# Sketched Gaussian Mechanism for Private Federated Learning

**Qiaobo Li**
Siebel School of Computing and Data Science
University of Illinois Urbana-Champaign
qiaobol2@illinois.edu

**Zhijie Chen**
Siebel School of Computing and Data Science
University of Illinois Urbana-Champaign
lucmon@illinois.edu

**Arindam Banerjee**
Siebel School of Computing and Data Science
University of Illinois Urbana-Champaign
arindamb@illinois.edu

## Abstract

Communication cost and privacy are two major considerations in federated learning (FL). For communication cost, gradient compression by sketching the clients' transmitted model updates is often used for reducing per-round communication. For privacy, the Gaussian mechanism (GM), which consists of clipping updates and adding Gaussian noise, is commonly used to guarantee client-level differential privacy. Existing literature on private FL analyzes privacy of sketching and GM in an isolated manner, illustrating that sketching provides privacy determined by the sketching dimension and that GM has to supply any additional desired privacy.

In this paper, we introduce the Sketched Gaussian Mechanism (SGM), which directly combines sketching and the Gaussian mechanism for privacy. Using Rényi-DP tools, we present a joint analysis of SGM's overall privacy guarantee, which is significantly more flexible and sharper compared to isolated analysis of sketching and GM privacy. In particular, we prove that the privacy level of SGM for a fixed noise magnitude is proportional to $1/\sqrt{b}$, where $b$ is the sketching dimension, indicating that (for moderate $b$) SGM can provide much stronger privacy guarantees than the original GM under the same noise budget. We demonstrate the application of SGM to FL with either gradient descent or adaptive server optimizers, and establish theoretical results on optimization convergence, which exhibits only a logarithmic dependence on the number of parameters $d$. Experimental results confirm that at the same privacy level, SGM based FL is at least competitive with non-sketching private FL variants and outperforms them in some settings. Moreover, using adaptive optimization at the server improves empirical performance while maintaining the privacy guarantees.

## 1 Introduction

Federated learning (FL) [44] is a widely used machine learning framework where a shared global model is trained under the coordination of a central server using data distributed across various clients. In FL, each client carries out training steps on its local data and sends only the model updates back to the server, which then refines the global model by suitably aggregating these local updates. Because the data remains on the clients, FL can offer enhanced privacy protection compared to traditional centralized learning. Nonetheless, FL faces two significant challenges: (1) it lacks a rigorous privacy guarantee (e.g., differential privacy (DP)) [17] and has been shown to be susceptible to various

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

inference attacks, leading to local information leakage during training [49, 50, 75, 83, 84]; and (2) it usually requires a high communication overhead due to the frequent communication between the server and the client [35]. Many recent advances in FL have been motivated by these two challenges.

Towards reducing communication costs, a key goal has been to reduce the communication cost per round by compressing local updates from clients to a lower dimension [30, 36], including sparsification [40, 70, 5], quantization [4, 41, 52], and sketching [59, 54, 28]. Among these approaches, sketching methods stand out for its simplicity, making it easier to integrate with existing FL methods. As an unbiased compressor, it does not require bias correction using error feedback mechanisms, unlike sparsification, which incurs additional memory costs [57, 60]. As a linear operation, sketching ensures the geometric properties are approximately preserved after compression [13] as opposed to quantization distorting the inner product structure, potentially slowing convergence [4].

Towards preserving privacy, two DP definitions are commonly considered in FL algorithm design: sample-level privacy and client-level privacy, where client-level privacy is the stricter guarantee in the sense that it ensures the output remains statistically indistinguishable when an entire client's dataset is altered, rather than merely hiding the inclusion or exclusion of individual data samples as sample-level privacy [81]. Various FL algorithms [21, 71, 63, 66] have focused on client-level privacy by adapting standard differential privacy techniques from centralized training (i.e., clipping gradients and adding Gaussian noise to the clipped values [1, 19]) to the federated learning setting.

Though privacy and communication-efficiency have mostly been studied independently, there have been some efforts towards solving these two challenges together. $D^2$P-FED[71] combines stochastic quantization and random rotation [44] techniques to lower communication costs, to the discrete Gaussian mechanism for privacy [9], along with secure aggregation [7] to mitigate noise magnitude. Alternatively, DPSFL [80] directly applies the standard Gaussian mechanism to the communication-efficient federated learning algorithm FetchSGD [54]. However, these methods merely select individual compression and privacy components and stitch them together, treating communication-efficiency and differential privacy as separate concerns rather than examining their intrinsic interplay.

[37] is one of the very few works to investigate the underlying relationship between differential privacy and communication efficiency in distributed learning. It demonstrates that the Count Sketch algorithm [10], despite being originally developed for communication efficiency, inherently satisfies a form of differential privacy for distributed learning algorithms. However, their results have several major limitations. First, when the privacy level $\epsilon$ provided by the Count Sketch mechanism falls short of the requirement, additional Laplacian or Gaussian noise must be injected, but there is no precise characterization of the amount of noise required. Second, their theoretical guarantees require arguably impractical assumptions, including the input (client gradients) to follow a Gaussian distribution and the sketching dimension $b \lesssim \sqrt{d}$ where $d$ is the input (gradient) dimension, which contradicts typical empirical configurations in sketched distributed learning, where $b$ is often a fixed fraction of $d$, e.g., $b = d/100$, to avoid derailing the optimization.

In this paper, we introduce the Sketched Gaussian Mechanism (SGM), which combines an isometric Gaussian sketching transform [59] with the classical Gaussian mechanism. Leveraging tools from Rényi differential privacy (RDP) [46] with the subsampling, post-processing, and composition theorems of differential privacy [18], we derive a tight upper bound on the overall privacy level $\epsilon$ of SGM. Concretely, with all other hyperparameters held fixed, $\epsilon = O(\frac{1}{\sqrt{b}\sigma_g^2})$, where $b$ denotes the sketch dimension and $\sigma_g^2$ the variance of the added Gaussian noise. Our new result for SGM **establishes a clear dependence of $\epsilon$ on both $b$ and $\sigma_g^2$**. Unlike prior work, our result imposes no restrictive upper limit on $b$, thereby covering practical regimes in which $b$ is a fixed fraction of the ambient dimension $d$. More importantly, this bound implies that for suitably large $b$, SGM achieves strictly stronger privacy guarantees than the standard Gaussian mechanism, demonstrating that the sketching operation itself confers inherent privacy benefits.

We further integrate SGM into a federated learning (FL) framework, referred to as Fed-SGM, supporting flexible choices of server optimizers to match practical deployment needs. We prove that this Fed-SGM satisfies client-level privacy guarantees. Moreover, by fully leveraging the fast-decaying spectrum of the deep-learning loss Hessian [77, 82, 76], which implies a small absolute intrinsic dimension [26], we establish rigorous optimization convergence bounds that scale only logarithmically in the ambient dimension $d$ and linearly in this absolute intrinsic dimension, ensuring scalability to high-dimensional problems.

We empirically validate our approach on deep learning models for both vision and language benchmarks. Across these tasks, federated SGM requires strictly less Gaussian noises than the standard DP-FedAvg algorithm [81] to achieve the same privacy guarantee, and consistently delivers comparable or even superior model accuracy. Furthermore, when we replace vanilla gradient descent on the server with an adaptive optimizer, we observe additional accuracy gains, highlighting the practical benefit of incorporating adaptive optimization into our privacy-preserving federated learning framework.

The remainder of the paper is organized as follows. Section 2 provides the description of our SGM and establishes the privacy guarantee. Section 3 introduces Fed-SGM, the application of SGM into federated learning setting, and establish optimization analysis of our framework. In Section 4, we present experimental evaluations that assess performances of our Fed-SGM and compare them with existing approaches. Finally, Section 5 concludes our contributions and highlight potential future directions. Due to space limit, we present our discussion on related works in Appendix A.

## 2 Privacy Guarantee of SGM

In this section, we will introduce the Sketched Gaussian Mechanism (SGM), and provide the privacy guarantee with two different kinds of analysis.

**Definition 2.1** (Sketched Gaussian Mechanism (SGM))**.** For any statistic $\theta(D) \in \mathbb{R}^d$ of the dataset $D$, the Sketched Gaussian Mechanism outputs $\mathcal{SG}(\theta; R, \xi) = R\theta + \xi$, in which $R \in \mathbb{R}^{b\times d}$ is a Gaussian sketching matrix with each entry sampled i.i.d. from $\mathcal{N}(0, \frac{1}{\sqrt{b}})$ and $\xi \in \mathbb{R}^b$ follows the Gaussian distribution $\mathcal{N}\left(0, \sigma_g^2 \mathbb{I}_b\right)$.

**Algorithm 1** Sketched Gaussian Mechanism

**Hyperparameters:** learning rate $\eta_t$, noise scale $\sigma_g$, clipping threshold $\tau$, number of iterations $T$.
**Inputs:** Examples $D = \{x_1, \ldots, x_n\}$, loss function $\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^n \mathcal{L}(\theta, x_i)$.
  **Initialize** $\theta_0$ randomly.
  **for** $t = 0, \ldots, T-1$ **do**
    Take a mini-batch $B_t$ of $m$ samples with sampling probability $q = \frac{m}{n}$
    **Computer gradient:** For each $i \in B_t$, $g_t(x_i) \leftarrow \nabla\mathcal{L}(\theta_t, x_i)$
    **Clip the gradient:** $\hat{g}_t(x_i) = \text{clip}(g_t(x_i), \tau) = g_t(x_i) \cdot \min\left\{1, \frac{\tau}{\|g_t(x_i)\|_2}\right\}$
    **Apply SGM:** For each $i \in B_t$, $\tilde{g}_t(x_i) = \mathcal{SG}(\hat{g}_t(x_i); R_t, \xi_{t,i})$
    **Aggregate:** $\bar{\tilde{g}}_t = \frac{1}{m}\sum_{i\in B_t} \tilde{g}_t(x_i)$
    **Update parameter:** $\theta_{t+1} \leftarrow \texttt{OPT}\left(\theta_t, \bar{\tilde{g}}_t, \eta_t\right)$
  **end for**
**Outputs:** final model $\theta_T$

Algorithm 1 outlines a standard application of SGM in training a model with parameter $\theta$ in a similar manner to the standard DP-SGD [1]. At each step of training, we compute the gradient for a random subset of examples, apply clipping and SGM to each gradient, and update the parameter with the optimizer `OPT` using the aggregated gradient.

Denote $\gamma_t(D) = \sum_{i\in B_t} \hat{g}_t(x_i)$, then $\|\gamma_t(D)\|_2 \le m\tau$. Notice that sketching with the random matrix $R_t$ is a linear operation, so we can rewrite the aggregated gradient $\bar{\tilde{g}}_t$ as:

$$\bar{\tilde{g}}_t = \frac{1}{m}\sum_{i\in B_t} \tilde{g}_t(x_i) = \frac{1}{m}\sum_{i\in B_t} \mathcal{SG}(\hat{g}_t(x_i); R_t, \xi_{t,i}) = \frac{1}{m}\sum_{i\in B_t} (R_t\hat{g}_t(x_i) + \xi_{t,i})$$

$$= \frac{1}{m}\left(R_t\left(\sum_{i\in B_t}\hat{g}_t(x_i)\right) + \sum_{i\in B_t}\xi_{t,i}\right) = \frac{1}{m}\left(R_t\gamma_t(D) + \sum_{i\in B_t}\xi_{t,i}\right) = \frac{\mathcal{SG}(\gamma_t(D); R_t, \xi_t)}{m}$$

where $\xi_t = \sum_{i\in B_t}\xi_{t,i}$ is a Gaussian vector with covariance matrix $m\sigma_g^2\mathbb{I}_b$. Therefore, the aggregated gradient $\bar{\tilde{g}}_t$ is also an output of SGM based on the examples.

We study the privacy guarantee of Algorithm 1 subject to the rigorous privacy guarantees of Differential Privacy (DP)[17], whose formal definition is given below.

**Definition 2.2.** A randomized algorithm $\mathcal{M}$ is $(\epsilon, \delta)$-differentially private if for any pair of datasets $D, D'$ differ in exactly one data point and for all event $\mathcal{Y}$ in the output range of $\mathcal{M}$, we have

$$\mathbb{P}\left\{\mathcal{M}(D) \in \mathcal{Y}\right\} \leq e^{\epsilon}\mathbb{P}\left\{\mathcal{M}(D') \in \mathcal{Y}\right\} + \delta$$

where the probability is taken over the randomness of $\mathcal{M}$.

### 2.1 Warm Up: Privacy Analysis by Moments Accountant

Compared to the standard Gaussian mechanism [18], SGM incorporates an additional sketching operation. By the Johnson–Lindenstrauss lemma [13], sketching can be interpreted as a distance-preserving embedding that approximates the original high-dimensional update in a lower-dimensional subspace. Under this viewpoint of sketching, we obtain the privacy guarantee for Algorithm 1 by an extension of the moments accountant analysis of Gaussian mechanism [1]. We state the main theorem below with the full proof in Appendix C.1.

**Theorem 2.1.** *There exists constants $c_1$ and $c_2$ so that given the sampling probability $q = \frac{m}{n}$ and the number of steps $T$, for any $\epsilon_p < c_1 q^2 T$, Algorithm 1 is $(\epsilon_p, \delta_p)$-differentially private for any $\delta_p > 0$ if we choose*

$$\sigma_g \geq c_2 \frac{\tau\sqrt{\left(1 + \frac{\log^{1.5}(2mT/\delta_p)}{\sqrt{b}}\right) mT\log(2/\delta_p)}}{n\epsilon_p} \,. \tag{1}$$

**Remark 2.1.** This privacy bound decreases monotonically in the sketching dimension $b$. However, since the dependence is of the form $(1 + O(1/\sqrt{b}))$, even for large $b$, the requisite Gaussian noise variance $\sigma_g^2$ remains asymptotically at par with that of standard DP-SGD [1]. Here, the analysis attributes the entire privacy solely to the Gaussian mechanism, treating sketching purely as an approximation and omitting any potential privacy that the sketching step might confer. □

### 2.2 Main Result: Privacy Analysis by Rényi Differential Privacy

To more precisely characterize the privacy guarantees of the Sketched Gaussian Mechanism and to investigate any privacy contributions imparted by the sketching step, we develop the following novel privacy guarantee of Algorithm 1.

**Theorem 2.2.** *There exists constants $c_3$ and $c_4$ so that given the sampling probability $q = \frac{m}{n}$ and the number of steps $T$, for any $\epsilon_p \leq c_3 q\sqrt{T}$, Algorithm 1 is $(\epsilon_p, \delta_p)$-differentially private for any $\delta_p > 0$ if we choose*

$$\sigma_g^2 \geq \frac{c_4 q\tau^2\sqrt{T}\log(2qT/\delta_p)}{\sqrt{b}\epsilon_p} \,. \tag{2}$$

**Remark 2.2.** In Theorem 2.2, for any fixed privacy target $\epsilon_p$, the required Gaussian noise variance $\sigma_g^2$ is a monotonically decreasing function of the sketch dimension $b$, so sketching to moderately high dimensions $b$ provably reduces the marginal variance of the Gaussian mechanism. Moreover, compared to Theorem 2.1, one attains the same privacy guarantee with strictly less noise once the sketch dimension satisfies $b \geq \Omega\left(\frac{(n\epsilon_p \log(2pT/\delta_p))^2}{T\log^2(1/\delta_p)}\right)$. □

We refer readers to Appendix C.2 for a detailed proof, and provide a high-level sketch here. Our analysis leverages tools from the Rényi Differential Privacy (RDP) framework [46].

**Definition 2.3** (Rényi divergence [53])**.** For two probability distributions $P$ and $Q$ defined over $\mathcal{R}$, the Rényi divergence of order $\alpha > 1$ is $D_\alpha\left(P\|Q\right) \triangleq \frac{1}{\alpha-1}\log\mathbb{E}_{x\sim Q}\left(\frac{P(x)}{Q(x)}\right)^\alpha$.

From the definition of SGM, $\mathcal{SG}(\gamma_t(D); R_t, \xi_t) \sim \mathcal{N}\left(0, \left(\frac{\|\gamma_t(D)\|^2}{b} + m\sigma_g^2\right)\mathbb{I}_b\right)$, which can be used to show the following result:

**Lemma 2.1.** *The Rényi divergence between $\mathcal{SG}(\gamma_t; R_t, \xi_t)$ for neighboring datasets $D, D'$ is*

$$D_\alpha\left(\mathcal{SG}(\gamma_t(D); R_t, \xi_t)\|\mathcal{SG}(\gamma_t(D'); R_t, \xi_t)\right) = bf_\alpha\left(\sqrt{\frac{\|\gamma_t(D')\|^2 + mb\sigma_g^2}{\|\gamma_t(D)\|^2 + mb\sigma_g^2}}\right) \,,$$

*where $f_\alpha(x) = \log x + \frac{1}{2(\alpha-1)}\log\frac{x^2}{\alpha x^2 + 1 - \alpha}$.*

Since this divergence purely depends on the ratio $\sqrt{\frac{\|\gamma_t(D')\|^2+mb\sigma_g^2}{\|\gamma_t(D)\|^2+mb\sigma_g^2}}$, we define the ratio sensitivity of the statistic analogous to the classical sensitivity measure in the Gaussian mechanism [18].

**Definition 2.4** (Ratio Sensitivity)**.** For any constant $c \geq 0$, define the ratio sensitivity of $\theta$ as

$$\text{rsens}_c(\theta) = \sup_{D,D'} \sqrt{\frac{\|\theta(D')\|^2 + c^2}{\|\theta(D)\|^2 + c^2}}, \tag{3}$$

where the supremum is over all neighboring datasets $D, D'$.

From the definition, a direct analysis shows

$$\sqrt{1 - \frac{2\tau^2}{b\sigma_g^2}} \leq \frac{1}{\text{rsens}_{\sqrt{mb}\sigma_g}(\gamma_t)} \leq 1 \leq \text{rsens}_{\sqrt{mb}\sigma_g}(\gamma_t) \leq \sqrt{1 + \frac{2\tau^2}{b\sigma_g^2}}. \tag{4}$$

Since $f_\alpha(x)$ is monotonically decreasing for $x \leq 1$ and increasing for $x \geq 1$, we can obtain the following bound on the Rényi divergence.

**Lemma 2.2.** *For any neighboring datasets $D, D'$,*

$$D_\alpha\left(\mathcal{SG}(\gamma_t(D)); R_t, \xi_t) \parallel \mathcal{SG}(\gamma_t(D')); R_t, \xi_t)\right) \leq b \max\left\{f_\alpha\left(\sqrt{1 + \frac{2\tau^2}{b\sigma_g^2}}\right), f_\alpha\left(\sqrt{1 - \frac{2\tau^2}{b\sigma_g^2}}\right)\right\} \leq \frac{\alpha^2\tau^4}{(\alpha-1)b\sigma_g^4}$$

We are now ready to analyze the privacy of SGM using Rényi Differential Privacy (RDP).

**Definition 2.5** ($(\alpha, \epsilon)$-RDP [46])**.** A randomized mechanism $f : \mathcal{D} \to \mathcal{R}$ is said to have $\epsilon$-Rényi differential privacy of order $\alpha$, or $(\alpha, \epsilon)$-RDP for short, if for any adjacent $D, D' \in \mathcal{D}$, it holds that $D_\alpha\left(f(D) \| f(D')\right) \leq \epsilon$.

Recall that RDP can be transformed into the standard $(\epsilon, \delta)$-DP.

**Lemma 2.3** (Relationship with $(\epsilon, \delta)$-DP [46])**.** *If $f$ is an $(\alpha, \epsilon)$-RDP mechanism, it also satisfies $\left(\epsilon + \frac{\log 1/\delta}{\alpha-1}, \delta\right)$-differential privacy for any $0 < \delta < 1$.*

Based on Lemma 2.2 and Lemma 2.3, we immediately have the RDP and DP result for SGM.

**Lemma 2.4.** *SGM on $\gamma_t$ is $(\alpha, \frac{\alpha^2\tau^4}{(\alpha-1)b\sigma_g^4})$-RDP, therefore $(\frac{\alpha^2\tau^4}{(\alpha-1)b\sigma_g^4} + \frac{\log(1/\delta)}{\alpha-1}, \delta)$-DP.*

Optimizing over the Rényi order $\alpha$ (in Lemma 2.4) recovers the optimal $(\epsilon, \delta)$-DP guarantee for SGM at each step for the entire dataset $D$. Subsequently, by invoking the subsampling lemma, the post-processing invariance, and the sequential composition theorem from [18], we can establish the complete privacy guarantee of Algorithm 1.

**Remark 2.3.** Theorem 2.2 shows that, holding all other parameters fixed, the privacy level $\epsilon_p$ is monotonically decreasing with the sketching dimension $b$. This can appear counterintuitive at first, so we outline the intuition. According to the previous calculation, $\mathcal{SG}(\gamma_t(D); R_t, \xi_t) \sim \mathcal{N}\left(0, \left(\frac{\|\gamma_t(D)\|^2}{b} + m\sigma_g^2\right)\mathbb{I}_b\right)$. As $b$ grows, the signal term $\frac{\|\gamma_t(D)\|^2}{b}$ vanishes and the noise term $m\sigma_g^2$ dominates, making the output distribution increasingly independent of $D$. Consequently, for any two neighboring datasets $D$ and $D'$, the distributions $\mathcal{SG}(\gamma_t(D); R_t, \xi_t)$ and $\mathcal{SG}(\gamma_t(D'); R_t, \xi_t)$ become increasingly similar, since both are essentially noise–driven. Lemma 2.2 formalizes this: for a fixed noise multiplier $\sigma_g/\tau$, the $\alpha$-Rényi divergence between these two distributions is bounded by $\frac{\alpha^2\tau^4}{(\alpha-1)b\sigma_g^4}$, so that larger $b$ strictly decreases the divergence. Therefore, increasing $b$ makes it harder to distinguish which dataset produces the sketch, which implies a higher privacy level.

**Remark 2.4.** For the original Gaussian mechanism, i.e., $\mathcal{G}(\gamma_t(D)) = \gamma_t(D) + \xi_t'$ where $\xi_t' \sim \mathcal{N}\left(0, \sigma_g^2\mathbb{I}_d\right)$, $\mathcal{G}(\gamma_t(D))$ and $\mathcal{G}(\gamma_t(D'))$ are Gaussian distributions with the same variance $\sigma_g^2$, but different means $\gamma_t(D)$ and $\gamma_t(D')$ respectively. So their Rényi divergence is $\frac{\alpha^2\left\|\gamma_t(D)-\gamma_t(D')\right\|^2}{2\sigma^2} \leq \frac{\alpha^2\tau^2}{2\sigma^2}$, which is fixed for any dimensions. In contrast, for the SGM, $\mathcal{SG}\left(\gamma_t(D)\right)$ and $\mathcal{SG}\left(\gamma_t(D')\right)$ are two Gaussians with the same mean of 0 but different variances $\frac{\|\gamma_t(D)\|_2^2}{b} + m\sigma_g^2$ and $\frac{\left\|\gamma_t(D')\right\|_2^2}{b} + m\sigma_g^2$, and their Rényi divergence is proportional to $O(\frac{1}{b})$ according to Lemma 2.2. This offers an intuitive justification for the additional $1/\sqrt{b}$ factor in Theorem 2.2's privacy bound, as compared to the privacy guarantee of the standard Gaussian mechanism [1]. □

**Algorithm 2** Fed-SGM

**Hyperparameters:** server learning rate $\eta_{\text{global}}$, local learning rate $\eta_{\text{local}}$, noise scale $\sigma_g$, clipping threshold $\tau$, 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** $\theta_0$ 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 $\theta_T$

## 3 Application in FL: Fed-SGM

In this section, we introduce Fed-SGM, the application of the SGM within a federated learning framework that simultaneously achieves communication efficiency and differential privacy guarantees. We then establish its client-level privacy bounds and derive optimization convergence results.
We consider a federated learning setting with $C$ clients indexed by $c \in [C]$. Each client $c$ holds a local dataset $\mathcal{D}_c = \{(x_{i,c}, y_{i,c})\}_{i=1}^{n_c}$ of size $n_c$ and defines its empirical loss by $\mathcal{L}_c(\theta) = \frac{1}{n_c}\sum_{i=1}^{n_c} \ell\big(\theta; (x_{i,c}, y_{i,c})\big)$, where $\theta \in \mathbb{R}^d$ is the model parameter and $\ell$ the per-example loss. The goal of the algorithm is to minimize the average empirical loss over clients, i.e.,$\mathcal{L}(\theta) := \frac{1}{C}\sum_{c=1}^{C} \mathcal{L}_c(\theta)$.

Algorithm 2 formalizes the Fed-SGM algorithm. At each communication round, 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 its local dataset to obtain an update $\Delta_{c,t}$.
2. Clips $\frac{\Delta_{c,t}}{\eta_{\text{local}}}$ with respect to the clipping threshold $\tau$.
3. Applies the SGM to the clipped update and transmits the resulting sketch to the server.

Upon receiving client sketches, the server computes their aggregate and broadcasts this aggregated sketch back to every client. Each client inverts the sketch to recover the aggregated update in the original ambient space, and then updates the global model parameters in a single step via the `GLOBAL_OPT` operator. In this paper, we study the case with gradient descent (GD) and AMSGrad as `GLOBAL_OPT`.

**Remark 3.1.** In the parameter-update step, we lift the aggregated sketch back to the ambient space using the transpose $R_t^\top$, rather than the least-squares estimator $R_t^\dagger$, i.e., the Moore–Penrose pseudo-inverse. This preference is motivated by the following points. First, since each entry of the sketching matrix $R_t \in \mathbb{R}^{b\times d}$ is drawn i.i.d. Gaussian with mean 0 and variance $\frac{1}{\sqrt{b}}$, so $\mathbb{E}\left[R_t^\top R_t\right] = \mathbb{I}_d$. Therefore, for any vector $g$, $\mathbb{E}\left[R_t^\top R_t g\right] = g$, which shows that $R_t^\top$ can recover $g$ in expectation,

implying $R_t^\top$ is a near-optimal desketching operator for our use. Second, computing $R_t^\dagger$ each round requires solving a least-squares system, typically via QR/SVD, with per-round cost on the order of $\mathcal{O}\left(b^2 d\right)$ plus the overhead of forming and inverting a Gram matrix. In contrast, the transpose map $R_t^\top$ is a single matrix–vector multiplication with cost $\mathcal{O}(bd)$, which is substantially cheaper.

### 3.1 Privacy Guarantee

As a direct application of Theorem 2.2, we can obtain the following client-level privacy guarantee of Algorithm 2.

**Theorem 3.1.** *There exists 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 $\epsilon_p \leq c_3 q\sqrt{T}$, Algorithm 2 is $(\epsilon_p, \delta_p)$-client-level private for any $\delta_p > 0$ if we choose*

$$\sigma_g^2 \geq \frac{c_4 q \tau^2 \sqrt{T} \log(2qT/\delta_p)}{\sqrt{b}\epsilon_p} \,. \tag{5}$$

### 3.2 Convergence Analysis

In this section, we analyze the optimization performance of Fed-SGM as specified in Algorithm 2. Section 3.2.1 introduces our assumptions on the loss-gradient bounds and Hessian spectrum, and justifies their empirical validity via prior work. Section 3.2.2 then establishes convergence guarantees for Fed-SGM when employing AMSGrad as the `GLOBAL_OPT`, with the corresponding gradient-descent analysis deferred to Appendix D.

#### 3.2.1 Assumptions

We begin by presenting a set of standard assumptions that are widely adopted in the literature on first-order stochastic methods.

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

According to the definition, $\nabla \mathcal{L}(\theta) = \frac{1}{C}\sum_{c=1}^{C} \nabla \mathcal{L}_c(\theta)$, so we can directly see that $\|\nabla \mathcal{L}(\theta)\|_2 \leq G$. Therefore, $G$ is the upper bound for both the client and global gradient norms. We also assume the stochastic noise from mini-batches is sub-Gaussian, which is widely adopted in first-order optimization [25, 48].

**Assumption 2** (Sub-Gaussian Noise)**.** *The stochastic noise $\|\nabla \mathcal{L}_c(x) - g_c(x)\|_2$ at each client is a $\sigma_s$-sub-Gaussian random variable, i.e. $\mathbb{P}(\|\nabla \mathcal{L}_c(x) - g_c(x)\|_2 \geq a) \leq 2\exp(-a^2/\sigma_s^2)$, for all $a \geq 0$.*

Besides, we also assume a special structure on the Hessian eigenspectrum $\{\lambda_i, v_i\}_{i=1}^d$ of the loss function $\ell$.

**Assumption 3** (Bounded Loss Hessian Eigenvalues)**.** *For each $c \in [C]$, the smoothness of the loss function $\mathcal{L}_c$, i.e. the largest eigenvalue of the loss Hessian $H_{\mathcal{L}_c}$ is bounded by $L$.*

This local smoothness assumption is commonly used in many federated learning analysis [55, 20]. Due to a similar reason as Assumption 1, this assumption also indicates the $L$-smoothness of the average loss function $\mathcal{L}$.

**Assumption 4** (Loss Hessian Eigenspetrum)**.** *Denote $\{\lambda_i\}_{i=1}^d$ the eigenvectors of the Hessian matrix $H_{\mathcal{L}}$ of the average loss function $\mathcal{L}$. Then we assume the absolute intrinsic dimension of $H_{\mathcal{L}}$ is bounded, i.e., $\frac{\sum_{i=1}^d |\lambda_i|}{\max_i \lambda_i} \leq \mathcal{I}$.*

The absolute intrinsic dimension considered here, $\frac{\sum_{i=1}^d |\lambda_i|}{\max_i \lambda_i}$, is close to the concept of intrinsic dimension proposed in [26], and the difference is that we consider absolute values of eigenvalues. A growing amount of empirical research suggests that the absolute intrinsic dimension of the Hessian in deep learning can be significantly smaller than the ambient dimension $d$. Studies by [22, 38, 42] indicate that the eigenspectrum undergoes a sharp decay in magnitude. Additionally, research by [56, 39] demonstrates that a substantial portion of the eigenspectrum is concentrated near zero. Further investigations by [76, 82] reveal that the eigenvalues follow a power-law distribution, suggesting that in such cases, the absolute intrinsic dimension remains a much smaller order than $d$.

#### 3.2.2 Optimization Analysis

In this section, we will demonstrate the optimization result of Algorithm 2 with AMSGrad as `GLOBAL_OPT`. We refer the reader to Appendix E for a more formal statement with the full analysis.

**Theorem 3.1.** [Informal version of Theorem E.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 Assumption 1-4, with learning rate $\eta = \eta_{\text{global}}\eta_{\text{local}}$, we have that with probability at least $1 - \Theta(\delta)$,

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

in which $E_s^{\text{AMSGrad}}$, $E_c^{\text{AMSGrad}}$ and $E_g^{\text{AMSGrad}}$ denote terms from sketching-based FedAvg algorithm, caused by clipping, and caused by Gaussian noises. Specifically, with $\tau_{K,G} = \min\{\tau, KG\}$,

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

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

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

**Remark 3.2.** Regarding the term $E_s^{\text{AMSGrad}}$, most of the optimization results for sketching-based algorithms are expectation results [59, 54], while ours is a high-probability optimization bound for the sketching-based FedAvg algorithm. In addition, previous optimization results in expectation either have a linear dependence on the ambient dimension $d$ [59], or get around the dependence by using `Top-k` components of the gradient vector with rely on heavy-hitter assumptions [54, 80]. As a contrast, our result $E_s^{\text{AMSGrad}}$ only has a logarithmic dimensional dependence from the need of high probability with no additional mechanisms.

**Remark 3.3.** When there is no clipping activated, i.e., $\tau \geq KG$, with learning rates $\eta_{\text{local}} = \mathcal{O}\left(\frac{\sqrt{\mathcal{I}}}{\sqrt{TK}}\right)$, $\eta = \mathcal{O}\left(\frac{1}{\sqrt{T\mathcal{I}K}}\right)$, we can get a the term $E_s^{\text{AMSGrad}}$ at the order of $\tilde{\mathcal{O}}\left(\sqrt{\frac{\mathcal{I}}{T}}\right)$. When we clip with a constant, i.e., $\tau = \mathcal{O}(1)$, by setting $\eta_{\text{local}} = \mathcal{O}\left(\frac{1}{\sqrt{NTK}}\right)$, $\eta = \mathcal{O}\left(\frac{1}{\sqrt{T\mathcal{I}}}\right)$, the order of $E_s^{\text{AMSGrad}}$ becomes $\tilde{\mathcal{O}}\left(\frac{1}{\sqrt{NT}} + \frac{\sqrt{\mathcal{I}}}{\sqrt{T}K}\right)$.

**Remark 3.4.** Regarding the term $E_c^{\text{AMSGrad}}$ caused by clipping, it is a reasonable term in the sense that it is monotonically decreasing with the clipping threshold $\tau$. In the extreme case, when $\tau = 0$, all the clipped updates will become 0, so that there is no training, then $E_c^{\text{AMSGrad}} \approx G^2$ is a natural bound of $\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\theta_t)\|_2^2 = \|\nabla\mathcal{L}(\theta_0)\|_2^2$ with Assumption 1; when $\tau \geq KG$, there is no clipping activated during training, which is aligned with the fact that $E_c^{\text{AMSGrad}} = \max\left\{0, \frac{\left(\epsilon+\tilde{\mathcal{O}}(\eta_{\text{local}})\right)G(KG-\tau)}{K\epsilon}\right\} = 0$.

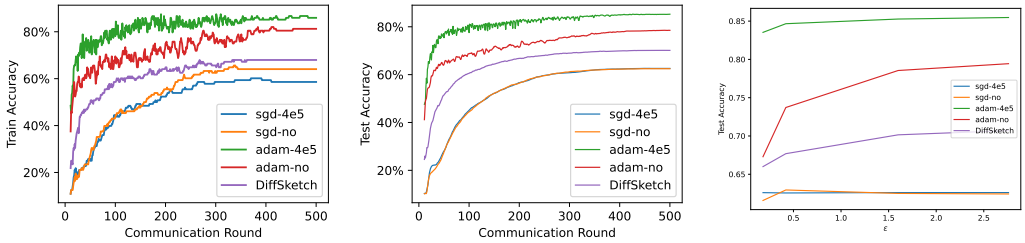
**Remark 3.5.** Regarding the term $E_g^{\text{AMSGrad}}$ caused by Gaussian noises, with the learning rates mentioned in Remark 3.3, $E_g^{\text{AMSGrad}} = \tilde{\mathcal{O}}\left(\frac{\sqrt{\mathcal{I}}\sigma_g^2}{N\sqrt{TK}}\right)$. Besides, $E_g^{\text{AMSGrad}}$ does not have any dependence on either the ambient dimension $d$ or the sketching dimension $b$.

**Remark 3.6.** From the proof in Appendix E, we can see that our analysis on AMSGrad can be generalized to any adaptive optimizers involved with first-order and second-order moments, including Adam [32], Yogi [78], etc.

## 4 Experiments

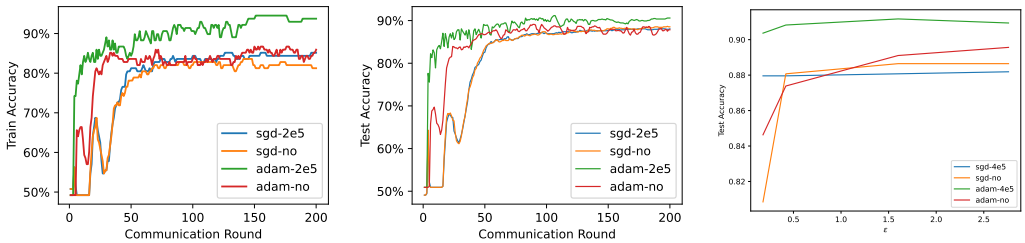
We conduct empirical evaluations of Fed-SGM on neural network training for both vision and language benchmarks. Across various privacy budgets $\epsilon_p$, we compare Fed-SGM using GD and Adam as the global optimizer against their unsketched counterparts—DP-FedAvg [81] for the GD variant and a matched unsketched Adam implementation, and additionally benchmark against DiffSketch [37]. The results are presented after a detailed description of the experimental setup.

**Datasets and Network Structure.** We adopt two different experiment settings including both vision and language tasks. For the vision task, We use the full EMNIST ByClass dataset, which comprises 814K training samples and 140K testing samples across 62 classes, representing the complete set of




(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-SGM with ADAM, Fed-SGM with GD, DP-FedAvg and its Adam variant of ResNet101 trained on EMNIST with $\epsilon_p = 1.6$. The X-axis is the number of communication rounds $T$, and the Y-axis is the train/test accuracy. (c) The trend of test accuracy over privacy levels. The X-axis is the $\epsilon_p$, and the Y-axis is the test accuracy. In all three subfigures, 'sgd' and 'adam' denote the selection of GD and Adam as the global optimizers, respectively; '4e5' signifies a sketching dimension of $4 \times 10^5$, and 'no' indicates that no sketching is applied.




(a) Training Accuracy, $\epsilon_p = 1.44$ (b) Test Accuracy, $\epsilon_p = 1.44$ (c) Test accuracy across privacy levels

Figure 2: (a)(b) Comparison of Fed-SGM with ADAM, Fed-SGM with GD, DP-FedAvg and its Adam variant of Bert finetuned on SST-2 with $\epsilon_p = 1.44$. The X-axis is the number of communication rounds $T$, and the Y-axis is the train/test accuracy. (c) The trend of test accuracy over privacy levels. The X-axis is the $\epsilon_p$, and the Y-axis is the test accuracy. In all three subfigures, 'sgd' and 'adam' denote the selection of GD and Adam as the global optimizers, respectively; '2e5' signifies a sketching dimension of $2 \times 10^5$, and 'no' indicates that no sketching is applied.

handwritten characters. We conduct experiments on ResNet101 [74] with a total of 42M parameter. For the language task, we use the SST-2 dataset from the GLUE benchmark [69], which comprises 67349 training samples and 1821 test samples across two sentiment classes. We finetune a BERT-Base model [14], comprising approximately 100M parameters.

**Parameter Setting** We deploy $C = 625$ clients in total, sampling $N = 4$ clients uniformly at random in each communication round. Each selected client executes $K = 18$ local SGD updates on mini-batches of size 64, with gradient clipping threshold $\tau = 1$. Sketching dimension and total rounds are chosen per task: for the vision task, we set $b = 4 \times 10^5$ (approximately $1\%$ compression rate) and run $T = 500$ communication rounds; for the language task, we use $b = 2 \times 10^5$ (approximately $0.2\%$ compression rate) and $T = 200$ rounds.

**Privacy Level and Noise calculation:** For privacy, we fix the parameter $\delta_p = 10^{-5}$. For both tasks, we consider noise scales $\sigma_g \in \{0.8, 1, 2, 4\}$ for the unsketched algorithms. By employing the Moments Accountant method [1, 8], we compute the cumulative privacy loss and obtain approximately $\{2.75, 1.60, 0.42, 0.18\}$ for the vision task and $\{2.45, 1.44, 0.35, 0.12\}$ for the language task, each corresponding to the respective noise scales. For Fed-SGM, at each privacy level, we compute the minimal noise scale $\sigma_g$ by calculating the RDP and minimizing over the Rényi order $\alpha$ numerically [46]. This yields the noise levels 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. Across all considered privacy budgets in both learning tasks, Fed-SGM attains the same privacy guarantee with strictly lower Gaussian noise variance than the corresponding unsketched algorithm.

**Experimental Results** We report our experimental results in Figure 1 and 2. Additional figures and ablation studies are presented in Appendix F. For each task, Figures 1a and 2a depict the training accuracy of the five algorithms at a fixed privacy budget, whereas Figures 1b and 2b show the corresponding test accuracy. Additionally, Figures 1c and 2c present a trend of the test accuracies of these algorithms under various privacy levels, providing deeper insight into their performance across different privacy constraints. Based on these figures, the following observations can be made:

- Regardless of the choice of global optimizer, Fed-SGM consistently matches or surpasses its unsketched counterpart, confirming the effectiveness of sketching within a differentially private FL framework. Notably, when Adam is employed, Fed-SGM even outperforms the non-sketching baseline. This improvement arises because Fed-SGM always requires a strictly lower Gaussian noise variance than the unsketched mechanism to attain the same privacy level. The resulting reduction in the amount of injected noise may compensate for any performance degradation typically associated with sketching.
- Irrespective of whether sketching is employed, the variant using Adam consistently outperforms the ones using gradient descent and the baseline DiffSketch. Notably, even the Adam variant with sketching surpasses the gradient descent variant without sketching, and this trend is maintained across different privacy levels. These results demonstrate the performance enhancement achieved by employing Adam.

**Remark 4.1.** For experiment results in Figure 1 and Figure 2, we use an IID split of the training dataset across all 625 clients. To assess robustness under heterogeneity, we rerun our image-classification experiments with a non-IID partition using a Dirichlet distribution with concentration parameter 0.05. Table 1 compares test accuracy for models with and without sketching, under both SGD and AMAGrad global optimizers. As expected, overall accuracy decreases under non-IID settings; however, AMSGrad variants still outperforms SGD counterparts, and sketching methods remains competitive with and sometimes outperforms non-sketching ones.

Table 1: Test accuracy under IID vs. Non-IID data.

| | GD, Sketching | GD, Non-Sketching | Adam, Sketching | Adam, Non-Sketching |
|---|---|---|---|---|
| IID | 62.98% | 63.34% | 85.09% | 78.60% |
| Non-IID | 40.71% | 42.31% | 52.55% | 51.42% |

**Remark 4.2.** We use Adam as a representative adaptive optimizer in our experiments. To assess sensitivity to the optimizer choice, we repeated the image-classification experiments with AMSGrad under the same settings as Figure 1. As shown in Table 2, the results with AMSGrad are comparable to those with Adam.

Table 2: Adam vs. AMSGrad as the adaptive global optimizer.

| | Sketching | Non-Sketching |
|---|---|---|
| Adam | 85.09% | 78.60% |
| AMSGrad | 85.04% | 78.10% |

## 5 Conclusion

In conclusion, we introduced the Sketched Gaussian Mechanism (SGM), which combines an isometric sketching transform with the classical Gaussian mechanism, and showed via Rényi differential privacy—together with subsampling, post-processing, and composition theorems—that its privacy loss satisfies $\epsilon = O\left(\frac{1}{\sqrt{b}\sigma_g^2}\right)$ without imposing any restrictive upper bound on the sketching dimension $b$, thereby demonstrating inherent privacy amplification from sketching. We then embed SGM into a federated learning framework (Fed-SGM) supporting arbitrary server optimizers and prove convergence guarantees that grow only logarithmically in the ambient dimension $d$ and linearly in the empirically small absolute intrinsic Hessian dimension $\mathcal{I}$. Empirical results on vision and language benchmarks confirm that Fed-SGM attains a fixed privacy budget with strictly less noise than unsketched DP-FedAvg, consistently matches or exceeds model accuracy, and benefits further from adaptive optimization. As a direction for future work, we note that our analysis is currently limited to isotropic Gaussian sketching matrices. It therefore remains to establish whether comparable privacy and convergence guarantees hold for more general classes of sketching transforms.

## Acknowledgments

The work was supported by the National Science Foundation (NSF) through awards IIS 21-31335, OAC 21-30835, DBI 20-21898, as well as a C3.ai research award. Compute support for the work was provided by the National Center for Supercomputing Applications (NCSA) and the Illinois Campus Cluster Program (ICCP).

## References

[1] 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.

[2] 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.

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

[4] 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.

[5] 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.

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

[7] 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, 2017.

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

[9] Clément L Canonne, Gautam Kamath, and Thomas Steinke. The discrete gaussian for differential privacy. *Advances in Neural Information Processing Systems*, 33:15676–15688, 2020.

[10] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *International Colloquium on Automata, Languages, and Programming*, pages 693–703. Springer, 2002.

[11] 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.

[12] 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.

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

[14] 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.

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

[16] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[17] 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.

[18] 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.

[19] 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.

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

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

[22] 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.

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

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

[25] 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.

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

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

[32] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[33] 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*, 2020.

[34] 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.

[35] 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, 2014.

[36] 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, 2014.

[37] Tian Li, Zaoxing Liu, Vyas Sekar, and Virginia Smith. Privacy for free: Communication-efficient learning with differential privacy using sketches. *arXiv preprint arXiv:1911.00972*, 2019.

[38] 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, 2020.

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

[40] 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.

[41] 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, 2023.

[42] 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*, 2023.

[43] Liangchen Luo, Yuanhao Xiong, Yan Liu, and Xu Sun. Adaptive gradient methods with dynamic bound of learning rate. *arXiv preprint arXiv:1902.09843*, 2019.

[44] 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.

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

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

[47] 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.

[48] 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.

[49] 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.

[50] 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.

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

[52] 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.

[53] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pages 547–562. University of California Press, 1961.

[54] 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.

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

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

[57] 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.

[58] 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.

[59] 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.

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

[61] 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.

[62] Tijmen Tieleman and G Hinton. Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. *Technical report*, 2017.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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*, 2018.

[70] 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, 2018.

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

[72] 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.

[73] 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.

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

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

[76] 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.

[77] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael W Mahoney. Pyhessian: Neural networks through the lens of the hessian. In *2020 IEEE international conference on big data (Big data)*, pages 581–590. IEEE, 2020.

[78] Manzil Zaheer, Sashank Reddi, Devendra Sachan, Satyen Kale, and Sanjiv Kumar. Adaptive methods for nonconvex optimization. *Advances in neural information processing systems*, 31, 2018.

[79] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[80] Meifan Zhang, Zhanhong Xie, and Lihua Yin. Private and communication-efficient federated learning based on differentially private sketches. *arXiv preprint arXiv:2410.05733*, 2024.

[81] 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.

[82] 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.

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

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

[85] 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.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly state our theoretical and experimental claims in our abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We mention the limitation of this research in Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide proofs for all theoretical results in this paper in the appendix, and also supply proof sketches and the associated intuition in the main paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We disclose all the information needed to reproduce the experimental results in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the link of the code in Appendix F.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We present the training and test details of our experiments in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Although we do not include error bars in the experiments, we repeat each experiment five times and report the average results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute resources of our experiments are provides in Appendix F.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We fully comply with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our work establishes a novel theoretical result on privacy in machine learning, thereby deepening the formal understanding of privacy guarantees within this domain.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We discuss the data and models used in our experiments and cite all original papers in Section 4.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide the link of the code in Appendix F.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We do not use LLM for the core method development of this research.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

# A Related Work

**Communication-Efficient Distributed Learning.** The substantial cost of transmitting model updates between clients and the central server has driven recent efforts to enhance communication efficiency in both distributed and federated learning. One widely adopted method, FedAvg [44], reduces communication frequency by allowing clients to conduct multiple local updates within each training round before syncing with the server. Another prevalent approach is compressing model updates before transmission, lowering the communication burden per round. These compression strategies generally fall into five categories: sparsification [3, 70, 40], quantization [61, 4, 73], low-rank factorization [67, 47, 68], sketching [54, 59, 29], and sparse subnetwork training [27, 33, 34]. While some of these techniques, such as certain quantization methods [4], naturally maintain unbiasedness, many introduce bias and require additional mechanisms to mitigate it for improved convergence [40, 54]. Another important characteristic is linearity, which guarantees that the geometric properties remain largely intact after compressing [13]. Among these techniques, sketching is particularly notable for its simplicity as a linear and unbiased transformation, allowing computations to be performed in the lower-dimensional space before reconstruction via desketching.

**Privacy in Federated Learning.** Differential Privacy (DP) [17] is the commonly used rigorous privacy guarantees in machine learning. In centralized training, the standard approach for ensuring DP follows a simple procedure of applying a clipping operation to the stochastic gradient, and introducing random Gaussian noise to the clipped gradient [1]. The clipping step plays a crucial role in enforcing DP, as the required noise variance is directly influenced by the chosen clipping threshold. [18]. Privacy mechanisms involving clipping are also widely applied in federated learning scenarios, but various requirements and factors result in different clipping operations. For sample-level privacy, clipping and injecting noise to every local update is proposed [65] while causing noticeable decline in performance. For client-level privacy, local models are clipped before transmission and perturbed bounded parameters [72, 66]. Later the mechanism to clip local updates instead was raised and turns out to have better numerical performance than model clipping [21, 71, 63].

**Sketching.** For decades, sketching has served as a core tool across various applications, predating the rise of deep learning in the 2010s [12, 23, 31]. It has been widely used in tasks such as low-rank approximation [64], graph sparsification [2], and least squares regression [15]. More recently, sketching has been increasingly employed in distributed and federated learning to compress model updates, thereby improving communication efficiency [29, 28, 54, 59, 24, 58]. Sketching-based frameworks have also been seamless integrated with secure aggregation and differential privacy mechanisms [11, 59, 45, 85, 6]. However, most of these approaches restrict their privacy analysis to the randomness injected after sketching and do not quantify any privacy amplification inherent to the sketching transformation itself. [37] is among the few to investigate this intrinsic privacy contribution of sketching.

**Adaptive Optimizers** [32] introduced Adam, an optimizer that has demonstrated rapid convergence and robustness to hyper-parameter choices. Adagrad [16] and RMSprop [62] update parameters using the gradient directly rather than relying on momentum. Additionally, Adadelta [79] modifies Adam's variance term to follow a non-decreasing update rule, and AdaBound [43] introduces both upper and lower bounds for this variance component. Most of these adaptive optimizers rely on first-order and second-order moment estimates, which is central to how these adaptive methods balance rapid progress during early training with more stable convergence later on [32]. [51] introduces a general federated optimization framework integrating adaptive server optimizers, demonstrating improved convergence rates and empirical performance in heterogeneous federated learning settings.

# B Sketching Guarantee

We have the following lemma of sketching guarantee from [59] for Gaussian sketching matrix.

**Lemma B.1** (Lemma D.24 from [59])**.** Let $R \in \mathbb{R}^{b\times d}$ denote a random Gaussian matrix . Then for any fixed vector $h \in \mathbb{R}^d$ and any fixed vector $g \in \mathbb{R}^d$, the following property holds:

$$\Pr_{R\sim\Pi}\left[|(g^\top R^\top Rh) - (g^\top h)| > \frac{\log^{1.5}(d/\delta)}{\sqrt{b}}\|g\|_2\|h\|_2\right] \leq \Theta(\delta).$$

## C Privacy Analysis

In this section, we will provide the proof of Theorem 2.1 and Theorem 2.2 on the privacy guarantee of Algorithm 1.

### C.1 Proof of Theorem 2.1

Following the analysis of [1], we provide a proof of Theorem 2.1. We restate the theorem first.

**Theorem C.1.** There exists constants $c_1$ and $c_2$ so that given the sampling probability $q = \frac{m}{n}$ and the number of steps $T$, for any $\varepsilon_p < c_1 q^2 T$, Algorithm 1 is $(\epsilon_p, \delta_p)$-differentially private for any $\delta_p > 0$ if we choose

$$\sigma_g \geq c_2 \frac{\tau \sqrt{\left(1 + \frac{\log^{1.5}(2mT/\delta_p)}{\sqrt{b}}\right) mT \log(2/\delta_p)}}{n\varepsilon_p}.$$

*Proof.* Since the process after getting $\bar{\tilde{g}}_t$ can be viewed as post-processing, and due to the post-processing property of differential privacy, we only need to explore the privacy guarantee till the aggregation step in Algorithm 1.

According to Lemma B.1, for each $t \in [T]$, $i \in L_t$, we have that

$$\mathbb{P}\left[\|R_t \hat{g}_t(x_i)\|^2 \geq \left(1 + \frac{\log^{1.5}(2mT/\delta_p)}{\sqrt{b}}\right) \|\hat{g}_t(x_i)\|^2\right] \leq \frac{\delta_p}{2mT}$$

Therefore, with probability at least $1 - \frac{\delta_p}{2}$, for all $t \in [T]$ and $i \in L_t$,

$$\|R_t \hat{g}_t(x_i)\| \leq \sqrt{1 + \frac{\log^{1.5}(2mT/\delta_p)}{\sqrt{b}}} \|\hat{g}_t(x_i)\| \leq \tau \sqrt{1 + \frac{\log^{1.5}(2mT/\delta_p)}{\sqrt{b}}}$$

Denote $\mathcal{E} = \left\{\|R_t \hat{g}_t(x_i)\| \leq \tau \sqrt{1 + \frac{\log^{1.5}(2mT/\delta_p)}{\sqrt{b}}}, \forall t \in [T], \forall i \in L_t\right\}$, then we have $\mathbb{P}(\mathcal{E}) \geq 1 - \frac{\delta_p}{2}$. There are two difference cases:

**Case 1:** When $\mathcal{E}$ happens, the function $f$ mentioned in Lemma 3 in [1] is still bounded by a constant. Therefore, we can still follow the Moments Accountant (MA) method in Lemma 3, Theorem 1, and Theorem 2 in [1] to get that Algorithm 1 is $\left(\varepsilon_p, \frac{\delta_p}{2}\right)$-differentially private for any $\delta_p > 0$ if

$$\sigma_g \geq c_3 \frac{\tau \sqrt{\left(1 + \frac{\log^{1.5}(2mT/\delta_p)}{\sqrt{b}}\right) mT \log\left(\frac{2}{\delta_p}\right)}}{n\varepsilon_p}$$

**Case 2:** When $\mathcal{E}$ does not happen, we have $\mathbb{P}(\mathcal{E}^c) \leq \frac{\delta_p}{2}$.

Combining these two cases, so we can get that Algorithm 1 is $(\varepsilon_p, \delta_p)$-differentially private with

$$\sigma_g \geq c_3 \frac{\tau \sqrt{\left(1 + \frac{\log^{1.5}(2mT/\delta_p)}{\sqrt{b}}\right) mT \log\left(\frac{2}{\delta_p}\right)}}{n\varepsilon_p}.$$

Then we finish the proof. □

### C.2 Proof of Theorem 2.2

We will establish our analysis in the framework of [46]. We will restate the concepts and proposition here.

**Definition C.1** (Rényi divergence)**.** For two probability distributions $P$ and $Q$ defined over $\mathcal{R}$, the Rényi divergence of order $\alpha > 1$ is

$$D_\alpha(P\|Q) \triangleq \frac{1}{\alpha - 1} \log \mathbb{E}_{x \sim Q}\left(\frac{P(x)}{Q(x)}\right)^\alpha.$$

**Definition C.2** ($(\alpha, \epsilon)$-RDP)**.** A randomized mechanism $f : \mathcal{D} \to \mathcal{R}$ is said to have $\epsilon$-Rényi differential privacy of order $\alpha$, or $(\alpha, \epsilon)$-RDP for short, if for any adjacent $D, D' \in \mathcal{D}$, it holds that

$$D_\alpha \left( f(D) \| f(D') \right) \leq \epsilon.$$

**Lemma C.2** (Lemma 2.1)**.**

$$D_\alpha \left( \mathcal{SG}(\gamma_t(D); R_t, \xi_t) \| \mathcal{SG}(\gamma_t(D'); R_t, \xi_t) \right) = b f_\alpha \left( \sqrt{\frac{\|\gamma_t(D')\|^2 + mb\sigma_g^2}{\|\gamma_t(D)\|^2 + mb\sigma_g^2}} \right)$$

where

$$f_\alpha(x) = \log x + \frac{1}{2(\alpha - 1)} \log \frac{x^2}{\alpha x^2 + 1 - \alpha}.$$

*Proof.* According to the definition of SGM,

$$\mathcal{SG}(\gamma_t(D); R_t, \xi_t) \sim \mathcal{N}\left(0, \left( \frac{\|\gamma_t(D)\|^2}{b} + m\sigma_g^2 \right) \mathbb{I}_b \right);$$

$$\mathcal{SG}(\gamma_t(D'); R_t, \xi_t) \sim \mathcal{N}\left(0, \left( \frac{\|\gamma_t(D')\|^2}{b} + m\sigma_g^2 \right) \mathbb{I}_b \right)$$

Denote

$$\sigma_1^2 \ = \ \tfrac{\|\gamma_t(D)\|_2^2}{b} + m\sigma_g^2, \qquad \sigma_2^2 \ = \ \tfrac{\|\gamma_t(D')\|_2^2}{b} + m\sigma_g^2.$$

We need to compute the order-$\alpha$ Rényi divergence between $P = \mathcal{N}(0, \sigma_1^2 I_b)$ and $Q = \mathcal{N}(0, \sigma_2^2 I_b)$. According to definition of multivariate Gaussian distribution, for $x \in \mathbb{R}^b$, we can write density functions

$$P(x) = (2\pi\sigma_1^2)^{-b/2} \exp\left( -\frac{\|x\|_2^2}{2\sigma_1^2} \right),$$

$$Q(x) = (2\pi\sigma_2^2)^{-b/2} \exp\left( -\frac{\|x\|_2^2}{2\sigma_2^2} \right).$$

Hence

$$\begin{aligned}
\mathbb{E}_{x\sim Q}\left[ \left( \frac{P(x)}{Q(x)} \right)^\alpha \right] &= \int_{\mathbb{R}^b} Q(x) \cdot \left( \frac{P(x)}{Q(x)} \right)^\alpha dx \\
&= \int_{\mathbb{R}^b} P(x)^\alpha Q(x)^{1-\alpha} dx \\
&= \int_{\mathbb{R}^b} \left( \left(2\pi\sigma_1^2\right)^{-\frac{b}{2}} \exp\left( -\frac{\|x\|_2^2}{2\sigma_1^2} \right) \right)^\alpha \cdot \left( \left(2\pi\sigma_2^2\right)^{-\frac{b}{2}} \exp\left( -\frac{\|x\|_2^2}{2\sigma_2^2} \right) \right)^{1-\alpha} dx \\
&= \left(2\pi\sigma_1^2\right)^{-\frac{\alpha b}{2}} \left(2\pi\sigma_2^2\right)^{-\frac{(1-\alpha)b}{2}} \int_{\mathbb{R}^b} \exp\left( -\frac{\|x\|_2^2}{2} \left( \frac{\alpha}{\sigma_1^2} + \frac{1-\alpha}{\sigma_2^2} \right) \right) dx \\
&\overset{(i)}{=} \left(2\pi\sigma_1^2\right)^{-\frac{\alpha b}{2}} \left(2\pi\sigma_2^2\right)^{-\frac{(1-\alpha)b}{2}} \cdot \left( \frac{\pi}{\frac{1}{2}\left( \frac{\alpha}{\sigma_1^2} + \frac{1-\alpha}{\sigma_2^2} \right)} \right)^{\frac{b}{2}} \\
&= \sigma_1^{-\alpha b} \sigma_2^{-(1-\alpha)b} \left( \frac{\alpha}{\sigma_1^2} + \frac{1-\alpha}{\sigma_2^2} \right)^{-\frac{b}{2}} \\
&= \left( \alpha\sigma_2^2 + (1-\alpha)\sigma_1^2 \right)^{-\frac{b}{2}} \sigma_2^{\alpha b} \sigma_1^{(1-\alpha)b},
\end{aligned}$$

where (1) uses the standard Gaussian integral $\int_{\mathbb{R}^b} e^{-a\|x\|_2^2/2} dx = (2\pi/a)^{b/2}$ for $a > 0$. Therefore, following the definition of Rényi divergence in Definition C.1,

$$D_\alpha \left( P \| Q \right) = \frac{1}{\alpha - 1} \log \mathbb{E}_{x\sim Q}\left[ \left( \frac{P(x)}{Q(x)} \right)^\alpha \right]$$

$$
\begin{aligned}
&= \frac{1}{\alpha-1}\log\left(\left(\alpha\sigma_2^2+(1-\alpha)\,\sigma_1^2\right)^{-\frac{b}{2}}\sigma_2^{\alpha b}\sigma_1^{(1-\alpha)b}\right)\\
&= \frac{1}{\alpha-1}\left(-\frac{1}{2}\log\left(\alpha\sigma_2^2+(1-\alpha)\,\sigma_1^2\right)^b+\alpha\log\sigma_2^b+(1-\alpha)\log\sigma_1^b\right)\\
&= \frac{1}{\alpha-1}\left(-\frac{1}{2}\log\left(\alpha\sigma_2^2+(1-\alpha)\,\sigma_1^2\right)^b+\log\sigma_2^b+(\alpha-1)\log\sigma_2^b-(\alpha-1)\log\sigma_1^b\right)\\
&= \frac{1}{2\,(\alpha-1)}\left(\log\left(\sigma_2^2\right)^b-\log\left(\alpha\sigma_2^2+(1-\alpha)\,\sigma_1^2\right)^b\right)+\left(\log\sigma_2^b-\log\sigma_1^b\right)\\
&= \log\left(\frac{\sigma_2}{\sigma_1}\right)^b+\frac{1}{2\,(\alpha-1)}\log\left(\frac{\sigma_2^2}{\alpha\sigma_2^2+(1-\alpha)\,\sigma_1^2}\right)^b\\
&= \log\left(\frac{\sigma_2}{\sigma_1}\right)^b+\frac{1}{2\,(\alpha-1)}\log\left(\frac{\frac{\sigma_2^2}{\sigma_1^2}}{\alpha\frac{\sigma_2^2}{\sigma_1^2}+1-\alpha}\right)^b
\end{aligned}
$$

Substituting $\sigma_1^2=\frac{\|\gamma_t(D)\|_2^2}{b}+m\sigma_g^2$ and $\sigma_2^2=\frac{\left\|\gamma_t(D')\right\|_2^2}{b}+m\sigma_g^2$ into the above equation, we can obtain that

$$
\begin{aligned}
&D_\alpha\left(\mathcal{SG}\left(\gamma_t(D);R_t,\xi_t\right)\|\mathcal{SG}\left(\gamma_t(D');R_t,\xi_t\right)\right)\\
=&D_\alpha\left(\mathcal{N}\left(0,\frac{\|\gamma_t(D)\|_2^2}{b}+m\sigma_g^2\right)\Big\|\mathcal{N}\left(0,\frac{\|\gamma_t(D')\|_2^2}{b}+m\sigma_g^2\right)\right)\\
=&\log\left(\frac{\sqrt{\frac{\|\gamma_t(D')\|_2^2}{b}+m\sigma_g^2}}{\sqrt{\frac{\|\gamma_t(D)\|_2^2}{b}+m\sigma_g^2}}\right)^b+\frac{1}{2\,(\alpha-1)}\log\left(\frac{\frac{\frac{\|\gamma_t(D')\|_2^2}{b}+m\sigma_g^2}{\frac{\|\gamma_t(D)\|_2^2}{b}+m\sigma_g^2}}{\alpha\frac{\frac{\|\gamma_t(D')\|_2^2}{b}+m\sigma_g^2}{\frac{\|\gamma_t(D)\|_2^2}{b}+m\sigma_g^2}+1-\alpha}\right)^b\\
=&b\log\sqrt{\frac{\|\gamma_t(D')\|^2+mb\sigma_g^2}{\|\gamma_t(D)\|^2+mb\sigma_g^2}}+\frac{b}{2(\alpha-1)}\log\frac{\|\gamma_t(D')\|^2+mb\sigma_g^2}{\|\gamma_t(D)\|^2+mb\sigma_g^2}\\
=&bf_\alpha\left(\sqrt{\frac{\|\gamma_t(D')\|^2+mb\sigma_g^2}{\|\gamma_t(D)\|^2+mb\sigma_g^2}}\right)
\end{aligned}
$$

with

$$
f_\alpha(x)=\log x+\frac{1}{2(\alpha-1)}\log\frac{x^2}{\alpha x^2+1-\alpha}.
$$

□

**Definition C.3** (Ratio Sensitivity)**.** For any constant $c\geq 0$, define the ratio sensitivity of $\theta$ as

$$
\text{rsens}_c\left(\theta\right)=\sup_{D,D'}\sqrt{\frac{\|\theta(D')\|^2+c^2}{\|\theta\left(D\right)\|^2+c^2}}
$$

where the supremum is over all neighboring datasets $D, D'$.

Now we prove the monotonicity of $f_\alpha$ and the bound on $\text{rsens}_{\sqrt{mb}\sigma_g}(\gamma_t)$.

**Lemma C.3** (Monotonicity of $f_\alpha$)**.** For any $(x,\alpha)$ such that $f_\alpha$ is well-defined, $f_\alpha(x)$ is monotonically decreasing with respect to $x$ for $x\leq 1$ and increasing for $x\geq 1$;

*Proof.* In fact,

$$
f'_\alpha(x)=\left(\log x+\frac{1}{2(\alpha-1)}\log\frac{x^2}{\alpha x^2+1-\alpha}\right)'
$$

$$
\begin{aligned}
&= (\log x)' + \frac{1}{2(\alpha-1)}\left(\log\frac{x^2}{\alpha x^2+1-\alpha}\right)' \\
&= \frac{1}{x} + \frac{1}{2(\alpha-1)}\frac{\alpha x^2+1-\alpha}{x^2}\cdot\left(\frac{x^2}{\alpha x^2+1-\alpha}\right)' \\
&= \frac{1}{x} + \frac{1}{2(\alpha-1)}\frac{\alpha x^2+1-\alpha}{x^2}\cdot\frac{2x\cdot\left(\alpha x^2+1-\alpha\right)-2\alpha x\cdot x^2}{\left(\alpha x^2+1-\alpha\right)^2} \\
&= \frac{1}{x} + \frac{1}{2(\alpha-1)}\frac{\alpha x^2+1-\alpha}{x^2}\cdot\frac{2x(1-\alpha)}{\left(\alpha x^2+1-\alpha\right)^2} \\
&= \frac{1}{x} - \frac{1}{x\left(\alpha x^2+1-\alpha\right)} \\
&= \frac{1}{x}\left(1-\frac{1}{\alpha x^2+1-\alpha}\right) \\
&= \frac{\alpha(x-1)(x+1)}{x(\alpha x^2+1-\alpha)}
\end{aligned}
$$

Therefore, $f'_\alpha \geq 0$ for $x \geq 1$ and $f'_\alpha \leq 0$ for $x \leq 1$. □

**Lemma C.4** (Bound on $\text{rsens}_{\sqrt{mb}\sigma_g}(\gamma_t)$)**.**

$$
\begin{aligned}
\sqrt{1-\frac{2\tau^2}{b\sigma_g^2}} \leq \frac{1}{\text{rsens}_{\sqrt{mb}\sigma_g}(\gamma_t)} &= \inf_{D,D'}\sqrt{\frac{\left\|\gamma_t(D')\right\|^2+mb\sigma_g^2}{\left\|\gamma_t(D)\right\|^2+mb\sigma_g^2}} \leq 1 \\
\leq \text{rsens}_{\sqrt{mb}\sigma_g}(\gamma_t) &= \sup_{D,D'}\sqrt{\frac{\left\|\gamma_t(D')\right\|^2+mb\sigma_g^2}{\left\|\gamma_t(D)\right\|^2+mb\sigma_g^2}} \leq \sqrt{1+\frac{2\tau^2}{b\sigma_g^2}}
\end{aligned}
$$

*Proof.* According to Definition C.3,

$$
\begin{aligned}
&\sup_{D,D'}\sqrt{\frac{\left\|\gamma_t(D')\right\|^2+mb\sigma_g^2}{\left\|\gamma_t(D)\right\|^2+mb\sigma_g^2}} = \text{rsens}_{\sqrt{mb}\sigma_g}(\gamma_t) \geq 1; \\
&\inf_{D,D'}\sqrt{\frac{\left\|\gamma_t(D')\right\|^2+mb\sigma_g^2}{\left\|\gamma_t(D)\right\|^2+mb\sigma_g^2}} = \frac{1}{\sup_{D,D'}\sqrt{\frac{\left\|\gamma_t(D')\right\|^2+mb\sigma_g^2}{\left\|\gamma_t(D)\right\|^2+mb\sigma_g^2}}} = \frac{1}{\text{rsens}_{\sqrt{mb}\sigma_g}(\gamma_t)} \leq 1.
\end{aligned}
$$

In addition,

$$
\begin{aligned}
\text{rsens}_{\sqrt{mb}\sigma_g}(\gamma_t) &= \sup_{D,D'}\sqrt{\frac{\left\|\gamma_t(D')\right\|^2+mb\sigma_g^2}{\left\|\gamma_t(D)\right\|^2+mb\sigma_g^2}} \\
&\leq \sup_{D,D'}\sqrt{1+\frac{\left|\left\|\gamma_t(D)\right\|^2-\left\|\gamma_t(D')\right\|^2\right|}{\left\|\gamma_t(D)\right\|^2+mb\sigma_g^2}} \\
&= \sup_{D,D'}\sqrt{1+\frac{\left(\left\|\gamma_t(D)\right\|+\left\|\gamma_t(D')\right\|\right)\left|\left\|\gamma_t(D)\right\|-\left\|\gamma_t(D')\right\|\right|}{\left\|\gamma_t(D)\right\|^2+mb\sigma_g^2}} \\
&\leq \sup_{D,D'}\sqrt{1+\frac{\left(\left\|\gamma_t(D)\right\|+\left\|\gamma_t(D')\right\|\right)\left|\left\|\gamma_t(D)-\gamma_t(D')\right\|\right|}{\left\|\gamma_t(D)\right\|^2+mb\sigma_g^2}} \\
&\leq \sqrt{1+\frac{(m\tau+m\tau)\cdot\tau}{mb\sigma_g^2}} = \sqrt{1+\frac{2\tau^2}{b\sigma_g^2}}
\end{aligned}
$$

and we can get the lower bound

$$\frac{1}{\text{rsens}_{\sqrt{mb}\sigma_g}(\gamma_t)} = \inf_{D,D'} \sqrt{\frac{\|\gamma_t(D')\|^2 + mb\sigma_g^2}{\|\gamma_t(D)\|^2 + mb\sigma_g^2}} \geq \sqrt{1 - \frac{2\tau^2}{b\sigma_g^2}}$$

with a similar calculation. □

Now we can obtain the upper bound of $D_\alpha\left(\mathcal{SG}(\gamma_t(D)); R_t, \xi_t \| \mathcal{SG}(\gamma_t(D')); R_t, \xi_t\right)$.

**Lemma C.5** (Lemma 2.2)**.** For any neighboring datasets $D, D'$,

$$\begin{aligned}
&D_\alpha\left(\mathcal{SG}(\gamma_t(D)); R_t, \xi_t \| \mathcal{SG}(\gamma_t(D')); R_t, \xi_t\right) \\
\leq & b \max\left\{ f_\alpha\left(\sqrt{1 + \frac{2\tau^2}{b\sigma_g^2}}\right), f_\alpha\left(\sqrt{1 - \frac{2\tau^2}{b\sigma_g^2}}\right)\right\} \leq \frac{\alpha^2\tau^4}{(\alpha-1)b\sigma_g^4}
\end{aligned}$$

*Proof.* According to Lemma C.2,

$$D_\alpha\left(\mathcal{SG}(\gamma_t(D); R_t, \xi_t) \| \mathcal{SG}(\gamma_t(D'); R_t, \xi_t)\right) = b f_\alpha\left(\sqrt{\frac{\|\gamma_t(D')\|^2 + mb\sigma_g^2}{\|\gamma_t(D)\|^2 + mb\sigma_g^2}}\right)$$

From Lemma C.4, we have

$$\sqrt{1 - \frac{2\tau^2}{b\sigma_g^2}} \leq \inf_{D,D'} \sqrt{\frac{\|\gamma_t(D')\|^2 + mb\sigma_g^2}{\|\gamma_t(D)\|^2 + mb\sigma_g^2}} \leq \sqrt{\frac{\|\gamma_t(D')\|^2 + mb\sigma_g^2}{\|\gamma_t(D)\|^2 + mb\sigma_g^2}} \leq \sup_{D,D'} \sqrt{\frac{\|\gamma_t(D')\|^2 + mb\sigma_g^2}{\|\gamma_t(D)\|^2 + mb\sigma_g^2}} \leq \sqrt{1 + \frac{2\tau^2}{b\sigma_g^2}}.$$

Based on the monotonicity of $f_\alpha$ in Lemma C.3,

$$\begin{aligned}
&D_\alpha\left(\mathcal{SG}(\gamma_t(D); R_t, \xi_t) \| \mathcal{SG}(\gamma_t(D'); R_t, \xi_t)\right) = b f_\alpha\left(\sqrt{\frac{\|\gamma_t(D')\|^2 + mb\sigma_g^2}{\|\gamma_t(D)\|^2 + mb\sigma_g^2}}\right) \\
\leq & b \max\left\{ f_\alpha\left(\sqrt{1 + \frac{2\tau^2}{b\sigma_g^2}}\right), f_\alpha\left(\sqrt{1 - \frac{2\tau^2}{b\sigma_g^2}}\right)\right\}
\end{aligned}$$

In addition,

$$\begin{aligned}
f_\alpha\left(\sqrt{1 + \frac{2\tau^2}{b\sigma_g^2}}\right) &= \frac{1}{2}\log\left(1 + \frac{2\tau^2}{b\sigma_g^2}\right) + \frac{1}{2(\alpha-1)}\log\frac{1 + \frac{2\tau^2}{b\sigma_g^2}}{\alpha \cdot \left(1 + \frac{2\tau^2}{b\sigma_g^2}\right) + 1 - \alpha} \\
&= \frac{1}{2}\log\left(1 + \frac{2\tau^2}{b\sigma_g^2}\right) + \frac{1}{2(\alpha-1)}\log\frac{1 + \frac{2\tau^2}{b\sigma_g^2}}{1 + \frac{2\alpha\tau^2}{b\sigma_g^2}} \\
&= \frac{1}{2}\log\left(1 + \frac{2\tau^2}{b\sigma_g^2}\right) + \frac{1}{2(\alpha-1)}\left(\log\left(1 + \frac{2\tau^2}{b\sigma_g^2}\right) - \log\left(1 + \frac{2\alpha\tau^2}{b\sigma_g^2}\right)\right) \\
&= \frac{1}{2(\alpha-1)}\left(\alpha\log\left(1 + \frac{2\tau^2}{b\sigma_g^2}\right) - \log\left(1 + \frac{2\alpha\tau^2}{b\sigma_g^2}\right)\right) \\
&\leq \frac{1}{2(\alpha-1)}\left(\alpha \cdot \frac{2\tau^2}{b\sigma_g^2} - \left(\frac{2\alpha\tau^2}{b\sigma_g^2} - \frac{1}{2}\left(\frac{2\alpha\tau^2}{b\sigma_g^2}\right)^2\right)\right) \\
&= \frac{\alpha^2\tau^4}{(\alpha-1)b^2\sigma_g^4}
\end{aligned}$$

And similarly we can also get that $f_\alpha\left(\sqrt{1 - \frac{2\tau^2}{b\sigma_g^2}}\right) \leq \frac{\alpha^2\tau^4}{(\alpha-1)b^2\sigma_g^4}$. Therefore,

$$D_\alpha\left(\mathcal{SG}(\gamma_t(D); R_t, \xi_t) \| \mathcal{SG}(\gamma_t(D'); R_t, \xi_t)\right)$$

$$\leq b \max\left\{ f_\alpha\left(\sqrt{1+\frac{2\tau^2}{b\sigma_g^2}}\right), f_\alpha\left(\sqrt{1-\frac{2\tau^2}{b\sigma_g^2}}\right)\right\} \leq \frac{\alpha^2\tau^4}{(\alpha-1)\, b\sigma_g^4}.$$

□

**Definition C.4** ($(\alpha,\epsilon)$-RDP [46])**.** A randomized mechanism $f:\mathcal{D}\to\mathcal{R}$ is said to have $\epsilon$-Rényi differential privacy of order $\alpha$, or $(\alpha,\epsilon)$-RDP for short, if for any adjacent $D, D'\in\mathcal{D}$, it holds that

$$D_\alpha\left(f(D)\|f(D')\right)\leq\epsilon.$$

And RDP can be transformed into the standard $(\epsilon,\delta)$-DP.

**Lemma C.6** (Relationship with $(\epsilon,\delta)$-DP [46])**.** If f is an $(\alpha,\epsilon)$-RDP mechanism, it also satisfies $\left(\epsilon+\frac{\log 1/\delta}{\alpha-1},\delta\right)$-differential privacy for any $0<\delta<1$.

So we immediately have the RDP and DP result of SGM from Lemma C.5 and Lemma C.6.

**Lemma C.7.** SGM on $\gamma_t$ is $(\alpha, \frac{\alpha^2\tau^4}{(\alpha-1)b\sigma_g^4})$-RDP, therefore $(\frac{\alpha^2\tau^4}{(\alpha-1)b\sigma_g^4}+\frac{\log(1/\delta)}{\alpha-1},\delta)$-DP.

Finally we can prove Theorem 2.2.

**Theorem C.8.** There exists constants $c_3$ and $c_4$ so that given the sampling probability $q=\frac{m}{n}$ and the number of steps $T$, for any $\epsilon_p\leq c_3 q\sqrt{T}$, Algorithm 1 is $(\epsilon_p,\delta_p)$-differentially private for any $\delta_p>0$ if we choose

$$\sigma_g^2\geq\frac{c_4 q\tau^2\sqrt{T}\log(2qT/\delta_p)}{\sqrt{b}\epsilon_p}\,. \tag{6}$$

*Proof.* According to Lemma 2.4, $\mathcal{SG}(\gamma_t;R_t;\xi_t)$ is $\left(\frac{\alpha^2\tau^4}{(\alpha-1)b\sigma_g^4}+\frac{\log(1/\delta_0)}{\alpha-1},\delta_0\right)$-DP. By taking the derivative, we can get the optimal choice $\alpha=1+\sqrt{1+\frac{b\sigma_g^4\log(1/\delta_0)}{\tau^4}}$, we can get that

$$\frac{\alpha^2\tau^4}{(\alpha-1)\, b\sigma_g^4}+\frac{\log(1/\delta_0)}{\alpha-1}=\frac{2\tau^4}{b\sigma_g^4}\left(1+\sqrt{1+\frac{b\sigma_g^4\log(1/\delta_0)}{\tau^4}}\right)$$

in the case when $b\geq c_0\max\left\{\frac{\tau^4}{\sigma_g^4\log(1/\delta_0)},\frac{\tau^4\log(1/\delta_0)}{\sigma_g^4}\right\}$, then $\frac{b\sigma_g^4\log(1/\delta_0)}{\tau^4}\geq c_0$, we have that

$$\begin{aligned}\frac{2\tau^4}{b\sigma_g^4}\left(1+\sqrt{1+\frac{b\sigma_g^4\log(1/\delta_0)}{\tau^4}}\right)&\leq\frac{2\tau^4}{b\sigma_g^4}\left(\sqrt{\frac{1}{c_0}\frac{b\sigma_g^4\log(1/\delta_0)}{\tau^4}}+\sqrt{\frac{1}{c_0}\frac{b\sigma_g^4\log(1/\delta_0)}{\tau^4}+\frac{b\sigma_g^4\log(1/\delta_0)}{\tau^4}}\right)\\&=\frac{2\left(1+\sqrt{c_0+1}\right)}{\sqrt{c_0}}\frac{\tau^2\sqrt{\log(1/\delta_0)}}{\sqrt{b}\sigma_g^2}:=c_1\frac{\tau^2\sqrt{\log(1/\delta_0)}}{\sqrt{b}\sigma_g^2}\end{aligned}$$

so we get that $M$ is $(\epsilon_0,\delta_0)$-DP with $\epsilon_0=c_1\frac{\tau^2\sqrt{\log(1/\delta_0)}}{\sqrt{b}\sigma_g^2}$.

Now we list the subsampling and composition properties of $(\epsilon,\delta)$-DP.

**Lemma C.9** (Sub-sampling of $(\epsilon,\delta)$-DP)**.** If $M$ is $(\epsilon,\delta)$-DP, then $M'=M\circ\texttt{Sample}_m$ obeys $(\epsilon',\delta')$-DP with $\epsilon'=\log(1+p(e^\epsilon-1))$ and $\delta'=p\delta$, in which $p=\frac{m}{n}$ is the sampling ratio.

**Lemma C.10** (Strong composition of $(\epsilon,\delta)$-DP)**.** For all $\epsilon,\delta,\delta'\geq 0$, the class of $(\epsilon,\delta)$-differentially private mechanisms satisfies $(\epsilon',k\delta+\delta')$-differential privacy under $k$-fold adaptive composition for:

$$\epsilon'=\sqrt{2k\log\left(1/\delta'\right)}\epsilon+k\epsilon\left(e^\epsilon-1\right)$$

Since $\epsilon_0=c_1\frac{\tau^2\sqrt{\log(1/\delta_0)}}{\sqrt{b}\sigma_g^2}\leq\frac{c_1}{\sqrt{c_0}}$, so according to Lemma C.9, $M\circ\texttt{Sample}_m$ satisfies $(\epsilon_1,p\delta_0)$-DP with

$$\epsilon_1=\log\left(1+p\left(e^{\epsilon_0}-1\right)\right)\leq c_2 p\epsilon_0=c_1c_2\frac{p\tau^2\sqrt{\log(1/\delta_0)}}{\sqrt{b}\sigma_g^2}$$

According to Lemma C.10, we can see that $T$-fold composition of $M \circ \texttt{Sample}_m$ satisfies $(\epsilon_2, pT\delta_0 + \delta')$-DP with

$$\begin{aligned}
\epsilon_2 &= \sqrt{2T\log(1/\delta')}\epsilon_1 + T\epsilon_1\left(e^{\epsilon_1} - 1\right) \\
&\leq \sqrt{2T\log(1/\delta')}\epsilon_1 + c_3 T\epsilon_1^2 \\
&\leq \sqrt{2T\log(1/\delta')} \cdot c_1 c_2 \frac{p\tau^2\sqrt{\log(1/\delta_0)}}{\sqrt{b}\sigma_g^2} + c_3 T \cdot \left(c_1 c_2 \frac{p\tau^2\sqrt{\log(1/\delta_0)}}{\sqrt{b}\sigma_g^2}\right)^2 \\
&= \frac{c_1 c_2\sqrt{2T\log(1/\delta')\log(1/\delta_0)}p\tau^2}{\sqrt{b}\sigma_g^2} + c_1^2 c_2^2 c_3 \frac{Tp^2\tau^4\log(1/\delta_0)}{b\sigma_g^4} \\
&\leq \frac{c_1 c_2 p\sqrt{2T\log(1/\delta')}}{\sqrt{c_0}} + \frac{c_1^2 c_2^2 c_3 p^2 T}{c_0}
\end{aligned}$$

By setting $\epsilon_2 = \epsilon$, and choosing $\delta_0 = \frac{\delta}{2pT}$, $\delta' = \frac{\delta}{2}$,

$$\sigma_g^2 = \frac{C'\sqrt{T}p\tau^2\log(2pT/\delta)}{\sqrt{b}\epsilon},$$

then we can obtain $(\epsilon, \delta)$-DP. □

## D 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 $\theta_{t-1}$ with desketched aggregated updates $\operatorname{desk}\left(\bar{\tilde{\Delta}}_{t-1}\right)$.

**Algorithm 3** GLOBAL_OPT (GD)

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

**Output:** model $\theta_T$.

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

Then we will introduce the optimization result of Algorithm 2 with Algorithm 3. We follow a similar analysis to [59], but we exploit the second order structure of the deep learning losses, which helps avoid picking up dimension dependence due to the sketching operation. We will state the formal result.

**Theorem D.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 1-4, with learning rate $\eta = \eta_{\text{global}}\eta_{\text{local}}$, we have that with probability at least $1 - 10\delta$,

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

in which

$$\alpha_1 = 1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}, \tau_{K,G} = \min\{\tau, KG\}$$

**Remark D.1.** To analyze the bound in terms of $\tau$, we consider all other parameters as fixed and distinguish between two principal regimes:

1. When $\tau \leq KG$: In this regime, clipping may be activated, and $\tau$ acts as an upper bound on the norm of each clipped update $\Delta_{c,t}$ according to our algorithm. The total bound in this case is given by $\frac{\mathcal{L}(\theta_0)-\mathcal{L}^*}{\eta KT} + \frac{2\log(2T/\delta)G^2}{\sqrt{NT}} + \frac{\eta_{\text{local}}LKG^2}{2} + \frac{\sqrt{2}G\log(2T/\delta)\sigma_s}{\sqrt{NTK}} + \max\left\{0, \frac{G(KG-\tau)}{K}\right\} + \frac{\sqrt{2}\log^2(NTd/\delta)G\tau}{\sqrt{bTK}} + \frac{\log^2(2T/\delta)G^2}{\eta TK} + \frac{2\eta\log^2(2T/\delta)\sigma_g^2}{NK} + \frac{2\eta\alpha_1^2\mathcal{I}L\tau^2}{K} + \frac{2\eta\sigma_g^2\mathcal{I}L\log^2(2dT/\delta)}{NK}$. We can observe that $\frac{\sqrt{2}\log^2(NTd/\delta)G\tau}{\sqrt{bT}K} + \frac{2\eta\alpha_1^2\mathcal{I}L\tau^2}{K}$ in the optimization bound of the sketching-based algorithm is monotonically increasing in $\tau$, while $\max\left\{0, \frac{G(KG-\tau)}{K}\right\}$ caused by clipping is monotonically decreasing in $\tau$. This trade-off highlights the need for a careful choice of $\tau$, as overly aggressive clipping introduces clipping bias, while too loose a threshold amplifies optimization error of sketching algorithms.

2. When $\tau \geq KG$: In this regime, the clipping operation becomes inactive under our algorithm, as all updates fall within the clipping threshold $\tau$. $KG$ effectively bounds the norm of each update $\Delta_{c,t}$, and the clipping no longer influences the computation. The total bound becomes $\frac{\mathcal{L}(\theta_0)-\mathcal{L}^*}{\eta KT} + \frac{2\log(2T/\delta)G^2}{\sqrt{NT}} + \frac{\eta_{\text{local}}LKG^2}{2} + \frac{\sqrt{2}G\log(2T/\delta)\sigma_s}{\sqrt{NTK}} + \frac{\sqrt{2}\log^2(NTd/\delta)G^2}{\sqrt{bT}} + \frac{\log^2(2T/\delta)G^2}{\eta TK} + \frac{2\eta\log^2(2T/\delta)\sigma_g^2}{NK} + 2\eta\alpha_1^2\mathcal{I}LKG^2 + \frac{2\eta\sigma_g^2\mathcal{I}L\log^2(2dT/\delta)}{NK}$. Consequently, the entire bound becomes independent of $\tau$ and remains constant as $\tau$ increases further.

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

$$\theta_{t+1} - \theta_t = -\eta_{\text{global}}\operatorname{desk}\left(\bar{\tilde{\Delta}}_t\right)$$

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

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

$$
\mathcal{L}(\theta_T)-\mathcal{L}(\theta_0)=\sum_{t=1}^{T}(\mathcal{L}(\theta_{t+1})-\mathcal{L}(\theta_t))=\underbrace{\sum_{t=1}^{T}\nabla\mathcal{L}(\theta_t)^\top(\theta_{t+1}-\theta_t)}_{T_1}+\underbrace{\frac{1}{2}\sum_{t=1}^{T}(\theta_{t+1}-\theta_t)^\top\hat{H}_{\mathcal{L},t}(\theta_{t+1}-\theta_t)}_{T_2} \tag{7}
$$

### D.1 Bounding $T_1$

For each term in $T_1$, we have

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

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

$$
\sum_{t=1}^{T}\nabla\mathcal{L}(\theta_t)^\top(\theta_{t+1}-\theta_t)=-\eta\underbrace{\sum_{t=1}^{T}\frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top\sum_{c\in\mathcal{C}_t}R_t^\top R_t\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)}_{S_1}-\eta\underbrace{\sum_{t=1}^{T}\frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top\sum_{c\in\mathcal{C}_t}R_t^\top\mathbf{z}_{c,t}}_{S_2} \tag{8}
$$

#### D.1.1 Bounding $S_1$

For each term in $S_1$, we have that

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

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

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

$$\begin{aligned}
&\sum_{t=1}^T \frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top \sum_{c\in\mathcal{C}_t} R_t^\top R_t \text{clip}\left(\sum_{k=1}^K g_{c,t,k}, \tau\right) \\
=&K\underbrace{\sum_{t=1}^T \left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{C}\sum_{c=1}^C \nabla\mathcal{L}_c(\theta_t)\right\rangle}_{Y_1} + K\underbrace{\sum_{t=1}^T \left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\nabla\mathcal{L}_c(\theta_t) - \frac{1}{C}\sum_{c=1}^C \nabla\mathcal{L}_c(\theta_t)\right\rangle}_{Y_2} \\
&+\underbrace{\sum_{t=1}^T \left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^K (\nabla\mathcal{L}_c(\theta_{c,t,k}) - \nabla\mathcal{L}_c(\theta_t))\right\rangle}_{Y_3} + \underbrace{\sum_{t=1}^T \left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^K (g_{c,t,k} - \nabla\mathcal{L}_c(\theta_{c,t,k}))\right\rangle}_{Y_4} \\
&+\underbrace{\sum_{t=1}^T \left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\left(\text{clip}\left(\sum_{k=1}^K g_{c,t,k}, \tau\right) - \sum_{k=1}^K g_{c,t,k}\right)\right\rangle}_{Y_5} \\
&+\underbrace{\sum_{t=1}^T \left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\left(R_t^\top R_t \text{clip}\left(\sum_{k=1}^K g_{c,t,k}, \tau\right) - \text{clip}\left(\sum_{k=1}^K g_{c,t,k}, \tau\right)\right)\right\rangle}_{Y_6}
\end{aligned} \tag{9}$$

#### D.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

$$\sum_{t=1}^T \left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{C}\sum_{c=1}^C \nabla\mathcal{L}_c(\theta_t)\right\rangle = \sum_{t=1}^T \|\nabla\mathcal{L}(\theta_t)\|_2^2 \tag{10}$$

#### D.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

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

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

$$Z_{T-1} = \sum_{t=1}^{T} \left\langle \nabla \mathcal{L}(\theta_t), \frac{1}{N} \sum_{c \in \mathcal{C}_t} \nabla \mathcal{L}_c(\theta_t) - \frac{1}{C} \sum_{c=1}^{C} \nabla \mathcal{L}_c(\theta_t) \right\rangle \geq - \frac{2\sqrt{T}\log(2T/\delta)G^2}{\sqrt{N}} \tag{11}$$

#### D.1.1.3 Bounding $Y_3$

For each term, we have

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

so

$$\sum_{t=1}^{T} \left\langle \nabla \mathcal{L}(\theta_t), \frac{1}{N} \sum_{c \in \mathcal{C}_t} \sum_{k=1}^{K} \left( \nabla \mathcal{L}_c(\theta_{c,t,k}) - \nabla \mathcal{L}_c(\theta_t) \right) \right\rangle \geq - \frac{\eta_{\text{local}} T L K^2 G^2}{2} \tag{12}$$

#### D.1.1.4 Bounding $Y_4$

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

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

According to the assumption, the stochastic noise $\|g_{c,t,k} - \nabla\mathcal{L}_c(\theta_{c,t,k})\|_2$ is a $\sigma_s$-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

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

Then denote $W_t = \sum_{\tau'=0}^{t}\left\langle \nabla\mathcal{L}(\theta'_\tau), \frac{1}{N}\sum_{c\in\mathcal{C}_{\tau'}}\sum_{k=1}^{K}(g_{c,\tau',k} - \nabla\mathcal{L}_c(\theta_{c,\tau',k}))\right\rangle$, 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 $t \in [T]$,

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

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 $1-2\delta$,

$$W_{T-1} = \sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}(g_{c,t,k} - \nabla\mathcal{L}_c(\theta_{c,t,k}))\right\rangle \geq -\frac{G\sqrt{2TK}\log(2T/\delta)\sigma_s}{\sqrt{N}} \quad (13)$$

#### D.1.1.5 Bounding $Y_5$

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

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

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

$$\sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\left(\text{clip}\left(\sum_{k=1}^{K} g_{c,t,k}, \tau\right) - \sum_{k=1}^{K} g_{c,t,k}\right)\right\rangle \geq -\max\left\{0, TG(KG-\tau)\right\} \quad (14)$$

**D.1.1.6 Bounding $Y_6$**

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

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

with $\tau_{K,G} = \min\{\tau, KG\}$. 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 $t \in [T]$,

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

Then by Azuma's inequality, we have

$$\mathbb{P}\left(U_{T-1} \leq -a\right) \leq \exp\left(-\frac{a^2}{2\cdot T\cdot\left(\frac{\log^{1.5}(NTd/\delta)G\tau_{K,G}}{\sqrt{b}}\right)^2}\right) = \exp\left(-\frac{ba^2}{2T\log^3(NTd/\delta)G^2\tau_{K,G}^2}\right)$$

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

$$W_{T-1} = \sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\left(R_t^\top R_t \text{clip}\left(\sum_{k=1}^{K} g_{c,t,k}, \tau\right) - \text{clip}\left(\sum_{k=1}^{K} g_{c,t,k}, \tau\right)\right)\right\rangle \geq -\frac{\log^2(NTd/\delta)\sqrt{2T}G\tau_{K,G}}{\sqrt{b}} \quad (15)$$

Substituting 10, 11, 12, 13, 14, 15 into 9, we have that with probability at least $1-6\delta$,

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

$$+\underbrace{\sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}(\nabla\mathcal{L}_c(\theta_{c,t,k})-\nabla\mathcal{L}_c(\theta_t))\right\rangle}_{Y_3}+\underbrace{\sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}(g_{c,t,k}-\nabla\mathcal{L}_c(\theta_{c,t,k}))\right\rangle}_{Y_4}$$

$$+\underbrace{\sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\left(\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)-\sum_{k=1}^{K}g_{c,t,k}\right)\right\rangle}_{Y_5}$$

$$+\underbrace{\sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\left(R_t^\top R_t\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)-\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)\right)\right\rangle}_{Y_6} \tag{16}$$

$$\geq K\sum_{t=1}^{T}\|\nabla\mathcal{L}(\theta_t)\|_2^2-\frac{2K\sqrt{T}\log(2T/\delta)G^2}{\sqrt{N}}-\frac{\eta_{\text{local}}TLK^2G^2}{2}-\frac{G\sqrt{2TK}\log(2T/\delta)\sigma_s}{\sqrt{N}}$$

$$-\max\{0,TG(KG-\tau)\}-\frac{\log^2(NTd/\delta)\sqrt{2T}G\tau_{K,G}}{\sqrt{b}} \tag{17}$$

### D.1.2 Bounding $S_2$

We first bound each term with a fixed $t\in[T]$ in $S_2$. Noticing that $\frac{1}{N}\sum_{c\in\mathcal{C}_t}\mathbf{z}_{c,t}\sim\mathcal{N}\left(0,\frac{\sigma_g^2}{N}\mathbb{I}\right)$, and $R_t\nabla\mathcal{L}(\theta_t)$ is a $\frac{\|\nabla\mathcal{L}(\theta_t)\|_2}{\sqrt{b}}$-sub-Gaussian random vector, so according to Bernstein inequality,

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

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

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

Then denote $X_t=\sum_{\tau'=0}^{t}\left\langle R_{\tau'}\nabla\mathcal{L}(\theta_{\tau'}),\frac{1}{N}\sum_{c\in\mathcal{C}_{\tau'}}\mathbf{z}_{c,\tau'}\right\rangle$, 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 $1-\delta$, for all $t\in[T]$,

$$|X_t-X_{t-1}|=\left|\left\langle R_t\nabla\mathcal{L}(\theta_t),\frac{1}{N}\sum_{c\in\mathcal{C}_t}\mathbf{z}_{c,t}\right\rangle\right|\leq\frac{\sigma_gG\log(2T/\delta)}{\sqrt{N}}$$

Then by Azuma's inequality, we have

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

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

$$X_{T-1}=\sum_{t=1}^{T}\left\langle R_t\nabla\mathcal{L}(\theta_t),\frac{1}{N}\sum_{c\in\mathcal{C}_t}\mathbf{z}_{c,t}\right\rangle\geq-\frac{\log^2(2T/\delta)\sqrt{2T}\sigma_gG}{\sqrt{N}} \tag{18}$$

Substituting 17 and 18 into 8, we have that with probability at least $1-8\delta$,

$$
\begin{aligned}
&\sum_{t=1}^{T} \nabla \mathcal{L}(\theta_t)^\top (\theta_{t+1}-\theta_t) \\
=& -\eta \underbrace{\sum_{t=1}^{T} \frac{1}{N} \nabla \mathcal{L}(\theta_t)^\top \sum_{c\in\mathcal{C}_t} R_t^\top R_t \text{clip}\left(\sum_{k=1}^{K} g_{c,t,k}, \tau\right)}_{S_1} - \eta \underbrace{\sum_{t=1}^{T} \frac{1}{N} \nabla \mathcal{L}(\theta_t)^\top \sum_{c\in\mathcal{C}_t} R_t^\top \mathbf{z}_{c,t}}_{S_2} \\
\leq& -\eta K \sum_{t=1}^{T} \|\nabla \mathcal{L}(\theta_t)\|_2^2 + \frac{2\eta K\sqrt{T}\log(2T/\delta)G^2}{\sqrt{N}} + \frac{\eta\eta_{\text{local}} TLK^2G^2}{2} + \frac{\eta G\sqrt{2TK}\log(2T/\delta)\sigma_s}{\sqrt{N}} \\
&+ \eta \max\{0, TG(KG-\tau)\} + \frac{\eta \log^2(NTd/\delta)\sqrt{2T}G\tau_{K,G}}{\sqrt{b}} + \frac{\eta\log^2(2T/\delta)\sqrt{2T}\sigma_g G}{\sqrt{N}} \\
\leq& -\eta K \sum_{t=1}^{T} \|\nabla \mathcal{L}(\theta_t)\|_2^2 + \frac{2\eta K\sqrt{T}\log(2T/\delta)G^2}{\sqrt{N}} + \frac{\eta\eta_{\text{local}} TLK^2G^2}{2} + \frac{\eta G\sqrt{2TK}\log(2T/\delta)\sigma_s}{\sqrt{N}} \\
&+ \eta \max\{0, TG(KG-\tau)\} + \frac{\eta \log^2(NTd/\delta)\sqrt{2T}G\tau_{K,G}}{\sqrt{b}} + \log^2(2T/\delta)G^2 + \frac{2\eta^2 T\log^2(2T/\delta)\sigma_g^2}{N}
\end{aligned} \tag{19}
$$

### D.2 Bounding $T_2$

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

$$
\begin{aligned}
(\theta_{t+1}-\theta_t)^\top \hat{H}_{\mathcal{L},t}(\theta_{t+1}-\theta_t) &= \eta_{\text{global}}^2 \left(\text{desk}\left(\bar{\tilde{\Delta}}_t\right)\right)^\top \left(\sum_{i=1}^{d} \lambda_i v_i v_i^\top\right) \left(\text{desk}\left(\bar{\tilde{\Delta}}_t\right)\right) \\
&= \eta_{\text{global}}^2 \sum_{i=1}^{d} \lambda_i \left| \left(\text{desk}\left(\bar{\tilde{\Delta}}_t\right)\right)^\top v_i \right|^2
\end{aligned} \tag{20}
$$

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

$$
\begin{aligned}
\left| \left(\text{desk}(\bar{\tilde{\Delta}}_t)\right)^\top v_i \right| &= \left| \left\langle \text{desk}(\bar{\tilde{\Delta}}_t), v_i \right\rangle \right| \\
&= \frac{1}{N} \left| \left\langle R_t^\top \sum_{c\in\mathcal{C}_t} \tilde{\Delta}_{c,t}, v_i \right\rangle \right| \\
&= \frac{1}{N} \left| \left\langle R_t^\top \sum_{c\in\mathcal{C}_\tau} \eta_{\text{local}} \left(\text{sk}\left(\text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}}, \tau\right)\right) + \mathbf{z}_{c,t}\right), v_i \right\rangle \right| \\
&\leq \frac{\eta_{\text{local}}}{N} \left| \left\langle \sum_{c\in\mathcal{C}_t} R_t^\top R_t \text{clip}\left(\sum_{k=1}^{K} g_{c,t,k}, \tau\right), v_i \right\rangle \right| + \frac{\eta_{\text{local}}}{N} \left| \left\langle \sum_{c\in\mathcal{C}_\tau} R_t^\top \mathbf{z}_{c,t}, v_i \right\rangle \right|
\end{aligned} \tag{21}
$$

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

$$
\begin{aligned}
&\left| \left\langle \sum_{c\in\mathcal{C}_t} R_t^\top R_t \text{clip}\left(\sum_{k=1}^{K} g_{c,t,k}, \tau\right), v_i \right\rangle \right| \\
\leq& \left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right) \|v_i\|_2 \sum_{c\in\mathcal{C}_\tau} \left\| \text{clip}\left(\sum_{k=1}^{K} g_{c,t,k}, \tau\right) \right\| \\
\leq& \left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right) N\tau_{K,G}
\end{aligned} \tag{22}
$$

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

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

so taking $a = \frac{\sigma_g \log(2dT/\delta)}{\sqrt{N}}$, we have that with probability at least $1-\frac{\delta}{dT}$,

$$\left|\left\langle \frac{1}{N}\sum_{c\in\mathcal{C}_t} R_t^\top \mathbf{z}_{c,t}, v_i\right\rangle\right| \leq \frac{\sigma_g \log(2dT/\delta)}{\sqrt{N}} \tag{23}$$

Substituting 22 and 23 into 21, we have that with probability at least $1-\frac{2\delta}{dT}$,

$$\begin{aligned}
\left|\left(\mathrm{desk}(\bar{\tilde{\Delta}}_t)\right)^\top v_i\right| &\leq \frac{\eta_{\text{local}}}{N}\left|\left\langle \sum_{c\in\mathcal{C}_t} R_t^\top R_t \mathrm{clip}\left(\sum_{k=1}^K g_{c,t,k}, \tau\right), v_i\right\rangle\right| + \frac{\eta_{\text{local}}}{N}\left|\left\langle \sum_{c\in\mathcal{C}_\tau} R_t^\top \mathbf{z}_{c,t}, v_i\right\rangle\right| \\
&\leq \eta_{\text{local}}\left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)\tau_{K,G} + \frac{\eta_{\text{local}}\sigma_g \log(2dT/\delta)}{\sqrt{N}}
\end{aligned}$$

which implies that

$$\begin{aligned}
\left|\left(\mathrm{desk}(\bar{\tilde{\Delta}}_t)\right)^\top v_i\right|^2 &\leq 2\left(\eta_{\text{local}}\left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)\tau_{K,G}\right)^2 + 2\left(\frac{\eta_{\text{local}}\sigma_g \log(2dT/\delta)}{\sqrt{N}}\right)^2 \\
&= 2\eta_{\text{local}}^2\left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)^2\tau_{K,G}^2 + \frac{2\eta_{\text{local}}^2\sigma_g^2\log^2(2dT/\delta)}{N}
\end{aligned} \tag{24}$$

Substituting 24 into 20, we have that with probability at least $1-\frac{2\delta}{T}$,

$$\begin{aligned}
(\theta_{t+1}-\theta_t)^\top \hat{H}_{\mathcal{L},t}(\theta_{t+1}-\theta_t) &= \eta_{\text{global}}^2 \sum_{i=1}^d \lambda_i \left|\left(\mathrm{desk}\left(\bar{\tilde{\Delta}}_t\right)\right)^\top v_i\right|^2 \\
&\leq \eta_{\text{global}}^2\left(2\eta_{\text{local}}^2\left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)^2\tau_{K,G}^2 + \frac{2\eta_{\text{local}}^2\sigma_g^2\log^2(2dT/\delta)}{N}\right)\sum_{i=1}^d|\lambda_i| \\
&= 2\eta^2\left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)^2 \mathcal{I}L\tau_{K,G}^2 + \frac{2\eta^2\sigma_g^2\mathcal{I}L\log^2(2dT/\delta)}{N}
\end{aligned}$$

By taking summation from 0 to $T-1$, we have that with probability at least $1-2\delta$,

$$\sum_{t=1}^T (\theta_{t+1}-\theta_t)^\top \hat{H}_{\mathcal{L},t}(\theta_{t+1}-\theta_t) \leq 2\eta^2\left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)^2 \mathcal{I}LT\tau_{K,G}^2 + \frac{2\eta^2\sigma_g^2\mathcal{I}LT\log^2(2dT/\delta)}{N} \tag{25}$$

Substituting 19 and 25 into 7, we have that with probability at least $1-10\delta$,

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

Since

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

we can get that with probability at least $1 - 10\delta$,

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

in which

$$\alpha_1 = 1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}, \tau_{K,G} = \min\left\{\tau, KG\right\},$$

then we finish the proof. □

## E Optimization Analysis with AMSGrad as GLOBAL_OPT

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

**Algorithm 4** GLOBAL_OPT (AMSGrad)

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

Then we demonstrate the optimization result of Algorithm 2 with Algorithm 4, which is the formal version of Theorem 3.1.

**Theorem E.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 1-4, with learning rate $\eta = \eta_{\text{global}}\eta_{\text{local}}$, we have that with probability at least $1-19\delta$,

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

in which

$$
\alpha_1 = 1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}, \alpha_2 = \eta_{\text{local}}\left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)\tau_{K,G} + \frac{\eta_{\text{local}}\sigma_g\log(2dT/\delta)}{\sqrt{N}} + \epsilon
$$

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

**Remark E.1.** To analyze the bound in terms of $\tau$, we consider all other parameters as fixed and distinguish between two principal regimes:

1. When $\tau \leq KG$: In this regime, clipping may be activated, and $\tau$ acts as an upper bound on the norm of each clipped update $\Delta_{c,t}$ according to our algorithm. The total bound in this case is given by $\frac{\alpha_2(\mathcal{L}(\theta_0)-\mathcal{L}^*)}{\eta KT} + \frac{2\alpha_2\log(2T/\delta)G^2}{\sqrt{NT}\epsilon} + \frac{\eta_{\text{local}}\alpha_2 LKG^2}{2\epsilon} + \frac{\sqrt{2}\alpha_2 G\log(2T/\delta)\sigma_s}{\sqrt{NTK}\epsilon} + \max\left\{0, \frac{\alpha_2 G(KG-\tau)}{K\epsilon}\right\} + \frac{\sqrt{2}\alpha_2\log^2(NTd/\delta)G\tau}{\sqrt{bT}K\epsilon} + \frac{\eta_{\text{local}}\alpha_1^2\alpha_2(2+\beta_1)\sqrt{1-\beta_2}G\tau^2\log(2dT/\delta)}{K\epsilon^2(1-\beta_1)} + \frac{\eta_{\text{local}}\alpha_2(1+2\beta_1)\sqrt{1-\beta_2}\sigma_g^2 G\log^2(2dT/\delta)}{NK\epsilon^2(1-\beta_1)} + \frac{\alpha_2\log^2(2T/\delta)G^2}{\eta TK\epsilon} + \frac{2\eta\alpha_2\log^2(2T/\delta)\sigma_g^2}{NK\epsilon} + \frac{2\eta\alpha_1^2\alpha_2(1+2\beta_1)(1+\beta_1)\mathcal{I}L\tau^2}{K\epsilon^2(1-\beta_1)^2} + \frac{2\eta\alpha_2(1+2\beta_1)(1+\beta_1)\sigma_g^2\mathcal{I}L\log^2(2dT/\delta)}{NK\epsilon^2(1-\beta_1)^2}$.
We can observe that $\frac{\sqrt{2}\alpha_2\log^2(NTd/\delta)G\tau}{\sqrt{bT}K\epsilon} + \frac{\eta_{\text{local}}\alpha_1^2\alpha_2(2+\beta_1)\sqrt{1-\beta_2}G\tau^2\log(2dT/\delta)}{K\epsilon^2(1-\beta_1)} + \frac{2\eta\alpha_1^2\alpha_2(1+2\beta_1)(1+\beta_1)\mathcal{I}L\tau^2}{K\epsilon^2(1-\beta_1)^2}$ in the optimization bound of the sketching-based algorithm is monotonically increasing in $\tau$, while $\max\left\{0, \frac{\alpha_2 G(KG-\tau)}{K\epsilon}\right\}$ caused by clipping is monotonically decreasing in $\tau$. This trade-off highlights the need for a careful choice of $\tau$, as overly aggressive

clipping introduces clipping bias, while too loose a threshold amplifies optimization error of sketching algorithms.

2. When $\tau \geq KG$: In this regime, the clipping operation becomes inactive under our algorithm, as all updates fall within the clipping threshold $\tau$. $KG$ effectively bounds the norm of each update $\Delta_{c,t}$, and the clipping no longer influences the computation. The total bound becomes $\frac{\alpha_2(\mathcal{L}(\theta_0)-\mathcal{L}^*)}{\eta KT} + \frac{2\alpha_2\log(2T/\delta)G^2}{\sqrt{NT}\epsilon} + \frac{\eta_{\text{local}}\alpha_2 LKG^2}{2\epsilon} + \frac{\sqrt{2}\alpha_2 G\log(2T/\delta)\sigma_s}{\sqrt{NTK}\epsilon} + \frac{\sqrt{2}\alpha_2\log^2(NTd/\delta)G^2}{\sqrt{bT}\epsilon} + \frac{\eta_{\text{local}}\alpha_1^2\alpha_2(2+\beta_1)\sqrt{1-\beta_2}KG^3\log(2dT/\delta)}{\epsilon^2(1-\beta_1)} + \frac{\eta_{\text{local}}\alpha_2(1+2\beta_1)\sqrt{1-\beta_2}\sigma_g^2 G\log^2(2dT/\delta)}{NK\epsilon^2(1-\beta_1)} + \frac{\alpha_2\log^2(2T/\delta)G^2}{\eta TK\epsilon} + \frac{2\eta\alpha_2\log^2(2T/\delta)\sigma_g^2}{NK\epsilon} + \frac{2\eta\alpha_1^2\alpha_2(1+2\beta_1)(1+\beta_1)\mathcal{I}LKG^2}{\epsilon^2(1-\beta_1)^2} + \frac{2\eta\alpha_2(1+2\beta_1)(1+\beta_1)\sigma_g^2\mathcal{I}L\log^2(2dT/\delta)}{NK\epsilon^2(1-\beta_1)^2}$. Consequently, the entire bound becomes independent of $\tau$ and remains constant as $\tau$ 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 $\gamma_t$ can be expressed as

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

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

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

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

$$\begin{aligned}
&\mathcal{L}(\gamma_T) - \mathcal{L}(\theta_0) = \mathcal{L}(\gamma_T) - \mathcal{L}(\gamma_0) = \sum_{t=1}^T (\mathcal{L}(\gamma_{t+1}) - \mathcal{L}(\gamma_t)) \\
&= \underbrace{\sum_{t=1}^T \nabla\mathcal{L}(\theta_t)^\top(\gamma_{t+1} - \gamma_t)}_{T_1} + \underbrace{\sum_{t=1}^T (\nabla\mathcal{L}(\gamma_t) - \nabla\mathcal{L}(\theta_t))^\top(\gamma_{t+1} - \gamma_t)}_{T_2} + \underbrace{\frac{1}{2}\sum_{t=1}^T (\gamma_{t+1} - \gamma_t)^\top \hat{H}_{\mathcal{L},t}(\gamma_{t+1} - \gamma_t)}_{T_3}
\end{aligned} \tag{26}$$

### E.1 Bounding $T_1$

For each term in $T_1$, we have

$$\nabla\mathcal{L}(\theta_t)^\top(\gamma_{t+1} - \gamma_t)$$

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

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

$$
\begin{aligned}
&\sum_{t=1}^T\nabla\mathcal{L}(\theta_t)^\top(\gamma_{t+1}-\gamma_t)\\
=&\underbrace{\sum_{t=1}^T\nabla\mathcal{L}(\theta_t)^\top\frac{\beta_1\eta_{\text{global}}}{1-\beta_1}\left(V_{t-1}^{-1/2}-V_t^{-1/2}\right)m_{t-1}}_{S_1}\\
&\underbrace{-\eta\sum_{t=1}^T\frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top\left(V_t^{-1/2}-V_{t-1}^{-1/2}\right)\sum_{c\in\mathcal{C}_t}R_t^\top R_t\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)}_{S_2}\\
&\underbrace{-\eta\sum_{t=1}^T\frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}R_t^\top R_t\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)}_{S_3}\underbrace{-\eta\sum_{t=1}^T\frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top V_t^{-1/2}\sum_{c\in\mathcal{C}_t}R_t^\top\mathbf{z}_{c,t}}_{S_4}
\end{aligned}
\tag{27}
$$

### E.1.1 Bounding $S_2$

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

$$
\begin{aligned}
&\left|\frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top\left(V_t^{-1/2}-V_{t-1}^{-1/2}\right)\sum_{c\in\mathcal{C}_t}R_t^\top R_t\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)\right|\\
=&\left|\frac{1}{N}\left\langle\nabla\mathcal{L}(\theta_t),\left(V_t^{-1/2}-V_{t-1}^{-1/2}\right)\sum_{c\in\mathcal{C}_t}R_t^\top R_t\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)\right\rangle\right|\\
\le&\frac{1}{N}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)\|\nabla\mathcal{L}(\theta_t)\|_2\left\|V_t^{-1/2}-V_{t-1}^{-1/2}\right\|_2\sum_{c\in\mathcal{C}_t}\left\|\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)\right\|_2\\
\le&G\tau_{K,G}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)\left\|V_t^{-1/2}-V_{t-1}^{-1/2}\right\|_2
\end{aligned}
\tag{28}
$$

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

$$
\begin{aligned}
\left|\left[V_t^{-1/2}-V_{t-1}^{-1/2}\right]_i\right|&=\left|\frac{1}{\sqrt{v_{t,i}}+\epsilon}-\frac{1}{\sqrt{v_{t-1,i}}+\epsilon}\right|\\
&=\left|\frac{\sqrt{v_{t-1,i}}-\sqrt{v_{t,i}}}{\left(\sqrt{v_{t,i}}+\epsilon\right)\left(\sqrt{v_{t-1,i}}+\epsilon\right)}\right|\\
&=\left|\frac{v_{t-1,i}-v_{t,i}}{\left(\sqrt{v_{t,i}}+\epsilon\right)\left(\sqrt{v_{t-1,i}}+\epsilon\right)\left(\sqrt{v_{t-1,i}}+\sqrt{v_{t,i}}\right)}\right|
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{(1-\beta_2)\left(\left[\left(\operatorname{desk}(\bar{\tilde{\Delta}}_t)\right)^2\right]_i - v_{t-1,i}\right)}{\left(\sqrt{v_{t,i}}+\epsilon\right)\left(\sqrt{v_{t-1,i}}+\epsilon\right)\left(\sqrt{v_{t-1,i}}+\sqrt{v_{t,i}}\right)} \\
&\leq \frac{(1-\beta_2)\left[\left(\operatorname{desk}(\bar{\tilde{\Delta}}_t)\right)^2\right]_i}{\epsilon^2\sqrt{(1-\beta_2)\left[\left(\operatorname{desk}(\bar{\tilde{\Delta}}_t)\right)^2\right]_i}} \\
&= \frac{\sqrt{1-\beta_2}}{\epsilon^2}\sqrt{\left[\left(\operatorname{desk}(\bar{\tilde{\Delta}}_t)\right)^2\right]_i} \\
&= \frac{\sqrt{1-\beta_2}}{\epsilon^2}\left|\left[\frac{\eta_{\text{local}}}{N}\sum_{c\in\mathcal{C}_t}\left(R_t^\top R_t \text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}},\tau\right)+R_t^\top \mathbf{z}_{c,t}\right)\right]_i\right| \\
&\leq \frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2}\left|\sum_{c\in\mathcal{C}_t}\left[R_t^\top R_t \text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}},\tau\right)\right]_i\right| + \frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2}\left|\left[\sum_{c\in\mathcal{C}_t}R_t^\top \mathbf{z}_{c,t}\right]_i\right|
\end{aligned} \tag{29}
$$

Therefore, we have

$$
\begin{aligned}
&\left\|V_t^{-1/2}-V_{t-1}^{-1/2}\right\|_2 = \max_i\left|\left[V_t^{-1/2}-V_{t-1}^{-1/2}\right]_i\right| \\
\leq &\max_i\left\{\frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2}\left|\sum_{c\in\mathcal{C}_t}\left[R_t^\top R_t \text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}},\tau\right)\right]_i\right| + \frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2}\left|\left[\sum_{c\in\mathcal{C}_t}R_t^\top \mathbf{z}_{c,t}\right]_i\right|\right\} \\
\leq &\max_i\left\{\frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2}\left|\sum_{c\in\mathcal{C}_t}\left[R_t^\top R_t \text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}},\tau\right)\right]_i\right|\right\} + \max_i\left\{\frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2}\left|\left[\sum_{c\in\mathcal{C}_t}R_t^\top \mathbf{z}_{c,t}\right]_i\right|\right\}
\end{aligned} \tag{30}
$$

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

$$
\left|\left[R_t^\top R_t \text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}},\tau\right)\right]_i\right| = \left|e_i R_t^\top R_t \text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}},\tau\right)\right| = \left|\left\langle R_t e_i, R_t \text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}},\tau\right)\right\rangle\right| \leq \left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)\tau_{K,G} \tag{31}
$$

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

$$
\begin{aligned}
&\max_i \frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2}\left|\sum_{c\in\mathcal{C}_t}\left[R_t^\top R_t \text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}},\tau\right)\right]_i\right| \\
\leq &\max_i \frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2}\sum_{c\in\mathcal{C}_t}\left|\left[R_t^\top R_t \text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}},\tau\right)\right]_i\right| \\
\leq &\frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2}\cdot N\cdot\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)\tau_{K,G} \\
= &\frac{\sqrt{1-\beta_2}\eta_{\text{local}}\tau_{K,G}}{\epsilon^2}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)
\end{aligned} \tag{32}
$$

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

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

so taking $a = \frac{\sigma_g \log(2dT/\delta)}{\sqrt{N}}$, we have that for each $i \in [d]$, with probability at least $1 - \frac{\delta}{Td}$,

$$\frac{1}{N}\left|\left[\sum_{c\in\mathcal{C}_t} R_t^\top \mathbf{z}_{c,t}\right]_i\right| \leq \frac{\sigma_g \log(2dT/\delta)}{\sqrt{N}} \tag{33}$$

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

$$\max_i \frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2}\left|\left[\sum_{c\in\mathcal{C}_t} R_t^\top \mathbf{z}_{c,t}\right]_i\right| \leq \frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{\epsilon^2}\cdot\frac{\sigma_g \log(2dT/\delta)}{\sqrt{N}} = \frac{\sqrt{1-\beta_2}\eta_{\text{local}}\sigma_g \log(2dT/\delta)}{\sqrt{N}\epsilon^2} \tag{34}$$

Substituting 32 and 34 into 30, we have that with probability at least $1 - \frac{2\delta}{T}$,

$$\begin{aligned}
&\left\|V_t^{-1/2} - V_{t-1}^{-1/2}\right\|_2 \\
\leq &\max_i\left\{\frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2}\left|\sum_{c\in\mathcal{C}_t}\left[R^\top R\text{clip}(\frac{\Delta_{c,t}}{\eta_{\text{local}}},\tau))\right]_i\right|\right\} + \max_i\left\{\frac{\sqrt{1-\beta_2}\eta_{\text{local}}}{N\epsilon^2}\left|\left[\sum_{c\in\mathcal{C}_t} R^\top \mathbf{z}_{c,t}\right]_i\right|\right\} \\
\leq &\frac{\sqrt{1-\beta_2}\eta_{\text{local}}\tau_{K,G}}{\epsilon^2}\left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right) + \frac{\sqrt{1-\beta_2}\eta_{\text{local}}\sigma_g \log(2dT/\delta)}{\sqrt{N}\epsilon^2}
\end{aligned} \tag{35}$$

Substituting 35 into 28, we have that with probability at least $1 - \frac{3\delta}{T}$,

$$\begin{aligned}
&\left|\frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top\left(V_t^{-1/2} - V_{t-1}^{-1/2}\right)\sum_{c\in\mathcal{C}_t} R_t^\top R_t\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)\right| \\
\leq &G\tau_{K,G}\left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)\left\|V_t^{-1/2} - V_{t-1}^{-1/2}\right\|_2 \\
\leq &G\tau_{K,G}\left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)\cdot\left(\frac{\sqrt{1-\beta_2}\eta_{\text{local}}\tau_{K,G}}{\epsilon^2}\left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right) + \frac{\sqrt{1-\beta_2}\eta_{\text{local}}\sigma_g \log(2dT/\delta)}{\sqrt{N}\epsilon^2}\right) \\
= &\frac{\eta_{\text{local}}\sqrt{1-\beta_2}G\tau_{K,G}^2}{\epsilon^2}\left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)^2 + \frac{\eta_{\text{local}}\sqrt{1-\beta_2}G\tau_{K,G}\sigma_g \log(2dT/\delta)}{\sqrt{N}\epsilon^2}\left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)
\end{aligned} \tag{36}$$

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

$$\begin{aligned}
&\left|\sum_{t=1}^T\frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top\left(V_t^{-1/2} - V_{t-1}^{-1/2}\right)\sum_{c\in\mathcal{C}_t} R_t^\top R_t\text{clip}\left(\sum_{k=1}^K g_{c,t,k},\tau\right)\right| \\
\leq &\frac{\eta_{\text{local}}\sqrt{1-\beta_2}TG\tau_{K,G}^2}{\epsilon^2}\left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)^2 + \frac{\eta_{\text{local}}\sqrt{1-\beta_2}TG\tau_{K,G}\sigma_g \log(2dT/\delta)}{\sqrt{N}\epsilon^2}\left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)
\end{aligned} \tag{37}$$

### E.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

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

$$
\begin{aligned}
&= \sum_{\tau'=0}^{t-1} \frac{\beta_1^{t-\tau'}\eta_{\text{global}}}{N} \nabla \mathcal{L}(\theta_t)^\top \left(V_{t-1}^{-1/2} - V_t^{-1/2}\right) R_{\tau'}^\top \sum_{c\in\mathcal{C}_{\tau'}} \tilde{\Delta}_{c,\tau'} \\
&= \sum_{\tau'=0}^{t-1} \frac{\beta_1^{t-\tau'}\eta_{\text{global}}}{N} \nabla \mathcal{L}(\theta_t)^\top \left(V_{t-1}^{-1/2} - V_t^{-1/2}\right) R_{\tau}^\top \sum_{c\in\mathcal{C}_{\tau'}} \eta_{\text{local}} \left(\text{sk}\left(\text{clip}\left(\frac{\Delta_{c,\tau'}}{\eta_{\text{local}}}, \tau\right)\right) + \mathbf{z}_{c,\tau'}\right) \\
&= \underbrace{\sum_{\tau'=0}^{t-1} \frac{\beta_1^{t-\tau'}\eta}{N} \nabla \mathcal{L}(\theta_t)^\top \left(V_{t-1}^{-1/2} - V_t^{-1/2}\right) \sum_{c\in\mathcal{C}_{\tau'}} R_{\tau'}^\top R_{\tau'} \text{clip}\left(\frac{\Delta_{c,\tau'}}{\eta_{\text{local}}}, \tau\right)}_{W_1} \\
&\quad + \underbrace{\sum_{\tau'=1}^{t-1} \frac{\beta_1^{t-\tau'}\eta}{N} \nabla \mathcal{L}(\theta_t)^\top \left(V_{t-1}^{-1/2} - V_t^{-1/2}\right) \sum_{c\in\mathcal{C}_{\tau'}} R_{\tau'}^\top \mathbf{z}_{c,\tau'}}_{W_2}
\end{aligned} \tag{38}
$$

#### E.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 36, with probability at least $1 - \frac{3\delta}{T}$,

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

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

$$
\begin{aligned}
&\sum_{\tau'=1}^{t-1} \frac{\beta_1^{t-\tau'}\eta}{N} \nabla \mathcal{L}(\theta_t)^\top \left(V_{t-1}^{-1/2} - V_t^{-1/2}\right) \sum_{c\in\mathcal{C}_{\tau'}} R_{\tau'}^\top R_{\tau'} \text{clip}\left(\frac{\Delta_{c,\tau'}}{\eta_{\text{local}}}, \tau\right) \\
\leq &\left(\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2} G \tau_{K,G}^2}{\epsilon^2} \left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)^2 + \frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2} G \tau_{K,G} \sigma_g \log(2dT/\delta)}{\sqrt{N}\epsilon^2} \left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)\right) \sum_{\tau'=0}^{t-1} \beta_1^{t-\tau'} \\
\leq &\frac{\eta\eta_{\text{local}}\beta_1\sqrt{1-\beta_2} G \tau_{K,G}^2}{\epsilon^2(1-\beta_1)} \left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)^2 + \frac{\eta\eta_{\text{local}}\beta_1\sqrt{1-\beta_2} G \tau_{K,G} \sigma_g \log(2dT/\delta)}{\sqrt{N}\epsilon^2(1-\beta_1)} \left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)
\end{aligned} \tag{39}
$$

#### E.1.2.2 Bounding $W_2$

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

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

Noticing that

$$
\frac{1}{N} \left(V_{t-1}^{-1/2} - V_t^{-1/2}\right) \sum_{c\in\mathcal{C}_{\tau'}} \mathbf{z}_{c,\tau'} \sim \mathcal{N}\left(0, \frac{\sigma_g^2}{N}\mathbb{I}\right),
$$

and $R_{\tau'} \left(V_{t-1}^{-1/2} - V_t^{-1/2}\right)^\top \nabla \mathcal{L}(\theta_t)$ is a $\frac{\left\|V_{t-1}^{-1/2} - V_t^{-1/2}\right\|_2 \|\nabla \mathcal{L}(\theta)_t\|_2}{\sqrt{b}}$-sub-Gaussian random vector, so according to Bernstein inequality,

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

$$
\begin{aligned}
\leq & 2\exp\left(-c\min\left(\frac{a^2}{b\cdot\frac{\sigma_g^2}{N}\cdot\frac{\left\|V_{t-1}^{-1/2}-V_t^{-1/2}\right\|_2^2\|\nabla\mathcal{L}(\theta)_t\|_2^2}{b}}, \frac{a}{\frac{\sigma_g}{\sqrt{N}}\cdot\frac{\left\|V_{t-1}^{-1/2}-V_t^{-1/2}\right\|_2\|\nabla\mathcal{L}(\theta)_t\|_2}{\sqrt{b}}}\right)\right) \\
= & 2\exp\left(-c\min\left(\frac{Na^2}{\sigma_g^2\left\|V_{t-1}^{-1/2}-V_t^{-1/2}\right\|_2^2\|\nabla\mathcal{L}(\theta)_t\|_2^2}, \frac{a\sqrt{bN}}{\sigma_g\left\|V_{t-1}^{-1/2}-V_t^{-1/2}\right\|_2\|\nabla\mathcal{L}(\theta)_t\|_2}\right)\right)
\end{aligned}
$$

so taking $a = \frac{\sigma_g\left\|V_{t-1}^{-1/2}-V_t^{-1/2}\right\|_2\|\nabla\mathcal{L}(\theta)_t\|_2\log(2T/\delta)}{\sqrt{N}}$, and combining with 35, we have that with probability at least $1-\frac{3\delta}{T}$,

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

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

$$
\begin{aligned}
& \sum_{\tau'=1}^{t-1}\frac{\beta_1^{t-\tau'}\eta}{N}\nabla\mathcal{L}(\theta_t)^\top\left(V_{t-1}^{-1/2}-V_t^{-1/2}\right)\sum_{c\in\mathcal{C}_{\tau'}}R_{\tau'}^\top\mathbf{z}_{c,\tau'} \\
\leq & \left(\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}\sigma_g G\tau_{K,G}\log(2T/\delta)}{\sqrt{N}\epsilon^2}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)+\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}\sigma_g^2 G\log(2T/\delta)\log(2dT/\delta)}{N\epsilon^2}\right)\sum_{\tau'=0}^{t-1}\beta_1^{t-\tau'} \\
\leq & \frac{\eta\eta_{\text{local}}\beta_1\sqrt{1-\beta_2}\sigma_g G\tau_{K,G}\log(2T/\delta)}{\sqrt{N}\epsilon^2(1-\beta_1)}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)+\frac{\eta\eta_{\text{local}}\beta_1\sqrt{1-\beta_2}\sigma_g^2 G\log(2T/\delta)\log(2dT/\delta)}{N\epsilon^2(1-\beta_1)}
\end{aligned}
\tag{40}
$$

Substituting 39 and 40 into 38, we can get that with probability at least $1-4\delta$,

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

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

$$\begin{aligned}
&\sum_{t=1}^{T} \nabla \mathcal{L}(\theta_t)^\top \frac{\beta_1 \eta_{\text{global}}}{1-\beta_1} \left(V_{t-1}^{-1/2} - V_t^{-1/2}\right) m_{t-1} \\
\leq & \frac{\eta \eta_{\text{local}} \beta_1 \sqrt{1-\beta_2} T G \tau_{K,G}^2}{\epsilon^2 (1-\beta_1)} \left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)^2 + \frac{2\eta \eta_{\text{local}} \beta_1 \sqrt{1-\beta_2} T G \tau_{K,G} \sigma_g \log(2dT/\delta)}{\sqrt{N} \epsilon^2 (1-\beta_1)} \left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right) \\
&+ \frac{\eta \eta_{\text{local}} \beta_1 \sqrt{1-\beta_2} T \sigma_g^2 G \log^2(2dT/\delta)}{N \epsilon^2 (1-\beta_1)} \qquad (41)
\end{aligned}$$

### E.1.3 Bounding $S_4$

We first bound each term with a fixed $t \in [T]$ in $S_4$. $\frac{1}{N} \nabla \mathcal{L}(\theta_t)^\top V_t^{-1/2} \sum_{c \in \mathcal{C}_t} R_t^\top \mathbf{z}_{c,t} = \left\langle R_t V_t^{-1/2} \nabla \mathcal{L}(\theta_t), \frac{1}{N} \sum_{c \in \mathcal{C}_t} \mathbf{z}_{c,t} \right\rangle$. Noticing that $\frac{1}{N} \sum_{c \in \mathcal{C}_t} \mathbf{z}_{c,t} \sim \mathcal{N}\left(0, \frac{\sigma_g^2}{N} \mathbb{I}\right)$, and $R_t V_t^{-1/2} \nabla \mathcal{L}(\theta_t)$ is a $\frac{\|\nabla \mathcal{L}(\theta_t)\|_2 \left\|V_t^{-1/2}\right\|}{\sqrt{b}}$-sub-Gaussian random vector, so according to Bernstein inequality,

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

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

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

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

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

Then by Azuma's inequality, we have

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

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

$$X_{T-1} = \sum_{t=1}^{T} \left\langle \nabla \mathcal{L}(\theta_t), \frac{1}{N} V_t^{-1/2} \sum_{c \in \mathcal{C}_t} R_t^\top \mathbf{z}_{c,t} \right\rangle \geq -\frac{\log^2(2T/\delta)\sqrt{2T}\sigma_g G}{\sqrt{N}\epsilon} \qquad (42)$$

### E.1.4 Bounding $S_3$

For each term in $S_1$, we have that

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

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

$$
\begin{aligned}
&\sum_{t=1}^{T}\frac{1}{N}\nabla\mathcal{L}(\theta_t)^\top V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}R_t^\top R_t\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)\\
=&K\underbrace{\sum_{t=1}^{T}\left\langle\nabla\mathcal{L}(\theta_t),V_{t-1}^{-1/2}\frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_t)\right\rangle}_{Y_1}+K\underbrace{\sum_{t=1}^{T}\left\langle\nabla\mathcal{L}(\theta_t),V_{t-1}^{-1/2}\left(\frac{1}{N}\sum_{c\in\mathcal{C}_t}\nabla\mathcal{L}_c(\theta_t)-\frac{1}{C}\sum_{i=1}^{C}\nabla\mathcal{L}_c(\theta_t)\right)\right\rangle}_{Y_2}\\
&+\underbrace{\sum_{t=1}^{T}\left\langle\nabla\mathcal{L}(\theta_t),\frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}(\nabla\mathcal{L}_c(\theta_{c,t,k})-\nabla\mathcal{L}_c(\theta_t))\right\rangle}_{Y_3}\\
&+\underbrace{\sum_{t=1}^{T}\left\langle\nabla\mathcal{L}(\theta_t),\frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}(g_{c,t,k}-\nabla\mathcal{L}_c(\theta_{c,t,k}))\right\rangle}_{Y_4}\\
&+\underbrace{\sum_{t=1}^{T}\left\langle\nabla\mathcal{L}(\theta_t),\frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\left(\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)-\sum_{k=1}^{K}g_{c,t,k}\right)\right\rangle}_{Y_5}\\
&+\underbrace{\sum_{t=1}^{T}\left\langle\nabla\mathcal{L}(\theta_t),\frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\left(R_t^\top R_t\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)-\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)\right)\right\rangle}_{Y_6}
\end{aligned}
\tag{43}
$$

#### E.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 31 and 33 into 29, we can see that with probability at least $1-\frac{2\delta}{d}$, for any $t\in[T]$,

$$\sqrt{\left[\left(\text{desk}(\tilde{\bar{\Delta}}_\tau)\right)^2\right]_i}=\left|\left[\frac{\eta_{\text{local}}}{N}\sum_{c\in\mathcal{C}_t}\left(R^\top R\text{clip}(\frac{\Delta_{c,t}}{\eta_{\text{local}}},\tau))+R^\top\mathbf{z}_{c,t}\right)\right]_i\right|$$

$$\leq \frac{\eta_{\text{local}}}{N}\left|\sum_{c\in\mathcal{C}_t}\left[R_t^\top R_t \text{clip}\left(\frac{\Delta_{c,t}}{\eta_{\text{local}}},\tau\right)\right]_i\right| + \frac{\eta_{\text{local}}}{N}\left|\left[\sum_{c\in\mathcal{C}_t} R_t^\top \mathbf{z}_{c,t}\right]_i\right|$$
$$\leq \eta_{\text{local}}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)\tau_{K,G} + \frac{\eta_{\text{local}}\sigma_g\log(2dT/\delta)}{\sqrt{N}}$$

Then according to the definition of second order moment, we have that with probability at least $1-\frac{2\delta}{d}$,

$$\sqrt{v_{t-1,i}} \leq \max_{\tau\leq t-1}\sqrt{\left[\left(\text{desk}(\bar{\tilde{\Delta}}_\tau)\right)^2\right]_i} \leq \eta_{\text{local}}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)\tau_{K,G} + \frac{\eta_{\text{local}}\sigma_g\log(2dT/\delta)}{\sqrt{N}}$$

so we can get that with probability at least $1-2\delta$, for all $i\in[d]$,

$$\left[V_{t-1}^{-1/2}\right]_i = \left(\sqrt{v_{t-1,i}}+\epsilon\right)^{-1} \geq \left(\eta_{\text{local}}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)\tau_{K,G} + \frac{\eta_{\text{local}}\sigma_g\log(2dT/\delta)}{\sqrt{N}}+\epsilon\right)^{-1}$$

which implies that with probability

$$\left\langle \nabla\mathcal{L}(\theta_t), V_{t-1}^{-1/2}\frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_t)\right\rangle$$
$$\geq \left(\eta_{\text{local}}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)\tau_{K,G} + \frac{\eta_{\text{local}}\sigma_g\log(2dT/\delta)}{\sqrt{N}}+\epsilon\right)^{-1}\nabla\mathcal{L}(\theta_t)^\top\frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_t)$$
$$= \left(\eta_{\text{local}}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)\tau_{K,G} + \frac{\eta_{\text{local}}\sigma_g\log(2dT/\delta)}{\sqrt{N}}+\epsilon\right)^{-1}\|\nabla\mathcal{L}(\theta_t)\|_2^2$$

so with probability at least $1-2\delta$,

$$\sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), V_{t-1}^{-1/2}\frac{1}{C}\sum_{c=1}^{C}\nabla\mathcal{L}_c(\theta_t)\right\rangle \geq \left(\eta_{\text{local}}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)\tau_{K,G} + \frac{\eta_{\text{local}}\sigma_g\log(2dT/\delta)}{\sqrt{N}}+\epsilon\right)^{-1}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\theta_t)\|_2^2 \tag{44}$$

#### E.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

$$\mathbb{P}\left(\left|\left\langle \nabla\mathcal{L}(\theta_t), V_{t-1}^{-1/2}\left(\frac{1}{N}\sum_{c\in\mathcal{C}_t}\nabla\mathcal{L}_c(\theta_t) - \frac{1}{C}\sum_{i=1}^{C}\nabla\mathcal{L}_c(\theta_t)\right)\right\rangle\right| \geq a\right)$$
$$\leq 2\exp\left(-\frac{2Nt^2}{\left(\frac{2G^2}{\epsilon}\right)^2}\right) = 2\exp\left(-\frac{N\epsilon^2t^2}{2G^4}\right)$$

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

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

Then denote $Z_t = \sum_{\tau'=0}^{t}\left\langle \nabla\mathcal{L}(\theta_{\tau'}), V_{t-1}^{-1/2}\left(\frac{1}{N}\sum_{c\in\mathcal{C}_{\tau'}}\nabla\mathcal{L}_c(\theta_{\tau'}) - \frac{1}{C}\sum_{i=1}^{C}\nabla\mathcal{L}_c(\theta_{\tau'})\right)\right\rangle$, 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 $t\in[T]$,

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

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

$$Z_{T-1} = \sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}\sum_{c\in\mathcal{C}_t}\nabla\mathcal{L}_c(\theta_t) - \frac{1}{C}\sum_{i=1}^{C}\nabla\mathcal{L}_c(\theta_t)\right\rangle \geq -\frac{2\log(2T/\delta)\sqrt{T}G^2}{\sqrt{N}\epsilon} \tag{45}$$

#### E.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

$$\sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}\left(\nabla\mathcal{L}_c(\theta_{c,t,k}) - \nabla\mathcal{L}_c(\theta_t)\right)\right\rangle \geq -\frac{\eta_{\text{local}}TLK^2G^2}{2\epsilon} \tag{46}$$

#### E.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 $\sigma_s$-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 $1-2\delta$,

$$W_{T-1} = \sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\sum_{k=1}^{K}(g_{c,t,k} - \nabla\mathcal{L}_c(\theta_{c,t,k}))\right\rangle \geq -\frac{G\sqrt{2TK}\log(2T/\delta)\sigma_s}{\sqrt{N}\epsilon} \tag{47}$$

#### E.1.4.5 Bounding $Y_5$

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

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

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

$$\sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\left(\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)-\sum_{k=1}^{K}g_{c,t,k}\right)\right\rangle \geq -\max\left\{0,\frac{TG(KG-\tau)}{\epsilon}\right\} \tag{48}$$

#### E.1.4.6 Bounding $Y_6$

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

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

Then denote $U_t=\sum_{\tau'=0}^{t}\left\langle \nabla\mathcal{L}(\theta_{\tau'}), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_{\tau'}}\left(R_{\tau'}^\top R_{\tau'}\text{clip}\left(\sum_{k=1}^{K}g_{c,\tau',k},\tau\right)-\text{clip}\left(\sum_{k=1}^{K}g_{c,\tau',k},\tau\right)\right)\right\rangle$, 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 $t\in[T]$,

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

Then by Azuma's inequality, we have

$$\mathbb{P}\left(U_{T-1}\leq -a\right)\leq\exp\left(-\frac{a^2}{2\cdot T\cdot\left(\frac{\log^{1.5}(NTd/\delta)G\tau_{K,G}}{\sqrt{b}\epsilon}\right)^2}\right)=\exp\left(-\frac{b\epsilon^2a^2}{2T\log^2(NTd/\delta)G^2\tau_{K,G}^2}\right)$$

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

$$\begin{aligned}
W_{T-1} &= \sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\left(R_t^\top R_t\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)-\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)\right)\right\rangle \\
&\geq -\frac{\log^2(NTd/\delta)\sqrt{2T}G\tau_{K,G}}{\sqrt{b}\epsilon}
\end{aligned} \tag{49}$$

Substituting 44, 45, 46, 47, 48, 49 into 43, we have that with probability at least $1-8\delta$,

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

$$
\begin{aligned}
&+\sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\left(\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)-\sum_{k=1}^{K}g_{c,t,k}\right)\right\rangle\\
&+\sum_{t=1}^{T}\left\langle \nabla\mathcal{L}(\theta_t), \frac{1}{N}V_{t-1}^{-1/2}\sum_{c\in\mathcal{C}_t}\left(R_t^\top R_t\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)-\text{clip}\left(\sum_{k=1}^{K}g_{c,t,k},\tau\right)\right)\right\rangle\\
\geq&\left(\eta_{\text{local}}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)\tau_{K,G}+\frac{\eta_{\text{local}}\sigma_g\log(2dT/\delta)}{\sqrt{N}}+\epsilon\right)^{-1}K\sum_{t=1}^{T}\|\nabla\mathcal{L}(\theta_t)\|_2^2-\frac{2K\sqrt{T}\log(2T/\delta)G^2}{\sqrt{N}\epsilon}\\
&-\frac{\eta_{\text{local}}TLK^2G^2}{2\epsilon}-\frac{G\sqrt{2TK}\log(2T/\delta)\sigma_s}{\sqrt{N}\epsilon}-\max\left\{0,\frac{TG(KG-\tau)}{\epsilon}\right\}-\frac{\log^2(NTd/\delta)\sqrt{2T}G\tau_{K,G}}{\sqrt{b}\epsilon}
\end{aligned}
\tag{50}
$$

Substituting 37, 41, 42, 50 into 27, we have that with probability at least $1-17\delta$,

$$
\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}R_t^\top R_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}R_t^\top R_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}R_t^\top\mathbf{z}_{c,t}\\
\leq&-\left(\eta_{\text{local}}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)\tau_{K,G}+\frac{\eta_{\text{local}}\sigma_g\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{\eta\log^2(NTd^2/\delta)\sqrt{2T}G\tau_{K,G}}{\sqrt{b}\epsilon}+\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}TG\tau_{K,G}^2}{\epsilon^2}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\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{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)\\
&+\frac{\eta\eta_{\text{local}}\beta_1\sqrt{1-\beta_2}TG\tau_{K,G}^2}{\epsilon^2(1-\beta_1)}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)^2\\
&+\frac{2\eta\eta_{\text{local}}\beta_1\sqrt{1-\beta_2}TG\tau_{K,G}\sigma_g\log(2dT/\delta)}{\sqrt{N}\epsilon^2(1-\beta_1)}\left(1+\frac{\log^{1.5}(CKTd/\delta)}{\sqrt{b}}\right)\\
&+\frac{\eta\eta_{\text{local}}\beta_1\sqrt{1-\beta_2}T\sigma_g^2G\log^2(2dT/\delta)}{N\epsilon^2(1-\beta_1)}+\frac{\eta\log^2(2T/\delta)\sqrt{2T}\sigma_gG}{\sqrt{N}\epsilon}\\
=&-\left(\eta_{\text{local}}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)\tau_{K,G}+\frac{\eta_{\text{local}}\sigma_g\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{\eta\log^2(NTd/\delta)\sqrt{2T}G\tau_{K,G}}{\sqrt{b}\epsilon}+\frac{\eta\eta_{\text{local}}\sqrt{1-\beta_2}TG\tau_{K,G}^2}{\epsilon^2(1-\beta_1)}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)^2\\
&+\frac{\eta\eta_{\text{local}}(1+\beta_1)\sqrt{1-\beta_2}TG\tau_{K,G}\sigma_g\log(2dT/\delta)}{\sqrt{N}\epsilon^2(1-\beta_1)}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{\eta\eta_{\text{local}}\beta_1\sqrt{1-\beta_2}T\sigma_g^2G\log^2(2dT/\delta)}{N\epsilon^2(1-\beta_1)}+\frac{\eta\log^2(2T/\delta)\sqrt{2T}\sigma_gG}{\sqrt{N}\epsilon}\\
&\leq-\left(\eta_{\text{local}}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)\tau_{K,G}+\frac{\eta_{\text{local}}\sigma_g\log(2dT/\delta)}{\sqrt{N}}+\epsilon\right)^{-1}\eta K\sum_{t=1}^{T}\|\nabla\mathcal{L}(\theta_t)\|_2^2\\
&+\frac{2\eta K\sqrt{T}\log(2T/\delta)G^2}{\sqrt{N}\epsilon}+\frac{\eta\eta_{\text{local}}TLK^2G^2}{2\epsilon}+\frac{\eta G\sqrt{2TK}\log(2T/\delta)\sigma_s}{\sqrt{N}\epsilon}+\eta\max\left\{0,\frac{TG(KG-\tau)}{\epsilon}\right\}\\
&+\frac{\eta\log^2(NTd/\delta)\sqrt{2T}G\tau_{K,G}}{\sqrt{b}\epsilon}+\frac{\eta\eta_{\text{local}}(2+\beta_1)\sqrt{1-\beta_2}TG\tau_{K,G}^2}{\epsilon^2(1-\beta_1)}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)^2\\
&+\frac{\eta\eta_{\text{local}}(1+2\beta_1)\sqrt{1-\beta_2}T\sigma_g^2G\log^2(2dT/\delta)}{N\epsilon^2(1-\beta_1)}\\
&+\frac{\log^2(2T/\delta)G^2}{\epsilon}+\frac{2\eta^2T\log^2(2T/\delta)\sigma_g^2}{N\epsilon}
\end{aligned}
\tag{51}
$$

### E.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

$$
\begin{aligned}
&\sum_{t=1}^{T}(\nabla\mathcal{L}(\gamma_t)-\nabla\mathcal{L}(\theta_t))^\top(\gamma_{t+1}-\gamma_t)\\
=&\frac{\beta_1}{(1-\beta_1)^2}\underbrace{\sum_{t=1}^{T}(\theta_t-\theta_{t-1})^\top\hat{H}'_{\mathcal{L},t}(\theta_{t+1}-\theta_t)}_{S_5}-\frac{\beta_1^2}{(1-\beta_1)^2}\underbrace{\sum_{t=1}^{T}(\theta_t-\theta_{t-1})^\top\hat{H}'_{\mathcal{L},t}(\theta_t-\theta_{t-1})}_{S_6}
\end{aligned}
\tag{52}
$$

#### E.2.1 Bounding $S_5$

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

$$
\begin{aligned}
(\theta_t-\theta_{t-1})^\top\hat{H}'_{\mathcal{L},t}(\theta_{t+1}-\theta_t)&=\eta_{\text{global}}^2\left(V_{t-1}^{-1/2}m_{t-1}\right)^\top\hat{H}'_{\mathcal{L},t}\left(V_t^{-1/2}m_t\right)\\
&=\eta_{\text{global}}^2\left(V_{t-1}^{-1/2}m_{t-1}\right)^\top\left(\sum_{i=1}^{d}\lambda_iv_iv_i^\top\right)\left(V_t^{-1/2}m_t\right)\\
&=\eta_{\text{global}}^2\sum_{i=1}^{d}\lambda_i\left|\left(V_{t-1}^{-1/2}m_{t-1}\right)^\top v_i\right|\left|\left(V_t^{-1/2}m_t\right)^\top v_i\right|
\end{aligned}
\tag{53}
$$

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

$$
\begin{aligned}
\left|\left(V_{t-1}^{-1/2}m_{t-1}\right)^\top v_i\right|&=(1-\beta_1)\left|\left(\sum_{\tau=1}^{t-1}\beta_1^{t-1-\tau}V_{t-1}^{-1/2}\operatorname{desk}(\bar{\tilde{\Delta}}_\tau)\right)^\top v_i\right|\\
&\leq(1-\beta_1)\sum_{\tau'=0}^{t-1}\beta_1^{t-1-\tau'}\left|\left(V_{t-1}^{-1/2}\operatorname{desk}(\bar{\tilde{\Delta}}_{\tau'})\right)^\top v_i\right|
\end{aligned}
$$

$$\leq \max_{\tau' \in [t-1]} \left| \left( V_{t-1}^{-1/2} \operatorname{desk}(\bar{\tilde{\Delta}}_{\tau'}) \right)^\top v_i \right| \tag{54}$$

Similarly,

$$\left| \left( V_t^{-1/2} m_t \right)^\top v_i \right| \leq \max_{\tau' \in [t]} \left| \left( V_t^{-1/2} \operatorname{desk}(\bar{\tilde{\Delta}}_{\tau'}) \right)^\top v_i \right| \tag{55}$$

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

$$\begin{aligned}
&\left| \left( V_{t-1}^{-1/2} \operatorname{desk}(\bar{\tilde{\Delta}}_{\tau'}) \right)^\top v_i \right| \\
=&\left| \left\langle \operatorname{desk}(\bar{\tilde{\Delta}}_{\tau'}), V_{t-1}^{-1/2} v_i \right\rangle \right| \\
=&\frac{1}{N} \left| \left\langle R_{\tau'}^\top \sum_{c \in \mathcal{C}_{\tau'}} \tilde{\Delta}_{c,\tau'}, V_{t-1}^{-1/2} v_i \right\rangle \right| \\
=&\frac{1}{N} \left| \left\langle R_{\tau'}^\top \sum_{c \in \mathcal{C}_{\tau'}} \eta_{\text{local}} \left( \operatorname{sk} \left( \operatorname{clip} \left( \frac{\Delta_{c,\tau'}}{\eta_{\text{local}}}, \tau \right) \right) + \mathbf{z}_{c,\tau'} \right), V_{t-1}^{-1/2} v_i \right\rangle \right| \\
\leq&\frac{\eta_{\text{local}}}{N} \left| \left\langle \sum_{c \in \mathcal{C}_\tau} R_{\tau'}^\top R_{\tau'} \operatorname{clip} \left( \sum_{k=1}^K g_{c,\tau',k}, \tau \right), V_{t-1}^{-1/2} v_i \right\rangle \right| + \frac{\eta_{\text{local}}}{N} \left| \left\langle \sum_{c \in \mathcal{C}_{\tau'}} R_{\tau'}^\top \mathbf{z}_{c,\tau'}, V_{t-1}^{-1/2} v_i \right\rangle \right|
\end{aligned} \tag{56}$$

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

$$\begin{aligned}
&\left| \left\langle \sum_{c \in \mathcal{C}_{\tau'}} R_{\tau'}^\top R_{\tau'} \operatorname{clip} \left( \sum_{k=1}^K g_{c,\tau',k}, \tau \right), V_{t-1}^{-1/2} v_i \right\rangle \right| \\
\leq& \left( 1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}} \right) \left\| V_{t-1}^{-1/2} v_i \right\|_2 \sum_{c \in \mathcal{C}_{\tau'}} \left\| \operatorname{clip} \left( \sum_{k=1}^K g_{c,\tau',k}, \tau \right) \right\| \\
\leq& \left( 1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}} \right) \frac{N \tau_{K,G}}{\epsilon}
\end{aligned} \tag{57}$$

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

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

so taking $a = \frac{\sigma_g \left\| V_{t-1}^{-1/2} \right\|_2 \log(2T/\delta)}{\sqrt{N}}$, and noticing that $\left\| V_{t-1}^{-1/2} \right\|_2 \leq \frac{1}{\epsilon}$, we have that with probability at least $1 - \frac{\delta}{Td}$,

$$\left| \left\langle \frac{1}{N} \sum_{c \in \mathcal{C}_{\tau'}} R_{\tau'}^\top \mathbf{z}_{c,\tau'}, V_{\tau'-1}^{-1/2} v_i \right\rangle \right| \leq \frac{\sigma_g \left\| V_{t-1}^{-1/2} \right\|_2 \log(2dT/\delta)}{\sqrt{N}} \leq \frac{\sigma_g \log(2dT/\delta)}{\sqrt{N}\epsilon} \tag{58}$$

Substituting 57 and 58 into 56, we have that with probability at least $1 - \frac{2\delta}{dT}$,

$$\begin{aligned}
&\left|\left(V_{t-1}^{-1/2}\operatorname{desk}(\bar{\tilde{\Delta}}_{\tau'})\right)^\top v_i\right| \\
\leq&\frac{\eta_{\text{local}}}{N}\left|\left\langle\sum_{c\in\mathcal{C}_\tau} R_{\tau'}^\top R_{\tau'}\text{clip}\left(\sum_{k=1}^K g_{c,\tau',k},\tau\right), V_{t-1}^{-1/2}v_i\right\rangle\right| + \frac{\eta_{\text{local}}}{N}\left|\left\langle\sum_{c\in\mathcal{C}_{\tau'}} R_{\tau'}^\top \mathbf{z}_{c,\tau'}, V_{t-1}^{-1/2}v_i\right\rangle\right| \\
\leq&\frac{\eta_{\text{local}}\tau_{K,G}}{\epsilon}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right) + \frac{\eta_{\text{local}}\sigma_g\log(2dT/\delta)}{\sqrt{N}\epsilon} \qquad (59)
\end{aligned}$$

Similarly, we can get that for $\tau' \in [t]$, with probability at least $1 - \frac{2\delta}{dT}$,

$$\left|\left(V_t^{-1/2}\operatorname{desk}(\bar{\tilde{\Delta}}_{\tau'})\right)^\top v_i\right| \leq \frac{\eta_{\text{local}}\tau_{K,G}}{\epsilon}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right) + \frac{\eta_{\text{local}}\sigma_g\log(2dT/\delta)}{\sqrt{N}\epsilon} \qquad (60)$$

Substituting 59 into 54, 60 and 55, we have that with probability at least $1 - \frac{2\delta}{d}$,

$$\begin{aligned}
\left|\left(V_{t-1}^{-1/2}m_{t-1}\right)^\top v_i\right| &\leq \max_{\tau'\in[t-1]}\left|\left(V_{t-1}^{-1/2}\operatorname{desk}(\bar{\tilde{\Delta}}_{\tau'})\right)^\top v_i\right| \\
&\leq \frac{\eta_{\text{local}}\tau_{K,G}}{\epsilon}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right) + \frac{\eta_{\text{local}}\sigma_g\log(2dT/\delta)}{\sqrt{N}\epsilon} \qquad (61) \\
\left|\left(V_t^{-1/2}m_t\right)^\top v_i\right| &\leq \max_{\tau'\in[t]}\left|\left(V_t^{-1/2}\operatorname{desk}(\bar{\tilde{\Delta}}_{\tau'})\right)^\top v_i\right| \\
&\leq \frac{\eta_{\text{local}}\tau_{K,G}}{\epsilon}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right) + \frac{\eta_{\text{local}}\sigma_g\log(2dT/\delta)}{\sqrt{N}\epsilon} \qquad (62)
\end{aligned}$$

Substituting 61 and 62 into 53, we have that with probability at least $1 - 2\delta$,

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

By taking summation from 0 to $T-1$, we have that with probability at least $1 - 2\delta$,

$$\sum_{t=1}^T\left(\theta_t - \theta_{t-1}\right)^\top \hat{H}'_{\mathcal{L},t}\left(\theta_{t+1} - \theta_t\right) \leq \frac{2\eta^2\mathcal{I}LT\tau_{K,G}^2}{\epsilon^2}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)^2 + \frac{2\eta^2\sigma_g^2\mathcal{I}LT\log^2(2dT/\delta)}{N\epsilon^2} \qquad (63)$$

#### E.2.2 Bounding $S_6$

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

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

$$= \eta_{\text{global}}^2 \sum_{i=1}^{d} \lambda_i \left| \left( V_{t-1}^{-1/2} m_{t-1} \right)^\top v_i \right|^2 \tag{64}$$

Substituting 61 into 64, we have that with probability at least $1 - 2\delta$,

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

By taking summation from 0 to $T-1$, we have that with probability at least $1 - 2\delta$,

$$\sum_{t=1}^{T} (\theta_t - \theta_{t-1})^\top \hat{H}'_{\mathcal{L},t} (\theta_t - \theta_{t-1}) \geq -\frac{2\eta^2 \mathcal{I} L T \tau_{K,G}^2}{\epsilon^2} \left( 1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}} \right)^2 - \frac{2\eta^2 \sigma_g^2 \mathcal{I} L T \log^2(2dT/\delta)}{N\epsilon^2} \tag{65}$$

Substituting 63 and 65 into 52, we have that with probability at least $1 - 2\delta$,

$$\begin{aligned}
&\sum_{t=1}^{T} (\nabla \mathcal{L}(\gamma_t) - \nabla \mathcal{L}(\theta_t))^\top (\gamma_{t+1} - \gamma_t) \\
=&\frac{\beta_1}{(1-\beta_1)^2} \sum_{t=1}^{T} (\theta_t - \theta_{t-1})^\top \hat{H}'_{\mathcal{L},t} (\theta_{t+1} - \theta_t) - \frac{\beta_1^2}{(1-\beta_1)^2} \sum_{t=1}^{T} (\theta_t - \theta_{t-1})^\top \hat{H}'_{\mathcal{L},t} (\theta_t - \theta_{t-1}) \\
\leq&\frac{\beta_1(1+\beta_1)}{(1-\beta_1)^2} \left( \frac{2\eta^2 \mathcal{I} L T \tau_{K,G}^2}{\epsilon^2} \left( 1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}} \right)^2 + \frac{2\eta^2 \sigma_g^2 \mathcal{I} L T \log^2(2dT/\delta)}{N\epsilon^2} \right) \\
=&\frac{2\eta^2 \beta_1(1+\beta_1) \mathcal{I} L T \tau_{K,G}^2}{\epsilon^2 (1-\beta_1)^2} \left( 1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}} \right)^2 + \frac{2\eta^2 \beta_1(1+\beta_1) \sigma_g^2 \mathcal{I} L T \log^2(2dT/\delta)}{N\epsilon^2 (1-\beta_1)^2}
\end{aligned} \tag{66}$$

### E.3 Bounding $T_3$

For each term in $T_3$, we have

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

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

$$\begin{aligned}
&\sum_{t=1}^{T} (\gamma_{t+1} - \gamma_t)^\top \hat{H}_{\mathcal{L},t} (\gamma_{t+1} - \gamma_t) \\
=&\frac{1}{(1-\beta_1)^2} \underbrace{\sum_{t=1}^{T} (\theta_{t+1} - \theta_t) \hat{H}_{\mathcal{L},t} (\theta_{t+1} - \theta_t)}_{S_7} - \frac{2\beta_1}{(1-\beta_1)^2} \underbrace{\sum_{t=1}^{T} (\theta_{t+1} - \theta_t) \hat{H}_{\mathcal{L},t} (\theta_t - \theta_{t-1})}_{S_8} \\
&+ \frac{\beta_1^2}{(1-\beta_1)^2} \underbrace{\sum_{t=1}^{T} (\theta_{t+1} - \theta_t) \hat{H}_{\mathcal{L},t} (\theta_{t+1} - \theta_t)}_{S_9}
\end{aligned} \tag{67}$$

### E.3.1 Bounding $S_7$

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

$$\begin{aligned}(\theta_{t+1}-\theta_t)\,\hat{H}_{\mathcal{L},t}\,(\theta_{t+1}-\theta_t) &= \eta_{\text{global}}^2 \left(V_t^{-1/2} m_t\right) \hat{H}_{\mathcal{L},t} \left(V_t^{-1/2} m_t\right) \\ &= \eta_{\text{global}}^2 \left(V_t^{-1/2} m_t\right) \left(\sum_{i=1}^d \lambda_i v_i v_i^\top\right) \left(V_t^{-1/2} m_t\right) \\ &= \eta_{\text{global}}^2 \sum_{i=1}^d \lambda_i \left| \left(V_t^{-1/2} m_t\right)^\top v_i \right|^2 \end{aligned} \tag{68}$$

Substituting 62 into 68, we have that with probability at least $1-2\delta$,

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

By taking summation from 0 to $T-1$, we have that with probability at least $1-2\delta$,

$$\sum_{t=1}^T (\theta_{t+1}-\theta_t)\,\hat{H}_{\mathcal{L},t}\,(\theta_{t+1}-\theta_t) \le \frac{2\eta^2 \mathcal{I} L T \tau^2}{\epsilon^2} \left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)^2 + \frac{2\eta^2 \sigma_g^2 \mathcal{I} L T \log^2(2dT/\delta)}{N\epsilon^2} \tag{69}$$

### E.3.2 Bounding $S_8$

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

$$\begin{aligned}(\theta_{t+1}-\theta_t)\,\hat{H}_{\mathcal{L}}\,(\theta_t-\theta_{t-1}) &= \eta_{\text{global}}^2 \left(V_t^{-1/2} m_t\right) \hat{H}_{\mathcal{L}} \left(V_{t-1}^{-1/2} m_{t-1}\right) \\ &= \eta_{\text{global}}^2 \left(V_t^{-1/2} m_t\right) \left(\sum_{i=1}^d \lambda_i v_i v_i^\top\right) \left(V_{t-1}^{-1/2} m_{t-1}\right) \\ &= \eta_{\text{global}}^2 \sum_{i=1}^d \lambda_i \left| \left(V_t^{-1/2} m_t\right)^\top v_i \right| \left| \left(V_{t-1}^{-1/2} m_{t-1}\right)^\top v_i \right| \end{aligned} \tag{70}$$

Substituting 61 and 62 into 70, we have that with probability at least $1-2\delta$,

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

By taking summation from 0 to $T-1$, we have that with probability at least $1-2\delta$,

$$\sum_{t=1}^T (\theta_{t+1}-\theta_t)\,\hat{H}_{\mathcal{L},t}\,(\theta_t-\theta_{t-1}) \ge -\frac{2\eta^2 \mathcal{I} L T \tau_{K,G}^2}{\epsilon^2} \left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)^2 - \frac{2\eta^2 \sigma_g^2 \mathcal{I} L T \log^2(2dT/\delta)}{N\epsilon^2} \tag{71}$$

### E.3.3 Bounding $S_9$

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

$$\begin{aligned}(\theta_t - \theta_{t-1})\,\hat{H}_{\mathcal{L},t}\,(\theta_t - \theta_{t-1}) &= \eta_{\text{global}}^2 \left(V_{t-1}^{-1/2} m_{t-1}\right) \hat{H}_{\mathcal{L},t} \left(V_{t-1}^{-1/2} m_{t-1}\right) \\ &= \eta_{\text{global}}^2 \left(V_{t-1}^{-1/2} m_{t-1}\right) \left(\sum_{i=1}^{d} \lambda_i v_i v_i^\top\right) \left(V_{t-1}^{-1/2} m_{t-1}\right) \\ &= \eta_{\text{global}}^2 \sum_{i=1}^{d} \lambda_i \left| \left(V_{t-1}^{-1/2} m_{t-1}\right)^\top v_i \right|^2 \end{aligned} \tag{72}$$

Substituting 61 into 72, we have that with probability at least $1 - 2\delta$,

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

By taking summation from 0 to $T-1$, we have that with probability at least $1 - 2\delta$,

$$\sum_{t=1}^{T} (\theta_t - \theta_{t-1})\,\hat{H}_{\mathcal{L},t}\,(\theta_t - \theta_{t-1}) \le \frac{2\eta^2 \mathcal{I} L T \tau_{K,G}^2}{\epsilon^2} \left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)^2 + \frac{2\eta^2 \sigma_g^2 \mathcal{I} L T \log^2(2dT/\delta)}{N\epsilon^2} \tag{73}$$

Substituting 69, 71, 73 into 67, we have that with probability at least $1 - 2\delta$,

$$\begin{aligned}&\sum_{t=1}^{T} (\gamma_{t+1} - \gamma_t)^\top \hat{H}_{\mathcal{L},t}\,(\gamma_{t+1} - \gamma_t) \\ =&\frac{1}{(1-\beta_1)^2} \sum_{t=1}^{T} (\theta_{t+1} - \theta_t)\,\hat{H}_{\mathcal{L},t}\,(\theta_{t+1} - \theta_t) - \frac{2\beta_1}{(1-\beta_1)^2} \sum_{t=1}^{T} (\theta_{t+1} - \theta_t)\,\hat{H}_{\mathcal{L},t}\,(\theta_t - \theta_{t-1}) \\ &+ \frac{\beta_1^2}{(1-\beta_1)^2} \sum_{t=1}^{T} (\theta_{t+1} - \theta_t)\,\hat{H}_{\mathcal{L},t}\,(\theta_{t+1} - \theta_t) \\ &\le \frac{(1+\beta_1)^2}{(1-\beta_1)^2} \left( \frac{2\eta^2 \mathcal{I} L T \tau_{K,G}^2}{\epsilon^2} \left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)^2 + \frac{2\eta^2 \sigma_g^2 \mathcal{I} L T \log^2(2dT/\delta)}{N\epsilon^2} \right) \\ &= \frac{2\eta^2 (1+\beta_1)^2 \mathcal{I} L T \tau_{K,G}^2}{\epsilon^2 (1-\beta_1)^2} \left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)^2 + \frac{2\eta^2 (1+\beta_1)^2 \sigma_g^2 \mathcal{I} L T \log^2(2dT/\delta)}{N\epsilon^2 (1-\beta_1)^2}\end{aligned} \tag{74}$$

Substituting 51, 66, 74 into 26, we have that with probability at least $1 - 19\delta$,

$$\begin{aligned}&\mathcal{L}(\gamma_T) - \mathcal{L}(\theta_0) \\ =&\sum_{t=1}^{T} \nabla \mathcal{L}(\theta_t)^\top (\gamma_{t+1} - \gamma_t) + \sum_{t=1}^{T} (\nabla \mathcal{L}(\gamma_t) - \nabla \mathcal{L}(\theta_t))^\top (\gamma_{t+1} - \gamma_t) + \frac{1}{2} \sum_{t=1}^{T} (\gamma_{t+1} - \gamma_t)^\top \hat{H}_{\mathcal{L},t} (\gamma_{t+1} - \gamma_t) \\ \le& - \left( \eta_{\text{local}} \left(1 + \frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right) \tau_{K,G} + \frac{\eta_{\text{local}} \sigma_g \log(2dT/\delta)}{\sqrt{N}} + \epsilon \right)^{-1} \eta K \sum_{t=1}^{T} \|\nabla \mathcal{L}(\theta_t)\|_2^2 \\ &+ \frac{2\eta K \sqrt{T} \log(2T/\delta) G^2}{\sqrt{N}\epsilon} + \frac{\eta \eta_{\text{local}} T L K^2 G^2}{2\epsilon} + \frac{\eta G \sqrt{2TK} \log(2T/\delta) \sigma_s}{\sqrt{N}\epsilon} + \eta \max\left\{0, \frac{TG(KG - \tau)}{\epsilon}\right\}\end{aligned}$$

$$
\begin{aligned}
&+\frac{\eta\log^2(NTd/\delta)\sqrt{2T}G\tau_{K,G}}{\sqrt{b}\epsilon}+\frac{\eta\eta_{\text{local}}(2+\beta_1)\sqrt{1-\beta_2}TG\tau_{K,G}^2}{\epsilon^2(1-\beta_1)}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)^2\\
&+\frac{\eta\eta_{\text{local}}(1+2\beta_1)\sqrt{1-\beta_2}T\sigma_g^2G\log^2(2dT/\delta)}{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}{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{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)^2+\frac{2\eta^2\beta_1(1+\beta_1)\sigma_g^2\mathcal{I}LT\log^2(2dT/\delta)}{N\epsilon^2(1-\beta_1)^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{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)^2+\frac{2\eta^2(1+\beta_1)^2\sigma_g^2\mathcal{I}LT\log^2(2dT/\delta)}{N\epsilon^2(1-\beta_1)^2}\\
=&-\left(\eta_{\text{local}}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)\tau_{K,G}+\frac{\eta_{\text{local}}\sigma_g\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{\eta\log^2(NTd/\delta)\sqrt{2T}G\tau_{K,G}}{\sqrt{b}\epsilon}+\frac{\eta\eta_{\text{local}}(2+\beta_1)\sqrt{1-\beta_2}TG\tau_{K,G}^2}{\epsilon^2(1-\beta_1)}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)^2\\
&+\frac{\eta\eta_{\text{local}}(1+2\beta_1)\sqrt{1-\beta_2}T\sigma_g^2G\log^2(2dT/\delta)}{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}{N\epsilon}\\
&+\frac{2\eta^2\left(1+2\beta_1\right)\left(1+\beta_1\right)\mathcal{I}LT\tau_{K,G}^2}{\epsilon^2(1-\beta_1)^2}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)^2+\frac{2\eta^2(1+2\beta_1)(1+\beta_1)\sigma_g^2\mathcal{I}LT\log^2(2dT/\delta)}{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 $1-19\delta$,

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(\theta_t)\|_2^2\leq&\frac{\alpha_2\left(\mathcal{L}(\theta_0)-\mathcal{L}^*\right)}{\eta KT}+\frac{2\alpha_2\log(2T/\delta)G^2}{\sqrt{NT}\epsilon}+\frac{\eta_{\text{local}}\alpha_2LKG^2}{2\epsilon}+\frac{\sqrt{2}\alpha_2G\log(2T/\delta)\sigma_s}{\sqrt{NTK}\epsilon}\\
&+\max\left\{0,\frac{\alpha_2G(KG-\tau)}{K\epsilon}\right\}+\frac{\sqrt{2}\alpha_2\log^2(NTd/\delta)G\tau_{K,G}}{\sqrt{bTK}\epsilon}+\frac{\eta_{\text{local}}\alpha_1^2\alpha_2(2+\beta_1)\sqrt{1-\beta_2}G\tau_{K,G}^2}{K\epsilon^2(1-\beta_1)}\\
&+\frac{\eta_{\text{local}}\alpha_2(1+2\beta_1)\sqrt{1-\beta_2}\sigma_g^2G\log^2(2dT/\delta)}{NK\epsilon^2(1-\beta_1)}+\frac{\alpha_2\log^2(2T/\delta)G^2}{\eta TK\epsilon}+\frac{2\eta\alpha_2\log^2(2T/\delta)\sigma_g^2}{NK\epsilon}\\
&+\frac{2\eta\alpha_1^2\alpha_2\left(1+2\beta_1\right)\left(1+\beta_1\right)\mathcal{I}L\tau_{K,G}^2}{K\epsilon^2(1-\beta_1)^2}+\frac{2\eta\alpha_2(1+2\beta_1)(1+\beta_1)\sigma_g^2\mathcal{I}L\log^2(2dT/\delta)}{NK\epsilon^2(1-\beta_1)^2}
\end{aligned}
$$

in which

$$
\alpha_1=1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}},\alpha_2=\eta_{\text{local}}\left(1+\frac{\log^{1.5}(NTd^2/\delta)}{\sqrt{b}}\right)\tau_{K,G}+\frac{\eta_{\text{local}}\sigma_g\log(2dT/\delta)}{\sqrt{N}}+\epsilon
$$

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

# F Additional Experimental Results

We will present more experiment results under different privacy levels. These figures further verifies our observations and conclusions in Section 4.[1]

## F.1 Vision Tasks

### F.1.1 ResNet101 on EMNIST

#### F.1.1.1 Different Privacy Level $\epsilon_p$

Training dynamics and test accuracies of Fed-SGM with ADAM, Fed-SGM with GD, DP-FedAvg and its Adam variant, and DiffSketch training ResNet101 on EMNIST with $\epsilon_p = \{2.75, 0.42, 0.18\}$ are presented in Figure 3-5. Furthermore, in Figure 6, with a fixed privacy level $\epsilon_p = 1.6$, we also show the comparison of training dynamics and test accuracies among sketching dimension $b \in \left\{4 \times 10^4, 4 \times 10^5, 4 \times 10^6, 4 \times 10^7\right\}$.

#### F.1.1.2 Ablation Studies

We conduct the following groups of ablation experiments, mirroring the setup of Figure 1 and varying only the clipping threshold $\tau$ and the number of local steps $K$.

- We compare our original setting ($\tau = 1$) with $\tau = 0.5$ and $\tau = 2$, holding all other hyperparameters constant. As shown in Table 3, in our current experimental setup, test accuracy increases as grows, indicating that a looser clipping bound allows larger gradient norms and yields better utility under our noise regime.

Table 3: Test accuracy (%) for different clipping thresholds $\tau$.

| | GD, Sketching | GD, Non-Sketching | Adam, Sketching | Adam, Non-Sketching |
|---|---|---|---|---|
| $\tau = 0.5$ | 62.91% | 62.89% | 85.03% | 78.42% |
| $\tau = 1$ | 62.98% | 63.34% | 85.09% | 78.60% |
| $\tau = 2$ | 63.18% | 63.57% | 85.33% | 78.75% |

- We compare our original setting ($K = 18$) with $K = 9$ and $K = 36$, keeping all other hyperparameters fixed. Table 4 shows that, in our current experimental setup, increasing allows each client to perform more optimization steps, which leads to better test accuracy.

Table 4: Test accuracy (%) for different local steps $K$.

| | GD, Sketching | GD, Non-Sketching | Adam, Sketching | Adam, Non-Sketching |
|---|---|---|---|---|
| $K = 9$ | 56.54% | 56.66% | 83.63% | 76.57% |
| $K = 18$ | 62.98% | 63.34% | 85.09% | 78.60% |
| $K = 36$ | 71.81% | 72.18% | 86.72% | 81.04% |

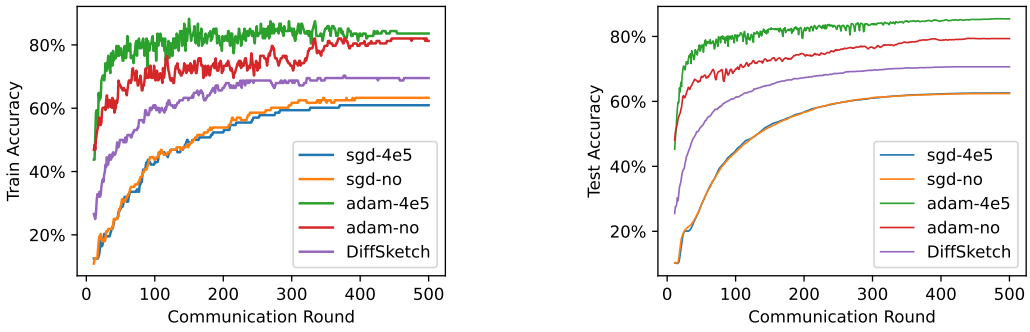
### F.1.2 ResNet50 on MNIST

Training dynamics and test accuracies of Fed-SGM with ADAM, Fed-SGM with GD, DP-FedAvg and its Adam variant training ResNet50 on EMNIST with $\epsilon_p = \{2.75, 0.42, 0.18\}$ are presented in Figure 7-10.

## F.2 Language Tasks

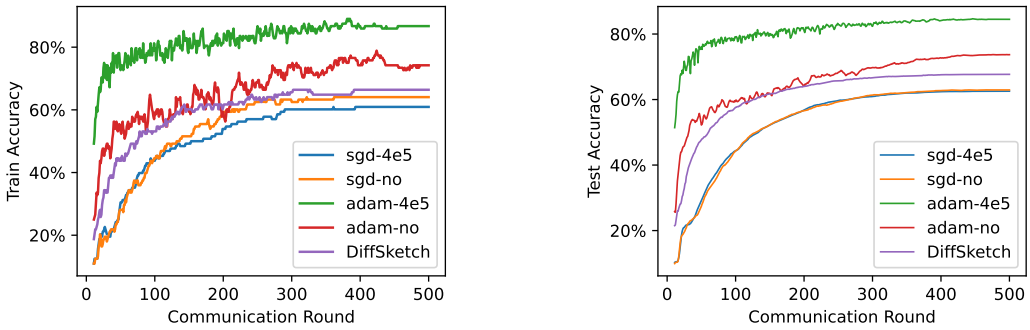
Training dynamics and test accuracies of Fed-SGM with ADAM, Fed-SGM with GD, DP-FedAvg and its Adam variant finetuning Bert on SST-2 with $\epsilon_p = \{2.45, 0.35, 0.12\}$ are presented in Figure 11-13.

[1]All experiments in this section and Section 4 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://github.com/lucmonl/mlopt/tree/main.



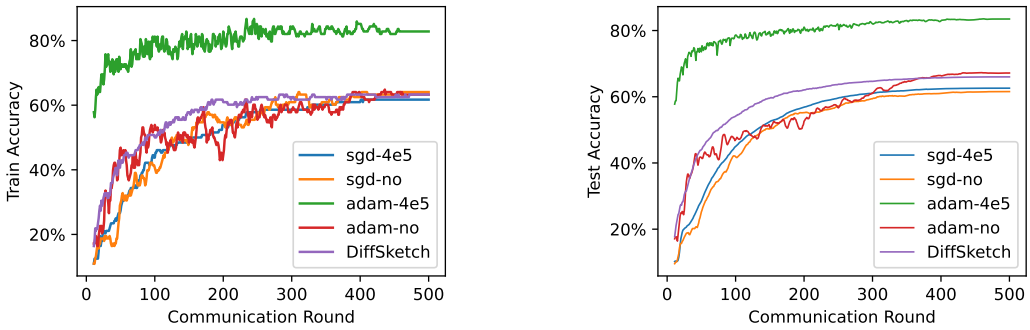

(a) Training Accuracy (b) Test Accuracy

Figure 3: Vision Task: EMNIST, $\epsilon_p = 2.75$



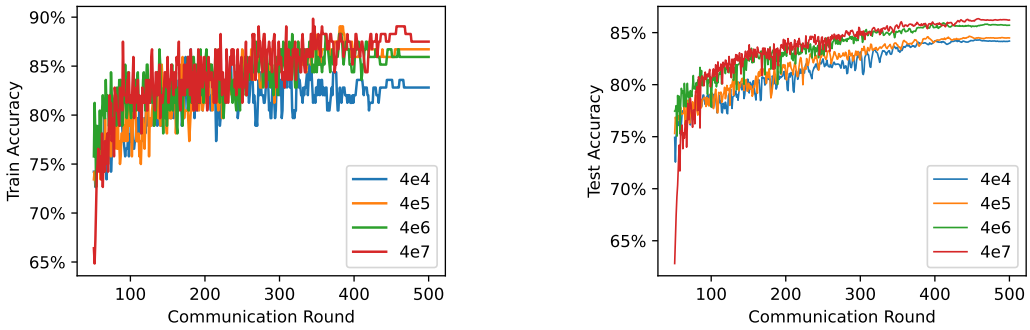

(a) Training Accuracy (b) Test Accuracy

Figure 4: Vision Task: EMNIST, $\epsilon_p = 0.42$



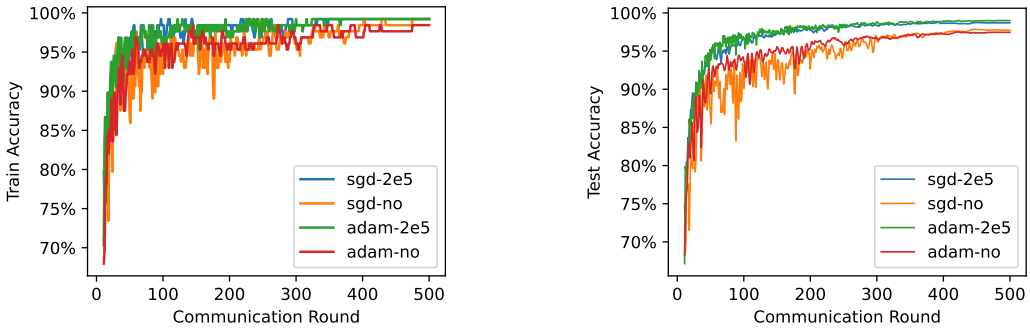

(a) Training Accuracy (b) Test Accuracy

Figure 5: Vision Task: EMNIST, $\epsilon_p = 0.18$



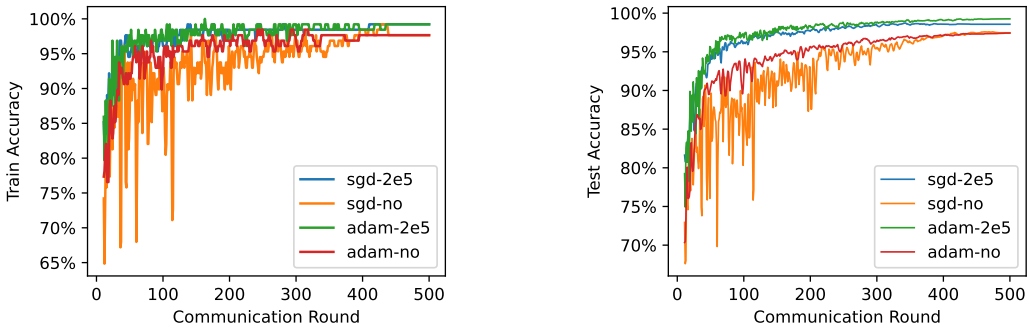

(a) Training Accuracy (b) Test Accuracy

Figure 6: Vision task: EMNIST, $\epsilon_p = 1, 60$, Comparison with Different Sketching Dimensions



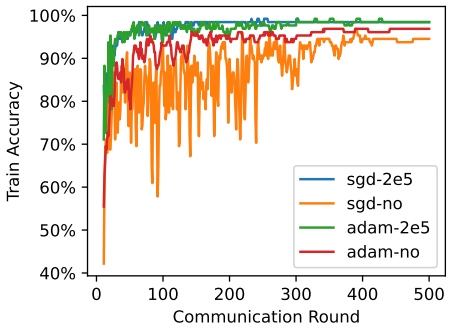

(a) Training Accuracy (b) Test Accuracy

Figure 7: Vision task: MNIST, $\epsilon_p = 2.75$

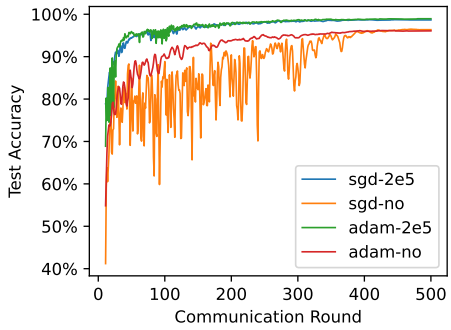



(a) Training Accuracy (b) Test Accuracy

Figure 8: Vision task: MNIST, $\epsilon_p = 1.60$

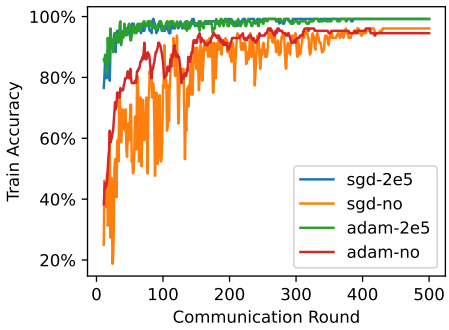



(a) Training Accuracy

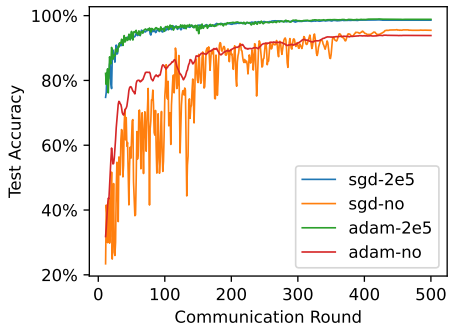



(b) Test Accuracy

Figure 9: Vision task: MNIST, $\epsilon_p = 0.42$

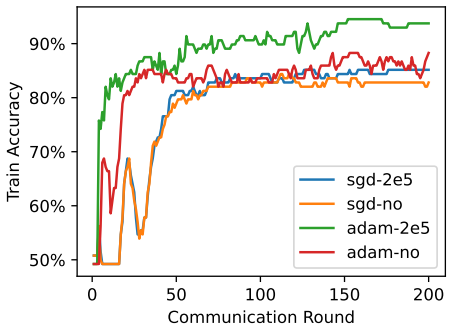



(a) Training Accuracy

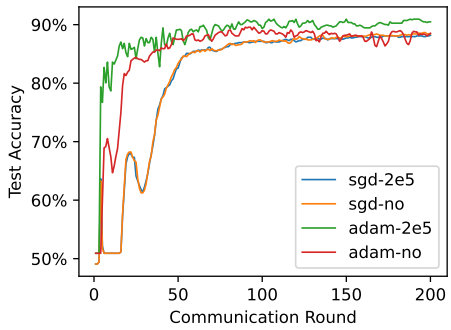



(b) Test Accuracy

Figure 10: Vision task: MNIST, $\epsilon_p = 0.18$

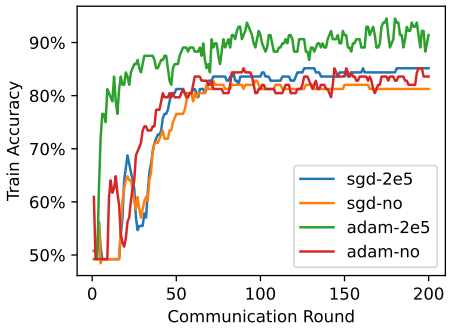



(a) Training Accuracy

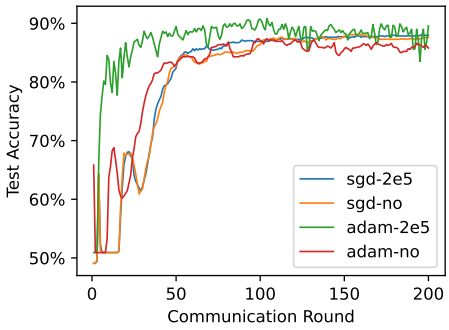



(b) Test Accuracy

Figure 11: Language task, $\epsilon_p = 2.45$

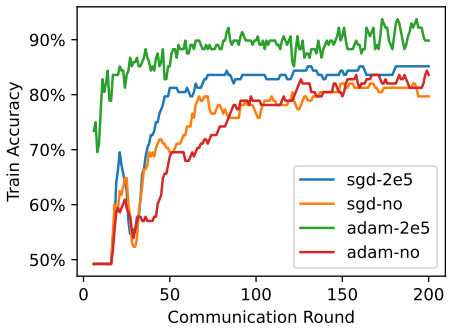



(a) Training Accuracy

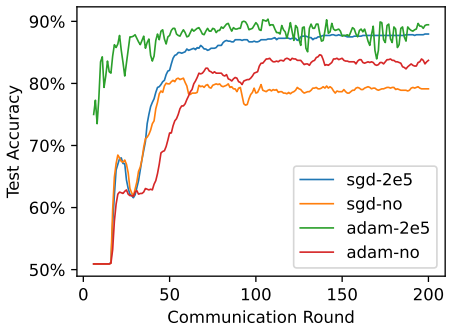



(b) Test Accuracy

Figure 12: Language task, $\epsilon_p = 0.35$




(a) Training Accuracy




(b) Test Accuracy

Figure 13: Language task, $\epsilon_p = 0.12$