

## A Limitations of the State-of-the-Art

### A.1 Limitation 1: Flawed Error Rate Analysis

**Gradient variance overlooked in saddle point escape.** The error rate bound for finding a DP-SOSP in [30] is fundamentally incorrect. Their analysis relies on Lemma 3.4 therein (adapted from [46, Lemma 12]), which claims that adding Gaussian noise at the same scale as the DP gradient estimation error suffices to reduce the function value with high probability, enabling escape from saddle points. This argument critically depends on proving that the region around a saddle point where SGD may get stuck is sufficiently narrow. Under this condition, perturbation along the escape direction ensures that the SGD sequence can escape with high probability.

However, the analysis neglects a key factor, which is the stochastic gradient variance. Their proof implicitly uses exact gradients of the population risk, which are unavailable to the algorithm. This is evidenced by the equation preceding equation (39) in [46]. Another indication of this oversight is their choice of step size  $\eta = 1/M$ . While valid for gradient descent with exact gradients, prior work [24] has shown that stochastic gradients require a smaller step size. The use of  $\eta = 1/M$  in [30] for population risk minimization reflects a failure to account for gradient stochasticity. This leads to an underestimated gradient complexity and an overestimated effective sample size per gradient estimate, which ultimately results in an overly optimistic error rate. A correct analysis must acknowledge that stochastic gradients increase estimation error, implying that the true error rate for finding a DP-SOSP is weaker than the one reported.

**Fixing the proof is insufficient, a new algorithm is necessary.** Although the analytical error can be identified, correcting the proof alone does not yield a satisfactory result. Any direct correction would only achieve a weaker  $(\alpha, \alpha^{2/5})$ -SOSP guarantee, rather than the desired  $\alpha$ -SOSP. In particular, the second-order accuracy would degrade to  $\tilde{O}(\alpha^{2/5})$  instead of the ideal  $\tilde{O}(\alpha^{1/2})$ .

This limitation arises because the algorithm in [30] can be viewed as a special case of perturbed gradient descent with bounded gradient inexactness as developed in [53], where the DP noise contributes to the perturbation. By invoking [53, Theorem 3], one only obtains an error rate bound with respect to a weaker class of SOSP where the second-order accuracy depends on  $\tilde{O}(\alpha^{2/5})$ .

The underlying reason is that both [53] and [31] rely on injecting additional noise to facilitate escape from saddle points, without considering the role of inherent DP Gaussian noise in the gradients. The excessive injected noise degrades the SOSP guarantee.

To fully resolve this issue, a new algorithmic design is required. In the setting of [53], where gradient perturbations stem from adversarial attacks, such degradation is unavoidable since the perturbations can hinder rather than assist escape. However, in the DP setting, the Gaussian noise is well-behaved and can naturally aid saddle point escape. By leveraging the inherent DP noise, it becomes possible to avoid the need for additional injected noise and to achieve  $\alpha$ -SOSP convergence as desired. Therefore, relying on the algorithmic designs of [53] or [31] is insufficient, and a new algorithm must be developed to achieve the desired guarantees.

### A.2 Limitation 2: Challenges of Private SOSP Selection

**Inapplicability of AboveThreshold in distributed learning.** The algorithm in [30] guarantees only the existence of an  $\alpha$ -SOSP among its iterates. To privately identify such a point, it applies the AboveThreshold mechanism to test whether candidate models satisfy the SOSP conditions by privately evaluating gradient norms and Hessian eigenvalues. While this procedure introduces negligible error in single-machine settings, it faces fundamental challenges in distributed learning.

According to [30, Lemma 4.5], for any  $x \in \mathbb{R}^d$  and a dataset  $S$  of size  $O(n)$ , with probability at least  $1 - \omega$ , the following holds:

$$\|\nabla F_{\mathcal{D}}(x) - \nabla \hat{f}_S(x)\| \leq O\left(\frac{G \log(d/\omega)}{\sqrt{n}}\right), \quad \|\nabla^2 F_{\mathcal{D}}(x) - \nabla^2 \hat{f}_S(x)\|_{\text{op}} \leq O\left(\frac{M \log(d/\omega)}{\sqrt{n}}\right).$$

This implies:

$$\|\nabla \hat{f}_S(x)\| \leq \|\nabla F_{\mathcal{D}}(x)\| + O\left(\frac{G \log \frac{d}{\omega}}{\sqrt{n}}\right), \quad \lambda_{\min}(\nabla^2 \hat{f}_S(x)) \geq \lambda_{\min}(\nabla^2 F_{\mathcal{D}}(x)) - O\left(\frac{M \log \frac{d}{\omega}}{\sqrt{n}}\right).$$

With these bounds, AboveThreshold can identify a DP-SOSP by setting appropriate thresholds. However, this procedure relies on centralized access to the dataset  $S$ .

In distributed learning, each client holds a local dataset  $S_i$ . To estimate global quantities, aggregation is required:

$$\|\nabla \hat{f}_S(x)\| \leq \frac{1}{m} \sum_{i=1}^m \|\nabla \hat{f}_{S_i}(x)\|, \quad \lambda_{\min}(\nabla^2 \hat{f}_S(x)) \geq \frac{1}{m} \sum_{i=1}^m \lambda_{\min}(\nabla^2 \hat{f}_{S_i}(x)).$$

Yet the learning algorithm guarantees only:

$$\|\nabla F_{\mathcal{D}}(x)\| \leq \frac{1}{m} \sum_{i=1}^m \|\nabla F_{\mathcal{D}_i}(x)\|, \quad \lambda_{\min}(\nabla^2 F_{\mathcal{D}}(x)) \geq \frac{1}{m} \sum_{i=1}^m \lambda_{\min}(\nabla^2 F_{\mathcal{D}_i}(x)),$$

This relationship does not provide an upper bound on  $\|\nabla \hat{f}_S(x)\|$  or a lower bound on  $\lambda_{\min}(\nabla^2 \hat{f}_S(x))$  solely from local empirical estimates. Therefore, it is infeasible to determine valid thresholds for AboveThreshold based only on local information. Any attempt to perform this selection would require clients to share their (noisy) gradients and Hessians with the server, which introduces substantial privacy, communication, and computation costs.

**Eliminating private model selection is essential in distributed learning.** A feasible method for private model selection in distributed learning would extend the centralized algorithm of [46, Algorithm 5]. Specifically, each client privately computes gradients and Hessians on additional local data beyond the training set, and the server aggregates these to estimate global quantities. However, this strategy has several drawbacks. It requires extra data outside the training process, increases communication overhead by transmitting high-dimensional gradients and Hessians, and incurs high computational costs. It also shifts the method from a first-order to a second-order algorithm.

Moreover, as shown in Section 6, sharing perturbed high-dimensional gradients and Hessians, rather than one-dimensional scalar queries as in AboveThreshold, introduces non-negligible additional error. This error accumulation degrades the accuracy guarantees provided by the learning algorithm. Unlike the single-machine case, private model selection in distributed learning incurs significant costs in accuracy, privacy, computation, and communication.

These challenges demonstrate the necessity of designing an algorithm that inherently outputs a DP-SOSP without relying on a private model selection procedure. Such a design avoids additional data consumption, computational burden, communication overhead, and deterioration of error guarantees.

## B Useful Facts for Analysis

### B.1 Probability Tools

**Definition 6** (Sub-Gaussian random vector [23, Definition 2]). A random vector  $v \in \mathbb{R}^d$  is  $\zeta$ -sub-Gaussian (or  $\text{SG}(\zeta)$ ), if there exists a positive constant  $\zeta$  such that

$$\mathbb{E}[\exp(\langle u, v - \mathbb{E}[v] \rangle)] \leq \exp\left(\frac{\|u\|_2^2 \zeta^2}{2}\right), \quad \forall u \in \mathbb{R}^d. \quad (10)$$

**Definition 7** (Norm-sub-Gaussian random vector [23, Definition 3]). A random vector  $v \in \mathbb{R}^d$  is  $\zeta$ -norm-sub-Gaussian (or  $\text{nSG}(\zeta)$ ), if there exists a positive constant  $\zeta$  such that

$$\mathbb{P}[\|v - \mathbb{E}[v]\| \geq t] \leq 2 \exp\left(-\frac{t^2}{2\zeta^2}\right), \quad \forall t \in \mathbb{R}. \quad (11)$$

Note that norm-sub-Gaussian random vectors (Definition 7) are more general than sub-Gaussian random vectors (Definition 6), as sub-Gaussian distributions require *isotropy*, whereas norm-sub-Gaussian distributions do not impose this condition.

**Lemma 10** ([23, Lemma 1]). A  $\text{SG}(r)$  random vector  $v \in \mathbb{R}^d$  is also  $\text{nSG}(2\sqrt{2} \cdot r\sqrt{d})$ .

We are interested in the properties of norm-subGaussian martingale difference sequences. Concretely, they are sequences satisfying the following properties.

**Condition 1.** Consider random vectors  $v_1, \dots, v_p \in \mathbb{R}^d$ , and corresponding filtrations  $\mathcal{F}_i = \sigma(v_1, \dots, v_i)$  for  $i \in [n]$ , such that  $v_i | \mathcal{F}_{i-1}$  is zero-mean nSG( $\zeta_i$ ) with  $\zeta_i \in \mathcal{F}_{i-1}$ . That is,

$$\mathbb{E}[v_i | \mathcal{F}_{i-1}] = 0, \quad \mathbb{P}[\|v_i\| \geq t | \mathcal{F}_{i-1}] \leq 2 \exp\left(-\frac{t^2}{2\zeta_i^2}\right), \quad \forall t \in \mathbb{R}, \forall i \in [p]. \quad (12)$$

**Lemma 11** (Hoeffding type inequality for norm-sub-Gaussian [23, Corollary 7]). Let random vectors  $v_1, \dots, v_p \in \mathbb{R}^d$ , and corresponding filtrations  $\mathcal{F}_i = \sigma(v_1, \dots, v_i)$  for  $i \in [k]$  satisfy condition 1 with fixed  $\{\zeta_i\}$ . Then for any  $\iota > 0$ , there exists an absolute constant  $C$  such that, with probability at least  $1 - 2d \cdot e^{-\iota}$ ,

$$\left\| \sum_{i=1}^p v_i \right\|_2 \leq C \cdot \sqrt{\sum_{i=1}^p \zeta_i^2} \cdot \iota. \quad (13)$$

Lemma 11 implies that the sum of norm-sub-Gaussian random vectors is still norm-sub-Gaussian.

**Corollary 3.** Let random vectors  $v_1, \dots, v_p \in \mathbb{R}^d$ , and corresponding filtrations  $\mathcal{F}_i = \sigma(v_1, \dots, v_i)$  for  $i \in [k]$  satisfy condition 1 with fixed  $\{\zeta_i\}$ . Then  $\sum_{i=1}^p v_i$  is nSG  $\left(C \cdot \sqrt{\log(d) \sum_{i=1}^k \zeta_i^2}\right)$ .

*Proof.* Let  $\zeta_+ := \sqrt{C \log(d) \sum_{i=1}^k \zeta_i}$ . According to Definition 7, we aim to show that, for any  $\omega \in (0, 1)$ , with probability at least  $1 - \omega$ ,  $\|\sum_{i=1}^p v_i\| \leq \sqrt{2\zeta_+^2 \ln \frac{2}{\omega}}$ . By Lemma 11, we have known that, with probability at least  $1 - \omega$ ,  $\|\sum_{i=1}^p v_i\| \leq C \cdot \sqrt{\sum_{i=1}^p \zeta_i^2 \ln \frac{2d}{\omega}}$ . Next, we show that  $\sqrt{2\zeta_+^2 \ln \frac{2}{\omega}} \geq C \cdot \sqrt{\sum_{i=1}^p \zeta_i^2 \ln \frac{2d}{\omega}}$ , which, by re-arranging the terms, is equivalent to show  $\zeta_+^2 \geq \frac{C^2}{2} (\sum_{i=1}^p \zeta_i^2) \frac{\log \frac{2d}{\omega}}{\log \frac{2}{\omega}}$ . This follows directly from the fact that  $\frac{\log \frac{2d}{\omega}}{\log \frac{2}{\omega}} \leq 2 \log d, \forall \omega \in (0, 1)$ .  $\square$

**Lemma 12** ([24, Lemma C.6]). Let random vectors  $v_1, \dots, v_p \in \mathbb{R}^d$ , and corresponding filtrations  $\mathcal{F}_i = \sigma(v_1, \dots, v_i)$  for  $i \in [k]$  satisfy condition 1, then for any  $\iota > 0$ , and  $B > b > 0$ , there exists an absolute constant  $C$  such that, with probability at least  $1 - 2d \log\left(\frac{B}{b}\right) \cdot e^{-\iota}$ ,

$$\sum_{i=1}^p \zeta_i^2 \geq B \quad \text{or} \quad \left\| \sum_{i=1}^p v_i \right\| \leq C \cdot \sqrt{\max\left\{\sum_{i=1}^p \zeta_i^2, b\right\}} \cdot \iota. \quad (14)$$

**Lemma 13** ([24, Lemma C.7]). Let random vectors  $v_1, \dots, v_p \in \mathbb{R}^d$ , and corresponding filtrations  $\mathcal{F}_i = \sigma(v_1, \dots, v_i)$  for  $i \in [k]$  satisfy condition 1 with fixed  $\zeta_1 = \zeta_2 = \dots = \zeta_p = \zeta$ , then there exists an absolute constant  $C$  such that, for any  $\iota > 0$ , with probability at least  $1 - e^{-\iota}$ ,

$$\sum_{i=1}^p \|v_i\|^2 \leq C \cdot \zeta^2 \cdot (p + \iota). \quad (15)$$

**Lemma 14** (Matrix Bernstein inequality [44, Theorem 1.4]). Consider a finite sequence  $\{\mathbf{M}_i\}_{i \in [k]}$  of independent, random, self-adjoint matrices with dimension  $d \times d$ . Assume that each random matrix satisfies  $\mathbb{E}[\mathbf{M}_i] = \mathbf{0}$ ,  $\|\mathbf{M}_i\|_2 \leq B$ , then for all  $t \geq 0$ , we have

$$\mathbb{P}\left[\left\| \sum_{i \in [k]} \mathbf{M}_i \right\|_2 \geq t\right] \leq d \exp\left(-\frac{t^2}{2(\sigma^2 + Bt/3)}\right), \quad (16)$$

where  $\sigma^2 = \left\| \sum_{i \in [k]} \mathbb{E}[\mathbf{M}_i^2] \right\|_2$ .

**Lemma 15** (Norm of symmetric matrices with sub-gaussian entries [45, Corollary 4.4.8]). Let  $\mathbf{M}$  be an  $d \times d$  symmetric random matrix whose entries  $\mathbf{M}_{i,j}$  on and above the diagonal are independent, mean zero, sub-gaussian random variables. Then, with probability at least  $1 - 4 \exp(-t^2)$ , for any  $t > 0$  we have

$$\|\mathbf{M}\|_2 \leq C \cdot \max_{i,j} \|\mathbf{M}_{i,j}\|_{\psi_2} \cdot (\sqrt{d} + t), \quad (17)$$

where  $C$  is a universal constant.

## B.2 Privacy Preliminaries

**Definition 8** (Gaussian Mechanism [12]). Given any input data  $D \in \mathcal{X}^n$  and a query function  $q : \mathcal{X}^n \rightarrow \mathbb{R}^d$ , the Gaussian mechanism  $\mathcal{M}_G$  is defined as  $q(D) + \nu$  where  $\nu \sim \mathcal{N}(0, \sigma_G^2 \mathbf{I}_d)$ . Let  $\Delta_2(q)$  be the  $\ell_2$ -sensitivity of  $q$ , i.e.,  $\Delta_2(q) := \sup_{D \sim D'} \|q(D) - q(D')\|_2$ . For any  $\sigma, \delta > 0$ ,  $\mathcal{M}_G$  guarantees  $(\frac{\Delta_2(q)}{\sigma_G} \sqrt{2 \log \frac{1.25}{\delta}}, \delta)$ -DP. That is, if we want the output of  $q$  to be  $(\epsilon, \delta)$ -DP for any  $0 < \epsilon, \delta < 1$ , then  $\sigma_G$  should be set to  $\frac{\Delta_2(q)}{\epsilon} \sqrt{2 \log \frac{1.25}{\delta}}$ .

**Lemma 16** (Adaptive Composition Theorem [12]). Given target privacy parameters  $0 < \epsilon < 1$  and  $0 < \delta < 1$ , to ensure  $(\epsilon, \delta)$ -DP over  $k$ -fold adaptive mechanisms, it suffices that each mechanism is  $(\epsilon', \delta')$ -DP, where  $\epsilon' = \frac{\epsilon}{2\sqrt{2k \ln(2/\delta)}}$  and  $\delta' = \frac{\delta}{2k}$ .

**Lemma 17** (Parallel Composition of DP [35]). Suppose there are  $n$   $(\epsilon, \delta)$ -differentially private mechanisms  $\{\mathcal{M}_i\}_{i=1}^n$  and  $n$  disjoint datasets denoted by  $\{D_i\}_{i=1}^n$ . Then the algorithm, which applies each  $\mathcal{M}_i$  on the corresponding  $D_i$ , preserves  $(\epsilon, \delta)$ -DP in total.

## C Omitted Proofs in Section 4

### C.1 Proof of Lemma 1

*Proof of Lemma 1.* We begin by introducing the following notations:

$$\hat{x}_t := x_t - x'_t, \quad (18)$$

$$\hat{\zeta}_t := \zeta_t - \zeta'_t, \quad (19)$$

$$\hat{\xi}_t := \xi_t - \xi'_t, \quad (20)$$

$$\Delta_t := \int_0^1 \nabla^2 F(y \cdot x_t + (1-y) \cdot x'_t) dy - \mathcal{H} \quad (21)$$

The proof strategy is to derive a contradiction by showing that if the model remains localized (i.e., stays within a radius  $\mathcal{R}$  around the saddle point) with high probability, then the coupling sequence must still diverge with non-negligible probability.

We first characterize the dynamics of  $\hat{x}_t$  in the following Lemma 18. At a high level, we decompose the difference of the coupling sequence  $x_t$  into three components: **(i)** a curvature-dependent term  $\mathcal{P}_h(t)$ , **(ii)** a stochastic gradient noise term  $\mathcal{P}_{sg}(t)$ , **(iii)** a perturbation-driven term  $\mathcal{P}_p(t)$ .

**Lemma 18** (Coupling Dynamics). For any  $t \geq 0$ , the difference between the two coupled iterates satisfies:

$$\hat{x}_t = -\underbrace{\eta \sum_{i=1}^t (\mathbf{I}_d - \eta \mathcal{H})^{t-i} \Delta_{i-1} \hat{x}_{i-1}}_{\mathcal{P}_h(t)} - \underbrace{\eta \sum_{i=1}^t (\mathbf{I}_d - \eta \mathcal{H})^{t-i} \hat{\zeta}_i}_{\mathcal{P}_{sg}(t)} - \underbrace{\eta \sum_{i=1}^t (\mathbf{I}_d - \eta \mathcal{H})^{t-i} \hat{\xi}_i}_{\mathcal{P}_p(t)}. \quad (22)$$

*Proof of Lemma 18.* By the update rule:

$$\hat{x}_t = x_t - x'_t \quad (23)$$

$$= \hat{x}_{t-1} - \eta [\nabla F(x_{t-1}) - \nabla F(x'_{t-1}) + \zeta_t - \zeta'_t + \xi_t - \xi'_t] \quad (24)$$

$$= \hat{x}_{t-1} - \eta [(\mathcal{H} + \Delta_{t-1}) \hat{x}_{t-1} + \hat{\zeta}_t + \hat{\xi}_t] \quad (25)$$

$$= (\mathbf{I}_d - \eta \mathcal{H}) \hat{x}_{t-1} - \eta [\Delta_{t-1} \hat{x}_{t-1} + \hat{\zeta}_t + \hat{\xi}_t]. \quad (26)$$

Unrolling the recursion with initial condition  $\hat{x}_0 = 0$  yields the desired result:

$$\hat{x}_t = (\mathbf{I}_d - \eta\mathcal{H})^t \hat{x}_0 - \eta \sum_{i=1}^t (\mathbf{I}_d - \eta\mathcal{H})^{t-i} (\Delta_{i-1} \hat{x}_{i-1} + \hat{\zeta}_i + \hat{\xi}_i) \quad (27)$$

$$= -\eta \sum_{i=1}^t (\mathbf{I}_d - \eta\mathcal{H})^{t-i} (\Delta_{i-1} \hat{x}_{i-1} + \hat{\zeta}_i + \hat{\xi}_i). \quad (28)$$

□

Let  $\mathcal{E}$  denote the event that both sequences remain localized:

$$\mathcal{E} := \{\forall t \leq \Gamma : \max\{\|x_t - \tilde{x}\|, \|x'_t - \tilde{x}'\|\} \leq \mathcal{R}\}.$$

We proceed by contradiction. Assume:

$$\mathbb{P}(\mathcal{E}) \geq \frac{3}{4}. \quad (29)$$

To derive a contradiction, we analyze the terms in (22), showing in Lemma 19 and Lemma 20 that the perturbation term  $\mathcal{P}_p(t)$  dominates, while the curvature and stochastic gradient terms remain controlled. Define:

$$\mathbf{a}(t) := \sqrt{\sum_{i=1}^t (1 + \eta\gamma)^{2(t-i)}}, \quad \mathbf{b}(t) := \frac{(1 + \eta\gamma)^t}{\sqrt{2\eta\gamma}}. \quad (30)$$

It has been verified in [24, Lemma 29] that  $\mathbf{a}(t) \leq \mathbf{b}(t)$  for all  $t \in \mathbb{N}$ .

**Lemma 19.** For all  $t \geq 0$ , the following hold:

$$\mathbb{P}[\|\mathcal{P}_p(t)\| \leq c\mathbf{b}(t)\eta r \cdot \sqrt{t}] \geq 1 - 2e^{-t} \quad (31)$$

$$\mathbb{P}\left[\|\mathcal{P}_p(t)\| \geq \frac{\mathbf{b}(\Gamma)\eta r}{10}\right] \geq \frac{2}{3} \quad (32)$$

The proof follows from standard Gaussian concentration and is omitted here; see [24, Lemma 30].

**Lemma 20.** For all  $t \geq 0$ , conditioned on  $\mathcal{E}$ , we have:

$$\mathbb{P}\left[\|\mathcal{P}_h(t) + \mathcal{P}_{sg}(t)\| \leq \frac{\mathbf{b}(t)\eta r}{20} \middle| \mathcal{E}\right] \geq 1 - 6d\Gamma \log\left(\frac{\mathcal{R}}{\eta r}\right) e^{-t} \quad (33)$$

*Proof of Lemma 20.* We prove the following strengthened claim for any  $t \leq \Gamma$  by induction:

$$\mathbb{P}\left[\forall i \leq t : \|\mathcal{P}_h(i) + \mathcal{P}_{sg}(i)\| \leq \frac{\mathbf{b}(i)\eta r}{20}, \|\mathcal{P}_p(i)\| \leq c\mathbf{b}(i)\eta r \sqrt{i} \middle| \mathcal{E}\right] \leq 1 - 6dt \log\left(\frac{\mathcal{R}}{\eta r}\right) e^{-t}. \quad (34)$$

For the base case of  $t = 0$ , the claim holds trivially as  $\mathcal{P}_h(0) = \mathcal{P}_{sg}(0) = 0$ . Suppose the claim holds for a step  $t < \Gamma$ , we then forward prove that the claim also holds for step  $t + 1 \leq \Gamma$ . Since for  $\forall i \leq t$ ,  $\|\mathcal{P}_p(i)\| \leq c\mathbf{b}(i)\eta r \sqrt{i}$ , we have

$$\|\hat{x}_i\| \leq \|\mathcal{P}_h(i) + \mathcal{P}_{sg}(i)\| + \|\mathcal{P}_p(i)\| \quad (35)$$

$$\leq \frac{\mathbf{b}(i)\eta r}{20} + c\mathbf{b}(i)\eta r \cdot \sqrt{i} \quad (36)$$

$$\leq 2c\mathbf{b}(i)\eta r \cdot \sqrt{i}. \quad (37)$$

Moreover, due to assumption (29) and the Hessian Lipschitz property, we have

$$\|\Delta_i\| = \int_0^1 \nabla^2 F(y \cdot x_i + (1-y) \cdot x'_i) dy \quad (38)$$

$$\leq \rho \max\{\|x_i - \tilde{x}\|, \|x'_i - \tilde{x}'\|\} \leq \rho \mathcal{R}. \quad (39)$$

With the above upper bounds on  $\|\hat{x}_i\|$  and  $\|\Delta_i\|$  for  $i \leq t$ , we immediately get for case  $t + 1$  from the definition of  $\mathcal{P}_h(\cdot)$  in (22) that

$$\|\mathcal{P}_h(t+1)\| \leq \eta\rho\mathcal{R} \sum_{i=1}^{t+1} (1+\eta\gamma)^{t+1-i} (2c\mathbf{b}(i)\eta r\sqrt{\iota}) \quad (40)$$

$$\leq 2\eta\rho\mathcal{R}\Gamma c\mathbf{b}(t+1)\eta r\sqrt{\iota} \leq \frac{\mathbf{b}(t+1)\eta r}{40}, \quad (41)$$

where the last inequality follows from  $2c\eta\rho\mathcal{R}\Gamma = \frac{2c}{s} \leq \frac{1}{40}$  for large enough  $s$  such that  $s \geq 80c$ .

Note that  $\hat{\zeta}_t|\mathcal{F}_{t-1} \sim \text{nSG}(M\|\hat{x}_t\|)$ , by applying Lemma 12 with  $B = [\mathbf{a}(t)]^2\eta^2M^2\mathcal{R}^2$  and  $b = [\mathbf{a}(t)]^2\eta^2M^2\eta^2r^2$  therein, we know that, with probability at least  $1 - 4d \log\left(\frac{\mathcal{R}}{\eta r}\right) e^{-\iota}$ , we have

$$\|\mathcal{P}_{sg}(t+1)\| \leq 2c\eta M\sqrt{\Gamma}\mathbf{b}(t)\eta r\sqrt{\iota}. \quad (42)$$

For large enough  $s$  such that  $s \geq (80c)^2$ , we have  $c\eta M\sqrt{\Gamma}\iota \leq \frac{2c}{\sqrt{s}} \leq \frac{1}{40}$ . Thus,

$$\|\mathcal{P}_{sg}(t+1)\| \leq c\eta M\sqrt{\Gamma}\mathbf{b}(t)\eta r\sqrt{\iota} \leq \frac{\mathbf{b}(t)\eta r}{40}. \quad (43)$$

By Lemma 19, we know that, for case  $t + 1$ , with probability at least  $1 - 2e^{-\iota}$ , we have

$$\|\mathcal{P}_p(t+1)\| \leq c\mathbf{b}(t+1)\eta r\sqrt{\iota} \quad (44)$$

By the union bound, with probability at least  $1 - \left(6dt \log\left(\frac{\mathcal{R}}{\eta r}\right) e^{-\iota} + 4d \log\left(\frac{\mathcal{R}}{\eta r}\right) e^{-\iota} + 2e^{-\iota}\right) \geq 1 - 6d(t+1) \log\left(\frac{\mathcal{R}}{\eta r}\right) e^{-\iota}$ ,

$$\|\mathcal{P}_h(t+1) + \mathcal{P}_{sg}(t+1)\| \leq \frac{\mathbf{b}(t)\eta r}{20} \leq \frac{\mathbf{b}(t+1)\eta r}{20}, \quad \|\mathcal{P}_p(t+1)\| \leq c\mathbf{b}(t+1)\eta r\sqrt{\iota}, \quad (45)$$

which concludes the proof.  $\square$

Now we complete the proof of Lemma 1. Choose  $\iota$  large enough such that

$$\iota \geq \log\left(36d\Gamma \log\left(\frac{\mathcal{R}}{\eta r}\right)\right), \quad (46)$$

which is promised by  $\mu \geq \frac{1}{s} \log\left(\frac{9d}{C^{\frac{1}{4}}\eta\sqrt{s\rho\psi}} \log\left(\frac{4C^{\frac{1}{4}}}{s\eta r} \sqrt{\frac{\psi}{\rho}}\right)\right)$  for sufficiently large numerical constant  $s$ . Then we have:

$$6d\Gamma \log\left(\frac{\mathcal{R}}{\eta r}\right) e^{-\iota} \leq \frac{2}{9}. \quad (47)$$

From Lemma 19, we have:

$$\mathbb{P}\left[\|\mathcal{P}_p(\Gamma)\| \geq \frac{\mathbf{b}(\Gamma)\eta r}{10}\right] \geq \frac{2}{3}, \quad (48)$$

and from Lemma 20,

$$\mathbb{P}\left[\|\mathcal{P}_h(\Gamma) + \mathcal{P}_{sg}(\Gamma)\| \leq \frac{\mathbf{b}(\Gamma)\eta r}{20}\right] \geq \frac{3}{4} \cdot \left(1 - 6d\Gamma \log\left(\frac{\mathcal{R}}{\eta r}\right) e^{-\iota}\right) \geq \frac{7}{12} \quad (49)$$

By the union bound, with probability at least  $1 - \left(1 - \frac{2}{3}\right) - \left(1 - \frac{7}{12}\right) = \frac{1}{4}$ , both events hold:

$$\|\mathcal{P}_p(\Gamma)\| \geq \frac{\mathbf{b}(\Gamma)\eta r}{10}, \quad \|\mathcal{P}_h(\Gamma) + \mathcal{P}_{sg}(\Gamma)\| \leq \frac{\mathbf{b}(\Gamma)\eta r}{20}. \quad (50)$$

Therefore, using the triangle inequality:

$$\max\{\|x_\Gamma - \tilde{x}\|, \|x'_\Gamma - \tilde{x}\|\} \quad (51)$$

$$\geq \frac{1}{2}\|\hat{x}_\Gamma\| \geq \frac{1}{2}[\|\mathcal{P}_p(\Gamma)\| - \|\mathcal{P}_h(\Gamma) + \mathcal{P}_{sg}(\Gamma)\|] \geq \frac{\mathbf{b}(\Gamma)\eta r}{40} = \frac{(1+\eta\gamma)^\Gamma\sqrt{\eta r}}{40\sqrt{2}} \quad (52)$$

$$\geq \frac{(1+\eta\sqrt{\rho\alpha})^\Gamma\sqrt{\eta r}}{40\sqrt{2}} \geq \frac{2^{\eta\sqrt{\rho\alpha}\Gamma}\sqrt{\eta r}}{40\sqrt{2}} = \frac{2^{\frac{\iota}{s}}\sqrt{\eta r}}{40\sqrt{2}} = \frac{2^\mu\sqrt{\eta r}}{40\sqrt{2}} > \mathcal{R}, \quad (53)$$

where the second last inequality is due to the fact  $1 + a > 2^a, \forall a \in (0, 1]$  and  $\eta\sqrt{\rho\alpha} \leq \frac{1}{\iota^2} \leq 1$ , and the last inequality is because  $\mu > \log \left( \frac{160\sqrt{2}C^{\frac{1}{4}}}{s\sqrt{\eta r}} \sqrt{\frac{\psi}{\rho}} \right)$ .

The above means that the localization event  $\mathcal{E}$  fails with probability at least  $1/4$ , i.e.,  $\mathbb{P}(\mathcal{E}) < \frac{3}{4}$ , which contradicts with our assumption (29). Therefore, the assumption (29) should be false, that is, with probability at least  $\frac{1}{4}$ ,  $\exists t \leq \Gamma, \max\{\|x_t - \tilde{x}\|, \|x'_t - \tilde{x}\|\} \geq \mathcal{R}$ , completing the proof.  $\square$

## C.2 Proof of Lemma 2

*Proof of Lemma 2.* The failure probability after  $Q$  independent repetitions is at most  $(7/8)^Q$ . Setting  $Q = \frac{26}{5} \log(1/\omega_0)$  yields  $(7/8)^Q \leq \omega_0$ , completing the proof.  $\square$

## C.3 Proof of Lemma 3

*Proof of Lemma 3.* For any  $t \geq 1$ , by  $M$ -smoothness of  $F$ , we have:

$$F(x_t) - F(x_{t-1}) \leq \langle \nabla F(x_{t-1}), x_t - x_{t-1} \rangle + \frac{M}{2} \|x_t - x_{t-1}\|^2 \quad (54)$$

$$\leq -\eta \langle \nabla F(x_{t-1}), \hat{g}_{t-1} \rangle + \frac{M}{2} \eta^2 \|\hat{g}_{t-1}\|^2 \quad (55)$$

$$\leq -\eta \langle \nabla F(x_{t-1}), \hat{g}_{t-1} \rangle + \frac{\eta}{2} \|\hat{g}_{t-1}\|^2 \quad (56)$$

$$\leq \frac{\eta}{2} \|\nu_t\|^2 - \frac{\eta}{2} \|\nabla F(x_{t-1})\|^2 - \frac{\eta}{2} \|\hat{g}_{t-1}\|^2 + \frac{\eta}{2} \|\hat{g}_{t-1}\|^2 \quad (57)$$

$$= -\frac{\eta}{2} \|\nabla F(x_{t-1})\|^2 + \frac{\eta}{2} \|\nu_t\|^2. \quad (58)$$

Summing from  $t_0 + 1$  to  $t_0 + t$ , we obtain:

$$F(x_{t_0+t}) - F(x_{t_0}) \leq -\frac{\eta}{2} \sum_{i=0}^{t-1} \|\nabla F(x_{t_0+i})\|^2 + \frac{\eta}{2} \sum_{i=1}^t \|\nu_{t_0+i}\|^2 \quad (59)$$

$\square$

## C.4 Proof of Corollary 2

*Proof of Corollary 2.* Note that

$$\frac{\eta}{2} \sum_{i=1}^t \|\nu_{t_0+i}\|^2 = \frac{\eta}{2} \sum_{i=1}^t \|\zeta_{t_0+i} + \xi_{t_0+i}\|^2 \leq \eta \sum_{i=1}^t (\|\zeta_{t_0+i}\|^2 + \|\xi_{t_0+i}\|^2) \quad (60)$$

By Lemma 13, since  $\zeta_i \sim \text{nSG}(\sigma)$ , with probability at least  $1 - e^{-\iota}$ :

$$\sum_{i=1}^t \|\zeta_{t_0+i}\|^2 \leq C \cdot \sigma^2(t + \iota). \quad (61)$$

Using Lemma 10, each  $\xi_i \sim \text{nSG}(2\sqrt{2}r\sqrt{d})$ , and applying Lemma 13 again, with probability at least  $1 - e^{-\iota}$ :

$$\sum_{i=1}^t \|\xi_{t_0+i}\|^2 \leq 8C \cdot r^2 d(t + \iota). \quad (62)$$

By the union bound, both bounds hold with probability at least  $1 - 2e^{-\iota}$ .  $\square$

### C.5 Proof of Lemma 4

*Proof of Lemma 4.* We begin with:

$$\|x_{t_0+\tau} - x_{t_0}\|^2 = \eta^2 \left\| \sum_{t=1}^{\tau} \nabla F(x_{t_0+t-1}) + \nu_{t_0+t} \right\|^2 \quad (63)$$

$$\leq 2\eta^2 \tau \sum_{t=1}^{\tau} (\|\nabla F(x_{t_0+t-1})\|^2 + \|\nu_{t_0+t}\|^2). \quad (64)$$

Following the same argument in the proof of corollary 2, with probability at least  $1 - 2e^{-\iota}$ ,

$$\sum_{t=1}^{\tau} \|\nu_{t_0+t}\|^2 \leq c \cdot \psi^2(\tau + \iota), \quad (65)$$

From corollary 2, with the same probability of  $1 - 2e^{-\iota}$ ,

$$\sum_{t=1}^{\tau} \|\nabla F(x_{t_0+t-1})\|^2 \leq \frac{2}{\eta} [F(x_{t_0}) - F(x_{t_0+\tau})] + c \cdot \psi^2(\tau + \iota). \quad (66)$$

Combining above results, we have, with probability at least  $1 - 2e^{-\iota}$ ,

$$\|x_{t_0+\tau} - x_{t_0}\|^2 \leq 4\eta\tau [F(x_{t_0}) - F(x_{t_0+\tau})] + 4c \cdot \eta^2 \tau \psi^2(\tau + \iota). \quad (67)$$

Re-arranging the terms above, we obtain

$$F(x_{t_0+\tau}) - F(x_{t_0}) \leq -\frac{1}{4\eta\tau} \|x_{t_0+\tau} - x_{t_0}\|^2 + c \cdot \eta\psi^2(\tau + \iota). \quad (68)$$

According to the criterion for successful escape, we have  $\|x_{t_0+\tau} - x_{t_0}\| \geq \mathcal{R}$ . Then

$$F(x_{t_0+\tau}) - F(x_{t_0}) \leq -\frac{1}{4\eta\tau} \|x_{t_0+\tau} - x_{t_0}\|^2 + c \cdot \eta\psi^2(\tau + \iota) \quad (69)$$

$$\leq -\frac{\mathcal{R}^2}{4\eta\Gamma} + c \cdot \eta\psi^2(\Gamma + \iota) \quad (70)$$

$$\leq -\frac{s}{4\iota^3} \sqrt{\frac{\alpha^3}{\rho}} + \frac{2c \cdot \psi^2 \iota}{s\sqrt{\rho\alpha}} \quad (71)$$

$$\leq -\frac{s}{8\iota^3} \sqrt{\frac{\alpha^3}{\rho}} = \Phi, \quad (72)$$

where the second to last inequality is from the fact that  $s\eta\sqrt{\rho\alpha} = \frac{\rho\alpha}{M^2 s \mu^2} < 1$ , and the last inequality follows from  $\alpha \geq 4\sqrt{C} s \mu^2 \psi$ .  $\square$

### C.6 Proof of Lemma 5

*Proof of Lemma 5.* By Corollary 3, for all  $t$ ,  $\nu_t \sim \text{nSG}(C\sqrt{\sigma^2 + r^2 d})$ . Since  $\mathbb{E}[\nu_t] = 0$ , by Definition 7, with probability at least  $1 - \frac{\omega}{2T}$ :

$$\|\nu_t\| \leq \sqrt{2C}\psi \sqrt{\log \frac{4T}{\omega}} \leq \chi. \quad (73)$$

Applying a union bound over  $t \in [T]$  gives the desired result: with probability at least  $1 - \omega/2$ ,  $\|\hat{g}_t - \nabla F(x_{t-1})\| \leq \chi$  for all  $t$ .  $\square$

### C.7 Proof of Lemma 6

*Proof of Lemma 6.* By Lemma 5, with probability at least  $1 - \omega/2$ , the gradient estimation error satisfies  $\|\hat{g}_t - \nabla F(x_{t-1})\| \leq \chi$  for all  $t \in [T]$ . We analyze two cases based on whether the algorithm is in the escape phase.

**Case 1: In escape phase.** When  $\|\hat{g}_t\| \leq 3\chi$ , the escape process is triggered, implying  $\|\nabla F(x_{t-1})\| \leq \alpha = 4\chi$ . The average function decrease per step during a successful escape is at least:

$$\frac{\Phi}{\Gamma} = \frac{s^2\alpha^2\eta}{8\iota^4} = \frac{2\chi^2\eta}{s^2\mu^4}. \quad (74)$$

**Case 2: Outside escape phase.** When  $\|\hat{g}_t\| > 3\chi$ , we have  $\|\nabla F(x_{t-1})\| \geq 2\chi$ . Each PSGD step yields at least:

$$\frac{\eta}{2}(2\chi)^2 = 2\chi^2\eta > \frac{2\chi^2\eta}{s^2\mu^4}. \quad (75)$$

Thus, in either case, the function value decreases by at least  $2\chi^2\eta/(s^2\mu^4)$  per step. Denoting  $U := F_0 - F^*$ , the number of effective descent steps is bounded by:

$$T_{\text{effective}} := \frac{Us^2\mu^4}{2\chi^2\eta}. \quad (76)$$

Next, consider the number of  $\alpha$ -strict saddle points encountered. Each successful escape yields function decrease of at least  $\Phi$ , so the total number of such escape phases is at most:

$$N_{\text{saddle}} := \frac{U}{\Phi} = \frac{8\iota^3U}{s} \sqrt{\frac{\rho}{\chi^3}}. \quad (77)$$

By Corollary 1, each  $\Gamma$ -descent succeeds with probability at least  $1/8$ , and we boost this to  $1 - \omega/2$  via the  $Q$  independent repetitions in every escape procedure. By Lemma 2 with failure probability  $\omega_0 = \frac{\omega}{2N_{\text{saddle}}}$ , we require:

$$Q = \frac{26}{5} \log \left( \frac{16\iota^3U}{s\omega} \sqrt{\frac{\rho}{\chi^3}} \right). \quad (78)$$

Hence, the total number of PSGD steps (including all  $\Gamma$ -descent repetitions) is at most:

$$T \leq T_{\text{effective}} \cdot Q = \frac{13Us^2\mu^4}{5\chi^2\eta} \log \left( \frac{16\iota^3U}{s\omega} \sqrt{\frac{\rho}{\chi^3}} \right) = \tilde{O} \left( \frac{U}{\eta\chi^2} \right). \quad (79)$$

□

## D Omitted Proofs in Section 5

### D.1 Proof of Lemma 7

*Proof of Lemma 7.* Let  $\tau(t)$  denote the most recent iteration (up to  $t$ ) at which oracle  $\mathcal{O}_1$  was used.

**Case 1:** If  $t = \tau(t)$ , then

$$\hat{g}_t = \mathcal{O}_1(x_{t-1}, \mathcal{B}_t) + \xi_t. \quad (80)$$

Let  $\zeta_t := \mathcal{O}_1(x_{t-1}, \mathcal{B}_t) - \nabla F(x_{t-1})$ , which is a zero-mean estimator with norm-subGaussian noise due to the  $G$ -Lipschitz condition:

$$\zeta_t \sim \text{nSG} \left( \frac{G\sqrt{\log d}}{\sqrt{b_1}} \right). \quad (81)$$

The noise term  $\xi_t$  is drawn from a Gaussian distribution:

$$\xi_t \sim \mathcal{N} \left( 0, c_1 \frac{G^2 \log(1/\delta)}{b_1^2 \epsilon^2} \mathbf{I}_d \right). \quad (82)$$

Thus, in this case, the oracle satisfies condition (2) with the desired bounds.

**Case 2:** If  $t > \tau(t)$ , then

$$\hat{g}_t = \mathcal{O}_1(x_{\tau(t)-1}, \mathcal{B}_{\tau(t)}) + \xi_{\tau(t)} + \sum_{i=\tau(t)+1}^t (\mathcal{O}_2(x_{i-1}, x_{i-2}, \mathcal{B}_i) + \xi_i). \quad (83)$$

Let  $\zeta_{\tau(t)} := \mathcal{O}_1(x_{\tau(t)-1}, \mathcal{B}_{\tau(t)}) - \nabla F(x_{\tau(t)-1})$  and define

$$\zeta'_i := \mathcal{O}_2(x_{i-1}, x_{i-2}, \mathcal{B}_i) - (\nabla F(x_{i-1}) - \nabla F(x_{i-2})). \quad (84)$$

Then

$$\hat{g}_t - \nabla F(x_{t-1}) = \zeta_{\tau(t)} + \sum_{i=\tau(t)+1}^t \zeta'_i + \xi_{\tau(t)} + \sum_{i=\tau(t)+1}^t \xi_i. \quad (85)$$

By the  $M$ -smoothness assumption, we have

$$\zeta'_i \sim \text{nSG} \left( \frac{M \|x_{i-1} - x_{i-2}\| \sqrt{\log d}}{\sqrt{b_2}} \right), \quad (86)$$

and the privacy noise is drawn from

$$\xi_i \sim \mathcal{N} \left( 0, c_2 \frac{M^2 \log(1/\delta)}{b_2^2 \epsilon^2} \|x_{i-1} - x_{i-2}\|^2 \mathbf{I}_d \right). \quad (87)$$

Since the algorithm ensures  $\text{drift}_t := \sum_{i=\tau(t)+1}^t \|x_{i-1} - x_{i-2}\|^2 \leq \kappa$ , we can bound the noise as follows:

– From Corollary 3, the total norm-subGaussian parameter becomes:

$$\sigma \leq O \left( \sqrt{\left[ \left( \frac{G \sqrt{\log d}}{\sqrt{b_1}} \right)^2 + \sum_{i=\tau(t)+1}^t \left( \frac{M \|x_{i-1} - x_{i-2}\| \sqrt{\log d}}{\sqrt{b_2}} \right)^2 \right] \cdot \log d} \right) \quad (88)$$

$$\leq O \left( \sqrt{\frac{G^2 \log^2 d}{b_1} + \frac{M^2 \log^2 d}{b_2} \kappa} \right). \quad (89)$$

– By the property of Gaussian, the total privacy noise magnitude satisfies:

$$r \leq O \left( \sqrt{\frac{G^2 \log \frac{1}{\delta}}{b_1^2 \epsilon^2} + \sum_{i=\tau(t)+1}^t \left( \frac{M^2 \log \frac{1}{\delta}}{b_2^2 \epsilon^2} \|x_{t-1} - x_{t-2}\|^2 \right)} \right) \quad (90)$$

$$\leq O \left( \sqrt{\frac{G^2 \log \frac{1}{\delta}}{b_1^2 \epsilon^2} + \frac{M^2 \log \frac{1}{\delta}}{b_2^2 \epsilon^2} \kappa} \right). \quad (91)$$

□

## D.2 Proof of Lemma 8

*Proof of Lemma 8.* By the  $M$ -smoothness assumption and using the fact  $\eta \leq \frac{1}{M}$ , we apply the standard descent lemma:

$$\begin{aligned} F(x_t) - F(x_{t-1}) &\leq \langle \nabla F(x_{t-1}), x_t - x_{t-1} \rangle + \frac{M}{2} \|x_t - x_{t-1}\|^2 \\ &\leq \langle \nabla F(x_{t-1}) - \hat{g}_t, -\eta \cdot \hat{g}_t \rangle - \eta \|\hat{g}_t\|^2 + \frac{\eta}{2} \|\hat{g}_t\|^2 \\ &\leq \eta \|\nabla F(x_{t-1}) - \hat{g}_t\| \|\hat{g}_t\|_2 - \frac{\eta}{2} \|\hat{g}_t\|^2. \end{aligned}$$

By Lemma 5, with probability at least  $1 - \omega/2$ , we have  $\|\nabla F(x_{t-1}) - \hat{g}_t\| \leq \chi$  for all  $t$ .

Now consider two cases:

**Case 1:** If  $\|\nabla F(x_{t-1})\| \geq 4\chi$ , then

$$\|\hat{g}_t\| \geq \|\nabla F(x_{t-1})\| - \chi \geq 3\chi \geq 3\|\nabla F(x_{t-1}) - \hat{g}_t\|, \quad (92)$$

yielding

$$F(x_t) - F(x_{t-1}) \leq -\frac{\eta}{6} \|\hat{g}_t\|^2. \quad (93)$$

**Case 2:** If  $\|\nabla F(x_{t-1})\| \leq 4\chi$ , then  $\|\hat{g}_t\| \leq 5\chi$ , and thus

$$F(x_t) - F(x_{t-1}) \leq 5\eta\chi^2. \quad (94)$$

Let  $\mathcal{T} = \{t_1, t_2, \dots, t_{|\mathcal{T}|}\}$  denote the set of iterations where model drift exceeds  $\kappa$ . For each pair of successive drift resets:

$$F(x_{t_{i+1}}) - F(x_{t_i}) \leq -\frac{1}{6\eta} \sum_{t=t_i+1}^{t_{i+1}} \eta^2 \|\hat{g}_t\|_2^2 + (t_{i+1} - t_i)5\eta\chi^2 \quad (95)$$

$$\leq -\frac{1}{6\eta} \text{drift}_{t_{i+1}} + (t_{i+1} - t_i)5\eta\chi^2 \leq -\frac{1}{6\eta}\kappa + (t_{i+1} - t_i)5\eta\chi^2. \quad (96)$$

Summing over  $i$ , we obtain:

$$F(x_{t_{|\mathcal{T}|}}) - F(x_{t_1}) \leq -\frac{|\mathcal{T}|}{6\eta}\kappa + 5T\eta\chi^2.$$

Since  $F(\cdot)$  is upper bounded by  $U$ , we must have:

$$-U \leq -\frac{|\mathcal{T}|\kappa}{6\eta} + 5T\eta\chi^2, \quad (97)$$

which yields:

$$|\mathcal{T}| \leq O\left(\frac{U\eta}{\kappa} + \frac{T\eta^2\chi^2}{\kappa}\right) = O\left(\frac{U\eta}{\kappa}\right),$$

using  $T = O(U/(\eta\chi^2))$ .  $\square$

### D.3 Proof of Theorem 2

*Proof of Theorem 2.* We first verify that the batch size settings  $b_1$  and  $b_2$  are feasible, i.e., the total number of data samples used remains  $O(n)$ . Recall from Lemma 8 that the number of rounds where drift exceeds the threshold is bounded by  $|\mathcal{T}| = O(U\eta/\kappa)$ , and the total number of steps is  $T = O(U/(\eta\chi^2))$ . Then:

$$b_1 \cdot |\mathcal{T}| + b_2 \cdot (T - |\mathcal{T}|) \leq b_1 \cdot |\mathcal{T}| + b_2 \cdot T \leq O(n), \quad (98)$$

under our settings of  $b_1 = \frac{n\kappa}{2U\eta}$  and  $b_2 = \frac{n\eta\chi^2}{2U}$ . This confirms feasibility.

Since each sample is used only once, the overall  $(\epsilon, \delta)$ -differential privacy guarantee follows directly from the Gaussian mechanism and the parallel composition theorem.

We now derive the convergence error  $\alpha$  via Theorem 1, which gives:

$$\alpha = O(\chi) = \tilde{O}(\psi) = \tilde{O}(\sqrt{\sigma^2 + r^2 d}), \quad (99)$$

where from Lemma 7:

$$\sigma^2 \leq \tilde{O}\left(\frac{G^2}{b_1} + \frac{M^2\kappa}{b_2}\right), \quad r^2 \leq \tilde{O}\left(\frac{G^2}{b_1^2\epsilon^2} + \frac{M^2\kappa}{b_2^2\epsilon^2}\right). \quad (100)$$

Substituting our settings  $b_1 = \frac{n\kappa}{2U\eta}$  and  $b_2 = \frac{n\eta\chi^2}{2U}$  into the expression, we get:

$$\alpha = \tilde{O}\left(\sqrt{\frac{G^2U\eta}{n\kappa} + \frac{G^2dU^2\eta^2}{n^2\epsilon^2\kappa^2} + \frac{M^2U\kappa}{n\eta\chi^2} + \frac{M^2dU^2\kappa}{n^2\epsilon^2\eta^2\chi^4}}\right) \quad (101)$$

$$= \tilde{O}\left(\sqrt{\frac{G^2U\sqrt{\rho\alpha}}{M^2n\kappa} + \frac{G^2dU^2\rho\alpha}{M^4n^2\epsilon^2\kappa^2} + \frac{M^4U\kappa}{\sqrt{\rho}n\alpha^{5/2}} + \frac{M^6dU^2\kappa}{\rho n^2\epsilon^2\alpha^5}}\right). \quad (102)$$

To isolate  $\alpha$ , we take the largest among the resulting bounds:

$$\alpha = \tilde{O}\left(\max\left\{\left(\frac{G^2U\sqrt{\rho}}{M^2n\kappa}\right)^{2/3}, \frac{G^2dU^2\rho}{M^4n^2\epsilon^2\kappa^2}, \left(\frac{M^4U\kappa}{n\sqrt{\rho}}\right)^{2/9}, \left(\frac{M^6dU^2\kappa}{\rho n^2\epsilon^2}\right)^{1/7}\right\}\right). \quad (103)$$

Now set:

$$\kappa = \max \left\{ \frac{G^{3/2} U^{1/2} \rho^{1/2}}{M^{5/2} n^{1/2}}, \frac{G^{14/15} d^{2/5} U^{4/5} \rho^{8/15}}{M^{34/15} (n\epsilon)^{4/5}} \right\}. \quad (104)$$

Substituting this into the above expression of  $\alpha$  yields:

$$\alpha = \tilde{O} \left( \left( \frac{GUM}{n} \right)^{1/3} + \frac{G^{2/15} U^{2/5} M^{8/15}}{\rho^{1/15}} \left( \frac{\sqrt{d}}{n\epsilon} \right)^{2/5} \right) = \tilde{O} \left( \frac{1}{n^{1/3}} + \left( \frac{\sqrt{d}}{n\epsilon} \right)^{2/5} \right). \quad (105)$$

□

## E Omitted Proofs in Section 6

### E.1 Proof of Lemma 9

*Proof of Lemma 9.* Let  $\tau(t)$  denote the most recent iteration at which oracle  $\mathcal{O}_1$  was queried before or at iteration  $t$ .

**Case 1:** If  $t = \tau(t)$ , then the global estimator is

$$\hat{g}_t = \frac{1}{m} \sum_{j=1}^m (\mathcal{O}_1(x_{t-1}, \mathcal{B}_{j,t}) + \xi_{j,t}). \quad (106)$$

Each  $\mathcal{O}_1(x_{t-1}, \mathcal{B}_{j,t})$  is an unbiased estimate of  $\nabla F_j(x_{t-1})$ . Let  $\zeta_{j,t} := \mathcal{O}_1(x_{t-1}, \mathcal{B}_{j,t}) - \nabla F_j(x_{t-1})$ , and define  $\zeta_t := \frac{1}{m} \sum_j \zeta_{j,t}$  and  $\xi_t := \frac{1}{m} \sum_j \xi_{j,t}$ . Then,

$$\hat{g}_t - \nabla F(x_{t-1}) = \zeta_t + \xi_t. \quad (107)$$

Since  $f$  is  $G$ -Lipschitz, we have  $\zeta_t \sim \text{nSG} \left( \frac{G\sqrt{\log d}}{\sqrt{mb_1}} \right)$ . Each  $\xi_{j,t} \sim \mathcal{N} \left( 0, c_1 \frac{G^2 \log(1/\delta)}{b_1^2 \epsilon^2} \mathbf{I}_d \right)$ , so their average satisfies:

$$\xi_t \sim \mathcal{N} \left( 0, c_1 \frac{G^2 \log(1/\delta)}{mb_1^2 \epsilon^2} \mathbf{I}_d \right). \quad (108)$$

Thus, in this case, the oracle satisfies condition (2) with the desired bounds.

**Case 2:** If  $t > \tau(t)$ , the global estimate is:

$$\hat{g}_t = \frac{1}{m} \sum_{j=1}^m \left( \mathcal{O}_1(x_{\tau(t)-1}, \mathcal{B}_{j,\tau(t)}) + \xi_{j,\tau(t)} + \sum_{i=\tau(t)+1}^t [\mathcal{O}_2(x_{i-1}, x_{i-2}, \mathcal{B}_{j,i}) + \xi_{j,i}] \right). \quad (109)$$

Let  $\zeta_{j,\tau} := \mathcal{O}_1(x_{\tau(t)-1}, \mathcal{B}_{j,\tau(t)}) - \nabla F_j(x_{\tau(t)-1})$ , and define:

$$\zeta'_{j,i} := \mathcal{O}_2(x_{i-1}, x_{i-2}, \mathcal{B}_{j,i}) - [\nabla F_j(x_{i-1}) - \nabla F_j(x_{i-2})]. \quad (110)$$

Then,

$$\hat{g}_t - \nabla F(x_{t-1}) = \zeta_{\tau(t)} + \sum_{i=\tau(t)+1}^t \zeta'_i + \xi_{\tau(t)} + \sum_{i=\tau(t)+1}^t \xi_i, \quad (111)$$

where  $\zeta_{\tau(t)} := \frac{1}{m} \sum_j \zeta_{j,\tau(t)}$ ,  $\zeta'_i := \frac{1}{m} \sum_j \zeta'_{j,i}$ , and similarly for  $\xi_{\tau(t)}$  and  $\xi_i$ . By the  $M$ -smoothness of  $f$ , we have:

$$\zeta'_i \sim \text{nSG} \left( \frac{M \|x_{i-1} - x_{i-2}\| \sqrt{\log d}}{\sqrt{mb_2}} \right), \quad \xi_i \sim \mathcal{N} \left( 0, c_2 \frac{M^2 \log(1/\delta)}{mb_2^2 \epsilon^2} \|x_{i-1} - x_{i-2}\|^2 \mathbf{I}_d \right). \quad (112)$$

Since the algorithm ensures that  $\text{drift}_t := \sum_{i=\tau(t)+1}^t \|x_{i-1} - x_{i-2}\|^2 \leq \kappa$ , we obtain:

$$\sigma = \tilde{O} \left( \sqrt{\frac{G^2 \log^2 d}{mb_1} + \frac{M^2 \log^2 d}{mb_2} \kappa} \right), \quad r = \tilde{O} \left( \sqrt{\frac{G^2 \log(1/\delta)}{mb_1^2 \epsilon^2} + \frac{M^2 \log(1/\delta)}{mb_2^2 \epsilon^2} \kappa} \right). \quad (113)$$

□

## E.2 Proof of Theorem 3

*Proof of Theorem 3.* We first verify that the total sample usage per client is  $O(n)$ . From Lemma 8, we have  $|\mathcal{T}| = O(U\eta/\kappa)$  and  $T = O(U/(\eta\chi^2))$ . Using the settings:

$$b_1 = \frac{n\kappa}{2U\eta}, \quad b_2 = \frac{n\eta\chi^2}{2U}, \quad (114)$$

the total number of samples used per client is:

$$b_1 \cdot |\mathcal{T}| + b_2 \cdot (T - |\mathcal{T}|) \leq b_1 \cdot |\mathcal{T}| + b_2 \cdot T = O(n). \quad (115)$$

Differential privacy guarantees follows from the Gaussian mechanism and parallel composition, since each data point is used at most once.

Now for the error analysis. By Theorem 1:

$$\alpha = O(\chi) = \tilde{O}(\psi) = \tilde{O}(\sqrt{\sigma^2 + r^2 d}). \quad (116)$$

From Lemma 9:

$$\alpha = \tilde{O} \left( \sqrt{\frac{G^2}{mb_1} + \frac{G^2 d}{mb_1^2 \epsilon^2} + \left( \frac{M^2}{mb_2} + \frac{M^2 d}{mb_2^2 \epsilon^2} \right) \cdot \kappa} \right). \quad (117)$$

Substitute the expressions for  $b_1, b_2$  into the bound and simplify, we get:

$$\alpha = \tilde{O} \left( \sqrt{\frac{G^2 U \eta}{mn\kappa} + \frac{G^2 d U^2 \eta^2}{mn^2 \epsilon^2 \kappa^2} + \frac{M^2 U \kappa}{mn\eta\chi^2} + \frac{M^2 d U^2 \kappa}{mn^2 \epsilon^2 \eta^2 \chi^4}} \right) \quad (118)$$

$$= \tilde{O} \left( \sqrt{\frac{G^2 U \sqrt{\rho} \alpha}{mM^2 n \kappa} + \frac{G^2 d U^2 \rho \alpha}{mn^2 \epsilon^2 M^4 \kappa^2} + \frac{M^4 U \kappa}{mn \rho^{\frac{1}{2}} \alpha^{\frac{5}{2}}} + \frac{M^6 d U^2 \kappa}{mn^2 \epsilon^2 \rho \alpha^5}} \right). \quad (119)$$

To isolate  $\alpha$ , we take the largest among the resulting bounds:

$$\alpha = \tilde{O} \left( \max \left\{ \left( \frac{G^2 U \sqrt{\rho}}{mM^2 n \kappa} \right)^{2/3}, \frac{G^2 d U^2 \rho}{mn^2 \epsilon^2 M^4 \kappa^2}, \left( \frac{M^4 U \kappa}{mn \sqrt{\rho}} \right)^{2/9}, \left( \frac{M^6 d U^2 \kappa}{mn^2 \epsilon^2 \rho} \right)^{1/7} \right\} \right).$$

Now set:

$$\kappa = \max \left\{ \frac{G^{3/2} \sqrt{\rho} U}{M^{5/2} \sqrt{mn}}, \frac{G^{14/15} d^{2/5} U^{4/5} \rho^{8/15}}{M^{34/15} (\sqrt{mn} \epsilon)^{4/5}} \right\} \quad (120)$$

Substituting this into the above expression of  $\alpha$  yields:

$$\alpha = \tilde{O} \left( \left( \frac{GUM}{mn} \right)^{1/3} + \frac{G^{2/15} U^{2/5} M^{8/15}}{\rho^{1/15}} \left( \frac{\sqrt{d}}{\sqrt{mn} \epsilon} \right)^{2/5} \right) = \tilde{O} \left( \frac{1}{(mn)^{1/3}} + \left( \frac{\sqrt{d}}{\sqrt{mn} \epsilon} \right)^{2/5} \right). \quad (121)$$

□

## E.3 Proof of Theorem 4

*Proof of Theorem 4.* The  $(\epsilon, \delta)$ -ICRL-DP guarantee follows directly from the Gaussian mechanism and the adaptive composition theorem, since each client adds independent Gaussian noise to both their gradient and Hessian estimates. Each local data point is used at most  $T$  times—once for each model iterate—and all messages sent to the server are privatized accordingly.

We now derive the error rate  $\alpha$  guarantee for the output  $x_o$ . Let  $\mathcal{S} := \bigsqcup_{j=1}^m S_j$  denote the full held-out evaluation dataset, and let  $x_p$  be an  $\alpha$ -SOSP in the input to Algorithm 4. Define the aggregate gradient noise and Hessian noise as

$$\theta_p := \frac{1}{m} \sum_{j=1}^m \theta_{j,p}, \quad \mathbf{H}_p := \frac{1}{m} \sum_{j=1}^m \mathbf{H}_{j,p}. \quad (122)$$

Let  $\sigma_1^2 = c_1 \frac{G^2 T \log(1/\delta)}{n^2 \epsilon^2}$  and  $\sigma_2^2 = c_2 \frac{M^2 d T \log(1/\delta)}{n^2 \epsilon^2}$  denote the variances of the noise added to the gradient and Hessian components, respectively.

**Gradient Estimation Error.** For any  $\mathcal{S}_j$  and  $x$ ,  $\nabla \hat{f}_{\mathcal{S}_j}(x) - \nabla F_j(x)$  is zero-mean and follows  $\text{nSG}\left(\frac{2G}{\sqrt{n}}\right)$ . By the  $G$ -Lipschitz assumption and norm-sub-Gaussian concentration (Lemma 11), we have with probability at least  $1 - \omega'/8$ :

$$\|\nabla F(x_p) - \nabla \hat{f}_{\mathcal{S}}(x_p)\| \leq O\left(\frac{G\sqrt{\log(d/\omega')}}{\sqrt{mn}}\right). \quad (123)$$

Also, since  $\theta_p \sim \mathcal{N}(0, \sigma_1^2/m)$ , standard Gaussian concentration (Lemma 10) gives, with probability at least  $1 - \omega'/8$ :

$$\|\theta_p\| \leq O\left(\frac{G\sqrt{dT \log(1/\delta) \log(1/\omega')}}{\sqrt{mn}\epsilon}\right). \quad (124)$$

**Hessian Estimation Error.** For any  $j \in [m]$  and  $z \in \mathcal{S}_j$ ,  $\mathbb{E}[\nabla^2 f(x_p; z) - \nabla^2 F_j(x_p)] = 0$ , and  $\|\nabla^2 f(x_p; z) - \nabla^2 F_j(x_p)\|_2 \leq 2M$  (due to  $M$ -smoothness). That is, each empirical Hessian term is  $2M$ -bounded in operator norm. Applying the matrix Bernstein inequality (Lemma 14), and using the assumption  $mn \geq \frac{4}{9} \log(8d/\omega')$ , we obtain with probability at least  $1 - \omega'/8$ :

$$\|\nabla^2 \hat{f}_{\mathcal{S}}(x_p) - \nabla^2 F(x_p)\| \leq O\left(M\sqrt{\frac{\log(d/\omega')}{mn}}\right). \quad (125)$$

For the added noise, since  $\mathbf{H}_p$  consists of symmetric Gaussian matrices with variance  $\sigma_2^2/m$ , Lemma 15 gives, with probability at least  $1 - \omega'/8$ :

$$\|\mathbf{H}_p\| \leq O\left(\frac{Md\sqrt{T \log(1/\delta) \log(1/\omega')}}{\sqrt{mn}\epsilon}\right). \quad (126)$$

**Verification for  $x_p$ .** Combining the above estimates and using a union bound, with probability at least  $1 - \omega'/2$ , we have:

$$\|\nabla \bar{F}(x_p)\|_2 \leq \|\nabla F(x_p)\|_2 + \|\nabla \bar{F}(x_p) - \nabla F(x_p)\|_2 \quad (127)$$

$$\leq \|\nabla F(x_p)\|_2 + \|\nabla \hat{f}_{\mathcal{S}}(x_p) - \nabla F(x_p)\|_2 + \|\theta_p\|_2 \quad (128)$$

$$\leq \alpha + (\text{estimation error}) \quad (129)$$

$$\leq O\left(\alpha + \frac{G \log(d/\omega')}{\sqrt{mn}} + \frac{G\sqrt{dT \log(1/\delta) \log(1/\omega')}}{\sqrt{mn}\epsilon}\right), \quad (130)$$

and

$$\lambda_{\min}(\nabla^2 \bar{F}(x_p)) \geq \lambda_{\min}(\nabla^2 F(x_p)) + \lambda_{\min}(\nabla^2 \bar{F}(x_p) - \nabla^2 F(x_p)) \quad (131)$$

$$\geq \lambda_{\min}(\nabla^2 F(x_p)) + \lambda_{\min}(\nabla^2 \hat{f}_{\mathcal{S}}(x_p) - \nabla^2 F(x_p)) + \lambda_{\min}(\mathbf{H}_p) \quad (132)$$

$$\geq -\sqrt{\rho\alpha} - \|\nabla^2 f(x_p; \mathcal{S}) - \nabla^2 F(x_p)\|_2 - \|\mathbf{H}_p\|_2 \quad (133)$$

$$\geq -(\sqrt{\rho\alpha} + (\text{estimation error})) \quad (134)$$

$$\geq -O\left(\sqrt{\rho\alpha} + M\sqrt{\frac{\log(d/\omega')}{mn}} + \frac{Md\sqrt{T \log(1/\delta) \log(1/\omega')}}{\sqrt{mn}\epsilon}\right). \quad (135)$$

Hence,  $x_p$  will be selected with probability at least  $1 - \omega'/2$ .

**Guarantee for Output  $x_o$ .** Let  $x_o$  be the output of Algorithm 4. By construction, it must satisfy:

$$\|\nabla F(x_o)\|_2 \leq \|\nabla \bar{F}(x_o)\|_2 + \|\nabla F(x_o) - \nabla \bar{F}(x_o)\|_2 \quad (136)$$

$$\leq \|\nabla \bar{F}(x_o)\|_2 + \|\nabla F(x_o) - \nabla \hat{f}_{\mathcal{S}}(x_o)\|_2 + \|\xi_o\|_2, \quad (137)$$

and

$$\lambda_{\min}(\nabla^2 F(x_o)) \geq \lambda_{\min}(\nabla^2 \bar{F}(x_o)) + \lambda_{\min}(\nabla^2 F(x_o) - \nabla^2 \bar{F}(x_o)) \quad (138)$$

$$\geq \lambda_{\min}(\nabla^2 \bar{F}(x_o)) - \|\nabla^2 F(x_o) - \nabla^2 \bar{F}(x_o)\|_2 \quad (139)$$

$$\geq \lambda_{\min}(\nabla^2 \bar{F}(x_o)) - \|\nabla^2 F(x_o) - \nabla^2 \hat{f}_S(x_o)\|_2 - \|H_o\|_2. \quad (140)$$

Using the same reasoning as above, applying the union bound again and using the fact that  $x_o$  is the output, we get that with probability at least  $1 - \omega'$ , the following hold:

$$\|\nabla F(x_o)\| \leq O\left(\alpha + \frac{G \log(d/\omega')}{\sqrt{mn}} + \frac{G\sqrt{dT \log(1/\delta) \log(1/\omega')}}{\sqrt{mn}\epsilon}\right), \quad (141)$$

and

$$\lambda_{\min}(\nabla^2 F(x_o)) \geq -O\left(\sqrt{\rho\alpha} + M\sqrt{\frac{\log(d/\omega')}{mn}} + \frac{Md\sqrt{T \log(1/\delta) \log(1/\omega')}}{\sqrt{mn}\epsilon}\right). \quad (142)$$

Finally, recalling that  $T = O(1/\alpha^{2.5})$ , and grouping the dependency on  $\alpha$ ,  $d$ ,  $m$ ,  $n$ , and  $\epsilon$ , we conclude that  $x_o$  is an  $\alpha'$ -SOSP with

$$\alpha' = \tilde{O}\left(\alpha + \frac{1}{mn} + \frac{1}{\sqrt{mn}} + \frac{\alpha}{\sqrt{mn}} + \frac{\sqrt{d}}{\sqrt{mn}\epsilon\alpha^{5/4}} + \frac{d}{\sqrt{mn}\epsilon\alpha^{3/4}} + \frac{d^2}{mn^2\epsilon^2\alpha^{5/2}}\right), \quad (143)$$

as claimed.  $\square$

## F Experiments

**Running Environments** All experiments were conducted with the following computing infrastructure:

- OS: Ubuntu 22.04.4 LTS
- CPU: AMD EPYC 7513 32-Core Processor
- CPU Memory: 503GB
- GPU: NVIDIA RTX A6000 GPU
- GPU Memory: 48GB
- Programming language: Python 3.11.8
- Deep learning framework: Pytorch 2.2.2 + cuda 12.1

**Tasks and Datasets** We conduct image classification tasks on two datasets: MNIST [27] and CIFAR-10 [25]. For each experiment, we set the number of training samples to  $n = 6000$  and vary the number of clients  $m$  in  $\{1, 2, 5, 10\}$ , where  $m = 1$  corresponds to the single-machine setting, while the others correspond to distributed learning scenarios. The test set consists of 10000 samples for both datasets.

**Models** We primarily use a fully connected (FC) neural network with one hidden layer containing 128 units and ReLU activation. The loss function is the standard cross-entropy loss. The model is initialized using Kaiming initialization [19], with biases set to zero by default. The FC network is mainly employed to verify our theoretical findings, such as the trends of performance variation under different parameter settings. In addition, we adopt a ResNet-18 architecture to demonstrate that our algorithm also attains strong practical performance when applied to deeper models.

**Algorithms** We compare our proposed algorithm, **Gauss-PSGD**, against multiple baselines:

- The method from [30], which serves as the primary baseline in our main experiments. This comparison highlights the superiority of Gauss-PSGD in achieving second-order convergence under differential privacy.

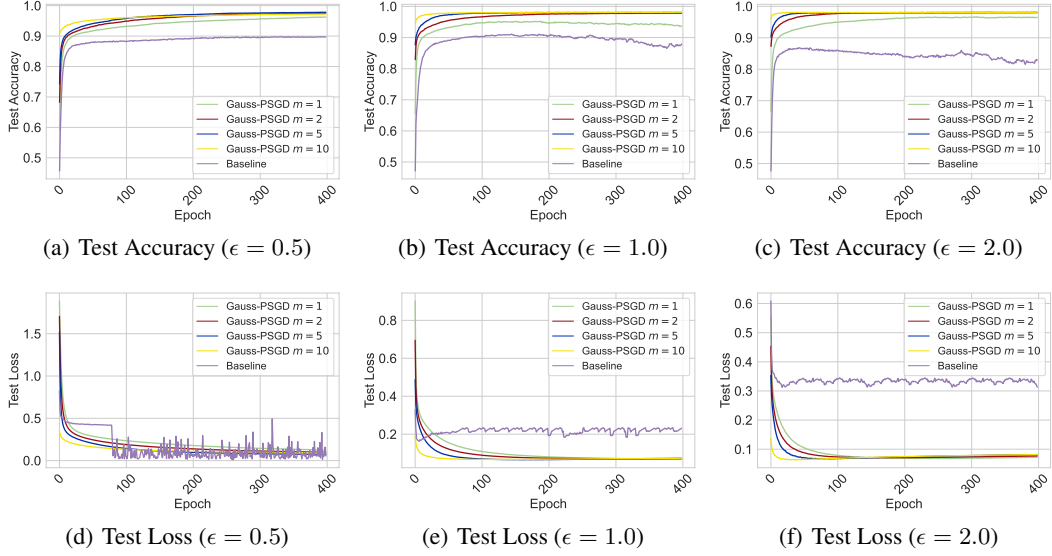


Figure 1: Comparison of learning performance for our Gauss-PSGD and the baseline method on **MNIST** dataset. **Top: Test accuracy** v.s. # epoch for varying privacy budget  $\epsilon \in \{0.5, 1.0, 2.0\}$ . **Bottom: Test loss** v.s. # epoch for varying privacy budget  $\epsilon \in \{0.5, 1.0, 2.0\}$ .

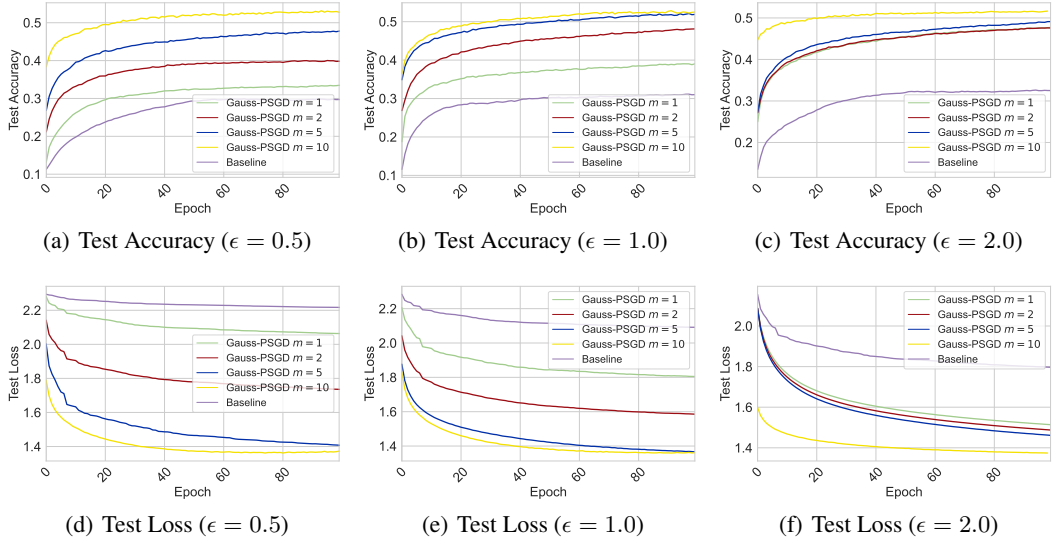


Figure 2: Comparison of learning performance for our Gauss-PSGD and the baseline method on **CIFAR-10** dataset. **Top: Test accuracy** v.s. # epoch for varying privacy budget  $\epsilon \in \{0.5, 1.0, 2.0\}$ . **Bottom: Test loss** v.s. # epoch for varying privacy budget  $\epsilon \in \{0.5, 1.0, 2.0\}$ .

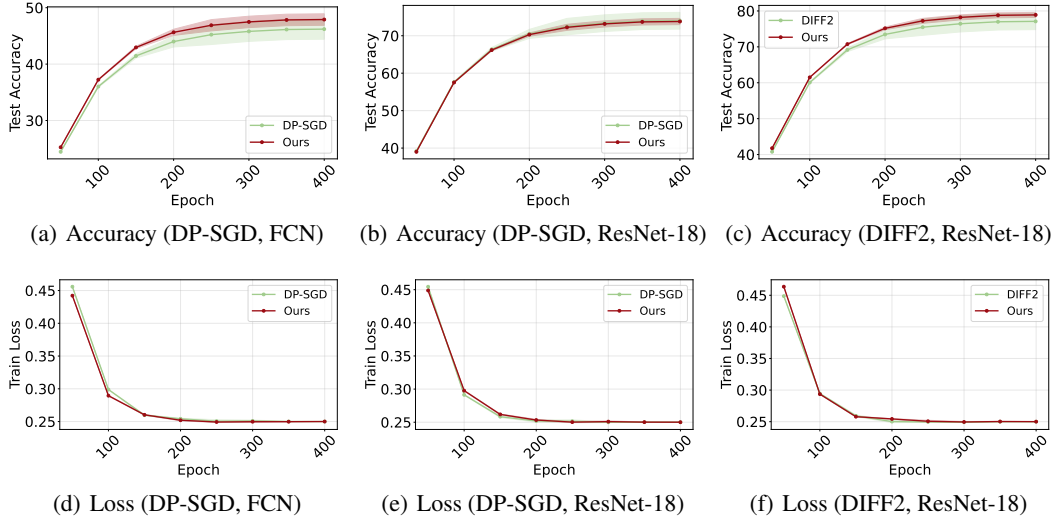


Figure 3: Comparison of Gauss-PSGD with baseline methods on the **CIFAR-10** dataset under a fixed privacy budget of  $\epsilon = 2$ . **Top:** Test accuracy over epochs. **Bottom:** Test loss over epochs. In the centralized setting ( $m = 1$ ), Gauss-PSGD is compared with DP-SGD using a fully connected network (FCN, left) and a ResNet-18 model (middle). In the distributed setting ( $m = 10$ ), Gauss-PSGD is compared with DIFF2 using the ResNet-18 model (right). Shaded areas indicate standard deviation over 5 independent runs

- Standard DP-SGD [1], used in centralized (single-client) settings. This comparison is designed to evaluate the benefit of incorporating second-order convergence.
- DIFF2 [37], a recent state-of-the-art differentially private federated learning (DP-FL) algorithm that employs the standard SPIDER variance reduction technique but achieves only first-order convergence. This comparison is intended to demonstrate the advantage of second-order convergence in distributed settings.

For Gauss-PSGD, we use the following empirical hyperparameters:

- Escape threshold  $\chi = 0.01$
- Model drift threshold  $\kappa = 0.1$
- Maximum escape steps  $\Gamma = 10$
- Maximum repeat number of escape  $Q = 3$

For all algorithms, we set the privacy parameters to  $\delta = 10^{-5}$  and vary  $\epsilon$  in  $\{0.5, 1.0, 2.0\}$ , corresponding to strong, medium, and weak privacy regimes, respectively. The learning rate is set to 0.001 for MNIST and 0.01 for CIFAR-10. In all experiments, we apply gradient clipping with a threshold of 1.0, selected via grid search.

**Evaluations** We evaluate the performance of the implemented algorithms using two criteria: test accuracy and test loss. Both metrics are analyzed over training epochs to assess convergence and generalization performance.

**Results** The experimental results on the MNIST and CIFAR-10 datasets are presented in Fig. 1 and Fig. 2, respectively. Each figure shows test accuracy (top row) and test loss (bottom row) over training epochs for different privacy budgets  $\epsilon \in \{0.5, 1.0, 2.0\}$ . Across all settings, Gauss-PSGD consistently outperforms the baseline method from [30]. Overall, the test accuracy improves as  $\epsilon$  increases, and the performance gap between Gauss-PSGD and the baseline widens in distributed settings ( $m > 1$ ), highlighting the collaborative synergy of distributed learning and the robustness of Gauss-PSGD in handling data heterogeneity. Additionally, Gauss-PSGD exhibits faster loss reduction in early training, suggesting improved convergence behavior.

To further evaluate the benefits of second-order convergence, we include additional comparisons in Fig. 3. In the centralized setting ( $m = 1$ ), we compare Gauss-PSGD with standard DP-SGD using both a simple fully connected network (Fig. 3 (a), (d)) and a deeper ResNet-18 model (Fig. 3 (b), (e)). In both cases, Gauss-PSGD achieves comparable or higher test accuracy and demonstrates reduced variance across runs.

In the distributed setting ( $m = 10$ ), we compare Gauss-PSGD with DIFF2, a recent differentially private federated learning algorithm designed for first-order convergence. The results (Fig. 3 (c), (f)) show that Gauss-PSGD achieves higher test accuracy and improved stability across runs, despite both methods being first-order in design. These comparisons further support the advantage of Gauss-PSGD’s second-order convergence behavior in both centralized and federated learning scenarios.

## G Broader Impact Statement

This paper advances the field of differentially private (DP) stochastic non-convex optimization by addressing key theoretical challenges in finding second-order stationary points (SOSP). Our contributions are particularly relevant for applications requiring strong privacy guarantees, including distributed learning with heterogeneous data. These advancements have practical implications for privacy-sensitive fields such as healthcare, finance, and large language models (LLMs), where data confidentiality is paramount.

By improving the efficiency and accuracy of DP optimization techniques, our work supports the development of machine learning systems that can operate on sensitive datasets without compromising privacy. This fosters greater trust in data-driven decision-making and encourages organizations to adopt privacy-preserving practices, enabling informed and responsible use of sensitive data.

Nevertheless, it is important to acknowledge the broader limitations inherent to DP-based learning algorithms, not just those specific to our work. Privacy-preserving methods often introduce trade-offs, such as reduced model accuracy compared to their non-private counterparts, which may impact decision-making in high-stakes applications.

Despite these challenges, we believe that advancing and responsibly applying privacy-preserving optimization techniques will have a positive societal impact. By enabling secure and ethical data analysis, our work contributes to the broader goal of building trustworthy AI/ML systems.

## H Limitation Discussion

One of the primary objective of this work is to rectify a key analytical error in [30] by presenting the correct error rates for DP stochastic non-convex optimization. Our proposed framework, Gauss-PSGD, is designed to be broadly applicable beyond the DP setting, offering a versatile optimization tool for general non-convex problems. Furthermore, this work makes the first attempt to extend DP-SOSP analysis to the distributed learning setting, establishing state-of-the-art utility guarantees.

To maintain consistency with prior work [30], we assume access to an unbiased gradient oracle. This assumption is fundamental in theoretical analysis and is also adopted by many recent studies in DP optimization and distributed learning, such as [2, 16]. However, it may not fully reflect the behavior of practical optimizers that employ biased and noisy gradients, particularly those using gradient clipping—a standard technique in DP implementations.

Nevertheless, our Gauss-PSGD framework can be extended to handle biased oracles induced by clipping. The main challenge lies in the analysis: incorporating clipping introduces bias, requiring a refined characterization of the descent dynamics. In particular, Lemma 3 (the descent lemma) must be adapted to reflect the bias–variance trade-off. Techniques for bias reduction in clipped DP learning—such as those developed in [52]—could offer a promising foundation for such an extension.

The saddle point escaping analysis (Lemma 1) can also be generalized. As shown in our proof, the key mechanism enabling escape is the injection of symmetric Gaussian noise, which drives the divergence in the coupling sequence. This mechanism remains valid under clipping, provided the Gaussian noise is appropriately calibrated. However, the number of steps required for escape may change due to the altered noise structure and bias, and a more delicate analysis would be required to quantify this behavior accurately.

We consider this as a promising direction for future work and leave its full exploration to subsequent studies.

## I Conclusion

In this work, we investigated the problem of finding second-order stationary points (SOSP) in differentially private (DP) stochastic non-convex optimization. We proposed a novel framework that leverages perturbed stochastic gradient descent (SGD) with Gaussian noise and introduces a novel criterion based on model drift distance to ensure provable saddle point escape and efficient convergence. By incorporating an adaptive SPIDER as the gradient oracle, we developed a new DP algorithm that rectifies existing error rates. Furthermore, we extended our approach to distributed learning scenarios with heterogeneous data, providing the first theoretical guarantees for finding DP-SOSP in such settings. Through rigorous analysis, we demonstrated that our framework not only avoids the pitfalls of private model selection but also remains effective in high-dimensional distributed learning environments.

Our work opens several promising directions for future research. A key challenge is bridging the gap between our upper bound and the existing DP lower bound for stochastic optimization, as established in [2]. The current lower bound is derived from convex loss functions and first-order stationary points, whereas finding DP-SOSP in non-convex optimization is inherently more difficult. We conjecture that the existing lower bound is not tight for the non-convex case. Establishing a tighter lower bound remains a critical open problem. Additionally, exploring whether our upper bounds can be further improved is another intriguing direction that warrants in-depth investigation.