model (Wu and He, 2018) with approximately 42M parameters. For language, we use



(a) Training Accuracy, $\epsilon_p = 1.60$ (b) Test Accuracy, $\epsilon_p = 1.60$ (c) Test accuracy across privacy levels

Figure 1: (a)(b) Comparison of Fed-DSSGM with Adam, Fed-SGM with Adam, Fed-SGM with GD, DP-FedAvg, and the unsketched Adam baseline on ResNet101 trained on EMNIST with $\epsilon_p = 1.6$. The x-axis shows the number of communication rounds T, and the y-axis shows the train/test accuracy. (c) Test accuracy across privacy levels. The x-axis shows ϵ_p, and the y-axis shows the test accuracy. In the legends, 'sgd' and 'adam' denote the choice of GD and Adam as the global optimizers, respectively; '4e5' indicates sketch dimensions $b_{\text{in}} = b_{\text{out}} = 4 \times 10^5$, and 'no' indicates that no sketching is applied.

SST-2 from the GLUE benchmark (Wang et al., 2018a), which contains 67,349 training samples and 1,821 test samples over two sentiment classes. We fine-tune a BERT-Base model (Devlin et al., 2019) with approximately 100M parameters.

Parameter Settings. We use a total of $C = 625$ clients and sample $N = 4$ clients uniformly at random in each communication round. Each selected client performs $K = 18$ local SGD steps with mini-batch size 64, and we set the clipping threshold to $\tau = 1$. The sketch dimension and number of communication rounds are task-dependent. For the vision task, we use $b_{\text{in}} = b_{\text{out}} = 4 \times 10^5$, corresponding to approximately a 1% compression rate, and run $T = 500$ rounds. For the language task, we set $b_{\text{in}} = b_{\text{out}} = 2 \times 10^5$, corresponding to approximately a 0.2% compression rate, and run $T = 200$ rounds. For each experimental setting, we repeat the experiment five times and report the average performance across runs.

Privacy Levels and Noise Calibration. For all experiments, we fix the privacy parameter $\delta_p = 10^{-5}$. For the unsketched baselines, we consider noise scales $\sigma_g \in \{0.8, 1, 2, 4\}$ for both tasks. Using the Moments Accountant method (Abadi et al., 2016; Bu et al., 2020), these correspond to cumulative privacy losses ϵ_p of approximately $\{2.75, 1.60, 0.42, 0.18\}$ on the vision task and $\{2.45, 1.44, 0.35, 0.12\}$ on the language task. For both Fed-DSSGM and Fed-SGM, we match these privacy levels by computing the minimum noise scale σ_g through the RDP-based analysis and numerical minimization over the R'enyi order α (Mironov, 2017). Under this calibration, the resulting noise levels are approximately $\{0.0883, 0.1013, 0.1588, 0.2265\}$ for the vision task and $\{0.0948, 0.1071, 0.1664, 0.2580\}$ for the language task. Hence, across all privacy budgets considered here, Fed-DSSGM and Fed-SGM achieve the same privacy guarantees as the corresponding unsketched methods while requiring substantially smaller Gaussian noise levels.

5.2 Model Performance

We present the vision-task results in Figure 1; due to space constraints, the corresponding language-task results are deferred to Appendix J. Additional figures are also provided in Appendix J. For the vision task, Figures 1a and 1b show the training and test accuracy, respectively, at a fixed privacy budget, while Figure 1c shows the trend of test accuracy across privacy levels. The main takeaway is that Fed-DSSGM with Adam performs nearly identically to Fed-SGM with Adam across communication rounds and privacy levels, and both achieve the strongest performance among the variants considered here. This shows that the additional sketching step introduced by decoupled sketching does not lead to noticeable extra performance degradation in practice. Meanwhile, both sketched Adam variants consistently outperform the corresponding unsketched Adam baseline under the same privacy budget.

5.3 Communication Cost Analysis

We next compare the communication overhead of the two practical solutions proposed in this work, with plaintext Fed-SGM serving as the baseline.

For the EMNIST configuration above ($N = 4$ and $b_{\text{in}} = b_{\text{out}} = b = 400{,}000$) and assuming 32-bit transmitted values, plaintext Fed-SGM transmits one b-dimensional sketch per client, resulting in a total per-round payload of $Nb \times 4$ bytes, or 6.40 MB. Fed-DSSGM instead transmits the b_{out}-dimensional outer sketch. Therefore, under the balanced setting $b_{\text{out}} = b$ used in our experiments, Fed-DSSGM incurs the same communication payload as plaintext Fed-SGM.

The cryptographic solution realizes Fed-SGM while keeping its shared privacy-amplifying sketch hidden from the server. The threshold-resilient construction in Protocol 3.1 is instantiated using standard honest-majority MPC based on Beaver triples and secure secret sharing (Damgard and Nielsen, 2007), with Shamir secret sharing (Shamir, 1979), the Naor–Pinkas–Reingold distributed PRF (Naor et al., 1999), and secure aggregation (Bonawitz et al., 2017b). For the parameter regime considered here, the seed-sharing cost is lower order relative to the sketch-dependent communication, whose dominant components consist of approximately $2b$ field elements per client for distributed pseudorandom generation, $2b$ for secure multiplication, and b for secure aggregation. This yields approximately $5Nb$ transmitted 32-bit field elements, or $20Nb$ bytes per round, which is approximately $5\times$ the communication volume of plaintext Fed-SGM. A complete step-by-step accounting is provided in Appendix J.2.

Table 1: Per-round communication volume under the EMNIST configuration with $N = 4$ and $b = b_{\text{in}} = b_{\text{out}} = 400{,}000$. Plaintext Fed-SGM serves as the baseline. The threshold-resilient MPC entry reports the dominant sketch-dependent communication and is therefore approximate.

Method	**Per-round Communication**	**Relative to Fed-SGM**
Fed-SGM (plaintext baseline)	6.40 MB	$1\times$
Fed-DSSGM	6.40 MB	$1\times$
Fed-SGM + threshold-resilient MPC	≈ 32.00 MB	$\approx 5\times$

The comparison highlights two different ways of making privacy-amplifying sketching practical. Fed-DSSGM avoids the shared-secret-sketch requirement altogether and, under the balanced setting used in our experiments, preserves the communication cost of plaintext Fed-SGM. The cryptographic approach instead retains the original Fed-SGM mechanism while protecting its shared sketch randomness and tolerating server–client collusion under our honest-majority assumption, at an approximately $4\times$ communication cost. Importantly, the threshold-resilient construction retains the same dominant $\Theta(Nb)$ scaling in the sketch dimension as plaintext Fed-SGM. Thus, the two proposed solutions offer different trade-offs between local computation and interactive cryptographic communication.

Remark 5.1. While Fed-DSSGM preserves the communication cost of Fed-SGM under the balanced setting, it incurs additional local computation. In particular, relative to Fed-SGM, Fed-DSSGM introduces one additional sketching/de-sketching stage. For dense sketches, this corresponds to two additional matrix-vector transformations involving a $b \times d$ sketching matrix, and hence an additional $\mathcal{O}(bd)$ computation per participating client and round. This overhead can be substantial when both the model dimension d and the sketch dimension b are large, and represents the main computational trade-off of the decoupled-sketching approach.

6 Conclusion

In conclusion, we studied the practical deployment challenge in Fed-SGM arising from the dual role of the per-round sketching matrix: it must remain hidden from the server to enable privacy amplification, yet also be shared across clients to support communication-efficient aggregation. To address this tension, we proposed two complementary solutions. First, we developed a threshold-resilient cryptographic protocol that allows clients to jointly generate the round sketching matrix while keeping it hidden from the server. The protocol is robust to server–client collusion under an honest-majority assumption. Second, we introduced Fed-DSSGM, which resolves the conflict statistically by decoupling the privacy and communication roles into separate sketching matrices. We showed that Fed-DSSGM preserves the same client-level privacy guarantee as Fed-SGM and attains the same-order optimization guarantee in the practically relevant regime where the sketching dimension is a constant fraction of the ambient dimension. Empirical results on vision and language benchmarks further show that Fed-DSSGM performs comparably to Fed-SGM, indicating that the additional decoupling step does not introduce noticeable extra degradation in practice. As a direction for future work, it remains to reduce the systems overhead of the cryptographic construction and to extend the privacy and optimization analysis beyond isotropic Gaussian sketching matrices.

REFERENCES

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

Michel Abdalla, Mihir Bellare, and Phillip Rogaway. The oracle diffie-hellman assumptions and an analysis of dhies. In *Topics in Cryptology—CT-RSA 2001: The Cryptographers' Track at RSA Conference 2001*, pages 143–158. Springer, 2001.

Kook Jin Ahn, Sudipto Guha, and Andrew McGregor. Graph sketches: sparsification, spanners, and subgraphs. In *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI symposium on Principles of Database Systems*, pages 5–14, 2012.

Alham Fikri Aji and Kenneth Heafield. Sparse communication for distributed gradient descent. *arXiv preprint arXiv:1704.05021*, 2017.

Daniel Alabi, Sainyam Galhotra, Shagufta Mehnaz, Zeyu Song, and Eugene Wu. Privacy and security in distributed data markets. In *Companion of the 2025 International Conference on Management of Data*, SIGMOD/PODS '25, 2025.

Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in neural information processing systems*, 30, 2017.

Leighton Pate Barnes, Huseyin A Inan, Berivan Isik, and Ayfer Özgür. rtop-k: A statistical estimation approach to distributed sgd. *IEEE Journal on Selected Areas in Information Theory*, 1(3):897–907, 2020.

Gilad Baruch, Moran Baruch, and Yoav Goldberg. A little is enough: Circumventing defenses for distributed learning. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8632–8642, 2019.

Raef Bassily, Kobbi Nissim, Uri Stemmer, and Abhradeep Guha Thakurta. Practical locally private heavy hitters. *Advances in Neural Information Processing Systems*, 30, 2017.

James Bell, Adrià Gascón, Badih Ghazi, Ravi Kumar, Pasin Manurangsi, Mariana Raykova, and Phillipp Schoppmann. Distributed, private, sparse histograms in the two-server model. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022*, pages 307–321. ACM, 2022. URL https://doi.org/10.1145/3548606.3559383.

James Henry Bell, Kallista A. Bonawitz, Adrià Gascón, Tancrède Lepoint, and Mariana Raykova. Secure single-server aggregation with (poly)logarithmic overhead. In *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020*, pages 1253–1269. ACM, 2020.

Michael Ben-Or, Shafi Goldwasser, and Avi Wigderson. Completeness theorems for non-cryptographic fault-tolerant distributed computation. In *Providing sound foundations for cryptography: on the work of Shafi Goldwasser and Silvio Micali*, pages 351–371. 2019.

Daniel J Bernstein. Curve25519: new diffie-hellman speed records. In *Public Key Cryptography - PKC 2006: 9th International Conference on Theory and Practice of Public-Key Cryptography*, pages 207–228. Springer, 2006.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, page 1175–1191, New York, NY, USA, 2017a. Association for Computing Machinery. ISBN 9781450349468. doi: 10.1145/3133956.3133982. URL https://doi.org/10.1145/3133956.3133982.

Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 1175–1191, 2017b.

Dan Boneh, Elette Boyle, Henry Corrigan-Gibbs, Niv Gilboa, and Yuval Ishai. Lightweight techniques for private heavy hitters. In *42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021*, pages 762–776. IEEE, 2021.

Dan Boneh, Elette Boyle, Henry Corrigan-Gibbs, Niv Gilboa, and Yuval Ishai. Arithmetic sketching. In *Advances in Cryptology - CRYPTO 2023 - 43rd Annual International Cryptology Conference, CRYPTO 2023, Santa Barbara, CA, USA, August 20-24, 2023, Proceedings, Part I*, volume 14081 of *Lecture Notes in Computer Science*, pages 171–202. Springer, 2023.

Zhiqi Bu, Jinshuo Dong, Qi Long, and Weijie J Su. Deep learning with gaussian differential privacy. *Harvard data science review*, 2020(23), 2020.

Wei-Ning Chen, Christopher A Choquette Choo, Peter Kairouz, and Ananda Theertha Suresh. The fundamental price of secure aggregation in differentially private federated learning. In *International Conference on Machine Learning*, pages 3056–3089. PMLR, 2022.

Graham Cormode and Shan Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*, 55(1):58–75, 2005.

Henry Corrigan-Gibbs and Dan Boneh. Prio: Private, robust, and scalable computation of aggregate statistics. In *14th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2017, Boston, MA, USA, March 27-29, 2017*, pages 259–282. USENIX Association, 2017. URL https://www.usenix.org/conference/nsdi17/technical-sessions/presentation/corrigan-gibbs.

Ronald Cramer, Ivan Bjerre Damgard, and Jesper Buus Nielsen. *Secure multiparty computation and secret sharing*. Cambridge University Press, 2015.

Ivan Damgard and Jesper Buus Nielsen. Scalable and unconditionally secure multiparty computation. In *Annual International Cryptology Conference*, pages 572–590. Springer, 2007.

Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

Leo de Castro and Antigoni Polychroniadou. Lightweight, maliciously secure verifiable function secret sharing. In *Advances in Cryptology - EUROCRYPT 2022 - 41st Annual International Conference on the Theory and Applications of Cryptographic Techniques, Trondheim, Norway, May 30 - June 3, 2022, Proceedings, Part I*, volume 13275 of *Lecture Notes in Computer Science*, pages 150–179. Springer, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Edgar Dobriban and Sifan Liu. Asymptotics for sketching in least squares regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 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.

Huang Fang, Xiaoyun Li, Chenglin Fan, and Ping Li. Improved convergence of differential private sgd with gradient clipping. In *The Eleventh International Conference on Learning Representations*, 2023.

Ilyas Fatkhullin, Alexander Tyurin, and Peter Richtárik. Momentum provably improves error feedback! *Advances in Neural Information Processing Systems*, 36, 2024.

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

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In *International Conference on Machine Learning*, pages 2232–2241. PMLR, 2019.

Oded Goldreich. *Foundations of cryptography: volume 2, basic applications*, volume 2. Cambridge university press, 2001.

Michael Greenwald and Sanjeev Khanna. Space-efficient online computation of quantile summaries. *ACM SIGMOD Record*, 30(2):58–66, 2001.

Farzin Haddadpour, Belhal Karimi, Ping Li, and Xiaoyun Li. Fedsketch: Communication-efficient and private federated learning via sketching. *arXiv preprint arXiv:2008.04975*, 2020.

Nicholas JA Harvey, Christopher Liaw, Yaniv Plan, and Sikander Randhawa. Tight analyses for non-smooth stochastic gradient descent. In *Conference on Learning Theory*, pages 1579–1613. PMLR, 2019.

Ilse CF Ipsen and Arvind K Saibaba. Stable rank and intrinsic dimension of real and complex matrices. *arXiv preprint arXiv:2407.21594*, 2024.

Berivan Isik, Francesco Pase, Deniz Gunduz, Tsachy Weissman, and Michele Zorzi. Sparse random networks for communication-efficient federated learning. *arXiv preprint arXiv:2209.15328*, 2022.

Nikita Ivkin, Daniel Rothchild, Enayat Ullah, Ion Stoica, Raman Arora, et al. Communication-efficient distributed sgd with sketching. *Advances in Neural Information Processing Systems*, 32, 2019.

Jiawei Jiang, Fangcheng Fu, Tong Yang, and Bin Cui. Sketchml: Accelerating distributed machine learning with data sketches. In *Proceedings of the 2018 International Conference on Management of Data*, pages 1269–1284, 2018.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

Daniel M Kane and Jelani Nelson. Sparser johnson-lindenstrauss transforms. *Journal of the ACM (JACM)*, 61(1):1–23, 2014.

Jonathan Katz and Yehuda Lindell. *Introduction to modern cryptography: principles and protocols*. Chapman and hall/CRC, 2007.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of statistics*, pages 1302–1338, 2000.

Omri Lev, Vishwak Srinivasan, Moshe Shenfeld, Katrina Ligett, Ayush Sekhari, and Ashia Wilson. The gaussian mixing mechanism: Renyi differential privacy via gaussian sketches. *Advances in Neural Information Processing Systems*, 38:50647–50679, 2026.

Ang Li, Jingwei Sun, Binghui Wang, Lin Duan, Sicheng Li, Yiran Chen, and Hai Li. Lotteryfl: Personalized and communication-efficient federated learning with lottery ticket hypothesis on non-iid datasets. *arXiv preprint arXiv:2008.03371*, 2020a.

Ang Li, Jingwei Sun, Xiao Zeng, Mi Zhang, Hai Li, and Yiran Chen. Fedmask: Joint computation and communication-efficient personalized federated learning via heterogeneous masking. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*, pages 42–55, 2021.

Mu Li, David G Andersen, Jun Woo Park, Alexander J Smola, Amr Ahmed, Vanja Josifovski, James Long, Eugene J Shekita, and Bor-Yiing Su. Scaling distributed machine learning with the parameter server. In *11th USENIX Symposium on operating systems design and implementation (OSDI 14)*, pages 583–598, 2014a.

Mu Li, David G Andersen, Alexander J Smola, and Kai Yu. Communication efficient distributed machine learning with the parameter server. *Advances in Neural Information Processing Systems*, 27, 2014b.

Qiaobo Li, Zhijie Chen, and Arindam Banerjee. Sketched gaussian mechanism for private federated learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Tian Li, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37:50–60, 05 2020b. doi: 10.1109/MSP.2020.2975749.

Xinyan Li, Qilong Gu, Yingxue Zhou, Tiancong Chen, and Arindam Banerjee. Hessian based analysis of sgd for deep nets: Dynamics and generalization. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 190–198. SIAM, 2020c.

Zhenyu Liao and Michael W Mahoney. Hessian eigenspectra of more realistic nonlinear models. *Advances in Neural Information Processing Systems*, 34:20104–20117, 2021.

Yujun Lin, Song Han, Huizi Mao, Yu Wang, and William J Dally. Deep gradient compression: Reducing the communication bandwidth for distributed training. *arXiv preprint arXiv:1712.01887*, 2017.

Heting Liu, Fang He, and Guohong Cao. Communication-efficient federated learning for heterogeneous edge devices based on adaptive gradient quantization. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2023a.

Hong Liu, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. *arXiv preprint arXiv:2305.14342*, 2023b.

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

Luca Melis, George Danezis, and Emiliano De Cristofaro. Efficient private statistics with succinct sketches. *arXiv preprint arXiv:1508.06110*, 2015.

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

Amirkeivan Mohtashami, Martin Jaggi, and Sebastian Stich. Masked training of neural networks with partial gradients. In *International Conference on Artificial Intelligence and Statistics*, pages 5876–5890. PMLR, 2022.

Wenlong Mou, Chris Junchi Li, Martin J Wainwright, Peter L Bartlett, and Michael I Jordan. On linear stochastic approximation: Fine-grained polyak-ruppert and non-asymptotic concentration. In *Conference on learning theory*, pages 2947–2997. PMLR, 2020.

Moni Naor, Benny Pinkas, and Omer Reingold. Distributed pseudo-random functions and kdcs. In *International conference on the theory and applications of cryptographic techniques*, pages 327–346. Springer, 1999.

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

Torben Pryds Pedersen. Non-interactive and information-theoretic secure verifiable secret sharing. In *Annual international cryptology conference*, pages 129–140. Springer, 1991.

Anastasia Pustozerova and Rudolf Mayer. Information leaks in federated learning. In *Proceedings of the network and distributed system security symposium*, volume 10, page 122, 2020.

Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization. In *International conference on artificial intelligence and statistics*, pages 2021–2031. PMLR, 2020.

Edo Roth, Daniel Noble, Brett Hemenway Falk, and Andreas Haeberlen. Honeycrisp: large-scale differentially private aggregation without a trusted core. In Tim Brecht and Carey Williamson, editors, *Proceedings of the 27th ACM Symposium on Operating Systems Principles, SOSP 2019, Huntsville, ON, Canada, October 27-30, 2019*, pages 196–210. ACM, 2019.

Guy N. Rothblum, Eran Omri, Junye Chen, and Kunal Talwar. Pine: efficient verification of a euclidean norm bound of a secret-shared vector. In *Proceedings of the 33rd USENIX Conference on Security Symposium*, SEC '24, USA, 2024. USENIX Association. ISBN 978-1-939133-44-1.

Daniel Rothchild, Ashwinee Panda, Enayat Ullah, Nikita Ivkin, Ion Stoica, Vladimir Braverman, Joseph Gonzalez, and Raman Arora. Fetchsgd: Communication-efficient federated learning with sketching. In *International Conference on Machine Learning*, pages 8253–8265. PMLR, 2020.

Mher Safaryan, Rustem Islamov, Xun Qian, and Peter Richtárik. Fednl: Making newton-type methods applicable to federated learning. *arXiv preprint arXiv:2106.02969*, 2021.

Levent Sagun, Leon Bottou, and Yann LeCun. Eigenvalues of the hessian in deep learning: Singularity and beyond. *arXiv preprint arXiv:1611.07476*, 2016.

Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns. In *Interspeech*, volume 2014, pages 1058–1062. Singapore, 2014.

Adi Shamir. How to share a secret. *Communications of the ACM*, 22(11):612–613, 1979.

Victor Shoup. Sequences of games: a tool for taming complexity in security proofs. *cryptology eprint archive*, 2004.

Mayank Shrivastava, Berivan Isik, Qiaobo Li, Sanmi Koyejo, and Arindam Banerjee. Sketching for distributed deep learning: A sharper analysis. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Zhao Song, Yitan Wang, Zheng Yu, and Lichen Zhang. Sketching for first order method: efficient algorithm for low-bandwidth channel and vulnerability. In *International Conference on Machine Learning*, pages 32365–32417. PMLR, 2023.

Standards for Efficient Cryptography Group (SECG). Sec 1: Elliptic curve cryptography. Technical report, Certicom Corp., 2009.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. *Advances in neural information processing systems*, 31, 2018.

Ananda Theertha Suresh, X Yu Felix, Sanjiv Kumar, and H Brendan McMahan. Distributed mean estimation with limited communication. In *International conference on machine learning*, pages 3329–3337. PMLR, 2017.

Aleksei Triastcyn and Boi Faltings. Federated learning with bayesian differential privacy. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2587–2596. IEEE, 2019.

Joel A Tropp, Alp Yurtsever, Madeleine Udell, and Volkan Cevher. Practical sketching algorithms for low-rank matrix approximation. *SIAM Journal on Matrix Analysis and Applications*, 38(4): 1454–1485, 2017.

Stacey Truex, Nathalie Baracaldo, Ali Anwar, Thomas Steinke, Heiko Ludwig, Rui Zhang, and Yi Zhou. A hybrid approach to privacy-preserving federated learning. In *Proceedings of the 12th ACM workshop on artificial intelligence and security*, pages 1–11, 2019.

Stacey Truex, Ling Liu, Ka-Ho Chow, Mehmet Emre Gursoy, and Wenqi Wei. Ldp-fed: Federated learning with local differential privacy. In *Proceedings of the third ACM international workshop on edge systems, analytics and networking*, pages 61–66, 2020.

Shay Vargaftik, Ran Ben-Basat, Amit Portnoy, Gal Mendelson, Yaniv Ben-Itzhak, and Michael Mitzenmacher. Drive: One-bit distributed mean estimation. *Advances in Neural Information Processing Systems*, 34:362–377, 2021.

Thijs Vogels, Sai Praneeth Karimireddy, and Martin Jaggi. Powersgd: Practical low-rank gradient compression for distributed optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018a.

Hongyi Wang, Scott Sievert, Shengchao Liu, Zachary Charles, Dimitris Papailiopoulos, and Stephen Wright. Atomo: Communication-efficient learning via atomic sparsification. *Advances in neural information processing systems*, 31, 2018b.

Lun Wang, Ruoxi Jia, and Dawn Song. D2p-fed: Differentially private federated learning with efficient communication. *arXiv preprint arXiv:2006.13039*, 2020.

Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE transactions on information forensics and security*, 15:3454–3469, 2020.

Wei Wen, Cong Xu, Feng Yan, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Terngrad: Ternary gradients to reduce communication in distributed deep learning. *Advances in neural information processing systems*, 30, 2017.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

Chulin Xie, Keli Huang, Pin-Yu Chen, and Bo Li. Dba: Distributed backdoor attacks against federated learning. In *International conference on learning representations*, 2019.

Zeke Xie, Qian-Yuan Tang, Yunfeng Cai, Mingming Sun, and Ping Li. On the power-law hessian spectrums in deep learning. *arXiv preprint arXiv:2201.13011*, 2022.

Xinwei Zhang, Xiangyi Chen, Mingyi Hong, Zhiwei Steven Wu, and Jinfeng Yi. Understanding clipping for federated learning: Convergence and client-level differential privacy. In *International Conference on Machine Learning, ICML 2022*, 2022.

Yushun Zhang, Congliang Chen, Tian Ding, Ziniu Li, Ruoyu Sun, and Zhi-Quan Luo. Why transformers need adam: A hessian perspective. *arXiv preprint arXiv:2402.16788*, 2024.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020.

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

Wennan Zhu, Peter Kairouz, Brendan McMahan, Haicheng Sun, and Wei Li. Federated heavy hitters discovery with differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pages 3837–3847. PMLR, 2020.

A Related Works

Communication-Efficient Distributed Learning. The high communication overhead of exchanging model updates between clients and a central server has motivated extensive work on communication-efficient distributed and federated optimization. A canonical example is FedAvg (McMahan et al., 2017), which decreases communication frequency by letting clients perform multiple local steps per round before synchronizing with the server. A complementary line of research reduces the per-round payload by compressing updates prior to transmission. Existing compression methods are commonly grouped into five families: sparsification (Aji and Heafield, 2017; Wang et al., 2018b; Lin et al., 2017), quantization (Suresh et al., 2017; Alistarh et al., 2017; Wen et al., 2017), low-rank factorization (Vargaftik et al., 2021; Mohtashami et al., 2022; Vogels et al., 2019), sketching (Rothchild et al., 2020; Song et al., 2023; Jiang et al., 2018), and sparse subnetwork training (Isik et al., 2022; Li et al., 2020a; 2021). These operators differ in key properties such as unbiasedness and linearity: while certain quantizers can be unbiased by design (Alistarh et al., 2017), many compressors are biased and therefore incorporate error-feedback or related corrections to recover favorable convergence behavior (Lin et al., 2017; Rothchild et al., 2020). Linearity is also desirable, as it approximately preserves geometric structure under dimensionality reduction (Dasgupta and Gupta, 2003). In this landscape, sketching is especially attractive because it offers a simple linear, unbiased mapping that enables computation in a reduced space, followed by reconstruction through desketching.

Privacy in Federated Learning. Differential privacy (DP) (Dwork et al., 2006) is a standard framework for providing rigorous privacy guarantees in machine learning. In the centralized setting, the canonical recipe is to clip per-iteration (stochastic) gradients and then add Gaussian noise to the clipped gradients (Abadi et al., 2016). Clipping is essential because it bounds sensitivity, and thus directly determines the noise scale required to achieve a target privacy level (Dwork et al., 2014). Clipping-based mechanisms have also been widely adopted in federated learning, though the appropriate clipping operation depends on the privacy granularity and system constraints. For sample-level privacy, clipping and injecting noise to every local update is proposed, but this can substantially degrade utility (Truex et al., 2019). For client-level privacy, early methods clip local models prior to communication and add noise to the bounded parameters (Wei et al., 2020; Truex et al., 2020). Subsequent work advocated clipping the transmitted updates instead of full models, which often yields improved numerical performance compared to model clipping (Geyer et al., 2017; Wang et al., 2020; Triastcyn and Faltings, 2019).

Sketching. Sketching has long been a foundational technique in data stream and randomized linear algebra, with a history that predates the deep learning era (Cormode and Muthukrishnan, 2005; Greenwald and Khanna, 2001; Kane and Nelson, 2014). It has been successfully applied to a broad range of problems, including low-rank approximation (Tropp et al., 2017), graph sparsification (Ahn et al., 2012), and least-squares regression (Dobriban and Liu, 2019). In recent years, sketching has also emerged as a practical primitive for communication-efficient distributed and federated learning, where it is used to compress transmitted updates (Jiang et al., 2018; Ivkin et al., 2019; Rothchild et al., 2020; Song et al., 2023; Haddadpour et al., 2020; Shrivastava et al., 2024). Moreover, sketching-based pipelines integrate naturally with secure aggregation and differential privacy techniques (Chen et al., 2022; Song et al., 2023; Melis et al., 2015; Zhu et al., 2020; Bassily et al., 2017). Recent work has also shown that the randomness of Gaussian sketching itself can provide stronger differential privacy guarantees, with refined analyses based on Rényi differential privacy (Lev et al., 2026). While this line of work focuses primarily on private regression problems, our setting concerns client-level privacy in iterative federated learning and the practical realization of privacy-amplifying sketch randomness under communication-efficient aggregation.

Cryptographic Protocols A large body of prior work studies cryptographic mechanisms for privacy-preserving aggregation and distributed learning, with the goal of protecting client updates from an untrusted or curious aggregator while maintaining scalability (Li et al., 2020b; Alabi et al., 2025). Secure aggregation protocols enable a server to learn only the sum of client updates, and nothing about individual contributions (Baruch et al., 2019; Roth et al., 2019). For example, the protocol of (Bonawitz et al., 2017a) introduced a practical secure aggregation system resilient to client dropouts, and has since been deployed in large-scale federated learning systems. Subsequent work has improved both asymptotic and concrete efficiency, achieving polylogarithmic communication per client and robustness to malicious behavior, as well as extensions to shuffling-based privacy models and outsourced MPC settings (Bell et al., 2020; Boneh et al., 2023; de Castro and Polychroniadou,

2022; Bell et al., 2022). These approaches provide strong cryptographic confidentiality, but typically assume that the aggregated values themselves are safe to release and do not address the interaction between aggregation and differential privacy mechanisms. Moreover, secure aggregation alone does not amplify privacy: the server learns the exact (or nearly exact) sum of clipped updates, and any DP guarantee must come solely from explicit noise addition (Rothblum et al., 2024). Systems such as Prio (Corrigan-Gibbs and Boneh, 2017), and related works, demonstrate that it is possible to compute rich aggregate statistics with strong cryptographic privacy and robustness against malicious clients, using secret-shared encodings and lightweight validity checks (Boneh et al., 2021). In contrast to these works, which use sketching primarily to enforce validity or compress cryptographic proofs, our work leverages sketching as an intrinsic source of privacy amplification whose secrecy must be maintained from the server. Our cryptographic protocols are complementary to prior secure aggregation and private measurement systems. Rather than replacing secure aggregation, we address a specific gap exposed by Fed-SGM: the need for clients to share a common sketch that remains hidden from an honest-but-curious server.

B Restatement of Fed-SGM from Li et al. (2025)

We restate the Fed-SGM algorithm from (Li et al., 2025) in Algorithm 1. At each communication round t, the server samples a subset of N clients uniformly at random without replacement from the population. Each selected client $c \in \mathcal{C}_t$ performs the following steps:

1. Computes a local update $\Delta_{c,t}$ by running stochastic gradient descent (SGD) on its local dataset $\mathcal{D}_c$.
2. Forms a clipped update $\hat{\Delta}_{c,t}(\mathcal{D}_c)$ with clipping threshold $\tau > 0$.
3. Applies the Sketched Gaussian Mechanism (SGM) to $\hat{\Delta}_{c,t}(\mathcal{D}_c)$, and transmits the resulting sketch to the server.

Upon receiving the sketches from participating clients, the server aggregates them and broadcasts the aggregated sketch to all clients. Each client then inverts the sketch to recover an estimate of the aggregated update in the original space, and updates the global model via a `GLOBAL_OPT` operator.

C Fed-SGM Privacy Degradation

In this section, we provide the formal proofs for the privacy analysis in Section 2, restating the theorems from the main text for convenience.

Lemma C.1 (Conditional Rényi Divergence with Known Sketch)**.** *Let $D_t := \{D_c\}_{c\in\mathcal{C}_t}$ be the collection of sampled client datasets at round t, and let D'_t be a neighboring round-level dataset obtained by replacing exactly one participating client dataset in D_t. Define $\gamma_t(D_t) := \sum_{c\in\mathcal{C}_t} \hat{\Delta}_{c,t}$, $\delta_t := \gamma_t(D_t) - \gamma_t(D'_t)$, and let $\mathbf{z}_t := \sum_{c\in\mathcal{C}_t} \mathbf{z}_{c,t}$. Conditioned on a fixed and known sketching matrix R_t, the Rényi divergence of order $\alpha > 1$ between the outputs satisfies*

$$D_\alpha(R_t\gamma_t(D_t) + \mathbf{z}_t \,\|\, R_t\gamma_t(D'_t) + \mathbf{z}_t) = \frac{\alpha}{2N\sigma_g^2}\|R_t\delta_t\|_2^2.$$

Proof. Conditioned on R_t, the two outputs $R_t\gamma_t(D_t) + \mathbf{z}_t$ and $R_t\gamma_t(D'_t) + \mathbf{z}_t$ are Gaussian with identical covariance matrix $N\sigma_g^2 I_b$, since $\mathbf{z}_t = \sum_{c\in\mathcal{C}_t} \mathbf{z}_{c,t} \sim \mathcal{N}(0, N\sigma_g^2 I_b)$. Their means differ by $R_t\big(\gamma_t(D_t) - \gamma_t(D'_t)\big) = R_t\delta_t$. Therefore, the Rényi divergence between these two Gaussians is $\frac{\alpha}{2N\sigma_g^2}\|R_t\delta_t\|_2^2$. □

Row-partition model. Fix a round t and suppose the $b \times d$ sketching matrix can be decomposed as

$$R_t \;=\; \begin{bmatrix} R_t^{\mathsf{kn}} \\ R_t^{\mathsf{un}} \end{bmatrix}, \qquad R_t^{\mathsf{kn}} \in \mathbb{R}^{b_{\mathsf{kn}}\times d}, \;\; R_t^{\mathsf{un}} \in \mathbb{R}^{b_{\mathsf{un}}\times d}, \;\; b_{\mathsf{kn}} + b_{\mathsf{un}} = b.$$

We interpret R_t^{kn} as the part of the sketch learned by the server (e.g., via collusion or side-channels), while R_t^{un} remains hidden.

Algorithm 1 Fed-SGM(Li et al., 2025)

Hyperparameters: server learning rate η_{global}, local learning rate η_{local}, noise scale σ_g, clipping threshold τ, number of rounds T.
Inputs: local datasets $D_c = \{(x_{i,c}, y_{i,c})\}_{i=1}^{n_c}$ and loss function $\mathcal{L}_c(\theta) = \frac{1}{n_c}\sum_{i=1}^{n_c} \ell\big(\theta, (x_{i,c}, y_{i,c})\big)$ for clients $c \in [C]$.

Initialize θ_0 randomly.
for $t = 0, \dots, T-1$ **do**
  Take a subset $\mathcal{C}_t$ of N clients with sampling probability $q = \frac{N}{C}$
  On Client Nodes:
  for $c \in \mathcal{C}_t$ **do**
    Assign local initialization: $\theta_{c,t,0} \leftarrow \theta_t$
    Local update:
    for $k = 1, \dots, K$ **do**
      $\theta_{c,t,k} \leftarrow \theta_{c,t,k-1} - \eta_{\text{local}} \cdot g_{c,t,k}$
    end for
    Compute update: $\Delta_{c,t} \leftarrow \theta_t - \theta_{c,t,K}$
    Clip the update: $\hat{\Delta}_{c,t} = \text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}}, \tau\right)$
    Apply SGM: $\tilde{\Delta}_{c,t} = \eta_{\text{local}} \mathcal{SG}(\hat{\Delta}_{c,t}(x_i); R_t, \mathbf{z}_{c,t})$
    Send $\tilde{\Delta}_{c,t}$ to the server
  end for
  On Server Node:
  Aggregate: $\bar{\tilde{\Delta}}_t = \frac{1}{N}\sum_{c\in\mathcal{C}_t} \tilde{\Delta}_{c,t}$
  Broadcast $\bar{\tilde{\Delta}}_t$ to the clients
  On Client Nodes:
  Update parameter: $\theta_{t+1} \leftarrow \texttt{GLOBAL_OPT}\left(\theta_t, R_t^\top \bar{\tilde{\Delta}}_t, \eta_{\text{global}}\right)$
end for
Outputs: final model θ_T

The single-round release for a clipped update $u(D) \in \mathbb{R}^d$ is

$$Y \;=\; R_t u(D) + Z \;=\; \begin{bmatrix} R_t^{\mathsf{kn}} u(D) \\ R_t^{\mathsf{un}} u(D) \end{bmatrix} + \begin{bmatrix} Z^{\mathsf{kn}} \\ Z^{\mathsf{un}} \end{bmatrix}, \qquad Z^{\mathsf{kn}} \sim \mathcal{N}(0, \sigma_g^2 I_{b_{\mathsf{kn}}}),\; Z^{\mathsf{un}} \sim \mathcal{N}(0, \sigma_g^2 I_{b_{\mathsf{un}}}),$$

with $Z^{\mathsf{kn}} \perp Z^{\mathsf{un}}$. We assume $\|u(D)\|_2 \le \tau$ for all D.

Theorem C.2 (Hybrid Rényi Divergence Bound: Known Rows + Hidden Rows)**.** *Let $D_t := \{D_c\}_{c\in\mathcal{C}_t}$ be the collection of sampled client datasets at round t, and let D_t' be a neighboring round-level dataset obtained by replacing exactly one participating client dataset in D_t. Define $\gamma_t(D_t) := \sum_{c\in\mathcal{C}_t} \hat{\Delta}_{c,t}$, $\delta_t := \gamma_t(D_t) - \gamma_t(D_t')$, and let $\mathbf{z}_t := \sum_{c\in\mathcal{C}_t} \mathbf{z}_{c,t}$. Assume the server knows $R_t^{\mathsf{kn}} \in \mathbb{R}^{b_{\mathsf{kn}}\times d}$ but not $R_t^{\mathsf{un}} \in \mathbb{R}^{b_{\mathsf{un}}\times d}$, where*

$$R_t = \begin{bmatrix} R_t^{\mathsf{kn}} \\ R_t^{\mathsf{un}} \end{bmatrix}, \qquad b_{\mathsf{kn}} + b_{\mathsf{un}} = b.$$

Then for any $\alpha > 1$, the round-t mechanism satisfies the following conditional Rényi divergence bound:

$$D_\alpha\left(R_t\gamma_t(D_t) + \mathbf{z}_t \,\|\, R_t\gamma_t(D_t') + \mathbf{z}_t\right) \le \underbrace{\frac{\alpha}{2N\sigma_g^2}\left\|R_t^{\mathsf{kn}}\delta_t\right\|_2^2}_{\text{mean-shift term (known sketch)}} + \underbrace{\frac{\alpha^2 b_{\mathsf{un}}\tau^4}{(\alpha-1)b^2\,\sigma_g^4}}_{\text{variance-ratio term (hidden sketch)}}.$$

Proof. Write the aggregate Gaussian noise as

$$\mathbf{z}_t = \begin{bmatrix} \mathbf{z}_t^{\mathsf{kn}} \\ \mathbf{z}_t^{\mathsf{un}} \end{bmatrix},$$

where $\mathbf{z}_t^{\mathsf{kn}}$ and $\mathbf{z}_t^{\mathsf{un}}$ are the components corresponding to the row partition of R_t. Then

$$R_t\gamma_t(D_t) + \mathbf{z}_t = \begin{bmatrix} R_t^{\mathsf{kn}}\gamma_t(D_t) + \mathbf{z}_t^{\mathsf{kn}} \\ R_t^{\mathsf{un}}\gamma_t(D_t) + \mathbf{z}_t^{\mathsf{un}} \end{bmatrix},$$

$$R_t\gamma_t(D'_t) + \mathbf{z}_t = \begin{bmatrix} R_t^{\mathsf{kn}}\gamma_t(D'_t) + \mathbf{z}_t^{\mathsf{kn}} \\ R_t^{\mathsf{un}}\gamma_t(D'_t) + \mathbf{z}_t^{\mathsf{un}} \end{bmatrix}.$$

By the subadditivity property of Rényi divergence under product structure,

$$\begin{aligned} & D_\alpha\left(R_t\gamma_t(D_t) + \mathbf{z}_t \,\|\, R_t\gamma_t(D'_t) + \mathbf{z}_t\right) \\ \le\ & D_\alpha\left(R_t^{\mathsf{kn}}\gamma_t(D_t) + \mathbf{z}_t^{\mathsf{kn}} \,\big\|\, R_t^{\mathsf{kn}}\gamma_t(D'_t) + \mathbf{z}_t^{\mathsf{kn}}\right) \\ & \quad + D_\alpha\left(R_t^{\mathsf{un}}\gamma_t(D_t) + \mathbf{z}_t^{\mathsf{un}} \,\|\, R_t^{\mathsf{un}}\gamma_t(D'_t) + \mathbf{z}_t^{\mathsf{un}}\right). \end{aligned} \tag{1}$$

For the known rows, Lemma C.1 gives

$$D_\alpha\left(R_t^{\mathsf{kn}}\gamma_t(D_t) + \mathbf{z}_t^{\mathsf{kn}} \,\big\|\, R_t^{\mathsf{kn}}\gamma_t(D'_t) + \mathbf{z}_t^{\mathsf{kn}}\right) = \frac{\alpha}{2N\sigma_g^2}\|R_t^{\mathsf{kn}}\delta_t\|_2^2. \tag{2}$$

For the hidden rows, note that R_t^{un} is a $\mathbb{R}^{b_{\mathsf{un}}\times d}$ Gaussian matrix with entries sampled i.i.d. from $\mathcal{N}(0, 1/b)$, and hence

$$\sqrt{\frac{b}{b_{\mathsf{un}}}}\, R_t^{\mathsf{un}}$$

is an isotropic Gaussian matrix. By Lemma 2.2 in Li et al. (2025),

$$D_\alpha\left(R_t^{\mathsf{un}}\gamma_t(D_t) + \mathbf{z}_t^{\mathsf{un}} \,\|\, R_t^{\mathsf{un}}\gamma_t(D'_t) + \mathbf{z}_t^{\mathsf{un}}\right) \le \frac{\alpha^2\left(\sqrt{\frac{b_{\mathsf{un}}}{b}}\tau\right)^4}{(\alpha-1)b_{\mathsf{un}}\sigma_g^4} = \frac{\alpha^2 b_{\mathsf{un}}\tau^4}{(\alpha-1)b^2\sigma_g^4}. \tag{3}$$

Substituting (2) and (3) into (1) gives the claim. □

RDP can be converted into the standard (ϵ, δ)-DP.

Lemma C.3 (Relationship with (ϵ, δ)-DP Mironov (2017))**.** *If f is an (α, ϵ)-RDP mechanism, then for any $0 < \delta < 1$, f also satisfies*

$$\left(\epsilon + \frac{\log(1/\delta)}{\alpha - 1}, \delta\right) \textit{-differential privacy}.$$

Corollary C.4 ((ϵ, δ)-DP from the Hybrid RDP Bound)**.** *Under the partial-knowledge scenario, the round-t aggregate mechanism is (ϵ_p, δ_p)-DP for any $\delta_p \in (0, 1)$ with*

$$\epsilon_p \le \inf_{\alpha>1}\left\{\frac{\alpha}{2N\sigma_g^2}\left\|R_t^{\mathsf{kn}}\delta_t\right\|_2^2 + \frac{\alpha^2 b_{\mathsf{un}}\tau^4}{(\alpha-1)b^2\,\sigma_g^4} + \frac{\log(1/\delta_p)}{\alpha-1}\right\}.$$

Proof. This is a direct application of Lemma C.3 on Theorem C.2. □

D Cryptographic Methods

In this section, we provide the formal cryptographic analysis underlying the construction in the main text. For completeness, we first present and analyze a simple public-key seed-sharing baseline and its vulnerability to single-client collusion. We then give the detailed specification and formal correctness and threshold-privacy analysis of Protocol 3.1.

D.1 Correctness and Security Against a Curious Server for the Baseline Protocol

In this section, we provide the complete formal specification and theoretical analysis of a simple public-key seed-sharing baseline. For completeness, we first detail its step-by-step execution in Protocol D.1. We then formalize the requisite cryptographic assumptions and present the game-based proofs of correctness and security against a semi-honest server.

Protocol D.1 (Seed-Sharing for Common Sketching Matrix). **Public parameters:** Encryption scheme $(\mathsf{KeyGen}, \mathsf{Enc}, \mathsf{Dec})$, aggregation rule $\oplus$, PRG PRG, security parameter λ, round identifier t.
Participants: Server and participating clients $\mathcal{C}_t$.

1. **Key generation (one-time or pre-provisioned).** Each client i locally generates a public–private key pair (pk_i, sk_i) using KeyGen and publishes pk_i. The private key sk_i is never shared.
2. **Client seed sampling.** Each client $i \in \mathcal{C}_t$ samples an independent random seed $s_{i,t} \leftarrow \{0,1\}^\lambda$.
3. **Encrypted seed exchange.** For each $i \in \mathcal{C}_t$ and each $j \in \mathcal{C}_t$, client i computes
$$c_{i\to j,t} \;=\; \mathsf{Enc}(pk_j, s_{i,t})$$
and sends the collection $\{c_{i\to j,t}\}_{j\in\mathcal{C}_t}$ to the server.
4. **Relay by server.** The server forwards to each client $j \in \mathcal{C}_t$ all ciphertexts $\{c_{i\to j,t}\}_{i\in\mathcal{C}_t}$ addressed to j.
5. **Seed reconstruction.** Each client $j \in \mathcal{C}_t$ decrypts
$$s_{i,t} = \mathsf{Dec}(sk_j, c_{i\to j,t}) \quad \forall i \in \mathcal{C}_t,$$
and computes the shared round seed
$$\mathsf{seed}_t \;=\; \bigoplus_{i\in\mathcal{C}_t} s_{i,t}.$$
6. **Sketch generation.** All clients deterministically generate the sketching matrix
$$R_t \;=\; \mathsf{PRG}(\mathsf{seed}_t).$$

Assumption D.1 (Perfect Correctness of PKE)**.** We assume the public-key encryption scheme $(\mathsf{KeyGen}, \mathsf{Enc}, \mathsf{Dec})$ is perfectly correct. That is, for all message spaces $\mathcal{M}$ and all $m \in \mathcal{M}$, if $(pk, sk) \leftarrow \mathsf{KeyGen}(1^\lambda)$, then $\Pr\big[\mathsf{Dec}(sk, \mathsf{Enc}(pk, m)) = m\big] = 1$(Katz and Lindell, 2007).

Assumption D.2 (IND-CPA Security)**.** We assume the encryption scheme provides indistinguishability under chosen-plaintext attack (IND-CPA) (Katz and Lindell, 2007).

Assumption D.3 (Secure PRG)**.** We assume $\mathsf{PRG} : \{0,1\}^\lambda \to \mathcal{R}$ is a secure pseudorandom generator, meaning its output is computationally indistinguishable from a uniform distribution over the target matrix space $\mathcal{R}$ (Goldreich, 2001).

Theorem D.4 (Protocol Correctness)**.** *If Assumption D.1 holds, and all participating clients in $\mathcal{C}_t$ honestly follow the protocol, then every client $j \in \mathcal{C}_t$ will deterministically output the exact same sketching matrix R_t.*

Proof. Let $\mathcal{C}_t$ be the set of participating clients. In Step 2, each client $i \in \mathcal{C}_t$ samples a seed $s_{i,t}$. In Step 3, client i computes the ciphertext
$$c_{i\to j,t} = \mathsf{Enc}(pk_j, s_{i,t})$$
for every client j.

Because the server faithfully relays the messages (Step 4), each client j receives the set $\{c_{i\to j,t}\}_{i\in\mathcal{C}_t}$. By Assumption D.1, the decryption in Step 5 perfectly reconstructs the original seeds:
$$s'_{i,t} = \mathsf{Dec}(sk_j, c_{i\to j,t}) = s_{i,t}, \qquad \forall i \in \mathcal{C}_t.$$

Consequently, every client $j \in \mathcal{C}_t$ computes the identical shared round seed:
$$\mathsf{seed}_t = \bigoplus_{i\in\mathcal{C}_t} s'_{i,t} = \bigoplus_{i\in\mathcal{C}_t} s_{i,t}.$$

Since PRG is a deterministic algorithm, the resulting sketching matrix
$$R_t = \mathsf{PRG}(\mathsf{seed}_t)$$
computed in Step 6 is identical for all clients in $\mathcal{C}_t$. This concludes the proof. □

Theorem D.5 (Matrix Pseudorandomness)**.** *Under Assumptions D.2 and D.3, the sketching matrix R_t is computationally indistinguishable from a truly random matrix $U \leftarrow \mathcal{R}$, given the view of a semi-honest server.*

Proof. We define the server's view, $\mathsf{View}_{\mathsf{Server}}$, as the collection of all public keys $\{pk_i\}_{i\in\mathcal{C}_t}$ and all relayed ciphertexts $\{c_{i\to j,t}\}_{i,j\in\mathcal{C}_t}$. We prove the theorem using a sequence of computationally indistinguishable hybrid games(Shoup, 2004).

- Game 0 (Real Protocol). This game corresponds to the actual execution of the protocol. The output of the game is
$$\big(\mathsf{View}^{(0)}_{\mathsf{Server}}, R_t\big),$$
where
$$R_t = \mathsf{PRG}\left(\bigoplus_{i\in\mathcal{C}_t} s_{i,t}\right).$$

- Game 1 (Simulated Ciphertexts). In this game, we modify the ciphertexts generated in Step 3. Instead of encrypting the true seeds $s_{i,t}$, the simulator replaces every ciphertext with an encryption of a dummy string of zeroes:
$$c'_{i\to j,t} = \mathsf{Enc}(pk_j, 0^\lambda).$$
The matrix R_t is still computed using the true underlying seeds $s_{i,t}$.

  By Assumption D.2, the PKE scheme is IND-CPA secure. Through a standard hybrid argument over the $|\mathcal{C}_t|^2$ ciphertexts, the server cannot distinguish between encryptions of $s_{i,t}$ and 0^λ. Therefore, Game 0 is computationally indistinguishable from Game 1:
$$\big(\mathsf{View}^{(0)}_{\mathsf{Server}}, R_t\big) \approx_c \big(\mathsf{View}^{(1)}_{\mathsf{Server}}, R_t\big).$$

- Game 2 (Random Matrix). Notice that in Game 1, the server's view $\mathsf{View}^{(1)}_{\mathsf{Server}}$ contains no information whatsoever about the true seeds $s_{i,t}$. Therefore, from the perspective of the server,
$$\mathsf{seed}_t = \bigoplus_{i\in\mathcal{C}_t} s_{i,t}$$
is a uniformly random value in $\{0,1\}^\lambda$.

  In Game 2, we replace the deterministic output
$$R_t = \mathsf{PRG}(\mathsf{seed}_t)$$
with a matrix U sampled uniformly at random from $\mathcal{R}$. Because seed_t is uniformly distributed and independent of $\mathsf{View}^{(1)}_{\mathsf{Server}}$, Assumption D.3 guarantees that the output of the PRG is computationally indistinguishable from U. Thus,
$$\big(\mathsf{View}^{(1)}_{\mathsf{Server}}, R_t\big) \approx_c \big(\mathsf{View}^{(1)}_{\mathsf{Server}}, U\big).$$

By the transitivity of computational indistinguishability, we conclude that
$$\big(\mathsf{View}_{\mathsf{Server}}, R_t\big) \approx_c \big(\mathsf{View}_{\mathsf{Server}}, U\big).$$
Hence, the server cannot distinguish the true sketching matrix from a random matrix. □

D.2 Collusion Attacks Against Protocol D.1

In this section, we formally show that the public-key seed-sharing baseline is not robust to server–client collusion.

Lemma D.6 (Seed Recovery under Single-Client Collusion)**.** *If the server colludes with any single participating client $j^\star \in \mathcal{C}_t$, then the adversary can reconstruct seed_t and hence the sketching matrix R_t with probability* 1.

Proof. By protocol design, each client $i \in \mathcal{C}_t$ encrypts its seed $s_{i,t}$ under the public key of every client $j \in \mathcal{C}_t$, including $j^\star$. Client $j^\star$ therefore decrypts and recovers the full collection $\{s_{i,t}\}_{i\in\mathcal{C}_t}$ using its private key. Since the aggregation rule (e.g., bitwise XOR) is public, the colluding client computes

$$\mathsf{seed}_t = \bigoplus_{i\in\mathcal{C}_t} s_{i,t}$$

and shares it with the server. The sketching matrix R_t is deterministically generated from seed_t, completing the attack. □

Protocol D.2 (Threshold-Resilient Shared Sketch Randomness). **Goal:** In round t, enable all participating clients $\mathcal{C}_t$ to use a common Gaussian sketching matrix R_t while ensuring that an adversary controlling the server and at most κ colluding clients learns no additional information about the round randomness or the honest clients' updates beyond what is implied by the final aggregate and the inputs of the corrupted clients.

Assumption: An honest majority among participants, i.e., $\kappa < |\mathcal{C}_t|/2$.

Public parameters: A computational security parameter λ, a finite field $\mathbb{F}$, a $(\kappa{+}1)$-out-of-$|\mathcal{C}_t|$ perfectly hiding verifiable secret sharing (VSS) scheme, a distributed PRG evaluation primitive dPRG, an honest-majority MPC protocol for secure multiplication, a public Gaussian sampler Gauss, and a secure aggregation protocol SecAgg.

Participants: Server and clients $\mathcal{C}_t$.

1. **Distributed generation of a secret-shared round seed.** Each client $i \in \mathcal{C}_t$ samples $s_{i,t} \leftarrow \{0,1\}^\lambda$ and VSS-shares it, producing shares $\{\langle s_{i,t}\rangle_j\}_{j\in\mathcal{C}_t}$. Each client j computes
$$\langle \mathsf{seed}_t\rangle_j \;:=\; \sum_{i\in\mathcal{C}_t} \langle s_{i,t}\rangle_j \in \mathbb{F}.$$
No client reconstructs seed_t.
2. **Distributed PRG expansion.** Clients jointly evaluate dPRG on the secret shares $\{\langle \mathsf{seed}_t\rangle_j\}$ to obtain additive shares of pseudorandom values $\{\langle w_{t,\ell}\rangle_j\}_{\ell=1}^{b}$, one for each sketch row.
3. **Implicit sketch definition.** The ℓ-th sketch row is defined as
$$(R_t)_{\ell,:} := \mathsf{Gauss}(w_{t,\ell}) \in \mathbb{R}^d,$$
but only in secret-shared form; client j holds $\langle (R_t)_{\ell,:}\rangle_j$.
4. **Local sketched update in shares.** Each client $j \in \mathcal{C}_t$ computes its clipped update $\hat{\Delta}_{j,t}$ and forms additive shares of the sketched update
$$\langle \hat{\Delta}^{\mathrm{sk}}_{j,t}\rangle := \langle R_t\rangle \hat{\Delta}_{j,t},$$
using honest-majority MPC.
5. **Noise addition.** Gaussian noise with variance σ_g^2 is added either locally or in shared form, yielding $\langle \tilde{\Delta}_{j,t}\rangle$.
6. **Secure aggregation.** Clients execute SecAgg to reveal only
$$\bar{\tilde{\Delta}}_t := \frac{1}{|\mathcal{C}_t|}\sum_{j\in\mathcal{C}_t} \tilde{\Delta}_{j,t}$$
to the server.

D.3 Correctness and Security Against Server and Colluding Clients for Protocol 3.1

In this section, we provide the formal correctness and security analysis of the Threshold-Resilient Shared Sketch Randomness protocol from the main text. To keep the appendix self-contained, we

first give its detailed specification in Protocol D.2, followed by the cryptographic assumptions and formal correctness and threshold-privacy guarantees.

Assumption D.7 (Correctness of Distributed Primitives)**.** We assume the underlying interactive protocols (VSS, dPRG, honest-majority MPC, and SecAgg) and local functions (Gauss) satisfy perfect correctness. Specifically, an honest execution of the VSS produces linear secret shares over $\mathbb{F}$ that allow exact reconstruction of the shared secret, the dPRG outputs additive shares that sum exactly to the deterministic PRG expansion, the honest-majority MPC computes valid secret shares of the exact matrix-vector product, and SecAgg reconstructs the exact sum of the provided inputs (Shamir, 1979; Pedersen, 1991; Ben-Or et al., 2019; Bonawitz et al., 2017b; Cramer et al., 2015).

Assumption D.8 (Secure VSS)**.** We assume the $(\kappa+1)$-out-of-$|\mathcal{C}_t|$ verifiable secret sharing (VSS) realization is perfectly hiding against a semi-honest adversary controlling the server and at most κ clients. In particular, the joint view of such an adversary during the VSS execution can be perfectly simulated without knowledge of the honest clients' shared secrets (Shamir, 1979; Pedersen, 1991).

Assumption D.9 (Secure Distributed PRG)**.** We assume the protocol for distributed PRG evaluation (dPRG) is securely realized against a semi-honest adversary controlling the server and at most κ clients. Specifically, there exists an efficient simulator that can generate a computationally indistinguishable joint adversarial view given only the corrupted parties' inputs and their shares of the PRG expansion (Naor et al., 1999; Cramer et al., 2015).

Assumption D.10 (Secure Multiplication)**.** We assume the honest-majority MPC used for local sketching securely realizes the multiplication of a secret-shared matrix by a client-private input vector against a semi-honest adversary controlling the server and at most κ clients. Specifically, the adversary learns nothing about the honest parties' updates beyond the resulting shares $\langle \hat{\Delta}^{\mathrm{sk}}_{j,t} \rangle$ (Ben-Or et al., 2019).

Assumption D.11 (Secure Aggregation)**.** We assume the SecAgg protocol securely computes the sum against a semi-honest adversary controlling the server and at most κ clients. In particular, the adversary learns no additional information about the honest parties' inputs beyond what is revealed by the final aggregate sum (Bonawitz et al., 2017b).

Theorem D.12 (Protocol Correctness)**.** *If all clients in $\mathcal{C}_t$ honestly follow Protocol D.2, the server deterministically obtains the correct noisy aggregate sketched update*

$$\bar{\tilde{\Delta}}_t = \frac{1}{|\mathcal{C}_t|} \sum_{j \in \mathcal{C}_t} \big(R_t \hat{\Delta}_{j,t} + \mathbf{z}_{j,t} \big),$$

where R_t is the target sketching matrix and $\mathbf{z}_{j,t}$ is the sampled Gaussian noise.

Proof. In Step 1, each client $i \in \mathcal{C}_t$ samples and shares a seed $s_{i,t}$ via VSS. By Assumption D.7, the VSS scheme guarantees perfect linearity and reconstructability over the finite field $\mathbb{F}$. Therefore, the local summation of these shares yields valid secret shares of the combined round seed:

$$\sum_{i \in \mathcal{C}_t} \langle s_{i,t} \rangle_j = \left\langle \sum_{i \in \mathcal{C}_t} s_{i,t} \right\rangle_j = \langle \mathsf{seed}_t \rangle_j.$$

In Steps 2 and 3, clients jointly evaluate dPRG and implicitly define the sketch via Gauss. By Assumption D.7, the perfect correctness of these distributed and local primitives ensures that the clients successfully obtain valid additive shares of the exact, deterministic target matrix R_t, without ever reconstructing it in the clear.

In Step 4, the clients execute an honest-majority MPC to compute

$$\langle \hat{\Delta}^{\mathrm{sk}}_{j,t} \rangle = \langle R_t \rangle \hat{\Delta}_{j,t}.$$

By Assumption D.7, the perfect correctness of this interactive operation guarantees that it results in exact, valid shares of the matrix-vector product $R_t \hat{\Delta}_{j,t}$.

Finally, in Steps 5 and 6, local Gaussian noise $\mathbf{z}_{j,t}$ is added to form the noised shares

$$\langle \tilde{\Delta}_{j,t} \rangle = \langle R_t \hat{\Delta}_{j,t} + \mathbf{z}_{j,t} \rangle,$$

and the clients execute SecAgg. By Assumption D.7, the SecAgg protocol perfectly reconstructs the exact sum of its inputs. Consequently, the server deterministically outputs the exact average:

$$\bar{\tilde{\Delta}}_t = \frac{1}{|\mathcal{C}_t|} \sum_{j \in \mathcal{C}_t} \tilde{\Delta}_{j,t} = \frac{1}{|\mathcal{C}_t|} \sum_{j \in \mathcal{C}_t} \big(R_t \hat{\Delta}_{j,t} + \mathbf{z}_{j,t} \big).$$

This matches the target aggregate distribution and concludes the proof. □

Theorem D.13 (Threshold Privacy)**.** *Under Assumptions D.8, D.9, D.10, and D.11, Protocol D.2 securely computes the aggregate sketched update against any adversary $\mathcal{A}$ controlling the server and $|M| \leq \kappa$ clients. Specifically, there exists an expected probabilistic polynomial-time simulator $\mathcal{S}$ such that the real view of $\mathcal{A}$ is computationally indistinguishable from the simulated view:*

$$\mathsf{View}_{\mathcal{A}}^{\mathsf{Real}} \approx_c \mathsf{View}_{\mathcal{A}}^{\mathsf{Ideal}(\mathcal{S})}.$$

Proof. The view of the adversary $\mathcal{A}$ consists of the internal states of the corrupted clients M, the messages received during the VSS, dPRG, and MPC phases, and the SecAgg messages observed by the server. We construct the simulator $\mathcal{S}$, which is given the inputs of the corrupted clients $\{\hat{\Delta}_{j,t}\}_{j \in M}$ and the final aggregate output $\bar{\tilde{\Delta}}_t$ from the ideal functionality. $\mathcal{S}$ generates a simulated view via the following indistinguishable hybrid games.

- Game 0 (Real Execution). This is the real execution of the protocol. The output is $\mathsf{View}_{\mathcal{A}}^{(0)} = \mathsf{View}_{\mathcal{A}}^{\mathsf{Real}}$.
- Game 1 (Simulating VSS). In Step 1, honest clients VSS-share their local seeds with the participating clients. Because $|M| \leq \kappa$, the adversary does not meet the reconstruction threshold. By Assumption D.8, the joint VSS view of the server and corrupted clients can be perfectly simulated without knowledge of the honest clients' seeds. We therefore replace the VSS executions of the honest clients by their perfectly simulated counterparts. The resulting view is identically distributed to that in Game 0, and hence
$$\mathsf{View}_{\mathcal{A}}^{(0)} \equiv \mathsf{View}_{\mathcal{A}}^{(1)}.$$
- Game 2 (Simulating dPRG and Secure Multiplication). In Steps 2 and 4, the clients engage in interactive protocols to evaluate the dPRG and compute the local sketched updates. By Assumptions D.9 and D.10, both the distributed PRG and the honest-majority MPC for secure multiplication are securely realized. The simulator $\mathcal{S}$ invokes the guaranteed sub-simulators for these respective protocols to generate a simulated joint view for the server and corrupted clients in M, without knowing the honest clients' inputs, the true matrix R_t, or the underlying shared seeds. Because both underlying protocols provide simulatable security, the simulated joint transcripts are computationally indistinguishable from the real execution:
$$\mathsf{View}_{\mathcal{A}}^{(1)} \approx_c \mathsf{View}_{\mathcal{A}}^{(2)}.$$
- Game 3 (Simulating Secure Aggregation). In Step 6, the server receives messages to reconstruct the final sum. The simulator $\mathcal{S}$, knowing the corrupted inputs $\{\hat{\Delta}_{j,t}\}_{j \in M}$ and the ideal output $\bar{\tilde{\Delta}}_t$, constructs simulated SecAgg messages on behalf of the honest clients. It structures these messages such that the resulting protocol output perfectly forces the server to output $\bar{\tilde{\Delta}}_t$. By Assumption D.11, the server's view of the SecAgg transcript reveals nothing but the sum, making this simulation indistinguishable from the real protocol run. Thus, $\mathsf{View}_{\mathcal{A}}^{(2)} \approx_c \mathsf{View}_{\mathcal{A}}^{(3)}$.

Since Game 3 relies strictly on the ideal outputs and corrupted inputs, we define $\mathsf{View}_{\mathcal{A}}^{\mathsf{Ideal}(\mathcal{S})} = \mathsf{View}_{\mathcal{A}}^{(3)}$. By the transitivity of indistinguishability, the adversary's view in the real protocol is computationally indistinguishable from the simulator's view. Consequently, the adversary learns no additional information about the honest clients' seeds, updates, or the underlying sketching matrix R_t beyond what is implied by the corrupted clients' inputs and the final aggregate. □

Algorithm 2 Fed-DSSGM

Hyperparameters: server learning rate η_{global}, local learning rate η_{local}, noise scale σ_g, clipping threshold τ, number of rounds T.
Inputs: local datasets $D_c = \{(x_{i,c}, y_{i,c})\}_{i=1}^{n_c}$ and loss function $\mathcal{L}_c(\theta) = \frac{1}{n_c}\sum_{i=1}^{n_c} \ell\big(\theta, (x_{i,c}, y_{i,c})\big)$ for clients $c \in [C]$.

Initialize θ_0 randomly.
for $t = 0, \dots, T-1$ **do**
  Take a subset $\mathcal{C}_t$ of N clients with sampling probability $q = \frac{N}{C}$
  On Client Nodes:
  for $c \in \mathcal{C}_t$ **do**
    Assign local initialization: $\theta_{c,t,0} \leftarrow \theta_t$
    Local update:
    for $k = 1, \dots, K$ **do**
      $\theta_{c,t,k} \leftarrow \theta_{c,t,k-1} - \eta_{\text{local}} \cdot g_{c,t,k}$
    end for
    Compute update: $\Delta_{c,t} \leftarrow \theta_t - \theta_{c,t,K}$
    Clip the update: $\hat{\Delta}_{c,t} = \text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}}, \tau\right)$
    Apply SGM with local sketching matrix: $\check{\Delta}_{c,t} = \eta_{\text{local}} \mathcal{SG}(\hat{\Delta}_{c,t}(x_i); R_{c,t}, \mathbf{z}_{c,t})$
    Apply De-sketching: $\breve{\Delta}_{c,t} = R_{c,t}^\top \check{\Delta}_{c,t}$
    Apply sketching with global sketching matrix: $\tilde{\Delta}_{c,t} = S_t \breve{\Delta}_{c,t}$
    Send $\tilde{\Delta}_{c,t}$ to the server
  end for
  On Server Node:
  Aggregate: $\bar{\tilde{\Delta}}_t = \frac{1}{N}\sum_{c \in \mathcal{C}_t} \tilde{\Delta}_{c,t}$
  Broadcast $\bar{\tilde{\Delta}}_t$ to the clients
  On Client Nodes:
  Update parameter:
  $\theta_{t+1} \leftarrow \texttt{GLOBAL_OPT}\left(\theta_t, S_t^\top \bar{\tilde{\Delta}}_t, \eta_{\text{global}}\right)$
end for
Outputs: final model θ_T

E FED-DSSGM ALGORITHM

Algorithm 2 presents Fed-DSSGM. At each communication round t, the server samples a subset of N clients uniformly at random without replacement. Each selected client c then:

1. performs local stochastic gradient descent (SGD) on $\mathcal{D}_c$ to obtain an update $\Delta_{c,t}$;
2. clips $\frac{\Delta_{c,t}}{\eta_{\text{local}}}$ to have ℓ_2-norm at most τ;
3. applies the sketched Gaussian mechanism (SGM) using a client-specific sketching matrix $R_{c,t}$, and maps the result back to the ambient dimension via $R_{c,t}^\top$;
4. applies a global sketching matrix S_t to reduce the dimension and transmits the resulting sketch to the server.

Upon receiving the client sketches, the server aggregates them and broadcasts the aggregated sketch to clients. Each client then applies the inverse mapping associated with the global sketching matrix to recover the aggregated update in the original ambient space, and updates the global model via the `GLOBAL_OPT` operator. The operator `GLOBAL_OPT` can instantiate a broad class of optimizers, including gradient descent (GD) and adaptive methods such as AMSGrad.

F SKETCHING GUARANTEE

We first present an auxiliary lemma giving a norm bound on the deviation introduced by a single sketching and de-sketching operation. We then use this result to establish a concentration guarantee under two Gaussian sketching operations.

Lemma F.1. *Let $R \in \mathbb{R}^{b\times d}$ have i.i.d. entries $\mathcal{N}(0, 1/b)$. Fix $v \in \mathbb{R}^d$ and $\delta \in (0, 1)$. Define*

$$B := R^\top R - I_d.$$

Then with probability at least $1 - \delta$,

$$\|Bv\| \le 8\|v\|\sqrt{\frac{d}{b}}\log(4/\delta).$$

Proof. Step 1: Rotate v to e_1 by rotational invariance. Write $R = \frac{1}{\sqrt{b}}G$ where G has i.i.d. $\mathcal{N}(0, 1)$ entries. Let Q be any orthogonal matrix with $Qv = \|v\|e_1$. Since $GQ \overset{d}{=} G$ (right rotational invariance of Gaussian matrices),

$$\begin{aligned}\|(R^\top R - I)v\| &= \left\|\left(\frac{1}{b}G^\top G - I\right)v\right\| \\ &= \left\|\left(\frac{1}{b}(GQ)^\top(GQ) - I\right)Qv\right\| \\ &\overset{d}{=} \|v\|\left\|\left(\frac{1}{b}G^\top G - I\right)e_1\right\|.\end{aligned}$$

So it suffices to prove the bound for $v = e_1$ and then multiply by $\|v\|$. From now on assume $v = e_1$.

Step 2: Decompose $(G^\top G)e_1$. Let $g_1 \in \mathbb{R}^b$ be the first column of G, and $H \in \mathbb{R}^{b\times(d-1)}$ be the remaining columns, so $G = [g_1 \;\; H]$. Then

$$G^\top G e_1 = \begin{bmatrix} g_1^\top g_1 \\ H^\top g_1 \end{bmatrix}.$$

Thus

$$\left(\frac{1}{b}G^\top G - I\right)e_1 = \begin{bmatrix} \frac{\|g_1\|^2}{b} - 1 \\ \frac{1}{b}H^\top g_1 \end{bmatrix}.$$

Hence, using $\sqrt{a^2 + b^2} \le |a| + |b|$,

$$\left\|\left(\frac{1}{b}G^\top G - I\right)e_1\right\| \le \left|\frac{\|g_1\|^2}{b} - 1\right| + \frac{1}{b}\|H^\top g_1\|. \tag{4}$$

So we bound these two terms with high probability.

Step 3: Chi-square tail bound we will use.

Theorem F.2. *(Laurent and Massart, 2000)[Laurent–Massart inequality] If $Z \sim \chi^2_m$, then for all $t \ge 0$,*

$$\Pr\left(Z - m \ge 2\sqrt{mt} + 2t\right) \le e^{-t}, \qquad \Pr\left(m - Z \ge 2\sqrt{mt}\right) \le e^{-t}. \tag{5}$$

Step 4: Bound the diagonal term $\left|\frac{\|g_1\|^2}{b} - 1\right|$. Apply (5) to $\|g_1\|^2 \sim \chi^2_b$, we can get that for any $t \ge 0$, with probability at least $1 - 2e^{-t}$,

$$\left|\|g_1\|^2 - b\right| \le 2\sqrt{bt} + 2t.$$

which implies that

$$\left|\frac{\|g_1\|^2}{b} - 1\right| \le 2\sqrt{\frac{t}{b}} + \frac{2t}{b}. \tag{6}$$

Step 5: Bound the off-diagonal term $\frac{1}{b}\|H^\top g_1\|$. We first condition on g_1. Each column h_j of H is i.i.d. $\mathcal{N}(0, I_b)$, independent of g_1. Then $h_j^\top g_1 \mid g_1 \sim \mathcal{N}(0, \|g_1\|^2)$ independently over j, so

$$H^\top g_1 \mid g_1 \sim \mathcal{N}\left(0, \|g_1\|^2 I_{d-1}\right).$$

Therefore

$$W = \frac{\|H^\top g_1\|^2}{\|g_1\|^2} \;\Big|\; g_1 \sim \chi^2_{d-1}.$$

Apply (5) to $W \sim \chi^2_{d-1}$: for any $t \geq 0$, with probability at least $1 - e^{-t}$,

$$W \leq (d-1) + 2\sqrt{(d-1)t} + 2t \leq (\sqrt{d-1} + \sqrt{2t})^2,$$

so

$$\|H^\top g_1\| \leq \|g_1\|(\sqrt{d-1} + \sqrt{2t}) \leq \|g_1\|(\sqrt{d} + \sqrt{2t}).$$

Thus

$$\frac{1}{b}\|H^\top g_1\| \leq \frac{\|g_1\|}{b}\left(\sqrt{d} + \sqrt{2t}\right). \tag{7}$$

Step 6: Bound $\|g_1\|$. Again $\|g_1\|^2 \sim \chi^2_b$. Using the upper tail in (5), with probability at least $1 - e^{-t}$,

$$\|g_1\|^2 \leq b + 2\sqrt{bt} + 2t \leq (\sqrt{b} + \sqrt{2t})^2,$$

hence

$$\|g_1\| \leq \sqrt{b} + \sqrt{2t}. \tag{8}$$

Plugging (8) into (7), we have

$$\frac{1}{b}\|H^\top g_1\| \leq \frac{\|g_1\|}{b}\left(\sqrt{d} + \sqrt{2t}\right) \leq \frac{\left(\sqrt{b} + \sqrt{2t}\right)\left(\sqrt{d} + \sqrt{2t}\right)}{b}. \tag{9}$$

Step 7: Put the pieces together. From (4),

$$\left\| \left(\frac{1}{b} G^\top G - I \right) e_1 \right\| \leq \left| \frac{\|g_1\|^2}{b} - 1 \right| + \frac{1}{b}\|H^\top g_1\|$$

Plugging (6) and (9), with probability at least $1 - 4e^{-t}$,

$$\left\| \left(\frac{1}{b} G^\top G - I \right) e_1 \right\| \leq 2\sqrt{\frac{t}{b}} + \frac{2t}{b} + \frac{\left(\sqrt{b} + \sqrt{2t}\right)\left(\sqrt{d} + \sqrt{2t}\right)}{b}$$

Now set $t = \log(4/\delta)$ so that $4e^{-t} = \delta$. Then with probability at least $1 - \delta$,

$$\begin{aligned}\left\| \left(\frac{1}{b} G^\top G - I \right) e_1 \right\| &\leq 2\sqrt{\frac{\log(4/\delta)}{b}} + \frac{2\log(4/\delta)}{b} + \frac{\left(\sqrt{b} + \sqrt{2\log(4/\delta)}\right)\left(\sqrt{d} + \sqrt{2\log(4/\delta)}\right)}{b} \\ &\leq 8\sqrt{\frac{d}{b}}\log(4/\delta).\end{aligned}$$

Returning to general v via Step 1, with probability at least $1 - \delta$,

$$\|(R^\top R - I)v\| \overset{d}{=} \|v\| \left\| \left(\frac{1}{b} G^\top G - I \right) e_1 \right\| \leq 8\|v\|\tilde{\mathcal{O}}\left(\sqrt{\frac{d}{b}}\right)$$

□

Using the above norm bound, we now establish the concentration guarantee for the composition of two independent Gaussian sketching and de-sketching operations.

Lemma F.3. *Let $S \in \mathbb{R}^{b_{out} \times d}$, $R \in \mathbb{R}^{b_{in} \times d}$ be two independent Gaussian sketching matrices, where the entries of S are sampled i.i.d. from $\mathcal{N}\left(0, \frac{1}{b_{out}}\right)$ and the entries of R are sampled i.i.d. from $\mathcal{N}\left(0, \frac{1}{b_{in}}\right)$. Then, for any fixed vectors $u, v \in \mathbb{R}^d$, with probability at least $1 - \delta$, the following property holds:*

$$\left|u^\top S^\top S R^\top R v - u^\top v\right| \leq \|u\|\|v\| \frac{10 \log^{1.5}(6d/\delta) \log(24/\delta) \sqrt{d}}{\sqrt{b_{in} b_{out}}}.$$

Proof. Let

$$C := S^\top S - I, \qquad D := R^\top R - I,$$

so that

$$S^\top S\, R^\top R - I = (I + C)(I + D) - I = C + D + CD.$$

Therefore

$$u^\top S^\top S R^\top R v - u^\top v = \underbrace{u^\top C v}_{T_C} + \underbrace{u^\top D v}_{T_D} + \underbrace{u^\top C (D v)}_{T_{CD}}.$$

For T_C and T_D, applying Lemma D.24 in (Song et al., 2023) to u and v, we can get that with probability at least $1 - \frac{\delta}{3}$,

$$|T_C| \leq \|u\|\|v\| \frac{\log^{1.5}(3d/\delta)}{\sqrt{b_{\text{out}}}}, \tag{10}$$

$$|T_D| \leq \|u\|\|v\| \frac{\log^{1.5}(3d/\delta)}{\sqrt{b_{\text{in}}}}. \tag{11}$$

For T_{CD}, since S and R are independent, we can condition on R, so that Dv is fixed with respect to the randomness of S. Applying Lemma D.24 in (Song et al., 2023) to u and Dv, we have that with probability $1 - \frac{\delta}{6}$,

$$|T_{CD}| \leq \|u\|\|Dv\| \frac{\log^{1.5}(6d/\delta)}{\sqrt{b_{\text{out}}}}, \tag{12}$$

and from Lemma F.1, with probability at least $1 - \frac{\delta}{6}$,

$$\|Dv\| = \|(R^\top R - I)v\| \leq 8\|v\| \sqrt{\frac{d}{b_{\text{in}}}} \log(24/\delta). \tag{13}$$

Combining (12) and (13), with probability at least $1 - \frac{\delta}{3}$,

$$\begin{aligned} |T_{CD}| &\leq \|u\|\|Dv\| \frac{\log^{1.5}(6d/\delta)}{\sqrt{b_{\text{out}}}} \\ &\leq \|u\| \cdot 8\|v\| \sqrt{\frac{d}{b_{\text{in}}}} \log(24/\delta) \cdot \frac{\log^{1.5}(6d/\delta)}{\sqrt{b_{\text{out}}}} \\ &= \|u\|\, \|v\|\, 8 \log^{1.5}(6d/\delta) \log(24/\delta) \sqrt{\frac{d}{b_{\text{in}} b_{\text{out}}}}. \end{aligned} \tag{14}$$

Taking (10), (11) and (14) together, with probability at least $1 - \delta$,

$$\begin{aligned} &\left|u^\top S^\top S R^\top R v - u^\top v\right| \\ =& |T_C + T_D + T_{CD}| \leq |T_C| + |T_D| + |T_{CD}| \\ \leq& \|u\|\|v\| \frac{\log^{1.5}(3d/\delta)}{\sqrt{b_{\text{out}}}} + \|u\|\|v\| \frac{\log^{1.5}(3d/\delta)}{\sqrt{b_{\text{in}}}} \\ &+ \|u\|\, \|v\|\, 8 \log^{1.5}(6d/\delta) \log(24/\delta) \sqrt{\frac{d}{b_{\text{in}} b_{\text{out}}}} \\ \leq& \|u\|\, \|v\|\, 10 \log^{1.5}(6d/\delta) \log(24/\delta) \sqrt{\frac{d}{b_{\text{in}} b_{\text{out}}}}, \end{aligned}$$

where the last inequality uses $b_{\text{in}} \leq d$ and $b_{\text{out}} \leq d$. □

G Assumptions for Optimization Analysis

We state several standard assumptions that are commonly used in the analysis of first-order stochastic optimization methods.

Assumption G.1 (Bounded Loss Gradients)**.** There exists a constant $G \geq 0$ such that, for all $\theta \in \mathbb{R}^p$ and all $c \in [C]$, we have $\|\nabla \mathcal{L}_c(\theta)\|_2 \leq G$.

Since $\nabla \mathcal{L}(\theta) = \frac{1}{C}\sum_{c=1}^{C} \nabla \mathcal{L}_c(\theta)$, it follows immediately that $\|\nabla \mathcal{L}(\theta)\|_2 \leq G$. Hence, the same constant G upper bounds both the client-level and the global gradient norms. We additionally impose a standard sub-Gaussian condition on the stochastic gradient noise arising from mini-batch sampling, as is typical in first-order optimization analyses (Harvey et al., 2019; Mou et al., 2020).

Assumption G.2 (Sub-Gaussian Noise)**.** For each client $c \in [C]$, the stochastic noise $|\nabla \mathcal{L}_c(x) - g_c(x)|_2$ is σ_s-sub-Gaussian; that is, for all $a \geq 0$,

$$\mathbb{P}\left(\|\nabla \mathcal{L}_c(x) - g_c(x)\|_2 \geq a\right) \leq 2\exp\left(-a^2/\sigma_s^2\right).$$

We further assume regularity of the curvature through conditions on the Hessian eigenspectrum. In particular, we require each local objective to be smooth.

Assumption G.3 (Bounded Loss Hessian Eigenvalues)**.** For every $c \in [C]$, the largest eigenvalue of the Hessian $H_{\mathcal{L}_c}$ is at most L, i.e., $\mathcal{L}_c$ is L-smooth.

This local smoothness condition is standard in federated learning theory (Safaryan et al., 2021; Fatkhullin et al., 2024), and by the same averaging argument as above, it also implies that the global objective $\mathcal{L}$ is L-smooth.

Finally, we assume that the Hessian of the averaged loss has bounded (absolute) intrinsic dimension. Let $\{\lambda_i\}_{i=1}^d$ denote the eigenvalues of $H\mathcal{L}$.

Assumption G.4 (Loss Hessian Eigenspectrum)**.** The absolute intrinsic dimension of $H_{\mathcal{L}}$ is bounded:

$$\frac{\sum_{i=1}^{d} |\lambda_i|}{\max_i \lambda_i} \leq \mathcal{I}.$$

The quantity $\sum_{i=1}^{d} |\lambda_i| / \max_i \lambda_i$ closely parallels the intrinsic-dimension notion of Ipsen and Saibaba (2024), with the modification that we take absolute values of the eigenvalues. Empirical evidence in deep learning suggests that this absolute intrinsic dimension can be far smaller than the ambient dimension d. Prior work reports a rapidly decaying Hessian spectrum (Ghorbani et al., 2019; Li et al., 2020c; Liu et al., 2023b), and also that many eigenvalues concentrate near zero (Sagun et al., 2016; Liao and Mahoney, 2021). Moreover, observations of power-law spectral behavior (Xie et al., 2022; Zhang et al., 2024) further indicate that, under such decay, the resulting absolute intrinsic dimension is typically of a smaller order than d.

H Optimization Analysis with GD as GLOBAL_OPT

First we write down Algorithm 3, which will be called as GLOBAL_OPT each round to do a one-step GD update of the global parameter θ_{t-1} with desketched aggregated updates $\mathrm{desk}\left(\bar{\tilde{\Delta}}_{t-1}\right)$.

Algorithm 3 GLOBAL_OPT (GD)

Inputs: model θ_{t-1}, desketched update $\mathrm{desk}\left(\bar{\tilde{\Delta}}_{t-1}\right)$.

Output: model θ_T.

One-step GD: $\theta_t \leftarrow \theta_{t-1} - \eta_{\text{global}} \cdot \mathrm{desk}(\bar{\tilde{\Delta}}_{t-1})$.

We next present the optimization guarantee for Algorithm 2 in conjunction with Algorithm 3. Our analysis follows a similar approach to (Song et al., 2023), but crucially leverages our new concentration lemma (Lemma F.3) for two sketches. We now state the formal result. "'latex

Theorem H.1. *Suppose $\{\theta_t\}_{t=0}^T$ is generated by Algorithm 2 with Algorithm 3 as* `GLOBAL_OPT`*. Denote $\mathcal{L}^*$ the minimum of the average empirical loss. Under Assumption G.1-G.4, with learning rate $\eta = \eta_{global}\eta_{local}$, we have that with probability at least $1 - 12\delta$,*

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\theta_t)\|_2^2 \leq \frac{\mathcal{L}(\theta_0)-\mathcal{L}^*}{\eta KT} + \frac{2\log(2T/\delta)G^2}{\sqrt{NT}} + \frac{\eta_{local}LKG^2}{2} + \frac{\sqrt{2}G\log(2T/\delta)\sigma_s}{\sqrt{NTK}} + \max\left\{0, \frac{G(KG-\tau)}{K}\right\} + \frac{15\sqrt{d}\log^3(24NTd/\delta)G\tau_{K,G}}{K\sqrt{Tb_{in}b_{out}}} + \frac{\log^2(2T/\delta)G^2}{\eta TK} + \frac{2\eta\log^2(2T/\delta)\sigma_g^2\alpha_2^2}{NK} + \frac{\eta\alpha_1^2\mathcal{I}L\tau_{K,G}^2}{K} + \frac{\eta\sigma_g^2\alpha_2^2\mathcal{I}L\log^2(2dT/\delta)}{NK},$$

in which

$$\alpha_1 = 1 + \frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{in}b_{out}}}, \alpha_2 = 1 + 8\sqrt{\frac{d}{b_{out}}}\log\left(\frac{4dT}{\delta}\right), \tau_{K,G} = \min\{\tau, KG\}.$$

Proof. According to the algorithm, we can write the update in the sync step as:

$$\begin{aligned}
\theta_{t+1} - \theta_t &= -\eta_{\text{global}} S_t^\top \bar{\tilde{\Delta}}_t \\
&= -\eta_{\text{global}} S_t^\top \left(\frac{1}{N}\sum_{c\in\mathcal{C}_t}\tilde{\Delta}_{c,t}\right) \\
&= -\eta_{\text{global}} S_t^\top \left(\frac{1}{N}\sum_{c\in\mathcal{C}_t} S_t\breve{\Delta}_{c,t}\right) \\
&= -\eta_{\text{global}} S_t^\top \left(\frac{1}{N}\sum_{c\in\mathcal{C}_t} S_t R_{c,t}^\top\breve{\Delta}_{c,t}\right) \\
&= -\eta_{\text{global}} S_t^\top \left(\eta_{\text{local}}\frac{1}{N}\sum_{c\in\mathcal{C}_t}\left(S_t R_{c,t}^\top\left(R_{c,t}\left(\text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}},\tau\right)\right)+\mathbf{z}_{c,t}\right)\right)\right) \\
&= -\eta\left(\frac{1}{N}\sum_{i\in\mathcal{C}_t}\left[S_t^\top S_t R_{c,t}^\top R_{c,t}\left(\text{clip}\left(\sum_{k=1}^{K} g_{c,t,k},\tau\right)\right)+S_t^\top S_t R_{c,t}^\top\mathbf{z}_{c,t}\right]\right) \\
&= -\frac{\eta}{N}\sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^{K} g_{c,t,k},\tau\right) - \frac{\eta}{N}\sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top\mathbf{z}_{c,t}
\end{aligned}$$

in which $\eta = \eta_{\text{global}}\eta_{\text{local}}$. By Taylor expansion, we have

$$\mathcal{L}(\theta_{t+1}) = \mathcal{L}(\theta_t) + \nabla\mathcal{L}(\theta_t)^\top(\theta_{t+1}-\theta_t) + \frac{1}{2}(\theta_{t+1}-\theta_t)^\top\hat{H}_{\mathcal{L},t}(\theta_{t+1}-\theta_t)$$

By taking summation from 0 to $T-1$, we can get that

$$\mathcal{L}(\theta_T) - \mathcal{L}(\theta_0) = \sum_{t=1}^{T}(\mathcal{L}(\theta_{t+1})-\mathcal{L}(\theta_t)) = \underbrace{\sum_{t=1}^{T}\nabla\mathcal{L}(\theta_t)^\top(\theta_{t+1}-\theta_t)}_{T_1} + \underbrace{\frac{1}{2}\sum_{t=1}^{T}(\theta_{t+1}-\theta_t)^\top\hat{H}_{\mathcal{L},t}(\theta_{t+1}-\theta_t)}_{T_2} \tag{15}$$

H.1 Bounding T_1

For each term in T_1, we have

$$\begin{aligned}
&\nabla\mathcal{L}(\theta_t)^\top(\theta_{t+1}-\theta_t) \\
=& -\nabla\mathcal{L}(\theta_t)^\top\left(\frac{\eta}{N}\sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^{K} g_{c,t,k},\tau\right) + \frac{\eta}{N}\sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top\mathbf{z}_{c,t}\right)
\end{aligned}$$

$$= -\frac{\eta}{N}\nabla\mathcal{L}(\theta_t)^\top \sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top R_{c,t} \text{clip}\left(\sum_{k=1}^K g_{c,t,k}, \tau\right) - \frac{\eta}{N}\nabla\mathcal{L}(\theta_t)^\top \sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top \mathbf{z}_{c,t}$$

By taking summation from 0 to $T-1$, we can get that

$$\begin{aligned}\sum_{t=1}^T \nabla\mathcal{L}(\theta_t)^\top(\theta_{t+1}-\theta_t) = &-\underbrace{\eta\sum_{t=1}^T \frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top \sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)}_{S_1} \\ &-\underbrace{\eta\sum_{t=1}^T \frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top \sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top \mathbf{z}_{c,t}}_{S_2} \end{aligned} \tag{16}$$

H.1.1 Bounding S_1

For each term in S_1, we have that

$$\begin{aligned}&\frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top \sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right) \\ =&K\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{C}\sum_{c=1}^C \nabla\mathcal{L}_c(\theta_t)\right\rangle + K\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\nabla\mathcal{L}_c(\theta_t) - \frac{1}{C}\sum_{c=1}^C\nabla\mathcal{L}_c(\theta_t)\right\rangle \\ &+\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^K(\nabla\mathcal{L}_c(\theta_{c,t,k}) - \nabla\mathcal{L}_c(\theta_t))\right\rangle + \left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^K(g_{c,t,k} - \nabla\mathcal{L}_c(\theta_{c,t,k}))\right\rangle \\ &+\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\left(\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right) - \sum_{k=1}^K g_{c,t,k}\right)\right\rangle \\ &+\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\left(S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right) - \text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)\right)\right\rangle\end{aligned}$$

By taking summation from 0 to $T-1$, we can get that

$$\begin{aligned}&\sum_{t=1}^T \frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top \sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right) \\ =&K\underbrace{\sum_{t=1}^T\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{C}\sum_{c=1}^C\nabla\mathcal{L}_c(\theta_t)\right\rangle}_{Y_1} + K\underbrace{\sum_{t=1}^T\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\nabla\mathcal{L}_c(\theta_t) - \frac{1}{C}\sum_{c=1}^C\nabla\mathcal{L}_c(\theta_t)\right\rangle}_{Y_2} \\ &+\underbrace{\sum_{t=1}^T\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^K(\nabla\mathcal{L}_c(\theta_{c,t,k}) - \nabla\mathcal{L}_c(\theta_t))\right\rangle}_{Y_3} + \underbrace{\sum_{t=1}^T\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^K(g_{c,t,k} - \nabla\mathcal{L}_c(\theta_{c,t,k}))\right\rangle}_{Y_4} \\ &+\underbrace{\sum_{t=1}^T\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\left(\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right) - \sum_{k=1}^K g_{c,t,k}\right)\right\rangle}_{Y_5} \\ &+\underbrace{\sum_{t=1}^T\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\left(S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right) - \text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)\right)\right\rangle}_{Y_6}\end{aligned} \tag{17}$$

H.1.1.1 Bounding Y_1

According to the definition,

$$\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_t) \right\rangle = \|\nabla\mathcal{L}(\theta_t)\|_2^2$$

so

$$\sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_t) \right\rangle = \sum_{t=1}^{T}\|\nabla\mathcal{L}(\theta_t)\|_2^2 \tag{18}$$

H.1.1.2 Bounding Y_2

We first bound each term with a fixed $t \in [T]$ in Y_2. According to the assumption, each $c \in \mathcal{C}_t$ is uniformly randomly selected from $[C]$, so by Hoeffding's inequality, we have

$$\begin{aligned}&\mathbb{P}\left(\left|\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\nabla\mathcal{L}_c(\theta_t) - \frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_t) \right\rangle\right| \geq a\right)\\ \leq& 2\exp\left(-\frac{2Na^2}{(2G^2)^2}\right) = 2\exp\left(-\frac{Na^2}{2G^4}\right)\end{aligned}$$

By selecting $a = \frac{\sqrt{2\log(2T/\delta)}G^2}{\sqrt{N}}$, we have that with probability at least $1 - \frac{\delta}{T}$,

$$\left|\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\nabla\mathcal{L}_c(\theta_t) - \frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_t) \right\rangle\right| \leq \frac{\sqrt{2\log(2T/\delta)}G^2}{\sqrt{N}}$$

Then denote $Z_t = \sum_{\tau'=0}^{t}\left\langle \nabla\mathcal{L}(\theta_{\tau'}), \frac{1}{N}\sum_{c\in\mathcal{C}_{\tau'}}\nabla\mathcal{L}_c(\theta_{\tau'}) - \frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_{\tau'}) \right\rangle$, we can see that Z_t is a martingale with respect to the selection each round, and from the above analysis, we have that with probability at least $1-\delta$, for all $t \in [T]$,

$$|Z_t - Z_{t-1}| = \left|\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\nabla\mathcal{L}_c(\theta_t) - \frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_t) \right\rangle\right| \leq \frac{\sqrt{2\log(2T/\delta)}G^2}{\sqrt{N}}$$

Then by Azuma's inequality, we have

$$\mathbb{P}\left(Z_{T-1} \leq -a\right) \leq \exp\left(-\frac{a^2}{2\cdot T\cdot\left(\frac{\sqrt{2\log(2T/\delta)}G^2}{\sqrt{N}}\right)^2}\right) = \exp\left(-\frac{Na^2}{4T\log(2T/\delta)G^4}\right)$$

By selecting $a = \frac{2\sqrt{T}\log(2T/\delta)G^2}{\sqrt{N}}$, we can get that with probability at least $1-2\delta$,

$$Z_{T-1} = \sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\nabla\mathcal{L}_c(\theta_t) - \frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_t) \right\rangle \geq -\frac{2\sqrt{T}\log(2T/\delta)G^2}{\sqrt{N}} \tag{19}$$

H.1.1.3 Bounding Y_3

For each term, we have

$$\begin{aligned}&\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\left(\nabla\mathcal{L}_c(\theta_{c,t,k}) - \nabla\mathcal{L}_c(\theta_t)\right) \right\rangle\\ =&\frac{1}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\left\langle \nabla\mathcal{L}(\theta_t), \hat{H}_{\mathcal{L}}^{c,t,k}\left(\theta_{c,t,k} - \theta_t\right) \right\rangle\end{aligned}$$

$$
\begin{aligned}
=&\frac{\eta_{\text{local}}}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\left\langle \nabla\mathcal{L}(\theta_t), \hat{H}_{\mathcal{L}}^{c,t,k}\sum_{\kappa=1}^{k} g_{c,t,\kappa}\right\rangle\\
\geq&-\frac{\eta_{\text{local}}}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\|\nabla\mathcal{L}(\theta_t)\|_2\cdot L\sum_{\kappa=1}^{k}\|g_{c,t,\kappa}\|_2\\
=&-\frac{\eta_{\text{local}}}{N}\cdot N\cdot G^2\cdot L\sum_{k=1}^{K}k\\
\geq&-\frac{\eta_{\text{local}}LK^2G^2}{2}
\end{aligned}
$$

so

$$
\sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\left(\nabla\mathcal{L}_c(\theta_{c,t,k})-\nabla\mathcal{L}_c(\theta_t)\right)\right\rangle \geq -\frac{\eta_{\text{local}}TLK^2G^2}{2} \tag{20}
$$

H.1.1.4 Bounding Y_4

We first bound each term with a fixed $t\in[T]$ in Y_4.

$$
\begin{aligned}
&\left|\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\left(g_{c,t,k}-\nabla\mathcal{L}_c(\theta_{c,t,k})\right)\right\rangle\right|\\
\leq&\frac{1}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\|\nabla\mathcal{L}(\theta_t)\|_2\cdot\|g_{c,t,k}-\nabla\mathcal{L}_c(\theta_{c,t,k})\|_2\\
=&\frac{G}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\|g_{c,t,k}-\nabla\mathcal{L}_c(\theta_{c,t,k})\|_2
\end{aligned}
$$

According to the assumption, the stochastic noise $\|g_{c,t,k}-\nabla\mathcal{L}_c(\theta_{c,t,k})\|_2$ is a σ_s-sub-Gaussian random variable, so by Hoeffding's inequality,

$$
\mathbb{P}\left(\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\|g_{c,t,k}-\nabla\mathcal{L}_c(\theta_{c,t,k})\|_2\geq a\right)\leq 2\exp\left(-\frac{a^2}{NK\sigma_s^2}\right)
$$

By selecting $a=\sqrt{NK\log(2T/\delta)}\sigma_s$, we have that with probability at least $1-\frac{\delta}{T}$,

$$
\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\|g_{c,t,k}-\nabla\mathcal{L}_c(\theta_{c,t,k})\|_2\leq\sqrt{NK\log(2T/\delta)}\sigma_s
$$

so

$$
\begin{aligned}
&\left|\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\left(g_{c,t,k}-\nabla\mathcal{L}_c(\theta_{c,t,k})\right)\right\rangle\right|\\
\leq&\frac{G}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\|g_{c,t,k}-\nabla\mathcal{L}_c(\theta_{c,t,k})\|_2\\
\leq&\frac{G}{N}\cdot\sqrt{NK\log(2T/\delta)}\sigma_s=\frac{G\sqrt{K\log(2T/\delta)}\sigma_s}{\sqrt{N}}
\end{aligned}
$$

Then denote $W_t=\sum_{\tau'=0}^{t}\left\langle \nabla\mathcal{L}(\theta'_\tau), \frac{1}{N}\sum_{c\in\mathcal{C}_{\tau'}}\sum_{k=1}^{K}\left(g_{c,\tau',k}-\nabla\mathcal{L}_c(\theta_{c,\tau',k})\right)\right\rangle$, we can see that W_t is a martingale with respect to the stochastic noise, and from the above analysis, we have that with probability at least $1-\delta$, for all $t\in[T]$,

$$
|W_t-W_{t-1}|=\left|\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\left(g_{c,t,k}-\nabla\mathcal{L}_c(\theta_{c,t,k})\right)\right\rangle\right|\leq\frac{G\sqrt{K\log(2T/\delta)}\sigma_s}{\sqrt{N}}
$$

Then by Azuma's inequality, we have

$$\mathbb{P}\left(W_{T-1} \leq -a\right) \leq \exp\left(-\frac{a^2}{2 \cdot T \cdot \left(\frac{G\sqrt{K\log(2T/\delta)}\sigma_s}{\sqrt{N}}\right)^2}\right) = \exp\left(-\frac{Na^2}{2TG^2K\log(2T/\delta)\sigma_s^2}\right)$$

By selecting $a = \frac{G\sqrt{2TK}\log(2T/\delta)\sigma_s}{\sqrt{N}}$, we can get that with probability at least $1-2\delta$,

$$W_{T-1} = \sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\left(g_{c,t,k} - \nabla\mathcal{L}_c(\theta_{c,t,k})\right)\right\rangle \geq -\frac{G\sqrt{2TK}\log(2T/\delta)\sigma_s}{\sqrt{N}} \quad (21)$$

H.1.1.5 Bounding Y_5

For each term, for $\tau \leq KG$, we have

$$\begin{aligned}
&\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\left(\text{clip}\left(\sum_{k=1}^{K} g_{c,t,k}, \tau\right) - \sum_{k=1}^{K} g_{c,t,k}\right)\right\rangle \\
\geq& -\frac{1}{N}\sum_{c\in\mathcal{C}_t}\left\|\nabla\mathcal{L}(\theta_t)\right\|_2 \left\|\sum_{c\in\mathcal{C}_t}\left(\text{clip}\left(\sum_{k=1}^{K} g_{c,t,k}, \tau\right) - \sum_{k=1}^{K} g_{c,t,k}\right)\right\|_2 \\
\geq& -G(KG-\tau)
\end{aligned}$$

for $\tau \geq KG$, we have

$$\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\left(\text{clip}\left(\sum_{k=1}^{K} g_{c,t,k}, \tau\right) - \sum_{k=1}^{K} g_{c,t,k}\right)\right\rangle = 0$$

so

$$\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\left(\text{clip}\left(\sum_{k=1}^{K} g_{c,t,k}, \tau\right) - \sum_{k=1}^{K} g_{c,t,k}\right)\right\rangle \geq -\max\left\{0, G(KG-\tau)\right\}$$

By taking summation from 0 to $T-1$, we can get that

$$\sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\left(\text{clip}\left(\sum_{k=1}^{K} g_{c,t,k}, \tau\right) - \sum_{k=1}^{K} g_{c,t,k}\right)\right\rangle \geq -\max\left\{0, TG(KG-\tau)\right\} \quad (22)$$

H.1.1.6 Bounding Y_6

We first bound each term with a fixed $t \in [T]$ in Y_6. According to Lemma F.3, we have that with probability at least $1-\frac{\delta}{T}$,

$$\begin{aligned}
&\left|\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\left(S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^{K} g_{c,t,k}, \tau\right) - \text{clip}\left(\sum_{k=1}^{K} g_{c,t,k}, \tau\right)\right)\right\rangle\right| \\
\leq& \frac{1}{N}\sum_{c\in\mathcal{C}_t}\frac{10\sqrt{d}\log^{1.5}(6NTd/\delta)\log(24NT/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\left\|\nabla\mathcal{L}(\theta_t)\right\|_2\left\|\text{clip}\left(\sum_{k=1}^{K} g_{c,t,k}, \tau\right)\right\|_2 \\
\leq& \frac{10\sqrt{d}\log^{1.5}(6NTd/\delta)\log(24NT/\delta)G\tau_{K,G}}{\sqrt{b_{\text{in}}b_{\text{out}}}}
\end{aligned}$$

with $\tau_{K,G} = \min\left\{\tau, KG\right\}$. Then denote

$$U_t = \sum_{\tau'=0}^{t}\left\langle \nabla\mathcal{L}(\theta_{\tau'}), \frac{1}{N}\sum_{c\in\mathcal{C}_{\tau'}}\left(R_{\tau'}^\top R_{\tau'}\text{clip}\left(\sum_{k=1}^{K} g_{c,\tau',k}, \tau\right) - \text{clip}\left(\sum_{k=1}^{K} g_{c,\tau',k}, \tau\right)\right)\right\rangle,$$

we can see that U_t is a martingale with respect to the sketching matrices, and from the above analysis, we have that with probability at least $1-\delta$, for all $t \in [T]$,

$$|U_t - U_{t-1}| = \left| \left\langle \nabla \mathcal{L}(\theta_t), \frac{1}{N} \sum_{c \in \mathcal{C}_t} \left(R_t^\top R_t \text{clip} \left(\sum_{k=1}^K g_{c,t,k}, \tau \right) - \text{clip} \left(\sum_{k=1}^K g_{c,t,k}, \tau \right) \right) \right\rangle \right|$$
$$\leq \frac{10\sqrt{d} \log^{1.5}(6NTd/\delta) \log(24NT/\delta) G \tau_{K,G}}{\sqrt{b_{\text{in}} b_{\text{out}}}}$$

Then by Azuma's inequality, we have

$$\mathbb{P}\left(U_{T-1} \leq -a\right) \leq \exp\left(- \frac{a^2}{2 \cdot T \cdot \left(\frac{10\sqrt{d} \log^{1.5}(6NTd/\delta) \log(24NT/\delta) G \tau_{K,G}}{\sqrt{b_{\text{in}} b_{\text{out}}}} \right)^2} \right)$$
$$= \exp\left(- \frac{b_{\text{in}} b_{\text{out}} a^2}{200 T d \log^3(6NTd/\delta) \log^2(24NT/\delta) G^2 \tau_{K,G}^2} \right)$$

By selecting $a = \frac{15\sqrt{d} \log^3(24NTd/\delta)\sqrt{T} G \tau_{K,G}}{\sqrt{b_{\text{in}} b_{\text{out}}}}$, we can get that with probability at least $1-2\delta$,

$$W_{T-1} = \sum_{t=1}^T \left\langle \nabla \mathcal{L}(\theta_t), \frac{1}{N} \sum_{c \in \mathcal{C}_t} \left(S_t^\top S_t R_{c,t}^\top R_{c,t} \text{clip} \left(\sum_{k=1}^K g_{c,t,k}, \tau \right) - \text{clip} \left(\sum_{k=1}^K g_{c,t,k}, \tau \right) \right) \right\rangle$$
$$\geq - \frac{15\sqrt{d} \log^3(24NTd/\delta)\sqrt{T} G \tau_{K,G}}{\sqrt{b_{\text{in}} b_{\text{out}}}} \tag{23}$$

Substituting (18), (19), (20), (21), (22), (23) into (17), we have that with probability at least $1-6\delta$,

$$\sum_{t=1}^T \frac{1}{N} \nabla \mathcal{L}(\theta_t)^\top \sum_{c \in \mathcal{C}_t} S_t^\top S_t R_{c,t}^\top R_{c,t} \text{clip} \left(\sum_{k=1}^K g_{c,t,k}, \tau \right)$$
$$= K \underbrace{\sum_{t=1}^T \left\langle \nabla \mathcal{L}(\theta_t), \frac{1}{C} \sum_{c=1}^C \nabla \mathcal{L}_c(\theta_t) \right\rangle}_{Y_1} + K \underbrace{\sum_{t=1}^T \left\langle \nabla \mathcal{L}(\theta_t), \frac{1}{N} \sum_{c \in \mathcal{C}_t} \nabla \mathcal{L}_c(\theta_t) - \frac{1}{C} \sum_{c=1}^C \nabla \mathcal{L}_c(\theta_t) \right\rangle}_{Y_2}$$
$$+ \underbrace{\sum_{t=1}^T \left\langle \nabla \mathcal{L}(\theta_t), \frac{1}{N} \sum_{c \in \mathcal{C}_t} \sum_{k=1}^K \left(\nabla \mathcal{L}_c(\theta_{c,t,k}) - \nabla \mathcal{L}_c(\theta_t) \right) \right\rangle}_{Y_3} + \underbrace{\sum_{t=1}^T \left\langle \nabla \mathcal{L}(\theta_t), \frac{1}{N} \sum_{c \in \mathcal{C}_t} \sum_{k=1}^K \left(g_{c,t,k} - \nabla \mathcal{L}_c(\theta_{c,t,k}) \right) \right\rangle}_{Y_4}$$
$$+ \underbrace{\sum_{t=1}^T \left\langle \nabla \mathcal{L}(\theta_t), \frac{1}{N} \sum_{c \in \mathcal{C}_t} \left(\text{clip} \left(\sum_{k=1}^K g_{c,t,k}, \tau \right) - \sum_{k=1}^K g_{c,t,k} \right) \right\rangle}_{Y_5}$$
$$+ \underbrace{\sum_{t=1}^T \left\langle \nabla \mathcal{L}(\theta_t), \frac{1}{N} \sum_{c \in \mathcal{C}_t} \left(S_t^\top S_t R_{c,t}^\top R_{c,t} \text{clip} \left(\sum_{k=1}^K g_{c,t,k}, \tau \right) - \text{clip} \left(\sum_{k=1}^K g_{c,t,k}, \tau \right) \right) \right\rangle}_{Y_6}$$
$$\geq K \sum_{t=1}^T \|\nabla \mathcal{L}(\theta_t)\|_2^2 - \frac{2K\sqrt{T} \log(2T/\delta) G^2}{\sqrt{N}} - \frac{\eta_{\text{local}} T L K^2 G^2}{2} - \frac{G\sqrt{2TK} \log(2T/\delta) \sigma_s}{\sqrt{N}}$$
$$- \max\left\{0, TG(KG - \tau)\right\} - \frac{15\sqrt{d} \log^3(24NTd/\delta)\sqrt{T} G \tau_{K,G}}{\sqrt{b_{\text{in}} b_{\text{out}}}} \tag{24}$$

H.1.2 BOUNDING S_2

We first bound each term with a fixed $t \in [T]$ in S_2. Noticing that conditioned on S_t, $\mathbf{z}_{c,t} \sim \mathcal{N}\left(0, \sigma_g^2 \mathbb{I}\right)$, and $R_{c,t} S_t^\top S_t \nabla \mathcal{L}(\theta_t)$ is a $\frac{\left\| S_t^\top S_t \nabla \mathcal{L}(\theta_t) \right\|_2}{\sqrt{b_{\text{in}}}}$-sub-Gaussian random vector, so according to

Bernstein inequality,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{c\in\mathcal{C}_t}\left\langle R_{c,t}S_t^\top S_t\nabla\mathcal{L}(\theta_t),\mathbf{z}_{c,t}\right\rangle\right|\geq a\right)\leq 2\exp\left(-\min\left(\frac{Na^2}{b_{\text{in}}\cdot\sigma_g^2\cdot\frac{\left\|S_t^\top S_t\nabla\mathcal{L}(\theta_t)\right\|_2^2}{b_{\text{in}}}},\frac{\sqrt{N}a}{\sigma_g\cdot\frac{\left\|S_t^\top S_t\nabla\mathcal{L}(\theta_t)\right\|_2}{\sqrt{b_{\text{in}}}}}\right)\right)$$
$$=2\exp\left(-\min\left(\frac{Na^2}{\sigma_g^2\left\|S_t^\top S_t\nabla\mathcal{L}(\theta_t)\right\|_2^2},\frac{a\sqrt{b_{\text{in}}N}}{\sigma_g\left\|S_t^\top S_t\nabla\mathcal{L}(\theta_t)\right\|_2}\right)\right)$$

so taking $a=\frac{\sigma_g\left\|S_t^\top S_t\nabla\mathcal{L}(\theta_t)\right\|_2\log(2T/\delta)}{\sqrt{N}}$, we have that with probability at least $1-\frac{\delta}{T}$,

$$\left|\frac{1}{N}\sum_{c\in\mathcal{C}_t}\left\langle R_{c,t}S_t^\top S_t\nabla\mathcal{L}(\theta_t),\mathbf{z}_{c,t}\right\rangle\right|\leq\frac{\sigma_g\left\|S_t^\top S_t\nabla\mathcal{L}(\theta_t)\right\|_2\log(2T/\delta)}{\sqrt{N}}$$

According to Lemma F.1, we can see that with probability at least $1-\frac{\delta}{T}$,

$$\left\|S_t^\top S_t\nabla\mathcal{L}(\theta_t)\right\|\leq\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4T}{\delta}\right)\right)\|\nabla\mathcal{L}(\theta_t)\|\leq\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4T}{\delta}\right)\right)G,$$

so with probability at least $1-\frac{2\delta}{T}$,

$$\left|\frac{1}{N}\sum_{c\in\mathcal{C}_t}\left\langle R_{c,t}S_t^\top S_t\nabla\mathcal{L}(\theta_t),\mathbf{z}_{c,t}\right\rangle\right|\leq\frac{\sigma_g\left\|S_t^\top S_t\nabla\mathcal{L}(\theta_t)\right\|_2\log(2T/\delta)}{\sqrt{N}}\leq\frac{\sigma_g\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4T}{\delta}\right)\right)G\log(2T/\delta)}{\sqrt{N}}.$$

Then denote $X_t=\sum_{\tau'=0}^{t}\frac{1}{N}\sum_{c\in\mathcal{C}_{\tau'}}\left\langle R_{c,\tau'}S_{\tau'}^\top S_{\tau'}\nabla\mathcal{L}(\theta_{\tau'}),\mathbf{z}_{c,\tau'}\right\rangle$, we can see that X_t is a martingale with respect to the Gaussian noise, and from the above analysis, we have that with probability at least $1-2\delta$, for all $t\in[T]$,

$$|X_t-X_{t-1}|=\left|\frac{1}{N}\sum_{c\in\mathcal{C}_t}\left\langle R_{c,t}S_t^\top S_t\nabla\mathcal{L}(\theta_t),\mathbf{z}_{c,t}\right\rangle\right|\leq\frac{\sigma_g\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4T}{\delta}\right)\right)G\log(2T/\delta)}{\sqrt{N}}$$

Then by Azuma's inequality, we have

$$\mathbb{P}\left(X_{T-1}\leq-a\right)\leq\exp\left(-\frac{a^2}{2\cdot T\cdot\left(\frac{\sigma_g\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4T}{\delta}\right)\right)G\log(2T/\delta)}{\sqrt{N}}\right)^2}\right)$$
$$=\exp\left(-\frac{Na^2}{2T\sigma_g^2G^2\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4T}{\delta}\right)\right)^2\log^2(2T/\delta)}\right)$$

By selecting $a=\frac{\log^2(2T/\delta)\sqrt{2T}\sigma_gG\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4T}{\delta}\right)\right)}{\sqrt{N}}$, we can get that with probability at least $1-3\delta$,

$$X_{T-1}=\sum_{t=1}^{T}\frac{1}{N}\sum_{c\in\mathcal{C}_t}\left\langle R_{c,t}S_t^\top S_t\nabla\mathcal{L}(\theta_t),\mathbf{z}_{c,t}\right\rangle\geq-\frac{\log^2(2T/\delta)\sqrt{2T}\sigma_gG\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4T}{\delta}\right)\right)}{\sqrt{N}}\tag{25}$$

Substituting (24) and (25) into (16), we have that with probability at least $1-9\delta$,

$$\sum_{t=1}^{T}\nabla\mathcal{L}(\theta_t)^\top(\theta_{t+1}-\theta_t)$$

$$
\begin{aligned}
&= -\eta K \sum_{t=1}^{T} \|\nabla \mathcal{L}(\theta_t)\|_2^2 + \frac{2\eta K\sqrt{T}\log(2T/\delta)G^2}{\sqrt{N}} + \frac{\eta\eta_{\text{local}} TLK^2G^2}{2} + \frac{\eta G\sqrt{2TK}\log(2T/\delta)\sigma_s}{\sqrt{N}} \\
&\quad + \eta\max\{0, TG(KG-\tau)\} + \frac{15\eta\sqrt{d}\log^3(24NTd/\delta)\sqrt{T}G\tau_{K,G}}{\sqrt{b_{\text{in}}b_{\text{out}}}} + \frac{\eta\log^2(2T/\delta)\sqrt{2T}\sigma_g G\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4T}{\delta}\right)\right)}{\sqrt{N}} \\
&\le -\eta K \sum_{t=1}^{T} \|\nabla \mathcal{L}(\theta_t)\|_2^2 + \frac{2\eta K\sqrt{T}\log(2T/\delta)G^2}{\sqrt{N}} + \frac{\eta\eta_{\text{local}} TLK^2G^2}{2} + \frac{\eta G\sqrt{2TK}\log(2T/\delta)\sigma_s}{\sqrt{N}} \\
&\quad + \eta\max\{0, TG(KG-\tau)\} + \frac{15\eta\sqrt{d}\log^3(24NTd/\delta)\sqrt{T}G\tau_{K,G}}{\sqrt{b_{\text{in}}b_{\text{out}}}} + \log^2(2T/\delta)G^2 \\
&\quad + \frac{2\eta^2 T\log^2(2T/\delta)\sigma_g^2\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4T}{\delta}\right)\right)^2}{N}
\end{aligned} \tag{26}
$$

H.2 Bounding T_2

We first bound each term in T_2 with a fixed $t \in [T]$.

$$
\begin{aligned}
(\theta_{t+1}-\theta_t)^\top \hat{H}_{\mathcal{L},t}(\theta_{t+1}-\theta_t) &= \eta_{\text{global}}^2 \left(\frac{1}{N}\sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top \check{\Delta}_{c,t}\right)^\top \left(\sum_{i=1}^{d}\lambda_i v_i v_i^\top\right)\left(\frac{1}{N}\sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top \check{\Delta}_{c,t}\right) \\
&= \eta_{\text{global}}^2 \sum_{i=1}^{d}\lambda_i \left|\left(\frac{1}{N}\sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top \check{\Delta}_{c,t}\right)^\top v_i\right|^2
\end{aligned} \tag{27}
$$

For each $i \in [d]$, we have

$$
\begin{aligned}
&\left|\left(\frac{1}{N}\sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top \check{\Delta}_{c,t}\right)^\top v_i\right| \\
=&\left|\left\langle \frac{1}{N}\sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top \check{\Delta}_{c,t}, v_i\right\rangle\right| \\
=&\frac{1}{N}\left|\left\langle \sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top \check{\Delta}_{c,t}, v_i\right\rangle\right| \\
=&\frac{1}{N}\left|\left\langle \sum_{c\in\mathcal{C}_t} \eta_{\text{local}} S_t^\top S_t R_{c,t}^\top \left(R_{c,t}\left(\text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}},\tau\right)\right)+\mathbf{z}_{c,t}\right), v_i\right\rangle\right| \\
\le&\frac{\eta_{\text{local}}}{N}\left|\left\langle \sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^{K} g_{c,t,k},\tau\right), v_i\right\rangle\right| + \frac{\eta_{\text{local}}}{N}\left|\left\langle \sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top \mathbf{z}_{c,t}, v_i\right\rangle\right|
\end{aligned} \tag{28}
$$

For the first term, according to Lemma F.3, with probability at least $1-\frac{\delta}{dT}$,

$$
\begin{aligned}
&\left|\left\langle \sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^{K} g_{c,t,k},\tau\right), v_i\right\rangle\right| \\
\le&\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\|v_i\|_2 \sum_{c\in\mathcal{C}_t}\left\|\text{clip}\left(\sum_{k=1}^{K} g_{c,t,k},\tau\right)\right\| \\
\le&\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)N\tau_{K,G}
\end{aligned} \tag{29}
$$

For the second term, $\left\langle \frac{1}{N}\sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top \mathbf{z}_{c,t}, v_i\right\rangle = \frac{1}{N}\sum_{c\in\mathcal{C}_t}\left\langle R_{c,t}S_t^\top S_t v_i, \mathbf{z}_{c,t}\right\rangle$. Noticing that $\mathbf{z}_{c,t} \sim \mathcal{N}\left(0,\sigma_g^2\mathbb{I}\right)$, and $R_{c,t}S_t^\top S_t v_i$ is a $\frac{\left\|S_t^\top S_t v_i\right\|_2}{\sqrt{b_{\text{in}}}}$-sub-Gaussian random vector, so according to

Bernstein inequality,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{c\in\mathcal{C}_t}\left\langle R_{c,t}S_t^\top S_t v_i, \mathbf{z}_{c,t}\right\rangle\right| \geq a\right) \leq 2\exp\left(-c\min\left(\frac{Na^2}{\sigma_g^2\cdot\left\|S_t^\top S_t v_i\right\|_2^2}, \frac{\sqrt{N}a}{\frac{\sigma_g\left\|S_t^\top S_t v_i\right\|_2}{\sqrt{b_{\text{in}}}}}\right)\right)$$
$$= 2\exp\left(-c\min\left(\frac{Na^2}{\sigma_g^2\left\|S_t^\top S_t v_i\right\|_2^2}, \frac{a\sqrt{b_{\text{in}}N}}{\sigma_g\left\|S_t^\top S_t v_i\right\|_2}\right)\right)$$

so taking $a = \frac{\sigma_g\left\|S_t^\top S_t v_i\right\|_2\log(2dT/\delta)}{\sqrt{N}}$, we have that with probability at least $1-\frac{\delta}{dT}$,

$$\left|\left\langle\frac{1}{N}\sum_{c\in\mathcal{C}_t}S_t^\top S_t R_{c,t}^\top\mathbf{z}_{c,t}, v_i\right\rangle\right| \leq \frac{\sigma_g\left\|S_t^\top S_t v_i\right\|_2\log(2dT/\delta)}{\sqrt{N}}$$

According to Lemma F.1, we can see that with probability at least $1-\frac{\delta}{dT}$,

$$\left\|S_t^\top S_t v_i\right\| \leq \left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)\|v_i\| = 1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)$$

Therefore, with probability at least $1-\frac{2\delta}{dT}$,

$$\left|\left\langle\frac{1}{N}\sum_{c\in\mathcal{C}_t}S_t^\top S_t R_{c,t}^\top\mathbf{z}_{c,t}, v_i\right\rangle\right| \leq \frac{\sigma_g\left\|S_t^\top S_t v_i\right\|_2\log(2dT/\delta)}{\sqrt{N}} \leq \frac{\sigma_g\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)\log(2dT/\delta)}{\sqrt{N}} \tag{30}$$

Substituting (29) and (30) into (28), we have that with probability at least $1-\frac{3\delta}{dT}$,

$$\begin{aligned}
&\left|\left(\frac{1}{N}\sum_{c\in\mathcal{C}_t}S_t^\top S_t R_{c,t}^\top\check{\Delta}_{c,t}\right)^\top v_i\right| \\
\leq&\frac{\eta_{\text{local}}}{N}\left|\left\langle\sum_{c\in\mathcal{C}_t}S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right), v_i\right\rangle\right| + \frac{\eta_{\text{local}}}{N}\left|\left\langle\sum_{c\in\mathcal{C}_t}S_t^\top S_t R_{c,t}^\top\mathbf{z}_{c,t}, v_i\right\rangle\right| \\
\leq&\eta_{\text{local}}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\tau_{K,G} + \frac{\eta_{\text{local}}\sigma_g\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)\log(2dT/\delta)}{\sqrt{N}}
\end{aligned}$$

which implies that

$$\begin{aligned}
&\left|\left(\frac{1}{N}\sum_{c\in\mathcal{C}_t}S_t^\top S_t R_{c,t}^\top\check{\Delta}_{c,t}\right)^\top v_i\right|^2 \\
\leq&2\eta_{\text{local}}^2\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2\tau_{K,G}^2 + \frac{2\eta_{\text{local}}^2\sigma_g^2\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2\log^2(2dT/\delta)}{N}
\end{aligned} \tag{31}$$

Substituting (31) into (27), we have that with probability at least $1-\frac{3\delta}{T}$,

$$\begin{aligned}
&(\theta_{t+1}-\theta_t)^\top\hat{H}_{\mathcal{L},t}(\theta_{t+1}-\theta_t) \\
=&\eta_{\text{global}}^2\sum_{i=1}^d\lambda_i\left|\left(\frac{1}{N}\sum_{c\in\mathcal{C}_t}S_t^\top S_t R_{c,t}^\top\check{\Delta}_{c,t}\right)^\top v_i\right|^2 \\
\leq&\eta_{\text{global}}^2\cdot 2\eta_{\text{local}}^2\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2\tau_{K,G}^2\cdot\sum_{i=1}^d|\lambda_i|
\end{aligned}$$

$$+\eta_{\text{global}}^2 \cdot \frac{2\eta_{\text{local}}^2\sigma_g^2\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2\log^2(2dT/\delta)}{N}\cdot\sum_{i=1}^{d}|\lambda_i|$$

$$=2\eta^2\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2\mathcal{I}L\tau_{K,G}^2+\frac{2\eta^2\sigma_g^2\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2\mathcal{I}L\log^2(2dT/\delta)}{N}$$

By taking summation from 0 to $T-1$, we have that with probability at least $1-3\delta$,

$$\sum_{t=1}^{T}(\theta_{t+1}-\theta_t)^\top\hat{H}_{\mathcal{L},t}(\theta_{t+1}-\theta_t) \tag{32}$$

$$\leq 2\eta^2\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2\mathcal{I}LT\tau_{K,G}^2+\frac{2\eta^2\sigma_g^2\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2\mathcal{I}LT\log^2(2dT/\delta)}{N} \tag{33}$$

Substituting (26) and (33) into (15), we have that with probability at least $1-12\delta$,

$$\begin{aligned}
&\mathcal{L}(\theta_T)-\mathcal{L}(\theta_0)\\
=&\sum_{t=1}^{T}\nabla\mathcal{L}(\theta_t)^\top(\theta_{t+1}-\theta_t)+\frac{1}{2}\sum_{t=1}^{T}(\theta_{t+1}-\theta_t)^\top\hat{H}_{\mathcal{L},t}(\theta_{t+1}-\theta_t)\\
\leq&-\eta K\sum_{t=1}^{T}\|\nabla\mathcal{L}(\theta_t)\|_2^2+\frac{2\eta K\sqrt{T}\log(2T/\delta)G^2}{\sqrt{N}}+\frac{\eta\eta_{\text{local}}TLK^2G^2}{2}+\frac{\eta G\sqrt{2TK}\log(2T/\delta)\sigma_s}{\sqrt{N}}\\
&+\eta\max\{0,TG(KG-\tau)\}+\frac{15\eta\sqrt{d}\log^3(24NTd/\delta)\sqrt{T}G\tau_{K,G}}{\sqrt{b_{\text{in}}b_{\text{out}}}}+\log^2(2T/\delta)G^2\\
&+\frac{2\eta^2T\log^2(2T/\delta)\sigma_g^2\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4T}{\delta}\right)\right)^2}{N}\\
&+\eta^2\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2\mathcal{I}LT\tau_{K,G}^2\\
&+\frac{\eta^2\sigma_g^2\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2\mathcal{I}LT\log^2(2dT/\delta)}{N}
\end{aligned}$$

Since

$$\mathcal{L}(\theta_0)-\mathcal{L}(\theta_T)\leq\mathcal{L}(\theta_0)-\mathcal{L}^*$$

we can get that with probability at least $1-12\delta$,

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\theta_t)\|_2^2\leq&\frac{\mathcal{L}(\theta_0)-\mathcal{L}^*}{\eta KT}+\frac{2\log(2T/\delta)G^2}{\sqrt{NT}}+\frac{\eta_{\text{local}}LKG^2}{2}+\frac{\sqrt{2}G\log(2T/\delta)\sigma_s}{\sqrt{NTK}}\\
&+\max\left\{0,\frac{G(KG-\tau)}{K}\right\}+\frac{15\sqrt{d}\log^3(24NTd/\delta)G\tau_{K,G}}{K\sqrt{Tb_{\text{in}}b_{\text{out}}}}+\frac{\log^2(2T/\delta)G^2}{\eta TK}\\
&+\frac{2\eta\log^2(2T/\delta)\sigma_g^2\alpha_2^2}{NK}+\frac{\eta\alpha_1^2\mathcal{I}L\tau_{K,G}^2}{K}+\frac{\eta\sigma_g^2\alpha_2^2\mathcal{I}L\log^2(2dT/\delta)}{NK}
\end{aligned}$$

in which

$$\alpha_1=1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}},\alpha_2=1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right),\tau_{K,G}=\min\{\tau,KG\},$$

then we finish the proof. □

I Optimization Analysis with AMSGrad as GLOBAL_OPT

Similar to Appendix H, we will state Algorithm 4, which will be used as GLOBAL_OPT to do a one-step adaptive update of θ_{t-1} with $\operatorname{desk}\left(\bar{\tilde{\Delta}}_{t-1}\right)$.

Algorithm 4 GLOBAL_OPT (AMSGrad)

Inputs: model θ_t, desketched update $\operatorname{desk}\left(\bar{\tilde{\Delta}}_t\right)$.
Output: model θ_{t+1}.
Update first moment estimate: $m_t = \beta_1 m_{t-1} + (1-\beta_1)\operatorname{desk}(\bar{\tilde{\Delta}}_t)$
Update second moment estimate: $\hat{v}_t = \beta_2 v_{t-1} + (1-\beta_2)\left(\operatorname{desk}(\bar{\tilde{\Delta}}_t)\right)^2$
Update maximum of past second moment estimates: $v_t = \max\left(\hat{v}_t, v_{t-1}\right)$
Update parameters: $\theta_{t+1} = \theta_t - \eta_{\text{global}}\frac{m_t}{\sqrt{v_t}+\epsilon} := \theta_t - \eta_{\text{global}} V_t^{-\frac{1}{2}} m_t$

Then we demonstrate the optimization result of Algorithm 2 with Algorithm 4, which is the formal version of Theorem 4.4.

Theorem I.1. *Suppose $\{\theta_t\}_{t=0}^T$ is generated by Algorithm 2 with Algorithm 4 as GLOBAL_OPT. Denote $\mathcal{L}^*$ the minimum of the average empirical loss. Under Assumption G.1-G.4, with learning rate $\eta = \eta_{global}\eta_{local}$, we have that with probability at least $1 - 22\delta$,*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\theta_t)\|_2^2 \le{}& \frac{\alpha_3\left(\mathcal{L}(\theta_0)-\mathcal{L}^*\right)}{\eta KT} + \frac{2\alpha_3\log(2T/\delta)G^2}{\sqrt{NT\epsilon}} + \frac{\eta_{local}\alpha_3 LKG^2}{2\epsilon} + \frac{\sqrt{2}\alpha_3 G\log(2T/\delta)\sigma_s}{\sqrt{NTK\epsilon}} \\ &+ \max\left\{0, \frac{\alpha_3 G(KG-\tau)}{K\epsilon}\right\} + \frac{15\alpha_3\log^3(24NTd/\delta)\sqrt{d}G\tau_{K,G}}{\sqrt{b_{in}b_{out}}\sqrt{T}K\epsilon} + \frac{\eta_{local}\alpha_1^2\alpha_3(2+\beta_1)\sqrt{1-\beta_2}G\tau_{K,G}^2}{K\epsilon^2(1-\beta_1)} \\ &+ \frac{\eta_{local}\alpha_2^2\alpha_3(1+2\beta_1)\sqrt{1-\beta_2}\sigma_g^2 G\log^2(2dT/\delta)}{NK\epsilon^2(1-\beta_1)} + \frac{\alpha_3\log^2(2T/\delta)G^2}{\eta TK\epsilon} + \frac{2\eta\alpha_2^2\alpha_3\log^2(2T/\delta)\sigma_g^2}{NK\epsilon} \\ &+ \frac{\eta\alpha_1^2\alpha_3\left(1+3\beta_1\right)\left(1+\beta_1\right)\mathcal{I}L\tau_{K,G}^2}{K\epsilon^2(1-\beta_1)^2} + \frac{\eta\alpha_2^2\alpha_3(1+3\beta_1)(1+\beta_1)\sigma_g^2\mathcal{I}L\log^2(2dT/\delta)}{NK\epsilon^2(1-\beta_1)^2}\end{aligned}$$

in which

$$\alpha_1 = 1 + \frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{in}b_{out}}}, \alpha_2 = 1 + 8\sqrt{\frac{d}{b_{out}}}\log\left(\frac{4dT}{\delta}\right),$$

$$\alpha_3 = \eta_{local}\left(1 + \frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{in}b_{out}}}\right)\tau_{K,G} + \frac{\eta_{local}\sigma_g\left(1+8\sqrt{\frac{d}{b_{out}}}\log\left(\frac{4dT}{\delta}\right)\right)\log(2dT/\delta)}{\sqrt{N}} + \epsilon$$

with $\tau_{K,G} = \min\{\tau, KG\}$.

Remark I.2. To analyze the dependence of the optimization bound on the clipping threshold τ, we consider all other parameters as fixed. For simplicity of exposition, in this remark we set $b_{\text{in}} = b_{\text{out}} = b$. Then

$$\alpha_1 = 1 + \frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{b}, \alpha_2 = 1 + 8\sqrt{\frac{d}{b}}\log\left(\frac{4dT}{\delta}\right),$$

$$\alpha_3 = \eta_{\text{local}}\alpha_1\tau_{K,G} + \frac{\eta_{\text{local}}\sigma_g\alpha_2\log(2dT/\delta)}{\sqrt{N}} + \epsilon,$$

where $\tau_{K,G} = \min\{\tau, KG\}$. We distinguish between two regimes.

1. When $\tau \le KG$: In this regime, $\tau_{K,G} = \tau$, and clipping may be activated. The sketching-related terms in the optimization bound depend on τ through $\frac{\alpha_3\sqrt{d}\tau}{b\sqrt{T}K}$, $\frac{\eta_{\text{local}}\alpha_1^2\alpha_3\tau^2}{K}$, $\frac{\eta\alpha_1^2\alpha_3\mathcal{I}\tau^2}{K}$, up to factors independent of τ and logarithmic terms. In addition, α_3 itself increases linearly with τ in this regime.

On the other hand, the clipping bias is represented by $\max\left\{0, \frac{\alpha_3 G(KG-\tau)}{K\epsilon}\right\}$, which vanishes as τ approaches KG. Therefore, the choice of τ reflects a trade-off between clipping bias and the optimization error contributed by the clipped and sketched updates.

2. When $\tau \geq KG$: In this regime, $\tau_{K,G} = KG$. Since the norm of each client update is bounded by KG, the clipping operation becomes inactive. Consequently, the clipping-bias term vanishes, while $\tau_{K,G}$ and α_3 no longer depend on τ. Hence, the entire optimization bound becomes independent of τ and remains constant as τ increases further.

Proof. Let

$$\gamma_t = \theta_t + \frac{\beta_1}{1-\beta_1}(\theta_t - \theta_{t-1}) = \frac{1}{1-\beta_1}\theta_t - \frac{\beta_1}{1-\beta_1}\theta_{t-1}$$

and set $\theta_{-1} = \theta_0$ so that $\gamma_0 = \theta_0$. Then, the update on γ_t can be expressed as

$$\begin{aligned}
&\gamma_{t+1} - \gamma_t \\
=&\frac{1}{1-\beta_1}(\theta_{t+1} - \theta_t) - \frac{\beta_1}{1-\beta_1}(\theta_t - \theta_{t-1}) \qquad (34)\\
=&-\frac{1}{1-\beta_1}\eta_{\text{global}} V_t^{-1/2}\cdot m_t + \frac{\beta_1}{1-\beta_1}\eta_{\text{global}} V_{t-1}^{-1/2}\cdot m_{t-1} \\
=&-\frac{1}{1-\beta_1}\eta_{\text{global}} V_t^{-1/2}\cdot (\beta_1 m_{t-1} + (1-\beta_1) S_t^\top \bar{\tilde{\Delta}}_t) + \frac{\beta_1}{1-\beta_1}\eta_{\text{global}} V_{t-1}^{-1/2}\cdot m_{t-1} \\
=&\frac{\beta_1\eta_{\text{global}}}{1-\beta_1}\left(V_{t-1}^{-1/2} - V_t^{-1/2}\right) m_{t-1} - \eta_{\text{global}} V_t^{-1/2} S_t^\top \bar{\tilde{\Delta}}_t \\
=&\frac{\beta_1\eta_{\text{global}}}{1-\beta_1}\left(V_{t-1}^{-1/2} - V_t^{-1/2}\right) m_{t-1} - \frac{\eta_{\text{global}}}{N} V_t^{-1/2} S_t^\top \sum_{c\in\mathcal{C}_t} \tilde{\Delta}_{c,t} \\
=&\frac{\beta_1\eta_{\text{global}}}{1-\beta_1}\left(V_{t-1}^{-1/2} - V_t^{-1/2}\right) m_{t-1} - \frac{\eta_{\text{global}}}{N} V_t^{-1/2} S_t^\top \sum_{c\in\mathcal{C}_t} S_t \breve{\Delta}_{c,t} \\
=&\frac{\beta_1\eta_{\text{global}}}{1-\beta_1}\left(V_{t-1}^{-1/2} - V_t^{-1/2}\right) m_{t-1} - \frac{\eta_{\text{global}}}{N} V_t^{-1/2} S_t^\top \sum_{c\in\mathcal{C}_t} S_t R_{c,t}^\top \breve{\Delta}_{c,t} \\
=&\frac{\beta_1\eta_{\text{global}}}{1-\beta_1}\left(V_{t-1}^{-1/2} - V_t^{-1/2}\right) m_{t-1} - \frac{\eta_{\text{global}}}{N} V_t^{-1/2} S_t^\top \left(\eta_{\text{local}} \sum_{c\in\mathcal{C}_t}\left(S_t R_{c,t}^\top R_{c,t}\left(\text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}}, \tau\right)\right) + \mathbf{z}_{c,t}\right)\right) \\
=&\frac{\beta_1\eta_{\text{global}}}{1-\beta_1}\left(V_{t-1}^{-1/2} - V_t^{-1/2}\right) m_{t-1} - \frac{\eta}{N} V_t^{-1/2} \sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top R_{c,t} \text{clip}\left(\sum_{k=1}^K g_{c,t,k}, \tau\right) - \frac{\eta}{N} V_t^{-1/2} \sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top \mathbf{z}_{c,t}
\end{aligned} \qquad (35)$$

in which $\eta = \eta_{\text{global}}\eta_{\text{local}}$. By Taylor expansion, we have

$$\begin{aligned}
\mathcal{L}(\gamma_{t+1}) &= \mathcal{L}(\gamma_t) + \nabla\mathcal{L}(\gamma_t)^\top(\gamma_{t+1} - \gamma_t) + \frac{1}{2}(\gamma_{t+1} - \gamma_t)^\top \hat{H}_{\mathcal{L},t}(\gamma_{t+1} - \gamma_t) \\
&= \mathcal{L}(\gamma_t) + \nabla\mathcal{L}(\theta_t)^\top(\gamma_{t+1} - \gamma_t) + (\nabla\mathcal{L}(\gamma_t) - \nabla\mathcal{L}(\theta_t))^\top(\gamma_{t+1} - \gamma_t) + \frac{1}{2}(\gamma_{t+1} - \gamma_t)^\top \hat{H}_{\mathcal{L},t}(\gamma_{t+1} - \gamma_t).
\end{aligned}$$

By taking summation from 0 to $T-1$, we can get that

$$\begin{aligned}
&\mathcal{L}(\gamma_T) - \mathcal{L}(\theta_0) = \mathcal{L}(\gamma_T) - \mathcal{L}(\gamma_0) = \sum_{t=1}^T (\mathcal{L}(\gamma_{t+1}) - \mathcal{L}(\gamma_t)) \\
=&\underbrace{\sum_{t=1}^T \nabla\mathcal{L}(\theta_t)^\top(\gamma_{t+1} - \gamma_t)}_{T_1} + \underbrace{\sum_{t=1}^T (\nabla\mathcal{L}(\gamma_t) - \nabla\mathcal{L}(\theta_t))^\top(\gamma_{t+1} - \gamma_t)}_{T_2} + \underbrace{\frac{1}{2}\sum_{t=1}^T (\gamma_{t+1} - \gamma_t)^\top \hat{H}_{\mathcal{L},t}(\gamma_{t+1} - \gamma_t)}_{T_3}
\end{aligned} \qquad (36)$$

I.1 BOUNDING T_1

For each term in T_1, according to (35), we have

$$\begin{aligned}
&\nabla\mathcal{L}(\theta_t)^\top(\gamma_{t+1}-\gamma_t)\\
=&\nabla\mathcal{L}(\theta_t)^\top\frac{\beta_1\eta_{\text{global}}}{1-\beta_1}\left(V_{t-1}^{-1/2}-V_t^{-1/2}\right)m_{t-1}-\nabla\mathcal{L}(\theta_t)^\top\frac{\eta}{N}V_t^{-1/2}\sum_{c\in\mathcal{C}_t}S_t^\top S_tR_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)\\
&-\nabla\mathcal{L}(\theta_t)^\top\frac{\eta}{N}V_t^{-1/2}\sum_{c\in\mathcal{C}_t}S_t^\top S_tR_{c,t}^\top\mathbf{z}_{c,t}\\
=&\nabla\mathcal{L}(\theta_t)^\top\frac{\beta_1\eta_{\text{global}}}{1-\beta_1}\left(V_{t-1}^{-1/2}-V_t^{-1/2}\right)m_{t-1}-\frac{\eta}{N}\nabla\mathcal{L}(\theta_t)^\top\left(V_t^{-1/2}-V_{t-1}^{-1/2}\right)\sum_{c\in\mathcal{C}_t}S_t^\top S_tR_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)\\
&-\frac{\eta}{N}\nabla\mathcal{L}(\theta_t)^\top V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}S_t^\top S_tR_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)-\frac{\eta}{N}\nabla\mathcal{L}(\theta_t)^\top V_t^{-1/2}\sum_{c\in\mathcal{C}_t}S_t^\top S_tR_{c,t}^\top\mathbf{z}_{c,t}
\end{aligned}$$

By taking summation from 0 to $T-1$, we can get that

$$\begin{aligned}
&\sum_{t=1}^T\nabla\mathcal{L}(\theta_t)^\top(\gamma_{t+1}-\gamma_t)\\
=&\underbrace{\sum_{t=1}^T\nabla\mathcal{L}(\theta_t)^\top\frac{\beta_1\eta_{\text{global}}}{1-\beta_1}\left(V_{t-1}^{-1/2}-V_t^{-1/2}\right)m_{t-1}}_{S_1}\\
&\underbrace{-\eta\sum_{t=1}^T\frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top\left(V_t^{-1/2}-V_{t-1}^{-1/2}\right)\sum_{c\in\mathcal{C}_t}S_t^\top S_tR_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)}_{S_2}\\
&\underbrace{-\eta\sum_{t=1}^T\frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}S_t^\top S_tR_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)}_{S_3}\underbrace{-\eta\sum_{t=1}^T\frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top V_t^{-1/2}\sum_{c\in\mathcal{C}_t}S_t^\top S_tR_{c,t}^\top\mathbf{z}_{c,t}}_{S_4}
\end{aligned}\tag{37}$$

I.1.1 BOUNDING S_2

We first bound each term with a fixed $t\in[T]$ in S_2. According to Lemma F.3, we have that with probability at least $1-\frac{\delta}{T}$,

$$\begin{aligned}
&\left|\frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top\left(V_t^{-1/2}-V_{t-1}^{-1/2}\right)\sum_{c\in\mathcal{C}_t}S_t^\top S_tR_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)\right|\\
=&\left|\frac{1}{N}\left\langle\nabla\mathcal{L}(\theta_t),\left(V_t^{-1/2}-V_{t-1}^{-1/2}\right)\sum_{c\in\mathcal{C}_t}S_t^\top S_tR_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)\right\rangle\right|\\
\le&\frac{1}{N}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\|\nabla\mathcal{L}(\theta_t)\|_2\left\|V_t^{-1/2}-V_{t-1}^{-1/2}\right\|_2\sum_{c\in\mathcal{C}_t}\left\|\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)\right\|_2\\
\le&G\tau_{K,G}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\left\|V_t^{-1/2}-V_{t-1}^{-1/2}\right\|_2
\end{aligned}\tag{38}$$

In addition, if $v_{t,i}=v_{t-1,i}$, then $\left[V_t^{-1/2}-V_{t-1}^{-1/2}\right]_i=0$. If $v_{t,i}>v_{t-1,i}$, then $\left[V_t^{-1/2}-V_{t-1}^{-1/2}\right]_i<0$, and

$$\left|\left[V_t^{-1/2}-V_{t-1}^{-1/2}\right]_i\right|$$

$$
\begin{aligned}
&= \left| \frac{1}{\sqrt{v_{t,i}} + \epsilon} - \frac{1}{\sqrt{v_{t-1,i}} + \epsilon} \right| \\
&= \left| \frac{\sqrt{v_{t-1,i}} - \sqrt{v_{t,i}}}{\left(\sqrt{v_{t,i}} + \epsilon\right)\left(\sqrt{v_{t-1,i}} + \epsilon\right)} \right| \\
&= \left| \frac{v_{t-1,i} - v_{t,i}}{\left(\sqrt{v_{t,i}} + \epsilon\right)\left(\sqrt{v_{t-1,i}} + \epsilon\right)\left(\sqrt{v_{t-1,i}} + \sqrt{v_{t,i}}\right)} \right| \\
&= \frac{(1-\beta_2)\left(\left[\left(S_t^\top \bar{\tilde{\Delta}}_t\right)^2\right]_i - v_{t-1,i}\right)}{\left(\sqrt{v_{t,i}} + \epsilon\right)\left(\sqrt{v_{t-1,i}} + \epsilon\right)\left(\sqrt{v_{t-1,i}} + \sqrt{v_{t,i}}\right)} \\
&\leq \frac{(1-\beta_2)\left[\left(S_t^\top \bar{\tilde{\Delta}}_t\right)^2\right]_i}{\epsilon^2 \sqrt{(1-\beta_2)\left[\left(S_t^\top \bar{\tilde{\Delta}}_t\right)^2\right]_i}} \\
&= \frac{\sqrt{1-\beta_2}}{\epsilon^2} \sqrt{\left[\left(S_t^\top \bar{\tilde{\Delta}}_t\right)^2\right]_i} \\
&= \frac{\sqrt{1-\beta_2}}{\epsilon^2} \left| \left[S_t^\top \left(\frac{\eta_{\text{local}}}{N} \sum_{c \in \mathcal{C}_t} \left(S_t R_{c,t}^\top \left(R_{c,t} \left(\text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}}, \tau \right) \right) + \mathbf{z}_{c,t} \right) \right) \right) \right]_i \right| \\
&\leq \frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2} \left| \sum_{c \in \mathcal{C}_t} \left[S_t^\top S_t R_{c,t}^\top R_{c,t} \text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}}, \tau \right) \right]_i \right| + \frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2} \left| \left[\sum_{c \in \mathcal{C}_t} S_t^\top S_t R_{c,t}^\top \mathbf{z}_{c,t} \right]_i \right|
\end{aligned} \tag{39}
$$

Therefore, we have

$$
\begin{aligned}
&\left\| V_t^{-1/2} - V_{t-1}^{-1/2} \right\|_2 = \max_i \left| \left[V_t^{-1/2} - V_{t-1}^{-1/2} \right]_i \right| \\
&\leq \max_i \left\{ \frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2} \left| \sum_{c \in \mathcal{C}_t} \left[S_t^\top S_t R_{c,t}^\top R_{c,t} \text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}}, \tau \right) \right]_i \right| + \frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2} \left| \left[\sum_{c \in \mathcal{C}_t} S_t^\top S_t R_{c,t}^\top \mathbf{z}_{c,t} \right]_i \right| \right\} \\
&\leq \max_i \left\{ \frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2} \left| \sum_{c \in \mathcal{C}_t} \left[S_t^\top S_t R_{c,t}^\top R_{c,t} \text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}}, \tau \right) \right]_i \right| \right\} + \max_i \left\{ \frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2} \left| \left[\sum_{c \in \mathcal{C}_t} S_t^\top S_t R_{c,t}^\top \mathbf{z}_{c,t} \right]_i \right| \right\}
\end{aligned} \tag{40}
$$

For the first term, according to Lemma F.3, for each $i \in [d]$, with probability at least $1 - \frac{\delta}{Td}$,

$$
\begin{aligned}
\left| \left[S_t^\top S_t R_{c,t}^\top R_{c,t} \text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}}, \tau \right) \right]_i \right| &= \left| e_i^\top S_t^\top S_t R_{c,t}^\top R_{c,t} \text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}}, \tau \right) \right| \\
&\leq \left(1 + \frac{10 \log^{1.5}(6NTd^2/\delta) \log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}} b_{\text{out}}}} \right) \tau_{K,G}
\end{aligned} \tag{41}
$$

then we can get that with probability at least $1 - \frac{\delta}{T}$,

$$
\begin{aligned}
&\max_i \left\{ \frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2} \left| \sum_{c \in \mathcal{C}_t} \left[S_t^\top S_t R_{c,t}^\top R_{c,t} \text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}}, \tau \right) \right]_i \right| \right\} \\
&\leq \max_i \left\{ \frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2} \sum_{c \in \mathcal{C}_t} \left| \left[S_t^\top S_t R_{c,t}^\top R_{c,t} \text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}}, \tau \right) \right]_i \right| \right\}
\end{aligned}
$$

$$\leq \frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2} \cdot N \cdot \left(1 + \frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\tau_{K,G}$$
$$= \frac{\sqrt{1-\beta_2}\eta_{\text{local}}\tau_{K,G}}{\epsilon^2}\left(1 + \frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right) \tag{42}$$

For the second term, $\frac{1}{N}\left[\sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top \mathbf{z}_{c,t}\right]_i = \frac{1}{N}\sum_{c\in\mathcal{C}_t} e_i^\top S_t^\top S_t R_{c,t}^\top \mathbf{z}_{c,t}$. Noticing that $\mathbf{z}_{c,t} \sim \mathcal{N}\left(0, \sigma_g^2\mathbb{I}\right)$, and $R_{c,t}S_t^\top S_t e_i$ is a $\frac{\left\|S_t^\top S_t e_i\right\|_2}{\sqrt{b_{\text{in}}}}$-sub-Gaussian random vector, so according to Bernstein inequality,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{c\in\mathcal{C}_t}\left\langle R_{c,t}S_t^\top S_t e_i, \mathbf{z}_{c,t}\right\rangle\right| \geq a\right) \leq 2\exp\left(-c\min\left(\frac{Na^2}{\sigma_g^2\cdot\left\|S_t^\top S_t e_i\right\|_2^2}, \frac{\sqrt{N}a}{\frac{\sigma_g\left\|S_t^\top S_t e_i\right\|_2}{\sqrt{b_{\text{in}}}}}\right)\right)$$
$$= 2\exp\left(-c\min\left(\frac{Na^2}{\sigma_g^2\left\|S_t^\top S_t e_i\right\|_2^2}, \frac{a\sqrt{b_{\text{in}}N}}{\sigma_g\left\|S_t^\top S_t e_i\right\|_2}\right)\right)$$

so taking $a = \frac{\sigma_g\left\|S_t^\top S_t e_i\right\|_2\log(2dT/\delta)}{\sqrt{N}}$, we have that with probability at least $1-\frac{\delta}{dT}$,

$$\left|\left\langle\frac{1}{N}\sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top \mathbf{z}_{c,t}, e_i\right\rangle\right| \leq \frac{\sigma_g\left\|S_t^\top S_t e_i\right\|_2\log(2dT/\delta)}{\sqrt{N}}$$

According to Lemma F.1, we can see that with probability at least $1-\frac{\delta}{dT}$,

$$\left\|S_t^\top S_t e_i\right\| \leq \left(1 + 8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)\|e_i\| = 1 + 8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)$$

Therefore, with probability at least $1-\frac{2\delta}{dT}$,

$$\left|\left\langle\frac{1}{N}\sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top \mathbf{z}_{c,t}, e_i\right\rangle\right| \leq \frac{\sigma_g\left\|S_t^\top S_t e_i\right\|_2\log(2dT/\delta)}{\sqrt{N}} \leq \frac{\sigma_g\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)\log(2dT/\delta)}{\sqrt{N}} \tag{43}$$

then we can get that with probability at least $1-\frac{2\delta}{T}$,

$$\max_i\left\{\frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2}\left|\left[\sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top \mathbf{z}_{c,t}\right]_i\right|\right\} \leq \frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{\epsilon^2}\cdot\frac{\sigma_g\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)\log(2dT/\delta)}{\sqrt{N}}$$
$$= \frac{\sqrt{1-\beta_2}\eta_{\text{local}}\sigma_g\log(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)}{\sqrt{N}\epsilon^2} \tag{44}$$

Substituting (42) and (44) into (40), we have that with probability at least $1-\frac{3\delta}{T}$,

$$\left\|V_t^{-1/2} - V_{t-1}^{-1/2}\right\|_2$$
$$\leq \max_i\left\{\frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2}\left|\sum_{c\in\mathcal{C}_t}\left[S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}}, \tau\right)\right]_i\right|\right\} + \max_i\left\{\frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2}\left|\left[\sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top \mathbf{z}_{c,t}\right]_i\right|\right\}$$
$$\leq \frac{\sqrt{1-\beta_2}\eta_{\text{local}}\tau_{K,G}}{\epsilon^2}\left(1 + \frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)$$
$$+ \frac{\sqrt{1-\beta_2}\eta_{\text{local}}\sigma_g\log(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)}{\sqrt{N}\epsilon^2} \tag{45}$$

Substituting (45) into (38), we have that with probability at least $1-\frac{4\delta}{T}$,

$$
\begin{aligned}
&\left|\frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top\left(V_t^{-1/2}-V_{t-1}^{-1/2}\right)\sum_{c\in\mathcal{C}_t}S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)\right| \\
\leq& G\tau_{K,G}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\left\|V_t^{-1/2}-V_{t-1}^{-1/2}\right\|_2 \\
\leq& G\tau_{K,G}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\frac{\sqrt{1-\beta_2}\eta_{\text{local}}\tau_{K,G}}{\epsilon^2}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right) \\
&+G\tau_{K,G}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\frac{\sqrt{1-\beta_2}\eta_{\text{local}}\sigma_g\log(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)}{\sqrt{N}\epsilon^2} \\
=&\frac{\eta_{\text{local}}\sqrt{1-\beta_2}G\tau_{K,G}^2}{\epsilon^2}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2 \\
&+\frac{\eta_{\text{local}}\sqrt{1-\beta_2}G\tau_{K,G}\sigma_g\log(2dT/\delta)}{\sqrt{N}\epsilon^2}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)
\end{aligned}
\tag{46}
$$

By taking summation from 0 to $T-1$, we can get that with probability at least $1-4\delta$,

$$
\begin{aligned}
&\left|\sum_{t=1}^{T}\frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top\left(V_t^{-1/2}-V_{t-1}^{-1/2}\right)\sum_{c\in\mathcal{C}_t}S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)\right| \\
\leq&\frac{\eta_{\text{local}}\sqrt{1-\beta_2}TG\tau_{K,G}^2}{\epsilon^2}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2 \\
&+\frac{\eta_{\text{local}}\sqrt{1-\beta_2}TG\tau_{K,G}\sigma_g\log(2dT/\delta)}{\sqrt{N}\epsilon^2}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)
\end{aligned}
\tag{47}
$$

I.1.2 Bounding S_1

We first bound each term with a fixed $t\in[T]$ in S_1. According to the definition of m_{t-1}, we have that

$$
\begin{aligned}
&\nabla\mathcal{L}(\theta_t)^\top\frac{\beta_1\eta_{\text{global}}}{1-\beta_1}\left(V_{t-1}^{-1/2}-V_t^{-1/2}\right)m_{t-1} \\
=&\nabla\mathcal{L}(\theta_t)^\top\frac{\beta_1\eta_{\text{global}}}{1-\beta_1}\left(V_{t-1}^{-1/2}-V_t^{-1/2}\right)\cdot\sum_{\tau'=0}^{t-1}(1-\beta_1)\beta_1^{t-1-\tau'}S_{\tau'}^\top\bar{\tilde{\Delta}}_{\tau'} \\
=&\nabla\mathcal{L}(\theta_t)^\top\beta_1\eta_{\text{global}}\left(V_{t-1}^{-1/2}-V_t^{-1/2}\right)\cdot\sum_{\tau'=0}^{t-1}\beta_1^{t-1-\tau'}S_{\tau'}^\top\left(\frac{1}{N}\sum_{c\in\mathcal{C}_{\tau'}}\tilde{\Delta}_{c,\tau'}\right) \\
=&\sum_{\tau'=0}^{t-1}\frac{\beta_1^{t-\tau'}\eta_{\text{global}}}{N}\nabla\mathcal{L}(\theta_t)^\top\left(V_{t-1}^{-1/2}-V_t^{-1/2}\right)S_{\tau'}^\top\sum_{c\in\mathcal{C}_{\tau'}}\tilde{\Delta}_{c,\tau'} \\
=&\sum_{\tau'=0}^{t-1}\frac{\beta_1^{t-\tau'}\eta_{\text{global}}}{N}\nabla\mathcal{L}(\theta_t)^\top\left(V_{t-1}^{-1/2}-V_t^{-1/2}\right)S_{\tau'}^\top\sum_{c\in\mathcal{C}_{\tau'}}S_{\tau'}\breve{\Delta}_{c,\tau'} \\
=&\sum_{\tau'=0}^{t-1}\frac{\beta_1^{t-\tau'}\eta_{\text{global}}}{N}\nabla\mathcal{L}(\theta_t)^\top\left(V_{t-1}^{-1/2}-V_t^{-1/2}\right)S_{\tau'}^\top\sum_{c\in\mathcal{C}_{\tau'}}S_{\tau'}R_{c,\tau'}^\top\breve{\Delta}_{c,\tau'}
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{\tau'=0}^{t-1} \frac{\beta_1^{t-\tau'}\eta_{\text{global}}}{N} \nabla \mathcal{L}(\theta_t)^\top \left(V_{t-1}^{-1/2} - V_t^{-1/2}\right) S_{\tau'}^\top \left(\eta_{\text{local}} \sum_{c\in\mathcal{C}_{\tau'}} \left(S_{\tau'} R_{c,\tau'}^\top R_{c,\tau'} \left(\text{clip}\left(\frac{\Delta_{c,\tau'}}{\eta_{\text{local}}}, \tau\right)\right) + \mathbf{z}_{c,\tau'}\right)\right) \\
&= \underbrace{\sum_{\tau'=0}^{t-1} \frac{\beta_1^{t-\tau'}\eta}{N} \nabla \mathcal{L}(\theta_t)^\top \left(V_{t-1}^{-1/2} - V_t^{-1/2}\right) \sum_{c\in\mathcal{C}_{\tau'}} S_{\tau'}^\top S_{\tau'} R_{c,\tau'}^\top R_{c,\tau'} \text{clip}\left(\frac{\Delta_{c,\tau'}}{\eta_{\text{local}}}, \tau\right)}_{W_1} \\
&\quad + \underbrace{\sum_{\tau'=0}^{t-1} \frac{\beta_1^{t-\tau'}\eta}{N} \nabla \mathcal{L}(\theta_t)^\top \left(V_{t-1}^{-1/2} - V_t^{-1/2}\right) \sum_{c\in\mathcal{C}_{\tau'}} S_{\tau'}^\top S_{\tau'} R_{c,\tau'}^\top \mathbf{z}_{c,\tau'}}_{W_2} \qquad (48)
\end{aligned}
$$

I.1.2.1 Bounding W_1

We first bound each term with a fixed $t \in [T]$ in W_1. By applying the same analysis as above, we can get the same bound as (46), with probability at least $1 - \frac{4\delta}{T}$,

$$
\begin{aligned}
&\frac{1}{N} \nabla \mathcal{L}(\theta_t)^\top \left(V_t^{-1/2} - V_{t-1}^{-1/2}\right) \sum_{c\in\mathcal{C}_{\tau'}} S_{\tau'}^\top S_{\tau'} R_{c,\tau'}^\top R_{c,\tau'} \text{clip}\left(\sum_{k=1}^{K} g_{c,\tau',k}, \tau\right) \\
\leq &\frac{\eta_{\text{local}}\sqrt{1-\beta_2} G \tau_{K,G}^2}{\epsilon^2} \left(1 + \frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}} b_{\text{out}}}}\right)^2 \\
&+ \frac{\eta_{\text{local}}\sqrt{1-\beta_2} G \tau_{K,G} \sigma_g \log(2dT/\delta)}{\sqrt{N}\epsilon^2} \left(1 + \frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}} b_{\text{out}}}}\right) \left(1 + 8\sqrt{\frac{d}{b_{\text{out}}}} \log\left(\frac{4dT}{\delta}\right)\right)
\end{aligned}
$$

By taking summation from 0 to $t-1$, we have that with probability at least $1 - 4\delta$,

$$
\begin{aligned}
&\sum_{\tau'=0}^{t-1} \frac{\beta_1^{t-\tau'}\eta}{N} \nabla \mathcal{L}(\theta_t)^\top \left(V_{t-1}^{-1/2} - V_t^{-1/2}\right) \sum_{c\in\mathcal{C}_{\tau'}} S_{\tau'}^\top S_{\tau'} R_{c,\tau'}^\top R_{c,\tau'} \text{clip}\left(\frac{\Delta_{c,\tau'}}{\eta_{\text{local}}}, \tau\right) \\
\leq &\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2} G \tau_{K,G}^2}{\epsilon^2} \left(1 + \frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}} b_{\text{out}}}}\right)^2 \cdot \sum_{\tau'=0}^{t-1} \beta_1^{t-\tau'} \\
&+ \frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2} G \tau_{K,G} \sigma_g \log(2dT/\delta)}{\sqrt{N}\epsilon^2} \left(1 + \frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}} b_{\text{out}}}}\right) \left(1 + 8\sqrt{\frac{d}{b_{\text{out}}}} \log\left(\frac{4dT}{\delta}\right)\right) \cdot \sum_{\tau'=0}^{t-1} \beta_1^{t-\tau'} \\
\leq &\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}\beta_1 G \tau_{K,G}^2}{\epsilon^2 (1-\beta_1)} \left(1 + \frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}} b_{\text{out}}}}\right)^2 \\
&+ \frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}\beta_1 G \tau_{K,G} \sigma_g \log(2dT/\delta)}{\sqrt{N}\epsilon^2 (1-\beta_1)} \left(1 + \frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}} b_{\text{out}}}}\right) \left(1 + 8\sqrt{\frac{d}{b_{\text{out}}}} \log\left(\frac{4dT}{\delta}\right)\right) \qquad (49)
\end{aligned}
$$

I.1.2.2 Bounding W_2

We first bound each term with a fixed $t \in [T]$ in W_2.

$$
\frac{1}{N} \nabla \mathcal{L}(\theta_t)^\top \left(V_{t-1}^{-1/2} - V_t^{-1/2}\right) \sum_{c\in\mathcal{C}_{\tau'}} S_{\tau'}^\top S_{\tau'} R_{c,\tau'}^\top \mathbf{z}_{c,\tau'} = \left\langle S_{\tau'} \left(V_{t-1}^{-1/2} - V_t^{-1/2}\right)^\top \nabla \mathcal{L}(\theta_t), \frac{1}{N} \sum_{c\in\mathcal{C}_{\tau'}} S_{\tau'} R_{c,\tau'}^\top \mathbf{z}_{c,\tau'} \right\rangle.
$$

Noticing that $\mathbf{z}_{c,t} \sim \mathcal{N}\left(0, \sigma_g^2 \mathbb{I}\right)$, and $R_{c,\tau'} S_{\tau'}^\top S_{\tau'} \left(V_t^{-1/2} - V_{t-1}^{-1/2}\right)^\top \nabla \mathcal{L}(\theta_t)$ is a $\frac{\left\| S_{\tau'}^\top S_{\tau'} \left(V_t^{-1/2} - V_{t-1}^{-1/2}\right)^\top \nabla \mathcal{L}(\theta_t) \right\|_2}{\sqrt{b_{\text{in}}}}$-sub-Gaussian random vector, so according to Bernstein inequality,

$$
\begin{aligned}
&\mathbb{P}\left(\left| \frac{1}{N} \sum_{c \in \mathcal{C}_{\tau'}} \left\langle R_{c,\tau'} S_{\tau'}^\top S_{\tau'} \left(V_t^{-1/2} - V_{t-1}^{-1/2}\right)^\top \nabla \mathcal{L}(\theta_t), \mathbf{z}_{c,\tau'} \right\rangle \right| \geq a \right) \\
\leq & 2 \exp\left(-c \min\left(\frac{Na^2}{\sigma_g^2 \cdot \left\| S_{\tau'}^\top S_{\tau'} \left(V_t^{-1/2} - V_{t-1}^{-1/2}\right)^\top \nabla \mathcal{L}(\theta_t) \right\|_2^2}, \frac{\sqrt{N} a}{\frac{\sigma_g \left\| S_{\tau'}^\top S_{\tau'} \left(V_t^{-1/2} - V_{t-1}^{-1/2}\right)^\top \nabla \mathcal{L}(\theta_t) \right\|_2}{\sqrt{b_{\text{in}}}}} \right) \right) \\
= & 2 \exp\left(-c \min\left(\frac{Na^2}{\sigma_g^2 \left\| S_{\tau'}^\top S_{\tau'} \left(V_t^{-1/2} - V_{t-1}^{-1/2}\right)^\top \nabla \mathcal{L}(\theta_t) \right\|_2^2}, \frac{a\sqrt{b_{\text{in}} N}}{\sigma_g \left\| S_{\tau'}^\top S_{\tau'} \left(V_t^{-1/2} - V_{t-1}^{-1/2}\right)^\top \nabla \mathcal{L}(\theta_t) \right\|_2} \right) \right)
\end{aligned}
$$

so taking $a = \frac{\sigma_g \left\| S_{\tau'}^\top S_{\tau'} \left(V_t^{-1/2} - V_{t-1}^{-1/2}\right)^\top \nabla \mathcal{L}(\theta_t) \right\|_2 \log(2dT/\delta)}{\sqrt{N}}$, we have that with probability at least $1 - \frac{\delta}{T}$,

$$
\left| \frac{1}{N} \sum_{c \in \mathcal{C}_{\tau'}} \left\langle R_{c,\tau'} S_{\tau'}^\top S_{\tau'} \left(V_t^{-1/2} - V_{t-1}^{-1/2}\right)^\top \nabla \mathcal{L}(\theta_t), \mathbf{z}_{c,\tau'} \right\rangle \right| \leq \frac{\sigma_g \left\| S_{\tau'}^\top S_{\tau'} \left(V_t^{-1/2} - V_{t-1}^{-1/2}\right)^\top \nabla \mathcal{L}(\theta_t) \right\|_2 \log(2dT/\delta)}{\sqrt{N}}
$$

According to Lemma F.1, and combining with (45), we can see that with probability at least $1 - \frac{4\delta}{T}$,

$$
\begin{aligned}
&\left\| S_{\tau'}^\top S_{\tau'} \left(V_t^{-1/2} - V_{t-1}^{-1/2}\right)^\top \nabla \mathcal{L}(\theta_t) \right\| \\
\leq & \left(1 + 8\sqrt{\frac{d}{b_{\text{out}}}} \log\left(\frac{4dT}{\delta}\right)\right) \left\| \left(V_t^{-1/2} - V_{t-1}^{-1/2}\right)^\top \nabla \mathcal{L}(\theta_t) \right\| \\
\leq & \left(1 + 8\sqrt{\frac{d}{b_{\text{out}}}} \log\left(\frac{4dT}{\delta}\right)\right) \left\| \nabla \mathcal{L}(\theta_t) \right\|_2 \left\| V_t^{-1/2} - V_{t-1}^{-1/2} \right\|_2 \\
\leq & \frac{\sqrt{1-\beta_2}\eta_{\text{local}} G \tau_{K,G}}{\epsilon^2} \left(1 + \frac{10 \log^{1.5}(6NTd^2/\delta) \log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}} b_{\text{out}}}}\right) \left(1 + 8\sqrt{\frac{d}{b_{\text{out}}}} \log\left(\frac{4dT}{\delta}\right)\right) \\
& + \frac{\sqrt{1-\beta_2}\eta_{\text{local}} G \sigma_g \log(2dT/\delta) \left(1 + 8\sqrt{\frac{d}{b_{\text{out}}}} \log\left(\frac{4dT}{\delta}\right)\right)^2}{\sqrt{N}\epsilon^2}
\end{aligned}
$$

Therefore, with probability at least $1 - \frac{5\delta}{T}$,

$$
\begin{aligned}
&\left| \frac{1}{N} \sum_{c \in \mathcal{C}_{\tau'}} \left\langle R_{c,\tau'} S_{\tau'}^\top S_{\tau'} \left(V_t^{-1/2} - V_{t-1}^{-1/2}\right)^\top \nabla \mathcal{L}(\theta_t), \mathbf{z}_{c,\tau'} \right\rangle \right| \\
\leq & \frac{\sigma_g \left\| S_{\tau'}^\top S_{\tau'} \left(V_t^{-1/2} - V_{t-1}^{-1/2}\right)^\top \nabla \mathcal{L}(\theta_t) \right\|_2 \log(2dT/\delta)}{\sqrt{N}} \\
\leq & \frac{\sigma_g \log(2dT/\delta)}{\sqrt{N}} \cdot \frac{\sqrt{1-\beta_2}\eta_{\text{local}} G \tau_{K,G}}{\epsilon^2} \left(1 + \frac{10 \log^{1.5}(6NTd^2/\delta) \log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}} b_{\text{out}}}}\right) \left(1 + 8\sqrt{\frac{d}{b_{\text{out}}}} \log\left(\frac{4dT}{\delta}\right)\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{\sigma_g \log(2dT/\delta)}{\sqrt{N}} \cdot \frac{\sqrt{1-\beta_2}\eta_{\text{local}} G\sigma_g \log(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{\sqrt{N}\epsilon^2} \\
\leq &\frac{\sqrt{1-\beta_2}\eta_{\text{local}} G\sigma_g \tau_{K,G}\log(2dT/\delta)}{\sqrt{N}\epsilon^2}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right) \\
&+\frac{\sqrt{1-\beta_2}\eta_{\text{local}} G\sigma_g^2 \log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2} \qquad (50)
\end{aligned}
$$

By taking summation from 0 to $t-1$, we have that with probability at least $1-5\delta$,

$$
\begin{aligned}
&\sum_{\tau'=0}^{t-1}\frac{\beta_1^{t-\tau'}\eta}{N}\nabla\mathcal{L}(\theta_t)^\top\left(V_{t-1}^{-1/2}-V_t^{-1/2}\right)\sum_{c\in\mathcal{C}_{\tau'}}S_{\tau'}^\top S_{\tau'}R_{c,\tau'}^\top \mathbf{z}_{c,\tau'} \\
\leq &\frac{\sqrt{1-\beta_2}\eta\eta_{\text{local}} G\sigma_g \tau_{K,G}\log(2dT/\delta)}{\sqrt{N}\epsilon^2}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)\sum_{\tau'=0}^{t-1}\beta_1^{t-\tau'} \\
&+\frac{\sqrt{1-\beta_2}\eta\eta_{\text{local}} G\sigma_g^2 \log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2}\cdot\sum_{\tau'=0}^{t-1}\beta_1^{t-\tau'} \\
\leq &\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}\beta_1 G\sigma_g \tau_{K,G}\log(2dT/\delta)}{\sqrt{N}\epsilon^2(1-\beta_1)}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right) \\
&+\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}\beta_1 G\sigma_g^2 \log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2(1-\beta_1)} \qquad (51)
\end{aligned}
$$

Substituting (49) and (51) into (48), we can get that with probability at least $1-6\delta$,

$$
\begin{aligned}
&\nabla\mathcal{L}(\theta_t)^\top\frac{\beta_1\eta_{\text{global}}}{1-\beta_1}\left(V_{t-1}^{-1/2}-V_t^{-1/2}\right)m_{t-1} \\
=&\sum_{\tau'=0}^{t-1}\frac{\beta_1^{t-\tau'}\eta}{N}\nabla\mathcal{L}(\theta_t)^\top\left(V_{t-1}^{-1/2}-V_t^{-1/2}\right)\sum_{c\in\mathcal{C}_{\tau'}}S_{\tau'}^\top S_{\tau'}R_{c,\tau'}^\top R_{c,\tau'}\text{clip}\left(\frac{\Delta_{c,\tau'}}{\eta_{\text{local}}},\tau\right) \\
&+\sum_{\tau'=0}^{t-1}\frac{\beta_1^{t-\tau'}\eta}{N}\nabla\mathcal{L}(\theta_t)^\top\left(V_{t-1}^{-1/2}-V_t^{-1/2}\right)\sum_{c\in\mathcal{C}_{\tau'}}S_{\tau'}^\top S_{\tau'}R_{c,\tau'}^\top \mathbf{z}_{c,\tau'} \\
\leq &\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}\beta_1 G\tau_{K,G}^2}{\epsilon^2(1-\beta_1)}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2 \\
&+\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}\beta_1 G\tau_{K,G}\sigma_g\log(2dT/\delta)}{\sqrt{N}\epsilon^2(1-\beta_1)}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right) \\
&+\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}\beta_1 G\sigma_g\tau_{K,G}\log(2dT/\delta)}{\sqrt{N}\epsilon^2(1-\beta_1)}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right) \\
&+\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}\beta_1 G\sigma_g^2\log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2(1-\beta_1)} \\
=&\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}\beta_1 G\tau_{K,G}^2}{\epsilon^2(1-\beta_1)}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2 \\
&+\frac{2\eta\eta_{\text{local}}\sqrt{1-\beta_2}\beta_1 G\tau_{K,G}\sigma_g\log(2dT/\delta)}{\sqrt{N}\epsilon^2(1-\beta_1)}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)
\end{aligned}
$$

$$+\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}\beta_1 G\sigma_g^2\log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2(1-\beta_1)}$$

By taking summation from 0 to $T-1$, we can get that with probability at least $1-8\delta$,

$$\begin{aligned}
&\sum_{t=1}^{T}\nabla\mathcal{L}(\theta_t)^\top\frac{\beta_1\eta_{\text{global}}}{1-\beta_1}\left(V_{t-1}^{-1/2}-V_t^{-1/2}\right)m_{t-1}\\
\leq&\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}\beta_1 TG\tau_{K,G}^2}{\epsilon^2(1-\beta_1)}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2\\
&+\frac{2\eta\eta_{\text{local}}\sqrt{1-\beta_2}\beta_1 TG\tau_{K,G}\sigma_g\log(2dT/\delta)}{\sqrt{N}\epsilon^2(1-\beta_1)}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)\\
&+\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}\beta_1 TG\sigma_g^2\log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2(1-\beta_1)}
\end{aligned}\tag{52}$$

I.1.3 Bounding S_4

We first bound each term with a fixed $t\in[T]$ in S_4.

$$\frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top V_t^{-1/2}\sum_{c\in\mathcal{C}_t}S_t^\top S_t R_{c,t}^\top\mathbf{z}_{c,t}=\frac{1}{N}\sum_{c\in\mathcal{C}_t}\left\langle R_{c,t}S_t^\top S_t V_t^{-1/2}\nabla\mathcal{L}(\theta_t),\mathbf{z}_{c,t}\right\rangle.$$

Noticing that $\mathbf{z}_{c,t}\sim\mathcal{N}\left(0,\sigma_g^2\mathbb{I}\right)$, and $R_{c,t}S_t^\top S_t V_t^{-1/2}\nabla\mathcal{L}(\theta_t)$ is a $\frac{\left\|S_t^\top S_t V_t^{-1/2}\nabla\mathcal{L}(\theta_t)\right\|}{\sqrt{b_{\text{in}}}}$-sub-Gaussian random vector, so according to Bernstein inequality,

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{N}\sum_{c\in\mathcal{C}_t}\left\langle R_{c,t}S_t^\top S_t V_t^{-1/2}\nabla\mathcal{L}(\theta_t),\mathbf{z}_{c,t}\right\rangle\right|\geq a\right)\\
\leq&2\exp\left(-c\min\left(\frac{Na^2}{\sigma_g^2\cdot\left\|S_t^\top S_t V_t^{-1/2}\nabla\mathcal{L}(\theta_t)\right\|_2^2},\frac{\sqrt{N}a}{\frac{\sigma_g\left\|S_t^\top S_t V_t^{-1/2}\nabla\mathcal{L}(\theta_t)\right\|_2}{\sqrt{b_{\text{in}}}}}\right)\right)\\
=&2\exp\left(-c\min\left(\frac{Na^2}{\sigma_g^2\left\|S_t^\top S_t V_t^{-1/2}\nabla\mathcal{L}(\theta_t)\right\|_2^2},\frac{a\sqrt{b_{\text{in}}N}}{\sigma_g\left\|S_t^\top S_t V_t^{-1/2}\nabla\mathcal{L}(\theta_t)\right\|_2}\right)\right)
\end{aligned}$$

so taking $a=\frac{\sigma_g\left\|S_t^\top S_t V_t^{-1/2}\nabla\mathcal{L}(\theta_t)\right\|_2\log(2T/\delta)}{\sqrt{N}}$, we have that with probability at least $1-\frac{\delta}{T}$,

$$\left|\frac{1}{N}\sum_{c\in\mathcal{C}_t}\left\langle R_{c,t}S_t^\top S_t V_t^{-1/2}\nabla\mathcal{L}(\theta_t),\mathbf{z}_{c,t}\right\rangle\right|\leq\frac{\sigma_g\left\|S_t^\top S_t V_t^{-1/2}\nabla\mathcal{L}(\theta_t)\right\|_2\log(2T/\delta)}{\sqrt{N}}$$

According to Lemma F.1, and noticing that $\left\|V_t^{-1/2}\right\|_2\leq\frac{1}{\epsilon}$, we can see that with probability at least $1-\frac{\delta}{T}$,

$$\begin{aligned}
\left\|S_t^\top S_t V_t^{-1/2}\nabla\mathcal{L}(\theta_t)\right\|&\leq\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)\left\|V_t^{-1/2}\nabla\mathcal{L}(\theta_t)\right\|_2\\
&\leq\frac{\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)G}{\epsilon}
\end{aligned}$$

Therefore, with probability at least $1-\frac{2\delta}{T}$,

$$\left|\frac{1}{N}\sum_{c\in\mathcal{C}_t}\left\langle R_{c,t}S_t^\top S_t V_t^{-1/2}\nabla\mathcal{L}(\theta_t),\mathbf{z}_{c,t}\right\rangle\right| \le \frac{\sigma_g\left(1+8\sqrt{\frac{d}{b_{\text{out}}}\log\left(\frac{4dT}{\delta}\right)}\right)\log(2T/\delta)G}{\sqrt{N}\epsilon}$$

Then denote

$$X_t = \sum_{\tau'=0}^{t}\frac{1}{N}\sum_{c\in\mathcal{C}_{\tau'}}\left\langle R_{c,\tau'}S_{\tau'}^\top S_{\tau'}V_{\tau'}^{-1/2}\nabla\mathcal{L}(\theta_{\tau'}),\mathbf{z}_{c,\tau'}\right\rangle,$$

we can see that X_t is a martingale with respect to the Gaussian noise, and from the above analysis, we have that with probability at least $1-2\delta$, for all $t\in[T]$,

$$\begin{aligned}|X_t - X_{t-1}| &= \left|\frac{1}{N}\sum_{c\in\mathcal{C}_t}\left\langle R_{c,t}S_t^\top S_t V_t^{-1/2}\nabla\mathcal{L}(\theta_t),\mathbf{z}_{c,t}\right\rangle\right| \\ &\le \frac{\sigma_g\left(1+8\sqrt{\frac{d}{b_{\text{out}}}\log\left(\frac{4dT}{\delta}\right)}\right)\log(2T/\delta)G}{\sqrt{N}\epsilon}\end{aligned}$$

Then by Azuma's inequality, we have

$$\begin{aligned}\mathbb{P}\left(X_{T-1}\le -a\right) &\le \exp\left(-\frac{a^2}{2\cdot T\cdot\left(\frac{\sigma_g\left(1+8\sqrt{\frac{d}{b_{\text{out}}}\log\left(\frac{4dT}{\delta}\right)}\right)\log(2T/\delta)G}{\sqrt{N}\epsilon}\right)^2}\right) \\ &= \exp\left(-\frac{N\epsilon^2a^2}{2T\sigma_g^2G^2\left(1+8\sqrt{\frac{d}{b_{\text{out}}}\log\left(\frac{4dT}{\delta}\right)}\right)^2\log^2(2T/\delta)}\right)\end{aligned}$$

By selecting $a=\frac{\log^2(2T/\delta)\sqrt{2T}\sigma_gG\left(1+8\sqrt{\frac{d}{b_{\text{out}}}\log\left(\frac{4dT}{\delta}\right)}\right)}{\sqrt{N}\epsilon}$, we can get that with probability at least $1-3\delta$,

$$\begin{aligned}X_{T-1} &= \sum_{t=1}^{T}\frac{1}{N}\sum_{c\in\mathcal{C}_t}\left\langle R_{c,t}S_t^\top S_t V_t^{-1/2}\nabla\mathcal{L}(\theta_t),\mathbf{z}_{c,t}\right\rangle \\ &\ge -\frac{\log^2(2T/\delta)\sqrt{2T}\sigma_gG\left(1+8\sqrt{\frac{d}{b_{\text{out}}}\log\left(\frac{4dT}{\delta}\right)}\right)}{\sqrt{N}\epsilon}\end{aligned} \tag{53}$$

I.1.4 Bounding S_3

For each term in S_3, we have that

$$\begin{aligned}&\frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}S_t^\top S_tR_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right) \\ =&K\left\langle\nabla\mathcal{L}(\theta_t),V_{t-1}^{-1/2}\frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_t)\right\rangle + K\left\langle\nabla\mathcal{L}(\theta_t),V_{t-1}^{-1/2}\left(\frac{1}{N}\sum_{c\in\mathcal{C}_t}\nabla\mathcal{L}_c(\theta_t)-\frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_t)\right)\right\rangle \\ &+\left\langle\nabla\mathcal{L}(\theta_t),\frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\left(\nabla\mathcal{L}_c(\theta_{c,t,k})-\nabla\mathcal{L}_c(\theta_t)\right)\right\rangle + \left\langle\nabla\mathcal{L}(\theta_t),\frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\left(g_{c,t,k}-\nabla\mathcal{L}_c(\theta_{c,t,k})\right)\right\rangle \\ &+\left\langle\nabla\mathcal{L}(\theta_t),\frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\left(\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)-\sum_{k=1}^{K}g_{c,t,k}\right)\right\rangle\end{aligned}$$

$$+\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\left(S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)-\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)\right)\right\rangle$$

By taking summation from 0 to $T-1$, we can get that

$$\begin{aligned}
&\sum_{t=1}^T \frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t} S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)\\
=&\underbrace{K\sum_{t=1}^T\left\langle \nabla\mathcal{L}(\theta_t), V_{t-1}^{-1/2}\frac{1}{C}\sum_{c=1}^C\nabla\mathcal{L}_c(\theta_t)\right\rangle}_{Y_1}+\underbrace{K\sum_{t=1}^T\left\langle \nabla\mathcal{L}(\theta_t), V_{t-1}^{-1/2}\left(\frac{1}{N}\sum_{c\in\mathcal{C}_t}\nabla\mathcal{L}_c(\theta_t)-\frac{1}{C}\sum_{c=1}^C\nabla\mathcal{L}_c(\theta_t)\right)\right\rangle}_{Y_2}\\
&+\underbrace{\sum_{t=1}^T\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^K\left(\nabla\mathcal{L}_c(\theta_{c,t,k})-\nabla\mathcal{L}_c(\theta_t)\right)\right\rangle}_{Y_3}\\
&+\underbrace{\sum_{t=1}^T\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^K\left(g_{c,t,k}-\nabla\mathcal{L}_c(\theta_{c,t,k})\right)\right\rangle}_{Y_4}\\
&+\underbrace{\sum_{t=1}^T\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\left(\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)-\sum_{k=1}^K g_{c,t,k}\right)\right\rangle}_{Y_5}\\
&+\underbrace{\sum_{t=1}^T\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\left(S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)-\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)\right)\right\rangle}_{Y_6}
\end{aligned}\tag{54}$$

I.1.4.1 Bounding Y_1

According to the definition of V_{t-1}, we have that

$$\left[V_{t-1}^{-1/2}\right]_i=\left[\left(\sqrt{v_{t-1}}+\epsilon\right)^{-1}\right]_i=\left(\sqrt{v_{t-1,i}}+\epsilon\right)^{-1}$$

Substituting (41) and (43) into (39), we can see that with probability at least $1-\frac{3\delta}{d}$, for any $t\in[T]$,

$$\begin{aligned}
&\sqrt{\left[\left(S_t^\top\bar{\tilde{\Delta}}_t\right)^2\right]_i}\\
=&\left|\left[S_t^\top\left(\frac{\eta_{\text{local}}}{N}\sum_{c\in\mathcal{C}_t}\left(S_t R_{c,t}^\top\left(R_{c,t}\left(\text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}},\tau\right)\right)+\mathbf{z}_{c,t}\right)\right)\right]_i\right|\\
\le&\frac{\eta_{\text{local}}}{N}\left|\sum_{c\in\mathcal{C}_t}\left[S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}},\tau\right)\right]_i\right|+\frac{\eta_{\text{local}}}{N}\left|\left[\sum_{c\in\mathcal{C}_t}S_t^\top S_t R_{c,t}^\top\mathbf{z}_{c,t}\right]_i\right|\\
\le&\eta_{\text{local}}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\tau_{K,G}+\frac{\eta_{\text{local}}\sigma_g\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)\log(2dT/\delta)}{\sqrt{N}}
\end{aligned}$$

Then according to the definition of second order moment, we have that with probability at least $1-\frac{3\delta}{d}$,

$$\sqrt{v_{t-1,i}}\le\max_{\tau\le t-1}\sqrt{\left[\left(S_\tau^\top\bar{\tilde{\Delta}}_\tau\right)^2\right]_i}$$

$$\leq \eta_{\text{local}} \left(1 + \frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\tau_{K,G} + \frac{\eta_{\text{local}}\sigma_g\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)\log(2dT/\delta)}{\sqrt{N}}$$

so we can get that with probability at least $1-3\delta$, for all $i \in [d]$,

$$\left[V_{t-1}^{-1/2}\right]_i = \left(\sqrt{v_{t-1,i}} + \epsilon\right)^{-1}$$

$$\geq \left(\eta_{\text{local}} \left(1 + \frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\tau_{K,G} + \frac{\eta_{\text{local}}\sigma_g\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)\log(2dT/\delta)}{\sqrt{N}} + \epsilon\right)^{-1}$$

which implies that with probability at least $1-3\delta$,

$$\left\langle \nabla\mathcal{L}(\theta_t), V_{t-1}^{-1/2}\frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_t)\right\rangle$$

$$= \left\langle \nabla\mathcal{L}(\theta_t), V_{t-1}^{-1/2}\nabla\mathcal{L}(\theta_t)\right\rangle$$

$$\geq \left(\eta_{\text{local}} \left(1 + \frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\tau_{K,G} + \frac{\eta_{\text{local}}\sigma_g\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)\log(2dT/\delta)}{\sqrt{N}} + \epsilon\right)^{-1}\|\nabla\mathcal{L}(\theta_t)\|_2^2$$

so with probability at least $1-3\delta$,

$$\sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), V_{t-1}^{-1/2}\frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_t)\right\rangle \tag{55}$$

$$\geq \left(\eta_{\text{local}} \left(1 + \frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\tau_{K,G} + \frac{\eta_{\text{local}}\sigma_g\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)\log(2dT/\delta)}{\sqrt{N}} + \epsilon\right)^{-1}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\theta_t)\|_2^2 \tag{56}$$

I.1.4.2 Bounding Y_2

We first bound each term with a fixed $t \in [T]$ in Y_2. According to the assumption, each $c \in \mathcal{C}_t$ is uniformly randomly selected from $[C]$, so by Hoeffding's inequality, we have

$$\mathbb{P}\left(\left|\left\langle \nabla\mathcal{L}(\theta_t), V_{t-1}^{-1/2}\left(\frac{1}{N}\sum_{c\in\mathcal{C}_t}\nabla\mathcal{L}_c(\theta_t) - \frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_t)\right)\right\rangle\right| \geq a\right)$$

$$\leq 2\exp\left(-\frac{2Na^2}{\left(\frac{2G^2}{\epsilon}\right)^2}\right) = 2\exp\left(-\frac{N\epsilon^2a^2}{2G^4}\right)$$

By selecting $a = \frac{\sqrt{2\log(2T/\delta)}G^2}{\sqrt{N}\epsilon}$, we have that with probability at least $1-\frac{\delta}{T}$,

$$\left|\left\langle \nabla\mathcal{L}(\theta_t), V_{t-1}^{-1/2}\left(\frac{1}{N}\sum_{c\in\mathcal{C}_t}\nabla\mathcal{L}_c(\theta_t) - \frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_t)\right)\right\rangle\right| \leq \frac{\sqrt{2\log(2T/\delta)}G^2}{\sqrt{N}\epsilon}$$

Then denote $Z_t = \sum_{\tau'=0}^{t}\left\langle \nabla\mathcal{L}(\theta_{\tau'}), V_{\tau'-1}^{-1/2}\left(\frac{1}{N}\sum_{c\in\mathcal{C}_{\tau'}}\nabla\mathcal{L}_c(\theta_{\tau'}) - \frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_{\tau'})\right)\right\rangle$, we can see that Z_t is a martingale with respect to the selection each round, and from the above analysis, we have that with probability at least $1-\delta$, for all $t \in [T]$,

$$|Z_t - Z_{t-1}| = \left|\left\langle \nabla\mathcal{L}(\theta_t), V_{t-1}^{-1/2}\left(\frac{1}{N}\sum_{c\in\mathcal{C}_t}\nabla\mathcal{L}_c(\theta_t) - \frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_t)\right)\right\rangle\right| \leq \frac{\sqrt{2\log(2T/\delta)}G^2}{\sqrt{N}\epsilon}$$

Then by Azuma's inequality, we have

$$\mathbb{P}\left(Z_{T-1} \leq -a\right) \leq \exp\left(-\frac{a^2}{2 \cdot T \cdot \left(\frac{\sqrt{2\log(2T/\delta)}G^2}{\sqrt{N}\epsilon}\right)^2}\right) = \exp\left(-\frac{N\epsilon^2 a^2}{4TG^4\log(2T/\delta)}\right)$$

By selecting $a = \frac{2\log(2T/\delta)\sqrt{T}G^2}{\sqrt{N}\epsilon}$, we can get that with probability at least $1-2\delta$,

$$Z_{T-1} = \sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), V_{t-1}^{-1/2}\left(\frac{1}{N}\sum_{c\in\mathcal{C}_t}\nabla\mathcal{L}_c(\theta_t) - \frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_t)\right)\right\rangle \geq -\frac{2\log(2T/\delta)\sqrt{T}G^2}{\sqrt{N}\epsilon} \tag{57}$$

I.1.4.3 Bounding Y_3

For each term, we have

$$\begin{aligned}
&\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\left(\nabla\mathcal{L}_c(\theta_{c,t,k}) - \nabla\mathcal{L}_c(\theta_t)\right)\right\rangle \\
=&\frac{1}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\left\langle \nabla\mathcal{L}(\theta_t), V_{t-1}^{-1/2}\hat{H}_{\mathcal{L}}^{c,t,k}\left(\theta_{c,t,k} - \theta_t\right)\right\rangle \\
=&\frac{\eta_{\text{local}}}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\left\langle \nabla\mathcal{L}(\theta_t), V_{t-1}^{-1/2}\hat{H}_{\mathcal{L}}^{c,t,k}\sum_{\kappa=1}^{k}g_{c,t,\kappa}\right\rangle \\
\geq&-\frac{\eta_{\text{local}}}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\left\|\nabla\mathcal{L}(\theta_t)\right\|_2\cdot\frac{L}{\epsilon}\sum_{\kappa=1}^{k}\left\|g_{c,t,\kappa}\right\|_2 \\
=&-\frac{\eta_{\text{local}}}{N}\cdot N\cdot G^2\cdot\frac{L}{\epsilon}\sum_{k=1}^{K}k \\
\geq&-\frac{\eta_{\text{local}}LK^2G^2}{2\epsilon}
\end{aligned}$$

so

$$\sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\left(\nabla\mathcal{L}_c(\theta_{c,t,k}) - \nabla\mathcal{L}_c(\theta_t)\right)\right\rangle \geq -\frac{\eta_{\text{local}}TLK^2G^2}{2\epsilon} \tag{58}$$

I.1.4.4 Bounding Y_4

We first bound each term with a fixed $t \in [T]$ in Y_4.

$$\begin{aligned}
&\left|\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\left(g_{c,t,k} - \nabla\mathcal{L}_c(\theta_{c,t,k})\right)\right\rangle\right| \\
\leq&\frac{1}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\left\|\nabla\mathcal{L}(\theta_t)\right\|_2\cdot\frac{1}{\epsilon}\left\|g_{c,t,k} - \nabla\mathcal{L}_c(\theta_{c,t,k})\right\|_2 \\
=&\frac{G}{N\epsilon}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\left\|g_{c,t,k} - \nabla\mathcal{L}_c(\theta_{c,t,k})\right\|_2
\end{aligned}$$

According to the assumption, the stochastic noise $\left\|g_{c,t,k} - \nabla\mathcal{L}_c(\theta_{c,t,k})\right\|_2$ is a σ_s-sub-Gaussian random variable, so by Hoeffding's inequality,

$$\mathbb{P}\left(\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\left\|g_{c,t,k} - \nabla\mathcal{L}_c(\theta_{c,t,k})\right\|_2 \geq a\right) \leq 2\exp\left(-\frac{a^2}{NK\sigma_s^2}\right)$$

By selecting $a = \sqrt{NK\log(2T/\delta)}\sigma_s$, we have that with probability at least $1 - \frac{\delta}{T}$,

$$\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\|g_{c,t,k} - \nabla\mathcal{L}_c(\theta_{c,t,k})\|_2 \leq \sqrt{NK\log(2T/\delta)}\sigma_s$$

so

$$\begin{aligned}
&\left|\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}(g_{c,t,k} - \nabla\mathcal{L}_c(\theta_{c,t,k}))\right\rangle\right| \\
\leq &\frac{G}{N\epsilon}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\|g_{c,t,k} - \nabla\mathcal{L}_c(\theta_{c,t,k})\|_2 \\
\leq &\frac{G}{N\epsilon}\cdot\sqrt{NK\log(2T/\delta)}\sigma_s = \frac{G\sqrt{K\log(2T/\delta)}\sigma_s}{\sqrt{N}\epsilon}
\end{aligned}$$

Then denote $W_t = \sum_{\tau'=0}^{t}\left\langle \nabla\mathcal{L}(\theta_{\tau'}), \frac{1}{N}V_{\tau'-1}^{-1/2}\sum_{c\in\mathcal{C}_{\tau'}}\sum_{k=1}^{K}(g_{c,\tau',k} - \nabla\mathcal{L}_c(\theta_{c,\tau',k}))\right\rangle$, we can see that W_t is a martingale with respect to the stochastic noise, and from the above analysis, we have that with probability at least $1-\delta$, for all $t \in [T]$,

$$|W_t - W_{t-1}| = \left|\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}(g_{c,t,k} - \nabla\mathcal{L}_c(\theta_{c,t,k}))\right\rangle\right| \leq \frac{G\sqrt{K\log(2T/\delta)}\sigma_s}{\sqrt{N}\epsilon}$$

Then by Azuma's inequality, we have

$$\mathbb{P}\left(W_{T-1} \leq -a\right) \leq \exp\left(-\frac{a^2}{2\cdot T\cdot\left(\frac{G\sqrt{K\log(2T/\delta)}\sigma_s}{\sqrt{N}\epsilon}\right)^2}\right) = \exp\left(-\frac{N\epsilon^2a^2}{2TG^2K\log(2T/\delta)\sigma_s^2}\right)$$

By selecting $a = \frac{G\sqrt{2TK}\log(2T/\delta)\sigma_s}{\sqrt{N}\epsilon}$, we can get that with probability at least $1-2\delta$,

$$W_{T-1} = \sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}(g_{c,t,k} - \nabla\mathcal{L}_c(\theta_{c,t,k}))\right\rangle \geq -\frac{G\sqrt{2TK}\log(2T/\delta)\sigma_s}{\sqrt{N}\epsilon} \tag{59}$$

I.1.4.5 Bounding Y_5

For each term, for $\tau \leq KG$, we have

$$\begin{aligned}
&\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\left(\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right) - \sum_{k=1}^{K}g_{c,t,k}\right)\right\rangle \\
\geq &-\frac{1}{N}\|\nabla\mathcal{L}(\theta_t)\|_2\left\|V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\left(\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right) - \sum_{k=1}^{K}g_{c,t,k}\right)\right\|_2 \\
\geq &-\frac{G(KG-\tau)}{\epsilon}
\end{aligned}$$

for $\tau \geq KG$, we have

$$\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\left(\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right) - \sum_{k=1}^{K}g_{c,t,k}\right)\right\rangle = 0$$

so

$$\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\left(\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right) - \sum_{k=1}^{K}g_{c,t,k}\right)\right\rangle \geq -\max\left\{0, \frac{G(KG-\tau)}{\epsilon}\right\}$$

By taking summation from 0 to $T-1$, we can get that

$$\sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\left(\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)-\sum_{k=1}^{K}g_{c,t,k}\right)\right\rangle \geq -\max\left\{0,\frac{TG(KG-\tau)}{\epsilon}\right\} \tag{60}$$

I.1.4.6 Bounding Y_6

We first bound each term with a fixed $t\in[T]$ in Y_6. According to Lemma F.3, we have that with probability at least $1-\frac{\delta}{T}$,

$$\begin{aligned}
&\left|\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\left(S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)-\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)\right)\right\rangle\right| \\
\leq&\frac{1}{N}\sum_{c\in\mathcal{C}_t}\frac{10\log^{1.5}(6NTd/\delta)\log(24NT/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\left\|V_{t-1}^{-1/2}\nabla\mathcal{L}(\theta_t)\right\|_2\left\|\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)\right\|_2 \\
\leq&\frac{10\log^{1.5}(6NTd/\delta)\log(24NT/\delta)\sqrt{d}G\tau_{K,G}}{\sqrt{b_{\text{in}}b_{\text{out}}}\epsilon}
\end{aligned}$$

Then denote $U_t=\sum_{\tau'=0}^{t}\left\langle \nabla\mathcal{L}(\theta_{\tau'}), \frac{1}{N}V_{\tau'-1}^{-1/2}\sum_{c\in\mathcal{C}_{\tau'}}\left(S_{\tau'}^\top S_{\tau'}R_{c,\tau'}^\top R_{c,\tau'}\text{clip}\left(\sum_{k=1}^{K}g_{c,\tau',k},\tau\right)-\text{clip}\left(\sum_{k=1}^{K}g_{c,\tau',k},\tau\right)\right)\right\rangle$, we can see that U_t is a martingale with respect to the sketching matrices, and from the above analysis, we have that with probability at least $1-\delta$, for all $t\in[T]$,

$$\begin{aligned}
|U_t-U_{t-1}| &= \left|\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\left(S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)-\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)\right)\right\rangle\right| \\
&\leq \frac{10\log^{1.5}(6NTd/\delta)\log(24NT/\delta)\sqrt{d}G\tau_{K,G}}{\sqrt{b_{\text{in}}b_{\text{out}}}\epsilon}
\end{aligned}$$

Then by Azuma's inequality, we have

$$\begin{aligned}
\mathbb{P}\left(U_{T-1}\leq -a\right) &\leq \exp\left(-\frac{a^2}{2\cdot T\cdot\left(\frac{10\log^{1.5}(6NTd/\delta)\log(24NT/\delta)\sqrt{d}G\tau_{K,G}}{\sqrt{b_{\text{in}}b_{\text{out}}}\epsilon}\right)^2}\right) \\
&= \exp\left(-\frac{b_{\text{in}}b_{\text{out}}\epsilon^2a^2}{200T\log^3(6NTd/\delta)\log^2(24NT/\delta)dG^2\tau_{K,G}^2}\right)
\end{aligned}$$

By selecting $a=\frac{15\sqrt{d}\log^3(24NTd/\delta)\sqrt{T}G\tau_{K,G}}{\sqrt{b_{\text{in}}b_{\text{out}}}\epsilon}$, we can get that with probability at least $1-2\delta$,

$$\begin{aligned}
U_{T-1} &= \sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\left(S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)-\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)\right)\right\rangle \\
&\geq -\frac{15\sqrt{d}\log^3(24NTd/\delta)\sqrt{T}G\tau_{K,G}}{\sqrt{b_{\text{in}}b_{\text{out}}}\epsilon}
\end{aligned} \tag{61}$$

Substituting (56), (57), (58), (59), (60), (61) into (54), we have that with probability at least $1-9\delta$,

$$\begin{aligned}
&\sum_{t=1}^{T}\frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right) \\
=&K\sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), V_{t-1}^{-1/2}\frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_t)\right\rangle + K\sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), V_{t-1}^{-1/2}\left(\frac{1}{N}\sum_{c\in\mathcal{C}_t}\nabla\mathcal{L}_c(\theta_t)-\frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_t)\right)\right\rangle \\
&+\sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\left(\nabla\mathcal{L}_c(\theta_{c,t,k})-\nabla\mathcal{L}_c(\theta_t)\right)\right\rangle
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\left(g_{c,t,k}-\nabla\mathcal{L}_c(\theta_{c,t,k})\right)\right\rangle\\
&+\sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\left(\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)-\sum_{k=1}^{K}g_{c,t,k}\right)\right\rangle\\
&+\sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\left(S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)-\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)\right)\right\rangle\\
\geq&\left(\eta_{\text{local}}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\tau_{K,G}+\frac{\eta_{\text{local}}\sigma_g\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)\log(2dT/\delta)}{\sqrt{N}}+\epsilon\right)^{-1}K\sum_{t=1}^{T}\|\nabla\mathcal{L}(\theta_t)\|_2^2\\
&-\frac{2K\sqrt{T}\log(2T/\delta)G^2}{\sqrt{N}\epsilon}-\frac{\eta_{\text{local}}TLK^2G^2}{2\epsilon}-\frac{G\sqrt{2TK}\log(2T/\delta)\sigma_s}{\sqrt{N}\epsilon}-\max\left\{0,\frac{TG(KG-\tau)}{\epsilon}\right\}\\
&-\frac{15\sqrt{d}\log^3(24NTd/\delta)\sqrt{T}G\tau_{K,G}}{\sqrt{b_{\text{in}}b_{\text{out}}}\epsilon}
\end{aligned}
\tag{62}
$$

Substituting (47), (52), (53), (62) into (37), we have that with probability at least $1-16\delta$,

$$
\begin{aligned}
&\sum_{t=1}^{T}\nabla\mathcal{L}(\theta_t)^\top(\gamma_{t+1}-\gamma_t)\\
=&\sum_{t=1}^{T}\nabla\mathcal{L}(\theta_t)^\top\frac{\beta_1\eta_{\text{global}}}{1-\beta_1}\left(V_{t-1}^{-1/2}-V_t^{-1/2}\right)m_{t-1}\\
&-\eta\sum_{t=1}^{T}\frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top\left(V_t^{-1/2}-V_{t-1}^{-1/2}\right)\sum_{c\in\mathcal{C}_t}S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)\\
&-\eta\sum_{t=1}^{T}\frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}S_t^\top S_t R_{c,t}^\top R_{c,t}\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)-\eta\sum_{t=1}^{T}\frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top V_t^{-1/2}\sum_{c\in\mathcal{C}_t}S_t^\top S_t R_{c,t}^\top\mathbf{z}_{c,t}\\
\leq&-\left(\eta_{\text{local}}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\tau_{K,G}+\frac{\eta_{\text{local}}\sigma_g\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)\log(2dT/\delta)}{\sqrt{N}}+\epsilon\right)^{-1}\eta K\sum_{t=1}^{T}\|\nabla\mathcal{L}(\theta_t)\|_2^2\\
&+\frac{2\eta K\sqrt{T}\log(2T/\delta)G^2}{\sqrt{N}\epsilon}+\frac{\eta\eta_{\text{local}}TLK^2G^2}{2\epsilon}+\frac{\eta G\sqrt{2TK}\log(2T/\delta)\sigma_s}{\sqrt{N}\epsilon}+\eta\max\left\{0,\frac{TG(KG-\tau)}{\epsilon}\right\}\\
&+\frac{15\eta\sqrt{d}\log^3(24NTd/\delta)\sqrt{T}G\tau_{K,G}}{\sqrt{b_{\text{in}}b_{\text{out}}}\epsilon}+\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}TG\tau_{K,G}^2}{\epsilon^2}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2\\
&+\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}TG\tau_{K,G}\sigma_g\log(2dT/\delta)}{\sqrt{N}\epsilon^2}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)\\
&+\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}\beta_1TG\tau_{K,G}^2}{\epsilon^2(1-\beta_1)}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2\\
&+\frac{2\eta\eta_{\text{local}}\sqrt{1-\beta_2}\beta_1TG\tau_{K,G}\sigma_g\log(2dT/\delta)}{\sqrt{N}\epsilon^2(1-\beta_1)}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)\\
&+\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}\beta_1TG\sigma_g^2\log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2(1-\beta_1)}+\frac{\eta\log^2(2T/\delta)\sqrt{2T}\sigma_gG\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)}{\sqrt{N}\epsilon}
\end{aligned}
$$

$$
\begin{aligned}
=&-\left(\eta_{\text{local}}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\tau_{K,G}+\frac{\eta_{\text{local}}\sigma_g\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)\log(2dT/\delta)}{\sqrt{N}}+\epsilon\right)^{-1}\eta K\sum_{t=1}^{T}\|\nabla\mathcal{L}(\theta_t)\|_2^2\\
&+\frac{2\eta K\sqrt{T}\log(2T/\delta)G^2}{\sqrt{N}\epsilon}+\frac{\eta\eta_{\text{local}}TLK^2G^2}{2\epsilon}+\frac{\eta G\sqrt{2TK}\log(2T/\delta)\sigma_s}{\sqrt{N}\epsilon}+\eta\max\left\{0,\frac{TG(KG-\tau)}{\epsilon}\right\}\\
&+\frac{15\eta\sqrt{d}\log^3(24NTd/\delta)\sqrt{T}G\tau_{K,G}}{\sqrt{b_{\text{in}}b_{\text{out}}}\epsilon}+\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}TG\tau_{K,G}^2}{\epsilon^2(1-\beta_1)}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2\\
&+\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}(1+\beta_1)TG\tau_{K,G}\sigma_g\log(2dT/\delta)}{\sqrt{N}\epsilon^2(1-\beta_1)}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)\\
&+\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}\beta_1TG\sigma_g^2\log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2(1-\beta_1)}+\frac{\eta\log^2(2T/\delta)\sqrt{2T}\sigma_gG\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)}{\sqrt{N}\epsilon}\\
\leq&-\left(\eta_{\text{local}}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\tau_{K,G}+\frac{\eta_{\text{local}}\sigma_g\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)\log(2dT/\delta)}{\sqrt{N}}+\epsilon\right)^{-1}\eta K\sum_{t=1}^{T}\|\nabla\mathcal{L}(\theta_t)\|_2^2\\
&+\frac{2\eta K\sqrt{T}\log(2T/\delta)G^2}{\sqrt{N}\epsilon}+\frac{\eta\eta_{\text{local}}TLK^2G^2}{2\epsilon}+\frac{\eta G\sqrt{2TK}\log(2T/\delta)\sigma_s}{\sqrt{N}\epsilon}+\eta\max\left\{0,\frac{TG(KG-\tau)}{\epsilon}\right\}\\
&+\frac{15\eta\sqrt{d}\log^3(24NTd/\delta)\sqrt{T}G\tau_{K,G}}{\sqrt{b_{\text{in}}b_{\text{out}}}\epsilon}+\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}(2+\beta_1)TG\tau_{K,G}^2}{\epsilon^2(1-\beta_1)}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\\
&+\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}(1+2\beta_1)TG\sigma_g^2\log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2(1-\beta_1)}\\
&+\frac{\log^2(2T/\delta)G^2}{\epsilon}+\frac{2\eta^2T\log^2(2T/\delta)\sigma_g^2\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon}
\end{aligned}
\tag{63}
$$

I.2 Bounding T_2

For each term in T_2, we have

$$
\begin{aligned}
&(\nabla\mathcal{L}(\gamma_t)-\nabla\mathcal{L}(\theta_t))^\top(\gamma_{t+1}-\gamma_t)\\
=&(\gamma_t-\theta_t)^\top\hat{H}'_{\mathcal{L},t}(\gamma_{t+1}-\gamma_t)\\
=&\frac{\beta_1}{1-\beta_1}(\theta_t-\theta_{t-1})^\top\hat{H}'_{\mathcal{L},t}\left(\frac{1}{1-\beta_1}(\theta_{t+1}-\theta_t)-\frac{\beta_1}{1-\beta_1}(\theta_t-\theta_{t-1})\right)\\
=&\frac{\beta_1}{(1-\beta_1)^2}(\theta_t-\theta_{t-1})^\top\hat{H}'_{\mathcal{L},t}(\theta_{t+1}-\theta_t)-\frac{\beta_1^2}{(1-\beta_1)^2}(\theta_t-\theta_{t-1})^\top\hat{H}'_{\mathcal{L},t}(\theta_t-\theta_{t-1})
\end{aligned}
$$

By taking summation from 0 to $T-1$, we can get that

$$
\begin{aligned}
&\sum_{t=1}^{T}(\nabla\mathcal{L}(\gamma_t)-\nabla\mathcal{L}(\theta_t))^\top(\gamma_{t+1}-\gamma_t)\\
=&\frac{\beta_1}{(1-\beta_1)^2}\underbrace{\sum_{t=1}^{T}(\theta_t-\theta_{t-1})^\top\hat{H}'_{\mathcal{L},t}(\theta_{t+1}-\theta_t)}_{S_5}-\frac{\beta_1^2}{(1-\beta_1)^2}\underbrace{\sum_{t=1}^{T}(\theta_t-\theta_{t-1})^\top\hat{H}'_{\mathcal{L},t}(\theta_t-\theta_{t-1})}_{S_6}
\end{aligned}
\tag{64}
$$

I.2.1 BOUNDING S_5

We first bound each term in S_5 with a fixed $t \in [T]$.

$$\begin{aligned}
(\theta_t - \theta_{t-1})^\top \hat{H}'_{\mathcal{L},t} (\theta_{t+1} - \theta_t) &= \eta_{\text{global}}^2 \left(V_{t-1}^{-1/2} m_{t-1}\right)^\top \hat{H}'_{\mathcal{L},t} \left(V_t^{-1/2} m_t\right) \\
&= \eta_{\text{global}}^2 \left(V_{t-1}^{-1/2} m_{t-1}\right)^\top \left(\sum_{i=1}^d \lambda_i v_i v_i^\top\right) \left(V_t^{-1/2} m_t\right) \\
&= \eta_{\text{global}}^2 \sum_{i=1}^d \lambda_i \left(\left(V_{t-1}^{-1/2} m_{t-1}\right)^\top v_i\right) \left(\left(V_t^{-1/2} m_t\right)^\top v_i\right) \\
&\le \eta_{\text{global}}^2 \sum_{i=1}^d \lambda_i \left|\left(V_{t-1}^{-1/2} m_{t-1}\right)^\top v_i\right| \left|\left(V_t^{-1/2} m_t\right)^\top v_i\right| \qquad (65)
\end{aligned}$$

For each $i \in [d]$, we have that

$$\begin{aligned}
\left|\left(V_{t-1}^{-1/2} m_{t-1}\right)^\top v_i\right| &= (1-\beta_1) \left|\left(\sum_{\tau'=0}^{t-1} \beta_1^{t-1-\tau'} V_{t-1}^{-1/2} S_{\tau'}^\top \bar{\tilde{\Delta}}_{\tau'}\right)^\top v_i\right| \\
&\le (1-\beta_1) \sum_{\tau'=0}^{t-1} \beta_1^{t-1-\tau'} \left|\left(V_{t-1}^{-1/2} S_{\tau'}^\top \bar{\tilde{\Delta}}_{\tau'}\right)^\top v_i\right| \\
&\le \max_{\tau' \in [t-1]} \left|\left(V_{t-1}^{-1/2} S_{\tau'}^\top \bar{\tilde{\Delta}}_{\tau'}\right)^\top v_i\right| \qquad (66)
\end{aligned}$$

Similarly,

$$\left|\left(V_t^{-1/2} m_t\right)^\top v_i\right| \le \max_{\tau' \in [t]} \left|\left(V_t^{-1/2} S_{\tau'}^\top \bar{\tilde{\Delta}}_{\tau'}\right)^\top v_i\right| \qquad (67)$$

For each $\tau' \in [t-1]$,

$$\begin{aligned}
&\left|\left(V_{t-1}^{-1/2} S_{\tau'}^\top \bar{\tilde{\Delta}}_{\tau'}\right)^\top v_i\right| \\
=&\left|\left\langle S_{\tau'}^\top \bar{\tilde{\Delta}}_{\tau'}, V_{t-1}^{-1/2} v_i \right\rangle\right| \\
=&\frac{1}{N} \left|\left\langle S_{\tau'}^\top \sum_{c \in \mathcal{C}_{\tau'}} \tilde{\Delta}_{c,\tau'}, V_{t-1}^{-1/2} v_i \right\rangle\right| \\
=&\frac{1}{N} \left|\left\langle S_{\tau'}^\top \sum_{c \in \mathcal{C}_{\tau'}} \eta_{\text{local}} \left(S_{\tau'} R_{c,\tau'}^\top \left(R_{c,\tau'} \text{clip}\left(\frac{\Delta_{c,\tau'}}{\eta_{\text{local}}}, \tau\right)\right) + \mathbf{z}_{c,\tau'}\right), V_{t-1}^{-1/2} v_i \right\rangle\right| \\
\le&\frac{\eta_{\text{local}}}{N} \left|\left\langle \sum_{c \in \mathcal{C}_{\tau'}} S_{\tau'}^\top S_{\tau'} R_{c,\tau'}^\top R_{c,\tau'} \text{clip}\left(\sum_{k=1}^K g_{c,\tau',k}, \tau\right), V_{t-1}^{-1/2} v_i \right\rangle\right| + \frac{\eta_{\text{local}}}{N} \left|\left\langle \sum_{c \in \mathcal{C}_{\tau'}} S_{\tau'}^\top S_{\tau'} R_{c,\tau'}^\top \mathbf{z}_{c,\tau'}, V_{t-1}^{-1/2} v_i \right\rangle\right| \qquad (68)
\end{aligned}$$

For the first term, according to Lemma F.3, with probability at least $1 - \frac{\delta}{Td}$,

$$\begin{aligned}
&\left|\left\langle \sum_{c \in \mathcal{C}_{\tau'}} S_{\tau'}^\top S_{\tau'} R_{c,\tau'}^\top R_{c,\tau'} \text{clip}\left(\sum_{k=1}^K g_{c,\tau',k}, \tau\right), V_{t-1}^{-1/2} v_i \right\rangle\right| \\
\le& \left(1 + \frac{10 \log^{1.5}(6NTd^2/\delta) \log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}} b_{\text{out}}}}\right) \left\|V_{t-1}^{-1/2} v_i\right\|_2 \sum_{c \in \mathcal{C}_{\tau'}} \left\|\text{clip}\left(\sum_{k=1}^K g_{c,\tau',k}, \tau\right)\right\| \\
\le& \left(1 + \frac{10 \log^{1.5}(6NTd^2/\delta) \log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}} b_{\text{out}}}}\right) \frac{N \tau_{K,G}}{\epsilon} \qquad (69)
\end{aligned}$$

For the second term, $\left\langle \frac{1}{N}\sum_{c\in\mathcal{C}_{\tau'}} S_{\tau'}^\top S_{\tau'} R_{c,\tau'}^\top \mathbf{z}_{c,\tau'}, V_{t-1}^{-1/2} v_i \right\rangle = \frac{1}{N}\sum_{c\in\mathcal{C}_{\tau'}} \left\langle R_{c,\tau'} S_{\tau'}^\top S_{\tau'} V_{t-1}^{-1/2} v_i, \mathbf{z}_{c,\tau'} \right\rangle$. Noticing that $\mathbf{z}_{c,\tau'} \sim \mathcal{N}\left(0, \sigma_g^2 \mathbb{I}\right)$, and $R_{c,\tau'} S_{\tau'}^\top S_{\tau'} V_{t-1}^{-1/2} v_i$ is a $\frac{\left\|S_{\tau'}^\top S_{\tau'} V_{t-1}^{-1/2} v_i\right\|_2}{\sqrt{b_{\text{in}}}}$-sub-Gaussian random vector, so according to Bernstein inequality,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{c\in\mathcal{C}_{\tau'}} \left\langle R_{c,\tau'} S_{\tau'}^\top S_{\tau'} V_{t-1}^{-1/2} v_i, \mathbf{z}_{c,\tau'} \right\rangle\right| \geq a\right) \leq 2\exp\left(-c\min\left(\frac{a^2}{\frac{\sigma_g^2 \cdot \left\|S_{\tau'}^\top S_{\tau'} V_{t-1}^{-1/2} v_i\right\|_2^2}{N}}, \frac{a}{\frac{\sigma_g \left\|S_{\tau'}^\top S_{\tau'} V_{t-1}^{-1/2} v_i\right\|_2}{\sqrt{b_{\text{in}} N}}}\right)\right)$$

$$= 2\exp\left(-c\min\left(\frac{Na^2}{\sigma_g^2 \left\|S_{\tau'}^\top S_{\tau'} V_{t-1}^{-1/2} v_i\right\|_2^2}, \frac{a\sqrt{b_{\text{in}} N}}{\sigma_g \left\|S_{\tau'}^\top S_{\tau'} V_{t-1}^{-1/2} v_i\right\|_2}\right)\right)$$

so taking $a = \frac{\sigma_g \left\|S_{\tau'}^\top S_{\tau'} V_{t-1}^{-1/2} v_i\right\|_2 \log(2dT/\delta)}{\sqrt{N}}$, and noticing that $\left\|V_{t-1}^{-1/2}\right\|_2 \leq \frac{1}{\epsilon}$, we have that with probability at least $1 - \frac{\delta}{Td}$,

$$\left|\frac{1}{N}\sum_{c\in\mathcal{C}_{\tau'}} \left\langle R_{c,\tau'} S_{\tau'}^\top S_{\tau'} V_{t-1}^{-1/2} v_i, \mathbf{z}_{c,\tau'} \right\rangle\right| \leq \frac{\sigma_g \left\|S_{\tau'}^\top S_{\tau'} V_{t-1}^{-1/2} v_i\right\|_2 \log(2dT/\delta)}{\sqrt{N}} \leq \frac{\sigma_g \left\|S_{\tau'}^\top S_{\tau'} v_i\right\|_2 \log(2dT/\delta)}{\sqrt{N}\epsilon}$$

According to Lemma F.1, we can see that with probability at least $1 - \frac{\delta}{dT}$,

$$\left\|S_{\tau'}^\top S_{\tau'} v_i\right\|_2 \leq \left(1 + 8\sqrt{\frac{d}{b_{\text{out}}}} \log\left(\frac{4dT}{\delta}\right)\right) \|v_i\|_2 = 1 + 8\sqrt{\frac{d}{b_{\text{out}}}} \log\left(\frac{4dT}{\delta}\right)$$

Therefore, with probability at least $1 - \frac{2\delta}{dT}$,

$$\left|\frac{1}{N}\sum_{c\in\mathcal{C}_{\tau'}} \left\langle R_{c,\tau'} S_{\tau'}^\top S_{\tau'} V_{t-1}^{-1/2} v_i, \mathbf{z}_{c,\tau'} \right\rangle\right| \leq \frac{\sigma_g \left\|S_{\tau'}^\top S_{\tau'} v_i\right\|_2 \log(2dT/\delta)}{\sqrt{N}\epsilon} \leq \frac{\sigma_g \log(2dT/\delta)\left(1 + 8\sqrt{\frac{d}{b_{\text{out}}}} \log\left(\frac{4dT}{\delta}\right)\right)}{\sqrt{N}\epsilon} \tag{70}$$

Substituting (69) and (70) into (68), we have that with probability at least $1 - \frac{3\delta}{dT}$,

$$\begin{aligned}
&\left|\left(V_{t-1}^{-1/2} S_{\tau'}^\top \bar{\tilde{\Delta}}_{\tau'}\right)^\top v_i\right| \\
\leq &\frac{\eta_{\text{local}}}{N}\left|\left\langle \sum_{c\in\mathcal{C}_{\tau'}} S_{\tau'}^\top S_{\tau'} R_{c,\tau'}^\top R_{c,\tau'} \text{clip}\left(\sum_{k=1}^K g_{c,\tau',k}, \tau\right), V_{t-1}^{-1/2} v_i \right\rangle\right| + \frac{\eta_{\text{local}}}{N}\left|\left\langle \sum_{c\in\mathcal{C}_{\tau'}} S_{\tau'}^\top S_{\tau'} R_{c,\tau'}^\top \mathbf{z}_{c,\tau'}, V_{t-1}^{-1/2} v_i \right\rangle\right| \\
\leq &\left(1 + \frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}} b_{\text{out}}}}\right)\frac{\eta_{\text{local}}\tau_{K,G}}{\epsilon} + \frac{\eta_{\text{local}}\sigma_g \log(2dT/\delta)\left(1 + 8\sqrt{\frac{d}{b_{\text{out}}}} \log\left(\frac{4dT}{\delta}\right)\right)}{\sqrt{N}\epsilon}
\end{aligned} \tag{71}$$

Similarly, we can get that for $\tau' \in [t]$, with probability at least $1 - \frac{3\delta}{dT}$,

$$\begin{aligned}
\left|\left(V_t^{-1/2} S_{\tau'}^\top \bar{\tilde{\Delta}}_{\tau'}\right)^\top v_i\right| \leq &\left(1 + \frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}} b_{\text{out}}}}\right)\frac{\eta_{\text{local}}\tau_{K,G}}{\epsilon} \\
&+ \frac{\eta_{\text{local}}\sigma_g \log(2dT/\delta)\left(1 + 8\sqrt{\frac{d}{b_{\text{out}}}} \log\left(\frac{4dT}{\delta}\right)\right)}{\sqrt{N}\epsilon}
\end{aligned} \tag{72}$$

Substituting (71) into (66), (72) and (67), we have that with probability at least $1 - \frac{6\delta}{d}$,

$$\left|\left(V_{t-1}^{-1/2} m_{t-1}\right)^\top v_i\right| \leq \max_{\tau'\in[t-1]}\left|\left(V_{t-1}^{-1/2} S_{\tau'}^\top \bar{\tilde{\Delta}}_{\tau'}\right)^\top v_i\right|$$

$$\leq \left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\frac{\eta_{\text{local}}\tau_{K,G}}{\epsilon}+\frac{\eta_{\text{local}}\sigma_g\log(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)}{\sqrt{N}\epsilon} \tag{73}$$

$$\left|\left(V_t^{-1/2}m_t\right)^\top v_i\right| \leq \max_{\tau'\in[t]}\left|\left(V_t^{-1/2}S_{\tau'}^\top\bar{\tilde{\Delta}}_{\tau'}\right)^\top v_i\right|$$

$$\leq \left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\frac{\eta_{\text{local}}\tau_{K,G}}{\epsilon}+\frac{\eta_{\text{local}}\sigma_g\log(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)}{\sqrt{N}\epsilon} \tag{74}$$

Substituting (73) and (74) into (65), we have that with probability at least $1-6\delta$,

$$\begin{aligned}
&\left(\theta_t-\theta_{t-1}\right)^\top\hat{H}'_{\mathcal{L},t}\left(\theta_{t+1}-\theta_t\right)\\
=&\eta_{\text{global}}^2\sum_{i=1}^d\lambda_i\left|\left(V_{t-1}^{-1/2}m_{t-1}\right)^\top v_i\right|\left|\left(V_t^{-1/2}m_t\right)^\top v_i\right|\\
\leq&\eta_{\text{global}}^2\sum_{i=1}^d|\lambda_i|\left(\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\frac{\eta_{\text{local}}\tau_{K,G}}{\epsilon}+\frac{\eta_{\text{local}}\sigma_g\log(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)}{\sqrt{N}\epsilon}\right)^2\\
\leq&\eta_{\text{global}}^2\sum_{i=1}^d|\lambda_i|\cdot 2\left(\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\frac{\eta_{\text{local}}\tau_{K,G}}{\epsilon}\right)^2\\
&+\eta_{\text{global}}^2\sum_{i=1}^d|\lambda_i|\cdot 2\left(\frac{\eta_{\text{local}}\sigma_g\log(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)}{\sqrt{N}\epsilon}\right)^2\\
=&\frac{2\eta^2\mathcal{I}L\tau_{K,G}^2}{\epsilon^2}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2+\frac{2\eta^2\sigma_g^2\mathcal{I}L\log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2}
\end{aligned}$$

By taking summation from 0 to $T-1$, we have that with probability at least $1-6\delta$,

$$\begin{aligned}
&\sum_{t=1}^T\left(\theta_t-\theta_{t-1}\right)^\top\hat{H}'_{\mathcal{L},t}\left(\theta_{t+1}-\theta_t\right)\\
\leq&\frac{2\eta^2\mathcal{I}LT\tau_{K,G}^2}{\epsilon^2}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2+\frac{2\eta^2\sigma_g^2\mathcal{I}LT\log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2}
\end{aligned} \tag{75}$$

I.2.2 Bounding S_6

We first bound each term in S_6 with a fixed $t\in[T]$.

$$\begin{aligned}
\left(\theta_t-\theta_{t-1}\right)^\top\hat{H}'_{\mathcal{L},t}\left(\theta_t-\theta_{t-1}\right)&=\eta_{\text{global}}^2\left(V_{t-1}^{-1/2}m_{t-1}\right)^\top\hat{H}'_{\mathcal{L},t}\left(V_{t-1}^{-1/2}m_{t-1}\right)\\
&=\eta_{\text{global}}^2\left(V_{t-1}^{-1/2}m_{t-1}\right)^\top\left(\sum_{i=1}^d\lambda_i v_i v_i^\top\right)\left(V_{t-1}^{-1/2}m_{t-1}\right)\\
&=\eta_{\text{global}}^2\sum_{i=1}^d\lambda_i\left|\left(V_{t-1}^{-1/2}m_{t-1}\right)^\top v_i\right|^2
\end{aligned} \tag{76}$$

Substituting (73) into (76), we have that with probability at least $1-3\delta$,

$$\left(\theta_t-\theta_{t-1}\right)^\top\hat{H}'_{\mathcal{L},t}\left(\theta_t-\theta_{t-1}\right)$$

$$
\begin{aligned}
&= \eta_{\text{global}}^2 \sum_{i=1}^{d} \lambda_i \left| \left(V_{t-1}^{-1/2} m_{t-1} \right)^\top v_i \right|^2 \\
&\geq -\eta_{\text{global}}^2 \sum_{i=1}^{d} |\lambda_i| \left(\left(1 + \frac{10 \log^{1.5}(6NTd^2/\delta) \log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}} b_{\text{out}}}} \right) \frac{\eta_{\text{local}} \tau_{K,G}}{\epsilon} + \frac{\eta_{\text{local}} \sigma_g \log(2dT/\delta) \left(1 + 8\sqrt{\frac{d}{b_{\text{out}}}} \log\left(\frac{4dT}{\delta}\right) \right)}{\sqrt{N}\epsilon} \right)^2 \\
&\geq -\frac{2\eta^2 \mathcal{I} L \tau_{K,G}^2}{\epsilon^2} \left(1 + \frac{10 \log^{1.5}(6NTd^2/\delta) \log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}} b_{\text{out}}}} \right)^2 - \frac{2\eta^2 \sigma_g^2 \mathcal{I} L \log^2(2dT/\delta) \left(1 + 8\sqrt{\frac{d}{b_{\text{out}}}} \log\left(\frac{4dT}{\delta}\right) \right)^2}{N\epsilon^2}
\end{aligned}
$$

By taking summation from 0 to $T-1$, we have that with probability at least $1 - 3\delta$,

$$
\begin{aligned}
&\sum_{t=1}^{T} (\theta_t - \theta_{t-1})^\top \hat{H}'_{\mathcal{L},t} (\theta_t - \theta_{t-1}) \\
&\geq -\frac{2\eta^2 \mathcal{I} L T \tau_{K,G}^2}{\epsilon^2} \left(1 + \frac{10 \log^{1.5}(6NTd^2/\delta) \log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}} b_{\text{out}}}} \right)^2 - \frac{2\eta^2 \sigma_g^2 \mathcal{I} L T \log^2(2dT/\delta) \left(1 + 8\sqrt{\frac{d}{b_{\text{out}}}} \log\left(\frac{4dT}{\delta}\right) \right)^2}{N\epsilon^2}
\end{aligned} \tag{77}
$$

Substituting (75) and (77) into (64), we have that with probability at least $1 - 6\delta$,

$$
\begin{aligned}
&\sum_{t=1}^{T} (\nabla \mathcal{L}(\gamma_t) - \nabla \mathcal{L}(\theta_t))^\top (\gamma_{t+1} - \gamma_t) \\
&= \frac{\beta_1}{(1-\beta_1)^2} \sum_{t=1}^{T} (\theta_t - \theta_{t-1})^\top \hat{H}'_{\mathcal{L},t} (\theta_{t+1} - \theta_t) - \frac{\beta_1^2}{(1-\beta_1)^2} \sum_{t=1}^{T} (\theta_t - \theta_{t-1})^\top \hat{H}'_{\mathcal{L},t} (\theta_t - \theta_{t-1}) \\
&\leq \frac{\beta_1(1+\beta_1)}{(1-\beta_1)^2} \left(\frac{2\eta^2 \mathcal{I} L T \tau_{K,G}^2}{\epsilon^2} \left(1 + \frac{10 \log^{1.5}(6NTd^2/\delta) \log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}} b_{\text{out}}}} \right)^2 + \frac{2\eta^2 \sigma_g^2 \mathcal{I} L T \log^2(2dT/\delta) \left(1 + 8\sqrt{\frac{d}{b_{\text{out}}}} \log\left(\frac{4dT}{\delta}\right) \right)^2}{N\epsilon^2} \right) \\
&= \frac{2\eta^2 \beta_1 (1+\beta_1) \mathcal{I} L T \tau_{K,G}^2}{\epsilon^2 (1-\beta_1)^2} \left(1 + \frac{10 \log^{1.5}(6NTd^2/\delta) \log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}} b_{\text{out}}}} \right)^2 \\
&\quad + \frac{2\eta^2 \beta_1 (1+\beta_1) \sigma_g^2 \mathcal{I} L T \log^2(2dT/\delta) \left(1 + 8\sqrt{\frac{d}{b_{\text{out}}}} \log\left(\frac{4dT}{\delta}\right) \right)^2}{N\epsilon^2 (1-\beta_1)^2}
\end{aligned} \tag{78}
$$

I.3 Bounding T_3

For each term in T_3, we have

$$
\begin{aligned}
&(\gamma_{t+1} - \gamma_t)^\top \hat{H}_{\mathcal{L},t} (\gamma_{t+1} - \gamma_t) \\
&= \left(\frac{1}{1-\beta_1} (\theta_{t+1} - \theta_t) - \frac{\beta_1}{1-\beta_1} (\theta_t - \theta_{t-1}) \right)^\top \hat{H}_{\mathcal{L},t} \left(\frac{1}{1-\beta_1} (\theta_{t+1} - \theta_t) - \frac{\beta_1}{1-\beta_1} (\theta_t - \theta_{t-1}) \right) \\
&= \frac{1}{(1-\beta_1)^2} (\theta_{t+1} - \theta_t)^\top \hat{H}_{\mathcal{L},t} (\theta_{t+1} - \theta_t) - \frac{2\beta_1}{(1-\beta_1)^2} (\theta_{t+1} - \theta_t)^\top \hat{H}_{\mathcal{L},t} (\theta_t - \theta_{t-1}) \\
&\quad + \frac{\beta_1^2}{(1-\beta_1)^2} (\theta_t - \theta_{t-1})^\top \hat{H}_{\mathcal{L},t} (\theta_t - \theta_{t-1})
\end{aligned}
$$

By taking summation from 0 to $T-1$, we can get that

$$
\sum_{t=1}^{T} (\gamma_{t+1} - \gamma_t)^\top \hat{H}_{\mathcal{L},t} (\gamma_{t+1} - \gamma_t)
$$

$$=\frac{1}{(1-\beta_1)^2}\underbrace{\sum_{t=1}^{T}\left(\theta_{t+1}-\theta_t\right)^\top \hat{H}_{\mathcal{L},t}\left(\theta_{t+1}-\theta_t\right)}_{S_7}-\frac{2\beta_1}{(1-\beta_1)^2}\underbrace{\sum_{t=1}^{T}\left(\theta_{t+1}-\theta_t\right)^\top \hat{H}_{\mathcal{L},t}\left(\theta_t-\theta_{t-1}\right)}_{S_8}$$
$$+\frac{\beta_1^2}{(1-\beta_1)^2}\underbrace{\sum_{t=1}^{T}\left(\theta_t-\theta_{t-1}\right)^\top \hat{H}_{\mathcal{L},t}\left(\theta_t-\theta_{t-1}\right)}_{S_9} \tag{79}$$

I.3.1 Bounding S_7

We first bound each term in S_7 with a fixed $t \in [T]$.

$$\begin{aligned}\left(\theta_{t+1}-\theta_t\right)^\top \hat{H}_{\mathcal{L},t}\left(\theta_{t+1}-\theta_t\right) &= \eta_{\text{global}}^2\left(V_t^{-1/2}m_t\right)^\top \hat{H}_{\mathcal{L},t}\left(V_t^{-1/2}m_t\right)\\ &= \eta_{\text{global}}^2\left(V_t^{-1/2}m_t\right)^\top\left(\sum_{i=1}^{d}\lambda_i v_i v_i^\top\right)\left(V_t^{-1/2}m_t\right)\\ &= \eta_{\text{global}}^2\sum_{i=1}^{d}\lambda_i\left|\left(V_t^{-1/2}m_t\right)^\top v_i\right|^2\end{aligned} \tag{80}$$

Substituting (74) into (80), we have that with probability at least $1-3\delta$,

$$\begin{aligned}&\left(\theta_{t+1}-\theta_t\right)^\top \hat{H}_{\mathcal{L},t}\left(\theta_{t+1}-\theta_t\right)\\ =&\eta_{\text{global}}^2\sum_{i=1}^{d}\lambda_i\left|\left(V_t^{-1/2}m_t\right)^\top v_i\right|^2\\ \le&\eta_{\text{global}}^2\sum_{i=1}^{d}|\lambda_i|\left(\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\frac{\eta_{\text{local}}\tau_{K,G}}{\epsilon}+\frac{\eta_{\text{local}}\sigma_g\log(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)}{\sqrt{N}\epsilon}\right)^2\\ \le&\frac{2\eta^2\mathcal{I}L\tau_{K,G}^2}{\epsilon^2}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2+\frac{2\eta^2\sigma_g^2\mathcal{I}L\log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2}\end{aligned}$$

By taking summation from 0 to $T-1$, we have that with probability at least $1-3\delta$,

$$\begin{aligned}&\sum_{t=1}^{T}\left(\theta_{t+1}-\theta_t\right)^\top \hat{H}_{\mathcal{L},t}\left(\theta_{t+1}-\theta_t\right)\\ \le&\frac{2\eta^2\mathcal{I}LT\tau_{K,G}^2}{\epsilon^2}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2+\frac{2\eta^2\sigma_g^2\mathcal{I}LT\log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2}\end{aligned} \tag{81}$$

I.3.2 Bounding S_8

We first bound each term in S_8 with a fixed $t \in [T]$.

$$\begin{aligned}\left(\theta_{t+1}-\theta_t\right)^\top \hat{H}_{\mathcal{L},t}\left(\theta_t-\theta_{t-1}\right) &= \eta_{\text{global}}^2\left(V_t^{-1/2}m_t\right)^\top \hat{H}_{\mathcal{L},t}\left(V_{t-1}^{-1/2}m_{t-1}\right)\\ &= \eta_{\text{global}}^2\left(V_t^{-1/2}m_t\right)^\top\left(\sum_{i=1}^{d}\lambda_i v_i v_i^\top\right)\left(V_{t-1}^{-1/2}m_{t-1}\right)\\ &= \eta_{\text{global}}^2\sum_{i=1}^{d}\lambda_i\left(\left(V_t^{-1/2}m_t\right)^\top v_i\right)\left(\left(V_{t-1}^{-1/2}m_{t-1}\right)^\top v_i\right)\end{aligned}$$

$$\leq \eta_{\text{global}}^2 \sum_{i=1}^{d} \lambda_i \left| \left(V_t^{-1/2} m_t \right)^\top v_i \right| \left| \left(V_{t-1}^{-1/2} m_{t-1} \right)^\top v_i \right| \tag{82}$$

Substituting (73) and (74) into (82), we have that with probability at least $1-6\delta$,

$$\begin{aligned}
&\left(\theta_{t+1}-\theta_t\right)^\top \hat{H}_{\mathcal{L},t}\left(\theta_t-\theta_{t-1}\right) \\
=&\eta_{\text{global}}^2 \sum_{i=1}^{d} \lambda_i \left| \left(V_t^{-1/2} m_t \right)^\top v_i \right| \left| \left(V_{t-1}^{-1/2} m_{t-1} \right)^\top v_i \right| \\
\geq&-\eta_{\text{global}}^2 \sum_{i=1}^{d} |\lambda_i| \left(\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}} \right) \frac{\eta_{\text{local}}\tau_{K,G}}{\epsilon} + \frac{\eta_{\text{local}}\sigma_g \log(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)}{\sqrt{N}\epsilon} \right)^2 \\
\geq&-\frac{2\eta^2\mathcal{I}L\tau_{K,G}^2}{\epsilon^2}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2 - \frac{2\eta^2\sigma_g^2\mathcal{I}L\log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2}
\end{aligned}$$

By taking summation from 0 to $T-1$, we have that with probability at least $1-6\delta$,

$$\begin{aligned}
&\sum_{t=1}^{T}\left(\theta_{t+1}-\theta_t\right)^\top \hat{H}_{\mathcal{L},t}\left(\theta_t-\theta_{t-1}\right) \\
\geq&-\frac{2\eta^2\mathcal{I}LT\tau_{K,G}^2}{\epsilon^2}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2 - \frac{2\eta^2\sigma_g^2\mathcal{I}LT\log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2}
\end{aligned} \tag{83}$$

I.3.3 Bounding S_9

We first bound each term in S_9 with a fixed $t \in [T]$.

$$\begin{aligned}
\left(\theta_t-\theta_{t-1}\right)^\top \hat{H}_{\mathcal{L},t}\left(\theta_t-\theta_{t-1}\right) &= \eta_{\text{global}}^2 \left(V_{t-1}^{-1/2} m_{t-1} \right)^\top \hat{H}_{\mathcal{L},t} \left(V_{t-1}^{-1/2} m_{t-1} \right) \\
&= \eta_{\text{global}}^2 \left(V_{t-1}^{-1/2} m_{t-1} \right)^\top \left(\sum_{i=1}^{d} \lambda_i v_i v_i^\top \right) \left(V_{t-1}^{-1/2} m_{t-1} \right) \\
&= \eta_{\text{global}}^2 \sum_{i=1}^{d} \lambda_i \left| \left(V_{t-1}^{-1/2} m_{t-1} \right)^\top v_i \right|^2
\end{aligned} \tag{84}$$

Substituting (73) into (84), we have that with probability at least $1-3\delta$,

$$\begin{aligned}
&\left(\theta_t-\theta_{t-1}\right)^\top \hat{H}_{\mathcal{L},t}\left(\theta_t-\theta_{t-1}\right) \\
=&\eta_{\text{global}}^2 \sum_{i=1}^{d} \lambda_i \left| \left(V_{t-1}^{-1/2} m_{t-1} \right)^\top v_i \right|^2 \\
\leq&\eta_{\text{global}}^2 \sum_{i=1}^{d} |\lambda_i| \left(\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}} \right) \frac{\eta_{\text{local}}\tau_{K,G}}{\epsilon} + \frac{\eta_{\text{local}}\sigma_g \log(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)}{\sqrt{N}\epsilon} \right)^2 \\
\leq&\frac{2\eta^2\mathcal{I}L\tau_{K,G}^2}{\epsilon^2}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2 + \frac{2\eta^2\sigma_g^2\mathcal{I}L\log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2}
\end{aligned}$$

By taking summation from 0 to $T-1$, we have that with probability at least $1-3\delta$,

$$\sum_{t=1}^{T}\left(\theta_t-\theta_{t-1}\right)^\top \hat{H}_{\mathcal{L},t}\left(\theta_t-\theta_{t-1}\right)$$

$$\leq \frac{2\eta^2 \mathcal{I} LT \tau_{K,G}^2}{\epsilon^2}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2 + \frac{2\eta^2\sigma_g^2\mathcal{I}LT\log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2} \tag{85}$$

Substituting (81), (83), (85) into (79), we have that with probability at least $1-6\delta$,

$$\begin{aligned}
&\sum_{t=1}^{T}(\gamma_{t+1}-\gamma_t)^\top \hat{H}_{\mathcal{L},t}(\gamma_{t+1}-\gamma_t)\\
=&\frac{1}{(1-\beta_1)^2}\sum_{t=1}^{T}(\theta_{t+1}-\theta_t)^\top \hat{H}_{\mathcal{L},t}(\theta_{t+1}-\theta_t) - \frac{2\beta_1}{(1-\beta_1)^2}\sum_{t=1}^{T}(\theta_{t+1}-\theta_t)^\top \hat{H}_{\mathcal{L},t}(\theta_t-\theta_{t-1})\\
&+\frac{\beta_1^2}{(1-\beta_1)^2}\sum_{t=1}^{T}(\theta_t-\theta_{t-1})^\top \hat{H}_{\mathcal{L},t}(\theta_t-\theta_{t-1})\\
\leq&\frac{(1+\beta_1)^2}{(1-\beta_1)^2}\cdot\frac{2\eta^2\mathcal{I}LT\tau_{K,G}^2}{\epsilon^2}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2\\
&+\frac{(1+\beta_1)^2}{(1-\beta_1)^2}\cdot\frac{2\eta^2\sigma_g^2\mathcal{I}LT\log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2}\\
=&\frac{2\eta^2(1+\beta_1)^2\mathcal{I}LT\tau_{K,G}^2}{\epsilon^2(1-\beta_1)^2}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2\\
&+\frac{2\eta^2(1+\beta_1)^2\sigma_g^2\mathcal{I}LT\log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2(1-\beta_1)^2}
\end{aligned} \tag{86}$$

Substituting (63), (78), (86) into (36), we have that with probability at least $1-22\delta$,

$$\begin{aligned}
&\mathcal{L}(\gamma_T)-\mathcal{L}(\theta_0)\\
=&\sum_{t=1}^{T}\nabla\mathcal{L}(\theta_t)^\top(\gamma_{t+1}-\gamma_t)+\sum_{t=1}^{T}(\nabla\mathcal{L}(\gamma_t)-\nabla\mathcal{L}(\theta_t))^\top(\gamma_{t+1}-\gamma_t)+\frac{1}{2}\sum_{t=1}^{T}(\gamma_{t+1}-\gamma_t)^\top\hat{H}_{\mathcal{L},t}(\gamma_{t+1}-\gamma_t)\\
\leq&-\left(\eta_{\text{local}}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\tau_{K,G}+\frac{\eta_{\text{local}}\sigma_g\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)\log(2dT/\delta)}{\sqrt{N}}+\epsilon\right)^{-1}\eta K\sum_{t=1}^{T}\|\nabla\mathcal{L}(\theta_t)\|_2^2\\
&+\frac{2\eta K\sqrt{T}\log(2T/\delta)G^2}{\sqrt{N}\epsilon}+\frac{\eta\eta_{\text{local}}TLK^2G^2}{2\epsilon}+\frac{\eta G\sqrt{2TK}\log(2T/\delta)\sigma_s}{\sqrt{N}\epsilon}+\eta\max\left\{0,\frac{TG(KG-\tau)}{\epsilon}\right\}\\
&+\frac{15\eta\sqrt{d}\log^3(24NTd/\delta)\sqrt{T}G\tau_{K,G}}{\sqrt{b_{\text{in}}b_{\text{out}}}\epsilon}+\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}(2+\beta_1)TG\tau_{K,G}^2}{\epsilon^2(1-\beta_1)}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2\\
&+\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}(1+2\beta_1)TG\sigma_g^2\log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2(1-\beta_1)}\\
&+\frac{\log^2(2T/\delta)G^2}{\epsilon}+\frac{2\eta^2T\log^2(2T/\delta)\sigma_g^2\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon}\\
&+\frac{2\eta^2\beta_1(1+\beta_1)\mathcal{I}LT\tau_{K,G}^2}{\epsilon^2(1-\beta_1)^2}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2\\
&+\frac{2\eta^2\beta_1(1+\beta_1)\sigma_g^2\mathcal{I}LT\log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2(1-\beta_1)^2}
\end{aligned}$$

$$
\begin{aligned}
&+\frac{\eta^2(1+\beta_1)^2\mathcal{I}LT\tau_{K,G}^2}{\epsilon^2(1-\beta_1)^2}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2\\
&+\frac{\eta^2(1+\beta_1)^2\sigma_g^2\mathcal{I}LT\log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2(1-\beta_1)^2}\\
=&-\left(\eta_{\text{local}}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)\tau_{K,G}+\frac{\eta_{\text{local}}\sigma_g\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)\log(2dT/\delta)}{\sqrt{N}}+\epsilon\right)^{-1}\eta K\sum_{t=1}^{T}\|\nabla\mathcal{L}(\theta_t)\|_2^2\\
&+\frac{2\eta K\sqrt{T}\log(2T/\delta)G^2}{\sqrt{N}\epsilon}+\frac{\eta\eta_{\text{local}}TLK^2G^2}{2\epsilon}+\frac{\eta G\sqrt{2TK}\log(2T/\delta)\sigma_s}{\sqrt{N}\epsilon}+\eta\max\left\{0,\frac{TG(KG-\tau)}{\epsilon}\right\}\\
&+\frac{15\eta\sqrt{d}\log^3(24NTd/\delta)\sqrt{T}G\tau_{K,G}}{\sqrt{b_{\text{in}}b_{\text{out}}}\epsilon}+\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}(2+\beta_1)TG\tau_{K,G}^2}{\epsilon^2(1-\beta_1)}\left(1+\frac{10\sqrt{d}\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2\\
&+\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}(1+2\beta_1)TG\sigma_g^2\log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2(1-\beta_1)}\\
&+\frac{\log^2(2T/\delta)G^2}{\epsilon}+\frac{2\eta^2T\log^2(2T/\delta)\sigma_g^2\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon}\\
&+\frac{\eta^2(1+3\beta_1)(1+\beta_1)\mathcal{I}LT\tau_{K,G}^2}{\epsilon^2(1-\beta_1)^2}\left(1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}}\right)^2\\
&+\frac{\eta^2(1+3\beta_1)(1+\beta_1)\sigma_g^2\mathcal{I}LT\log^2(2dT/\delta)\left(1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right)\right)^2}{N\epsilon^2(1-\beta_1)^2}
\end{aligned}
$$

Since

$$\mathcal{L}(\theta_0)-\mathcal{L}(\gamma_T)\leq\mathcal{L}(\theta_0)-\mathcal{L}^*$$

we can get that with probability at least $1-22\delta$,

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\theta_t)\|_2^2\leq&\frac{\alpha_3\left(\mathcal{L}(\theta_0)-\mathcal{L}^*\right)}{\eta KT}+\frac{2\alpha_3\log(2T/\delta)G^2}{\sqrt{NT}\epsilon}+\frac{\eta_{\text{local}}\alpha_3LKG^2}{2\epsilon}+\frac{\sqrt{2}\alpha_3G\log(2T/\delta)\sigma_s}{\sqrt{NTK}\epsilon}\\
&+\max\left\{0,\frac{\alpha_3G(KG-\tau)}{K\epsilon}\right\}+\frac{15\alpha_3\log^3(24NTd/\delta)\sqrt{d}G\tau_{K,G}}{\sqrt{b_{\text{in}}b_{\text{out}}}\sqrt{T}K\epsilon}+\frac{\eta_{\text{local}}\alpha_1^2\alpha_3(2+\beta_1)\sqrt{1-\beta_2}G\tau_{K,G}^2}{K\epsilon^2(1-\beta_1)}\\
&+\frac{\eta_{\text{local}}\alpha_2^2\alpha_3(1+2\beta_1)\sqrt{1-\beta_2}\sigma_g^2G\log^2(2dT/\delta)}{NK\epsilon^2(1-\beta_1)}+\frac{\alpha_3\log^2(2T/\delta)G^2}{\eta TK\epsilon}+\frac{2\eta\alpha_2^2\alpha_3\log^2(2T/\delta)\sigma_g^2}{NK\epsilon}\\
&+\frac{\eta\alpha_1^2\alpha_3\left(1+3\beta_1\right)\left(1+\beta_1\right)\mathcal{I}L\tau_{K,G}^2}{K\epsilon^2(1-\beta_1)^2}+\frac{\eta\alpha_2^2\alpha_3(1+3\beta_1)(1+\beta_1)\sigma_g^2\mathcal{I}L\log^2(2dT/\delta)}{NK\epsilon^2(1-\beta_1)^2}
\end{aligned}
$$

in which

$$
\begin{aligned}
&\alpha_1=1+\frac{10\log^{1.5}(6NTd^2/\delta)\log(24NTd/\delta)\sqrt{d}}{\sqrt{b_{\text{in}}b_{\text{out}}}},\alpha_2=1+8\sqrt{\frac{d}{b_{\text{out}}}}\log\left(\frac{4dT}{\delta}\right),\\
&\alpha_3=\eta_{\text{local}}\alpha_1\tau_{K,G}+\frac{\eta_{\text{local}}\sigma_g\alpha_2\log(2dT/\delta)}{\sqrt{N}}+\epsilon
\end{aligned}
$$

with $\tau_{K,G}=\min\{\tau,KG\}$, then we finish the proof. □

J Additional Experimental Results

J.1 Model Performance

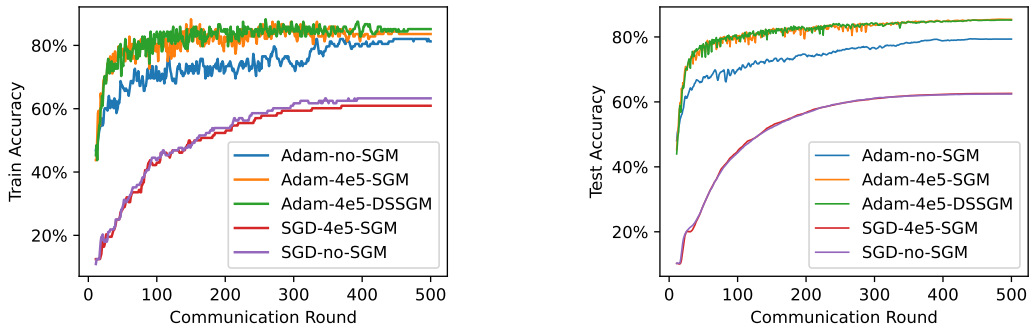
We will present more experiment results under different privacy levels. These figures further verifies our observations and conclusions in Section 5.[1]

J.1.1 Vision Task

Training dynamics and test accuracies of Fed-DSSGM with Adam, Fed-SGM with Adam, Fed-SGM with GD, DP-FedAvg, and the unsketched Adam baseline on ResNet101 trained on EMNIST with $\epsilon_p \in \{2.75, 0.42, 0.18\}$ are shown in Figures 2–4.

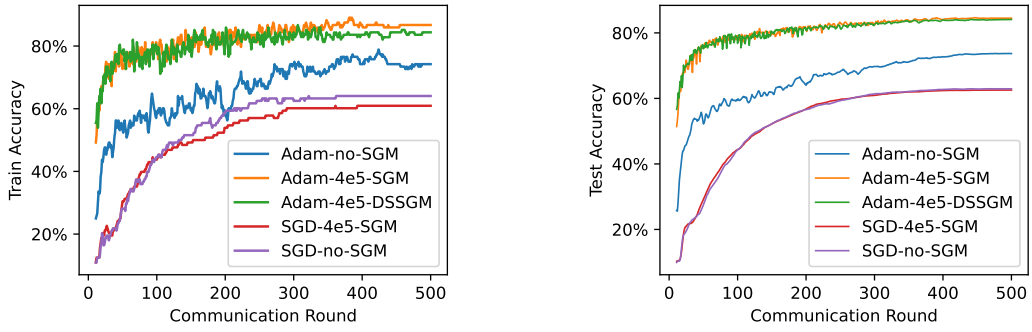
J.1.2 Language Task

Training dynamics and test accuracies of Fed-DSSGM with Adam, Fed-SGM with Adam, and the unsketched Adam baseline on BERT trained on SST-2 at privacy levels $\epsilon_p \in 2.45, 1.44, 0.35, 0.12$ are shown in Figures 5–8.



(a) Training Accuracy (b) Test Accuracy

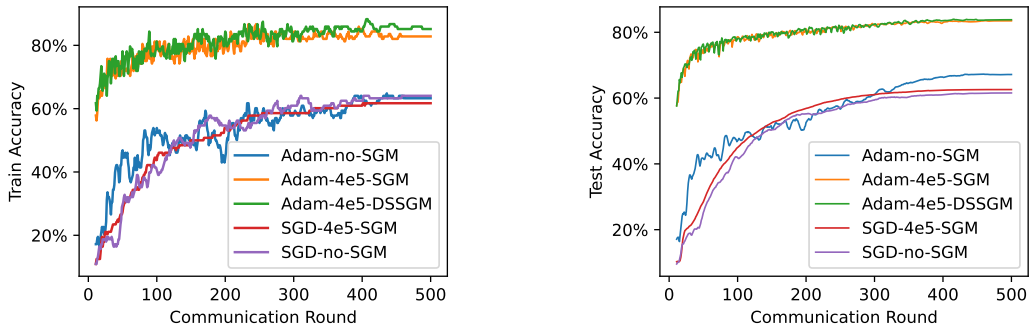
Figure 2: Vision Task: EMNIST, $\epsilon_p = 2.75$



(a) Training Accuracy (b) Test Accuracy

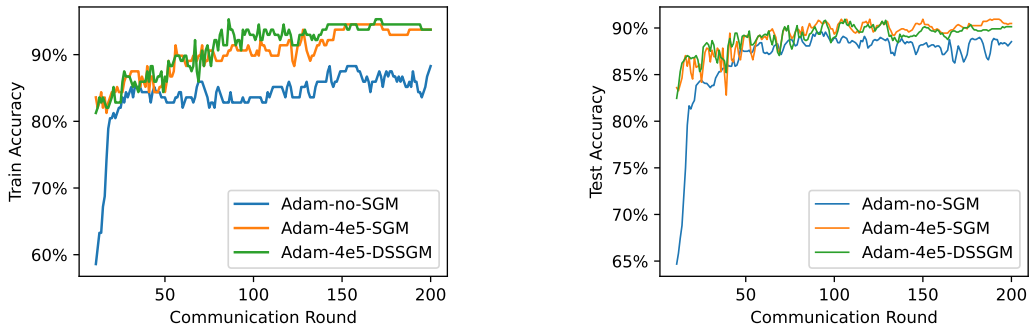
Figure 3: Vision Task: EMNIST, $\epsilon_p = 0.42$

[1]All experiments in this section and Section 5 are conducted on a computing cluster with an AMD EPYC 7713 64-core processor and an NVIDIA A100 Tensor Core GPU, and the code is provided at https://anonymous.4open.science/r/mlopt-D4E3.



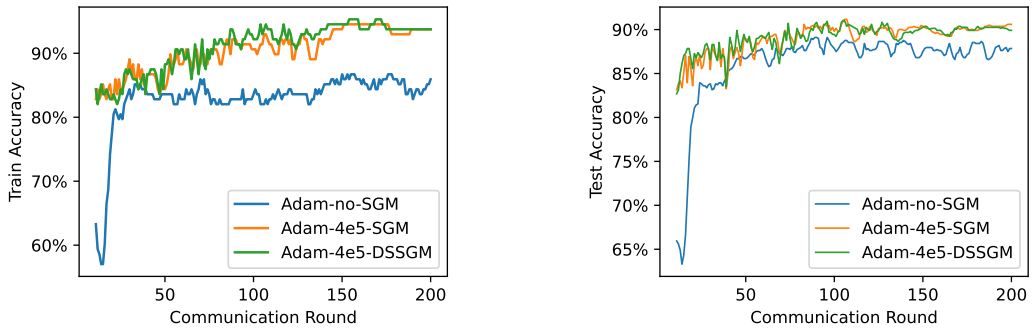
(a) Training Accuracy (b) Test Accuracy

Figure 4: Vision Task: EMNIST, $\epsilon_p = 0.18$



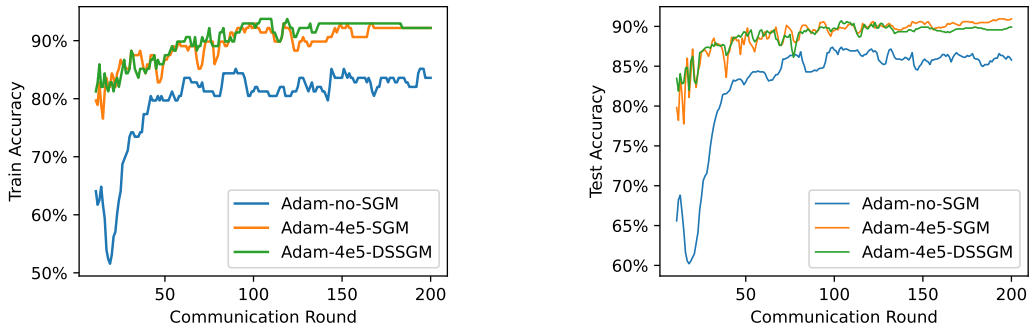
(a) Training Accuracy (b) Test Accuracy

Figure 5: Language Task: SST-2, $\epsilon_p = 2.45$



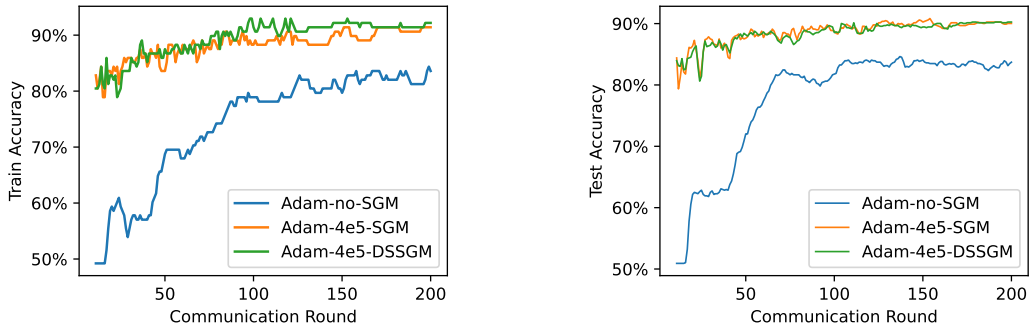
(a) Training Accuracy (b) Test Accuracy

Figure 6: Language Task: SST-2, $\epsilon_p = 1.44$



(a) Training Accuracy (b) Test Accuracy

Figure 7: Language Task: SST-2, $\epsilon_p = 0.35$



(a) Training Accuracy (b) Test Accuracy

Figure 8: Language Task: SST-2, $\epsilon_p = 0.12$

J.2 Communication-Cost Accounting

In this section, we provide detailed communication-cost accounting for the cryptographic constructions considered in this work, including the public-key seed-sharing baseline and the threshold-resilient realization. We measure the total volume of information transmitted through the server in one communication round. Let N denote the number of participating clients and b the sketch dimension. For the concrete calculations below, we use the EMNIST setting from our experiments, where $b = 400{,}000$, and represent transmitted sketch values and MPC field elements using 32 bits (4 bytes).

Plaintext Fed-SGM baseline. In plaintext Fed-SGM, each participating client transmits one b-dimensional sketched update to the server. The total client-to-server payload is therefore

$$\mathrm{Comm}_{\text{Fed-SGM}} = Nb \times 4 \text{ bytes}.$$

For $b = 400{,}000$, this gives

$$\mathrm{Comm}_{\text{Fed-SGM}} = \begin{cases} 6.40 \text{ MB}, & N = 4, \\ 160.00 \text{ MB}, & N = 100. \end{cases}$$

Public-key seed sharing. Protocol D.1 establishes a common server-hidden seed before the standard Fed-SGM update transmission. We instantiate the encrypted seed exchange using the Elliptic Curve Integrated Encryption Scheme (ECIES) (Abdalla et al., 2001; Standards for Efficient Cryptography Group (SECG), 2009) over Curve25519 (Bernstein, 2006). For a $\lambda = 256$-bit (32-byte) seed, our concrete ECIES encoding uses a 32-byte encrypted seed payload, a 33-byte SEC1-compressed ephemeral point, and a 16-byte MAC authentication tag, giving $32 + 33 + 16 = 81$ bytes per encrypted seed.

Each of the N participating clients encrypts its seed separately for each of the other $N-1$ clients. The server therefore relays $N(N-1)$ ciphertexts, and the additional seed-establishment communication is

$$\mathrm{Comm}_{\text{PK,extra}} = 81N(N-1) \text{ bytes}.$$

Hence the total per-round communication of the public-key realization is

$$\mathrm{Comm}_{\text{PK}} = 4Nb + 81N(N-1) \text{ bytes}.$$

For $N = 4$, the server relays $4 \times 3 = 12$ ciphertexts, giving

$$12 \times 81 = 972 \text{ bytes} \approx 0.97 \text{ KB}.$$

Thus,

$$\mathrm{Comm}_{\text{PK}} \approx 6.40 \text{ MB} + 0.97 \text{ KB}.$$

For $N = 100$, the additional ciphertext volume is

$$81 \times 100 \times 99 = 801{,}900 \text{ bytes} \approx 0.80 \text{ MB},$$

giving a total of approximately 160.80 MB per round. Since this additional term scales as $\mathcal{O}(N^2\lambda)$ and is independent of b, it is negligible relative to the $\mathcal{O}(Nb)$ sketched-update transmission when $b \gg N$.

Threshold-resilient MPC. Protocol 3.1 keeps the shared sketch hidden throughout its generation and application. For the concrete accounting, we instantiate the protocol using standard N-party honest-majority MPC based on Beaver triples and secure secret sharing (Damgard and Nielsen, 2007). Its communication consists of the following components.

1. **Seed sharing.** Each client distributes shares of its local 256-bit seed using Pedersen-style verifiable secret sharing (Shamir, 1979; Pedersen, 1991). We represent the seed as two 128-bit chunks and apply a $(\kappa+1)$-out-of-N VSS instance to each chunk. In each instance, the dealer sends two 32-byte scalar shares to each of the other $N-1$ participants and disseminates $\kappa+1$ coefficient commitments, each represented by a 33-byte compressed group element. The total server-relayed seed-sharing traffic is therefore

$$\mathrm{Comm}_{\text{VSS}} = 2N(N-1)\big(2 \cdot 32 + 33(\kappa+1)\big) \text{ bytes}.$$

   This term is independent of b and is therefore lower order in the large-sketch regime considered here.

2. **Distributed pseudorandom generation.** Clients execute the Naor–Pinkas–Reingold distributed PRF (Naor et al., 1999) to generate the shared pseudorandom values defining the sketch without reconstructing the common seed. Under the MPC realization used in our accounting, this requires approximately two transmitted field elements per generated sketch coordinate. The communication is therefore

$$\mathrm{Comm}_{\mathrm{dPRG}} \approx 2Nb \quad \text{field elements},$$

or

$$8Nb \text{ bytes}.$$

3. **Secure sketch application.** Client j securely combines the secret-shared sketch $\langle R_t \rangle$ with its private clipped update $\hat{\Delta}_{j,t}$ to obtain

$$\langle \hat{\Delta}^{\mathrm{sk}}_{j,t} \rangle = \langle R_t \rangle \hat{\Delta}_{j,t}.$$

In the Beaver-triple-based MPC realization (Damgard and Nielsen, 2007), each batched secure multiplication requires opening two masked values. Thus, under our row-wise accounting, each participating client transmits approximately two field elements per sketch coordinate, giving

$$\mathrm{Comm}_{\mathrm{Mult}} \approx 2Nb \quad \text{field elements},$$

or

$$8Nb \text{ bytes}.$$

4. **Secure aggregation.** After noise addition, each client contributes its final b-dimensional noisy share to secure aggregation (Bonawitz et al., 2017b). This contributes another

$$\mathrm{Comm}_{\mathrm{SecAgg}} = Nb \quad \text{field elements},$$

or

$$4Nb \text{ bytes}.$$

Combining the three sketch-dependent components gives

$$2Nb + 2Nb + Nb = 5Nb$$

transmitted field elements. With 32-bit field elements, this corresponds to

$$\mathrm{Comm}_{\mathrm{TR,sketch}} = 20Nb \text{ bytes}.$$

Including the lower-order seed-sharing traffic, the communication accounting is

$$\mathrm{Comm}_{\mathrm{TR}} = 20Nb + 2N(N-1)\big(2 \cdot 32 + 33(\kappa+1)\big) \text{ bytes}.$$

Because the second term is independent of b, for the large sketch dimensions considered in our experiments the dominant cost is $16Nb$ bytes, i.e., approximately four times the plaintext Fed-SGM communication.

For $N = 4$ and $b = 400{,}000$, the honest-majority requirement $\kappa < N/2$ gives $\kappa = 1$, and

$$20Nb = 20 \times 4 \times 400{,}000 = 32{,}000{,}000 \text{ bytes} = 32.00 \text{ MB},$$

while the seed-sharing term contributes only

$$2 \times 4 \times 3\big(2 \cdot 32 + 2 \cdot 33\big) = 3{,}120 \text{ bytes} = 3.12 \text{ KB}.$$

Thus the total volume under this accounting is

$$32{,}000{,}000 + 3{,}120 = 32{,}003{,}120 \text{ bytes} = 32.00312 \text{ MB},$$

which rounds to 32.00 MB.

Summary. The two cryptographic realizations introduce qualitatively different communication overheads. Public-key seed sharing adds

$$81N(N-1) = \mathcal{O}(N^2\lambda)$$

bytes independently of the sketch dimension. The threshold-resilient realization instead increases the dominant sketch-dependent traffic from

$$4Nb \text{ bytes} \quad \text{to} \quad 20Nb \text{ bytes},$$

while also incurring the b-independent VSS seed-sharing cost

$$2N(N-1)\big(2 \cdot 32 + 33(\kappa+1)\big) \text{ bytes}.$$

For the large sketch dimensions considered in our experiments, these seed-establishment terms are negligible relative to the $\Theta(Nb)$ sketch-dependent communication. Thus, in this regime, the dominant communication of both cryptographic realizations retains the same linear dependence on the sketch dimension as plaintext Fed-SGM, while the threshold-resilient construction incurs an approximately $4\times$ constant-factor increase in the dominant sketch-dependent traffic.

AI USE STATEMENT

In this work, we used generative AI tools to assist with polishing the writing, improving clarity, and refining the organization of the manuscript.

We have reviewed all AI-assisted work. We take full responsibility for the final content of this work.