
Convergent Differential Privacy Analysis for General Federated Learning

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 The powerful cooperation of federated learning (FL) and differential privacy (DP)
2 provides a promising paradigm for the large-scale private clients. However, existing
3 analyses in FL-DP mostly rely on the composition theorem and cannot tightly quan-
4 tify the privacy leakage challenges, which is tight for a few communication rounds
5 but yields an arbitrarily loose and divergent bound eventually. This also implies
6 a counterintuitive judgment, suggesting that FL-DP may not provide adequate
7 privacy support during long-term training under constant-level noisy perturbations,
8 yielding discrepancy between the theoretical and experimental results. To further
9 investigate the convergent privacy and reliability of the FL-DP framework, in this
10 paper, we comprehensively evaluate the worst privacy of two classical methods
11 under the non-convex and smooth objectives based on the f -DP analysis. With
12 the aid of the shifted interpolation technique, we successfully prove that privacy in
13 Noisy-FedAvg has a tight convergent bound. Moreover, with the regularization
14 of the proxy term, privacy in Noisy-FedProx has a stable constant lower bound.
15 Our analysis further demonstrates a solid theoretical foundation for the reliability
16 of privacy in FL-DP. Meanwhile, our conclusions can also be losslessly converted
17 to other classical DP analytical frameworks, e.g. (ϵ, δ) -DP and Rényi-DP (RDP),
18 to provide more fine-grained understandings for the FL-DP frameworks.

19 1 Introduction

20 Since McMahan et al. [2017] proposes the FedAvg method as a general FL framework, it has been
21 widely developed into a collaborative training standard with privacy protection attributes, which
22 successfully avoids *direct leakage* of sensitive data. As research on privacy progresses, researchers
23 have found that standard FL frameworks still face a threat from *indirect leakage*. Attackers can
24 potentially recover local private data through reverse inference by persistently stealing model states
25 via model (gradient) inversion attacks [Geiping et al., 2020] or distinguish whether individuals are
26 involved in the training via membership inference attacks [Nasr et al., 2019]. To further strengthen
27 the reliability of FL, DP [Dwork, 2006, Dwork et al., 2014, Abadi et al., 2016] has naturally been
28 incorporated into the FL framework, yielding FL-DP [Wei et al., 2020]. As a primary technique, the
29 noisy perturbation is widely applied in various advanced FL methods to further enhance its security.

30 However, the theoretical analysis of the FL-DP framework, especially in evaluating the privacy
31 levels, is currently unable to provide a comprehensive understanding of its proper application. Most
32 of the previous works are built upon the foundational lemma of privacy amplification by iteration,
33 directly resulting in divergent privacy bound as the training communication round T becomes large.
34 This implies an inference that contradicts intuition and empirical studies, which is, that the FL-DP
35 framework may completely lose its privacy protection attributes as $T \rightarrow \infty$. Such a conclusion is
36 almost unacceptable for FL-DP. Therefore, establishing a precise and tight analysis is a crucial target.

Table 1: The worst privacy of the Noisy-FedAvg and Noisy-FedProx methods in our analysis. V is the norm of clip gradient. K, T are local training interval and communication round. σ is the variance of the noise. The trade-off function $T_G(\cdot)$ ^[a] is defined in Definition 4. μ, c and z are constants.

	Lr ^[b]	Worst Privacy	Convergent? on $T \rightarrow \infty$	Convergent? on $K \rightarrow \infty$
Noisy FedAvg	C	$T_G\left(\frac{2\mu VK}{\sqrt{m}\sigma} \sqrt{\frac{(1+\mu L)^K + 1}{(1+\mu L)^K - 1} \frac{(1+\mu L)^{KT} - 1}{(1+\mu L)^{KT} + 1}}\right)$		
	CD	$T_G\left(\frac{2cV \ln(K+1)}{\sqrt{m}\sigma} \sqrt{\frac{(1+K)^{c\mu L} + 1}{(1+K)^{c\mu L} - 1} \frac{(1+K)^{c\mu LT} - 1}{(1+K)^{c\mu LT} + 1}}\right)$	✓	✗
	SD	$T_G\left(\frac{2\mu VK}{\sqrt{m}\sigma} \sqrt{2 - \frac{1}{T}}\right)$		
	ID	$T_G\left(\frac{2zV}{\sqrt{m}\sigma} \sqrt{2 - \frac{1}{T}}\right)$		
Noisy FedProx		$T_G\left(\frac{2V}{\sqrt{m}\alpha\sigma} \sqrt{\frac{2\alpha-L}{L}} \sqrt{\frac{\alpha^T - (\alpha-L)^T}{\alpha^T + (\alpha-L)^T}}\right)$	✓	✓

[a] For the trade-off function $T_G(s)$, smaller s means stronger privacy.

[b] Learning rate decaying policy. C: constant learning rate; CD: cyclically decaying; SD: stage-wise decaying; ID: iteratively decaying. More details are stated in Theorem 3 4.

Notably, significant progress has been made in characterizing convergent privacy in the noisy gradient descent method in RDP analysis [Chourasia et al., 2021, Ye and Shokri, 2022, Altschuler and Talwar, 2022, Altschuler et al., 2024]. However, due to the challenges and intricacies of the analytical techniques adopted, similar results have not yet successfully been extended to the FL-DP. The multi-step local updates on heterogeneous datasets lead to biased local models, posing significant obstacles to the analysis. Recently, analyses based on f -DP [Dong et al., 2022] have brought a promising resolution to this challenge. This information-theoretically lossless definition naturally evaluates privacy by the Type I / II error trade-off curve of the hypothesis testing problem about whether a given individual is in the training dataset. Combined with shifted interpolation techniques [Bok et al., 2024], it successfully recovers tighter convergent privacy for strongly convex and convex objectives in noisy gradient descent methods. This may make it possible to quantify convergent privacy in FL-DP and may offer novel understandings about impacts of some key hyperparameters.

In this paper, we investigate the privacy of two classic DP-FL methods, i.e. Noisy-FedAvg and Noisy-FedProx and successfully evaluate their *worst privacy* in the f -DP analysis, as shown in Table 1. For the Noisy-FedAvg method, we investigate four typical learning rate decay strategies and provide the coefficients corresponding to each case to ensure a tighter privacy lower bound. We also prove that its iterative privacy on non-convex and smooth objectives could not diverge w.r.t. the number of communication rounds T , i.e., a convergent privacy. To the best of our knowledge, this contributes the first convergent privacy analysis in FL-DP methods for non-convex functions. Furthermore, by exploring the decay properties of the proximal term in Noisy-FedProx, we prove that its worst privacy can converge to a general constant lower bound. Our analysis successfully challenges the long-standing belief that privacy budgets of FL-DP have to increase as training processes and provides reliable guarantees for its privacy protection ability. At the same time, the exploration from the proximal term provides a promising solution, suggesting that a well-designed local regularization term can achieve a win-win solution for both optimization and privacy in FL-DP.

2 Related Work

Federated Learning. FL is a classic learning paradigm that protects local privacy. Since McMahan et al. [2017] proposes the basic framework, it has been widely studied in several communities. As its foundational study, the local-SGD [Stich, 2019, Lin et al., 2019, Woodworth et al., 2020, Gorbunov et al., 2021] method fully demonstrates the efficiency of local training. Based on this, FL further considers the impacts of heterogeneous private datasets and communication bottlenecks [Wang et al., 2020, Chen et al., 2021, Kairouz et al., 2021]. To address these two basic issues, a series of studies have explored these processes from different perspectives. One approach involves proposing better

optimization algorithms by defining concepts such as client drift [Karimireddy et al., 2020] and heterogeneity similarity [Mendieta et al., 2022], specifically targeting and resolving the additional error terms they cause. This mainly includes the natural application and expansion of variance-reduction optimizers [Jhunjhunwala et al., 2022, Malinovsky et al., 2022, Li et al., 2023], the flexible implementation of the advanced primal-dual methods [Zhang et al., 2021c, Wang et al., 2022, Sun et al., 2023b, Mishchenko et al., 2022, Grudzień et al., 2023, Acar et al., Sun et al., 2023a], and the additional deployment of the momentum-based correction [Liu et al., 2020, Khanduri et al., 2021, Das et al., 2022, Sun et al., 2023c, 2024]. Upgraded optimizers allow the aggregation frequency to largely decrease while maintaining convergence. Another approach primarily focuses on sparse training and quantization to reduce communication bits [Reisizadeh et al., 2020, Shlezinger et al., 2020, Dai et al., 2022]. Additionally, research based on data domain and feature domain has also made significant contributions to the FL community [Yao et al., 2019, Zhang et al., 2021a, Xu et al., 2021].

FL-DP. DP is a natural privacy-preserving framework with theoretical foundations [Dwork et al., 2006b,a, Dwork, 2006]. As one of the main algorithms for differential privacy, noise perturbation has achieved great success in deep learning [Abadi et al., 2016, Zhao et al., 2019, Arachchige et al., 2019, Wu et al., 2020]. Combining this, FL-DP adds noise before transmitting their variables, i.e. client-level noises [Geyer et al., 2017] and server-level noises [Wei et al., 2020]. Since there is no fundamental difference between the analysis of them, in this paper, we mainly consider client-level noises. One major research direction involves conducting noise testing on widely developed federated optimization algorithms [Zhu et al., 2021, Noble et al., 2022, Lowy et al., 2023, Zhang and Tang, 2022, Yang and Wu, 2023], and evaluating the performance of different methods under DP noises through convergence analysis and privacy analysis. Another research direction involves injecting noise into real-world systems to address practical challenges, which primarily focuses on personalized scenarios [Hu et al., 2020, Yang et al., 2021, 2023, Wei et al., 2023], decentralized scenarios [Wittkopp and Acker, 2020, Chen et al., 2022, Gao et al., 2023, Shi et al., 2023], and adaptive or asymmetric update scenarios [Girgis et al., 2021, Wu et al., 2022, He et al., 2023]. FL-DP has been extensively tested across various scales of tasks and has successfully validated its robust local privacy protection capabilities. At the same time, the theoretical analysis of FL-DP has been progressing systematically and in tandem. Based on various DP relaxations, they provide a comparison of privacy performance by analyzing concepts such as privacy budgets, and further understand the specific attributes of privacy algorithms [Rodríguez-Barroso et al., 2020, Wei et al., 2021, Kim et al., 2021, Zheng et al., 2021, Ling et al., 2024, Jiao et al., 2024]. Theoretical advancements in DP have revolutionized how we could quantify and safeguard privacy, offering unprecedented precision and robustness.

3 Preliminaries

Notations. In the subsequent content, we use italics for scalars and denote the integer set from 1 to a by $[a]$. All sequences of variables are represented in subscript, e.g. $w_{i,k,t}$. For arithmetic operators, unless specifically stated otherwise, the calculations are performed element-wise. Other symbols used in this paper will be explicitly defined when they are first introduced.

3.1 General FL-DP framework

We consider the general finite-sum minimization problem in the classical federated learning:

$$w^* \in \arg \min_w f(w) \triangleq \frac{1}{m} \sum_{i \in \mathcal{I}} f_i(w), \quad (1)$$

where $f_i(w) = \mathbb{E}_{\varepsilon \sim \mathcal{D}_i} [f_i(w, \varepsilon)]$ denotes the local population risk. $w \in \mathbb{R}^d$ denotes d -dim learnable parameters. $\varepsilon \sim \mathcal{D}_i$ denotes that the private dataset on client i is sampled from distribution $\mathcal{D}_i$. We consider the general heterogeneity, i.e. $\mathcal{D}_i$ can differ from $\mathcal{D}_j$ if $i \neq j$, leading to $f_i(w) \neq f_j(w)$.

In our analysis, we consider the FL-DP framework with the classical client-level Gaussian noises. The FL training process remains consistent with standard training procedures. The local clients enhance local privacy by adding isotropic Gaussian noises to the uploaded model parameters, i.e. $n_i \sim \mathcal{N}(0, \sigma^2 I_d)$. Then the global server aggregates the noisy parameters as the global model w_{t+1} . Due to the page limitation, details of the algorithmic implementation are deferred to the Appendix A.

Noisy-FedAvg: we consider that each local client performs a fundamental gradient descent as follows:

$$w_{i,k+1,t} = w_{i,k,t} - \eta_{k,t} g_{i,k,t}, \quad (2)$$

119 where $g_{i,k,t} = \nabla f_i(w_{i,k,t}, \varepsilon) / \max\{1, \frac{\|\nabla f_i(w_{i,k,t}, \varepsilon)\|}{V}\}$, and V is a constant coefficient.

120 **Noisy-FedProx:** The vanilla local training in FedProx is based on solving the following surrogate:

$$\min_w f_i(w) + \frac{\alpha}{2} \|w - w_t\|^2. \quad (3)$$

121 To generally compare with Noisy-FedAvg, we consider an iterative form of gradient descent as:

$$w_{i,k+1,t} = w_{i,k,t} - \eta_{k,t} [g_{i,k,t} + \alpha(w_{i,k,t} - w_t)]. \quad (4)$$

122 3.2 DP and f -DP

123 **Definition 1** We denote heterogeneous datasets on the client i by $\mathcal{S}_i = \{\varepsilon_{ij}\}$ and let the union of all
124 local datasets be $\mathcal{C} = \{\mathcal{S}_i\}$. We say two unions are adjacent datasets if they only differ by one data
125 sample. For instance, there exists the union $\mathcal{C}' = \{\mathcal{S}'_i\}$. $(\mathcal{C}, \mathcal{C}')$ are adjacent datasets if there exists
126 the index pair (i^*, j^*) such that all other data samples are the same except for $\varepsilon_{i^*j^*} \neq \varepsilon'_{i^*j^*}$.

127 **Definition 2** A randomized mechanism $\mathcal{M}$ is (ϵ, δ) -DP if for any event E the following satisfies:

$$P(\mathcal{M}(\mathcal{C}) \in E) \leq e^\epsilon P(\mathcal{M}(\mathcal{C}') \in E) + \delta. \quad (5)$$

128 Definition 2 is the widely used (ϵ, δ) -DP, which is a lossy relaxation in the DP analysis since its
129 probabilistic gaps. To bridge the discrepancy of precise DP definitions, statistic analysis demonstrates
130 that DP could be naturally deduced by hypothesis-testing problems [Wasserman and Zhou, 2010,
131 Kairouz et al., 2015]. From the perspective of attackers, DP means the difficulty in distinguishing $\mathcal{C}$
132 and $\mathcal{C}'$ under the mechanism $\mathcal{M}$. They can generally consider the following problem:

133 Given $\mathcal{M}$, is the underlying union $\mathcal{C} (H_0)$ or $\mathcal{C}' (H_1)$?

134 To exactly quantify the difficulty of its answer, Dong et al. [2022] propose that distinguishing these
135 two hypotheses could be best delineated by the optimal trade-off between the possible type I and type
136 II errors. Specifically, by considering rejection rules $0 \leq \chi \leq 1$, type I and type II errors can be:

$$E_I = \mathbb{E}_{\mathcal{M}(\mathcal{C})} [\chi], \quad E_{II} = 1 - \mathbb{E}_{\mathcal{M}(\mathcal{C}')} [\chi], \quad (6)$$

137 Here, we abuse $\mathcal{M}(\mathcal{C})$ to represent its probability distribution. To measure the fine-grained relation-
138 ships between these two testing errors, f -DP is introduced.

139 **Definition 3 (Trade-off function)** For any two probability distributions P and Q , the trade-off
140 function is defined as: $T(P; Q)(\gamma) = \inf \{1 - \mathbb{E}_Q [\chi] \mid \mathbb{E}_P [\chi] \leq \gamma\}$, where the infimum is taken
141 over all measurable rejection rules.

142 $T(P; Q)(\gamma)$ is convex, continuous, and non-increasing. For any possible rejection rules, it satisfies
143 $T(P; Q)(\gamma) \leq 1 - \gamma$. It functions as the clear boundary between the achievable and unachievable
144 selections of type I and type II errors, essentially distinguishing the difficulties between these two
145 hypotheses. This relevant statistical property provides a stricter definition of privacy, which mitigates
146 the excessive relaxation of privacy based on composition analysis in existing approaches.

147 **Definition 4 (f -DP and GDP)** A mechanism $\mathcal{M}$ is f -DP if $T(\mathcal{M}(\mathcal{C}), \mathcal{M}(\mathcal{C}'))(\gamma) \geq f(\gamma)$ for all
148 possible adjacent datasets $\mathcal{C}$ and $\mathcal{C}'$. When f measures two Gaussian distributions, namely Gaussian-
149 DP (GDP), denoted as $T_G(\mu)(\gamma) \triangleq T(\mathcal{N}(0, 1), \mathcal{N}(\mu, 1))(\gamma)$ for $\mu \geq 0$.

150 According to the definition, the explicit representation of GDP is $T_G(\mu)(\gamma) = \Phi(\Phi^{-1}(1 - \gamma) - \mu)$
151 where Φ denotes the standard Gaussian CDF. Any single sampling mechanism that introduces Gaus-
152 sian noises can be considered as an exact GDP, which monotonically decreases when μ increases.

153 4 Convergent Privacy

154 In this section, we primarily demonstrate how to provide the worst privacy in FL-DP and its convergent
155 bound. Generally, we assume that local objectives satisfy smoothness with a constant L ,

156 **Assumption 1** Each local objective function $f_i(\cdot)$ satisfies L -smoothness, i.e.,

$$\|\nabla f_i(w_1) - \nabla f_i(w_2)\| \leq L \|w_1 - w_2\|. \quad (7)$$

Table 2: Specific formulation of ρ_t and γ_t in Theorem 2.

	Learning rate	ρ_t	γ_t
Noisy-FedAvg	μ	$(1 + \mu L)^K$	$\frac{2\mu V}{m} K$
	$\frac{\mu}{k+1}$	$(1 + K)^{c\mu L}$	$\frac{2cV}{m} \ln(K + 1)$
	$\frac{\mu}{t+1}$	$\left(1 + \frac{\mu L}{t+1}\right)^K$	$\frac{2\mu V}{m} \frac{K}{t+1}$
	$\frac{\mu}{tK+k+1}$	$\left(\frac{t+2}{t+1}\right)^{z\mu L}$	$\frac{2zV}{m} \ln\left(\frac{t+2}{t+1}\right)$
Noisy-FedProx	non-increase	$\frac{\alpha}{\alpha-L}$	$\frac{2V}{m\alpha}$

Remark 2.1 The result in Eq.(11) aligns with the intuition of designing the splitting operators. It can be observed that the coefficient ρ_t is consistently greater than 1, which is a typical characteristic of non-convexity. It also implies that the sensitivity upper bound tends to diverge as $t \rightarrow \infty$. However, in Eq.(10), the parameters $0 \leq \lambda_t \leq 1$ can efficiently scale the sensitivity terms. By carefully selecting the optimal λ_t values, it can ultimately achieve a convergent privacy lower bound.

4.3 Minimization Problem on t_0 and Its Relaxation

According to Eq.(10) and the sensitivity bound in Eq.(11), we denote the weighted accumulation of the sensitivity term as $\mathcal{H}(\lambda_t, t_0)$, where λ_t and t_0 are both to-be-optimized parameters. Therefore, we can provide the tight bound of the privacy by solving the minimization of the following problem:

$$\mathcal{H}_* = \min_{\lambda_t, t_0} \mathcal{H}(\lambda_t, t_0) \triangleq \sum_{t=t_0}^{T-1} \lambda_{t+1}^2 (\rho_t \|w_t - \tilde{w}_t\| + \gamma_t)^2. \quad (12)$$

If t_0 is very small, it means that the introduced stability gap will also be very small. However, consequently, the sensitivity terms will extremely increase due to the accumulation over $T - t_0$ rounds. Conversely, although the accumulated error is small, it remains divergent due to the unbounded global sensitivity term. To avoid this uncertain analysis, we have to make a compromise. Because t_0 is an integer belonging to $[0, T - 1]$, its optimal selection certainly exists when T is given. Therefore, we consider a relaxed and simple problem instead, i.e. under $t_0 = 0$,

$$\mathcal{H}_0 = \min_{\lambda_t} \mathcal{H}(\lambda_t, 0) = \sum_{t=0}^{T-1} \lambda_{t+1}^2 (\rho_t \|w_t - \tilde{w}_t\| + \gamma_t)^2. \quad (13)$$

Its advantage lies in the fact that when $t_0 = 0$, the sensitivity error is 0, avoiding its divergence. Compared to the optimal solution $\mathcal{H}_*$, it satisfies $\mathcal{H}_0 \geq \mathcal{H}_*$. More importantly, the solution of $\mathcal{H}_0$ eliminates the influence of t_0 , allowing us to obtain an effective solution to the minimization problem by directly minimizing the λ_t terms. The lower bound in Theorem 1 will be replaced by:

$$T(w_T; w'_T) \geq T_G \left(\frac{\sqrt{m\mathcal{H}_*}}{\sigma} \right) \geq T_G \left(\frac{\sqrt{m\mathcal{H}_0}}{\sigma} \right). \quad (14)$$

Although this is a relaxation of the privacy lower bound, our subsequent proof confirms that $\mathcal{H}_0$ can still achieve convergent into a constant form, which means local privacy can still achieve convergence.

4.4 Convergent Privacy

In this part, we demonstrate our convergent privacy analysis. By solving Eq.(13) under corresponding ρ_t and γ_t , we provide the worst privacy for the Noisy-FedAvg and Noisy-FedProx methods.

Theorem 3 Let $f_i(w)$ be a L -smooth and non-convex local objective and local updates be performed as shown in Eq.(2). Under perturbations of isotropic noises $n_i \sim \mathcal{N}(0, \sigma^2 I_d)$, the worst privacy of the Noisy-FedAvg method achieves:

(a) under constant learning rates $\eta_{k,t} = \mu$:

$$T(w_T; w'_T) \geq T_G \left(\frac{2\mu V K}{\sqrt{m}\sigma} \sqrt{\frac{(1 + \mu L)^K + 1}{(1 + \mu L)^K - 1} \frac{(1 + \mu L)^{KT} - 1}{(1 + \mu L)^{KT} + 1}} \right). \quad (15)$$

215 (b) under cyclically decaying $\eta_{k,t} = \frac{\mu}{k+1}$:

$$T(w_T; w'_T) \geq T_G \left(\frac{2cV \ln(K+1)}{\sqrt{m}\sigma} \sqrt{\frac{(1+K)^{c\mu L} + 1}{(1+K)^{c\mu L} - 1} \frac{(1+K)^{c\mu LT} - 1}{(1+K)^{c\mu LT} + 1}} \right). \quad (16)$$

216 (c) under stage-wise decaying $\eta_{k,t} = \frac{\mu}{t+1}$:

$$T(w_T; w'_T) > T_G \left(\frac{2\mu V K}{\sqrt{m}\sigma} \sqrt{2 - \frac{1}{T}} \right). \quad (17)$$

217 (d) under continuously decaying $\eta_{k,t} = \frac{\mu}{tK+k+1}$:

$$T(w_T; w'_T) > T_G \left(\frac{2zV}{\sqrt{m}\sigma} \sqrt{2 - \frac{1}{T}} \right). \quad (18)$$

218 **Remark 3.1** Theorem 3 provides the worst-case privacy analysis for the *Noisy-FedAvg* method.
 219 Its privacy is primarily affected by the clipping norm V , the local interval K , the scale m , and
 220 the noise intensity σ . A larger gradient clipping norm V always results in larger gaps. The local
 221 interval K determines the sensitivity of the entire local process, which is primarily influenced by the
 222 learning rate strategy. m in our proof represents the client scale; in fact, the number of data samples
 223 is also proportional to m . An increased m will largely reduce the sensitivity, yielding improvements
 224 in privacy. The impact of noise intensity σ is also very intuitive. Infinite noise can provide perfect
 225 privacy, while zero noise offers no privacy. Constant-level noise can still achieve convergent privacy.

226 **Theorem 4** Let $f_i(w)$ be a L -smooth and non-convex local objective and local updates be performed
 227 as shown in Eq.(4). Let the proximal coefficient $\alpha > L$ and $\eta < \frac{1}{\alpha-L}$, under perturbations of
 228 isotropic noises $n_i \sim \mathcal{N}(0, \sigma^2 I_d)$, the worst privacy of the *Noisy-FedProx* method achieves:

$$T(w_T; w'_T) \geq T_G \left(\frac{2V}{\sqrt{m}\alpha\sigma} \sqrt{\frac{2\alpha-L}{L} \left(1 - \frac{2}{\left(\frac{\alpha}{\alpha-L} \right)^T + 1} \right)} \right), \quad (19)$$

229 **Remark 4.1** Aside from the influence of standard coefficients, due to the correction of the regulariza-
 230 tion term, its privacy is no longer affected by the local interval K , even with a constant learning rate,
 231 which becomes a significant advantage of the *Noisy-FedProx* method. Specifically, when $\alpha > L$,
 232 increasing α significantly improves the worst privacy, which achieves $\mathcal{O}\left(\frac{V}{\sigma\sqrt{m\alpha L}}\right)$ distinguishability
 233 in GDP. Therefore, the selection of α is a delicate trade-off between optimization and privacy. By
 234 selecting a proper $\alpha > L$, it enables a win-win outcome for both optimization and privacy.

235 **Theoretical comparisons.** Table 3 demonstrates the comparison between existing theoretical results
 236 and ours of the *Noisy-FedAvg* method. Existing analyses are mostly based on the DP relaxations of
 237 (ϵ, δ) -DP and RDP [Mironov, 2017]. Apart from the looseness in their DP definition, an important
 238 weakness is that privacy amplification on composition is entirely loose. For instance, the general
 239 amplification in (ϵ, δ) -DP indicates, the composition of an (ϵ_1, δ_1) -DP and an (ϵ_2, δ_2) -DP leads to
 240 an $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$ -DP. Similarly, the composition of a (ζ, ϵ_1) -RDP and a (ζ, ϵ_2) -RDP results in a
 241 $(\zeta, \epsilon_1 + \epsilon_2)$ -RDP. This simple parameter addition mechanism directly leads to a linear amplification
 242 of the privacy budget. Therefore, in previous works, when achieving specific DP guarantees, it is
 243 often required that the noise intensity σ^2 is proportional to the communication rounds T (or TK).
 244 Wei et al. [2020] prove a double-noisy single-step local training on both client and server sides is
 245 possible to achieve the privacy amplification of $\mathcal{O}(T^2)$ rate. Shi et al. [2021] further consider the
 246 local intervals K . Zhang et al. [2021b] and Noble et al. [2022] elevate the theoretical results to
 247 $\mathcal{O}(TK)$. Subsequent research further indicates that the impact of the interval K can be eliminated to
 248 achieve $\mathcal{O}(T)$ rate via sparsified perturbation [Hu et al., 2023, Cheng et al., 2022], and algorithmic
 249 improvements [Fukami et al., 2024]. However, these conclusions all indicate that the condition for
 250 achieving constant privacy guarantees is to continually increase the noise intensity. Bastianello et al.
 251 [2024] provide constant privacy under β -strongly convex objectives.

Table 3: Comparisons with the existing theoretical results in FL-DP. We losslessly transfer our results into (ϵ, δ) -DP and RDP results. In (ϵ, δ) -DP, we compare the requirement of noise variance corresponding to achieving (ϵ, δ) -DP. In (ζ, ϵ) -RDP, we directly compare the privacy budget term $\delta(\zeta)$. We mainly focus on the privacy changes on T and K . $\Omega(\cdot)$, $\mathcal{O}(\cdot)$, and $o(\cdot)$ correspond to the lower, upper bound, and not tight upper bound of the complexity, respectively.

	(ϵ, δ) -DP	(ζ, ϵ) -RDP	when $T, K \rightarrow \infty$
Wei et al. [2020]	$\sigma = \mathcal{O}\left(\frac{V}{\epsilon m}\sqrt{T^2 - mL^2}\right)$	-	
Shi et al. [2021]	$\sigma = \mathcal{O}\left(\frac{V\sqrt{\log(\frac{1}{\delta})}}{\epsilon}T\sqrt{K}\right)$	-	
Zhang et al. [2021b]	$\sigma = \mathcal{O}\left(\frac{V\sqrt{\log(\frac{1}{\delta})}}{\epsilon m}\sqrt{T + mK}\right)$	-	
Noble et al. [2022]	$\sigma = \Omega\left(\frac{V\sqrt{\log(\frac{2T}{\delta})}}{\epsilon\sqrt{m}}\sqrt{TK}\right)$	-	$\sigma \rightarrow \infty$ on non-convex
Cheng et al. [2022]	$\sigma = \Omega\left(\frac{V\sqrt{\log(\frac{1}{\delta})}}{\epsilon}\sqrt{T}\right)$	-	
Zhang and Tang [2022]	-	$\epsilon = \Omega\left(\frac{\zeta V^2}{\sigma^2}TK\right)$	
Hu et al. [2023]	$\sigma = \Omega\left(\frac{V\sqrt{\epsilon + 2\log(\frac{1}{\delta})}}{\epsilon}\sqrt{T}\right)$	-	
Fukami et al. [2024]	$\sigma = \Omega\left(\frac{V(1+\sqrt{1+\epsilon})\sqrt{\log(e+\frac{1}{\delta})}}{\epsilon}\sqrt{T}\right)$	-	
Bastianello et al. [2024]	-	$\epsilon = \mathcal{O}\left(\frac{\zeta LV^2}{\beta^2 \sigma^2}(1 - e^{-\beta T})\right)$	convergent on β -strongly convex
Ours (Noisy-FedAvg)	$\sigma = o\left(\frac{V\sqrt{(\Phi^{-1}(\delta))^2 + 4\epsilon}}{\epsilon\sqrt{m}}\sqrt{2 - \frac{1}{T}}\right)$	$\epsilon = \mathcal{O}\left(\frac{\zeta V^2}{m\sigma^2}\left(2 - \frac{1}{T}\right)\right)$	convergent on non-convex

5 Empirical Validation

Setups. We conduct experiments on MNIST [LeCun et al., 1998] and CIFAR-10 [Krizhevsky et al., 2009] with the LeNet-5 [LeCun et al., 1998] and ResNet-18 [He et al., 2016] models. We follow the widely used standard federated learning experimental setups to introduce heterogeneity by the Dirichlet splitting. The heterogeneity level is set high (Dir-0.1 splitting).

Accuracy. Table 4 shows the comparison on Noisy-FedAvg. Our theory precisely addresses this misconception and rigorously provides its privacy protection performance. It can be observed that as the number of clients increases, the impact of noise gradually diminishes. We have previously explained this principle: for the globally averaged model, the more noise involved in the averaging process, the closer it gets to the noise mean, which is akin to the situation without noise interference. When we adjust the intensity from $\sigma = 10^{-3}$ to 10^{-1} , the accuracy decreases by 5.57% and 1.62% on $m = 20$ and 100 respectively on the MNIST and 14.19% and 11% on the CIFAR-10. The local interval K does not significantly affect noise, and the accuracy drops consistently. K primarily affects global sensitivity and higher aggregation frequency usually means better performance.

Sensitivity in Noisy-FedAvg. We mainly study the impact from the scale m , local interval K , and clipping norm V , as shown in Fig. 2. The first figure clearly demonstrates the impact of the scale m on sensitivity, which corresponds to the worst privacy bound $\mathcal{O}\left(\frac{1}{\sqrt{m}}\right)$. More clients generally imply stronger global privacy. The second figure shows evident that although increasing K can raise the sensitivity during the process, it does not alter the upper bound of sensitivity after optimization converges. This is entirely consistent with our analysis, indicating that the privacy lower bound exists and is unaffected by T and K . The third figure indicates that the sensitivity will be affected by the V , which corresponds to the worst privacy bound $\mathcal{O}(V)$.

Table 4: Comparison of the accuracy under different experimental settings. We select the scale m from $[50, 100]$. Each client holds 600 heterogeneous data samples of MNIST or 500 heterogeneous data samples of CIFAR-10. For each scale, we test two settings of the local interval $K = 50, 100$, and 200, respectively. Throughout the entire process, we fix $TK = 30000$. “-” means the training loss diverges. Each result is repeated 5 times to compute its mean and variance.

		Noisy Intensity	$m = 50$			$m = 100$		
			$K = 50$	$K = 100$	$K = 200$	$K = 50$	$K = 100$	$K = 200$
MNIST LeNet-5	$\sigma = 1.0$		-	-	-	-	-	-
	$\sigma = 10^{-1}$		95.40 $\pm$ 0.18	95.42 $\pm$ 0.15	95.21 $\pm$ 0.11	97.32 $\pm$ 0.14	97.50 $\pm$ 0.11	97.42 $\pm$ 0.18
	$\sigma = 10^{-2}$		98.33 $\pm$ 0.12	98.02 $\pm$ 0.15	97.88 $\pm$ 0.12	98.71 $\pm$ 0.10	97.97 $\pm$ 0.08	97.72 $\pm$ 0.12
	$\sigma = 10^{-3}$		98.41 $\pm$ 0.07	98.23 $\pm$ 0.03	98.00 $\pm$ 0.07	98.94 $\pm$ 0.04	98.50 $\pm$ 0.06	98.01 $\pm$ 0.10
CIFAR-10 ResNet-18	$\sigma = 1.0$		-	-	-	-	-	-
	$\sigma = 10^{-1}$		53.76 $\pm$ 0.25	53.38 $\pm$ 0.23	53.49 $\pm$ 0.21	62.02 $\pm$ 0.28	61.33 $\pm$ 0.25	61.11 $\pm$ 0.17
	$\sigma = 10^{-2}$		70.11 $\pm$ 0.22	69.08 $\pm$ 0.12	66.63 $\pm$ 0.16	74.34 $\pm$ 0.29	72.87 $\pm$ 0.19	70.74 $\pm$ 0.15
	$\sigma = 10^{-3}$		70.98 $\pm$ 0.11	69.81 $\pm$ 0.20	67.98 $\pm$ 0.03	75.38 $\pm$ 0.19	74.44 $\pm$ 0.12	72.11 $\pm$ 0.06

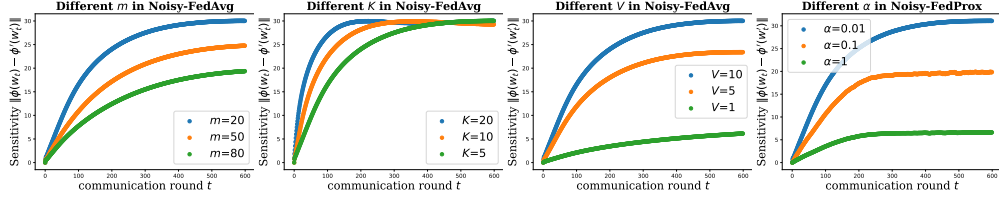


Figure 2: Sensitivity studies on Noisy-FedAvg and Noisy-FedProx. The general setups are $m = 20$, $K = 5$, and $V = 10$. In each group, we keep all other parameters fixed to ensure fairness.

274 **Sensitivity in Noisy-FedProx.** As
275 shown in Fig. 2 (the fourth figure),
276 the larger α means smaller global sensi-
277 tivity. This is consistent with our
278 analysis, which states that the lower
279 bound of privacy performance is given
280 by $\mathcal{O}\left(\frac{1}{\sqrt{\alpha}}\right)$. When we select $\alpha =$
281 0, it degrades to the Noisy-FedAvg
282 method. In fact, based on the comparison, we can see that when α is sufficiently small, i.e. $\alpha = 0.01$,
283 its global sensitivity is almost at the same level as Noisy-FedAvg. In Table 5, we present a compar-
284 ison between them. Although the proximal term provides limited improvement in accuracy, selecting
285 an appropriate α significantly reduces global sensitivity. This implies that the privacy performance of
286 Noisy-FedProx is far superior to that of Noisy-FedAvg. While achieving similar performance, the
287 regularization proxy term can significantly reduce the global sensitivity of the output model, thereby
288 enhancing privacy. This conclusion also demonstrates the superiority on privacy of a series of FL-DP
289 optimization methods based on training with this regularization approach.

Table 5: Performance and sensitivity ($T = 600$).

	Accuracy	Sensitivity
Noisy-FedAvg	60.67	31.33
Noisy-FedProx $\alpha = 0.01$	60.69	30.97
Noisy-FedProx $\alpha = 0.1$	60.94	18.52
Noisy-FedProx $\alpha = 1$	56.33	6.34

6 Summary

291 To our best knowledge, this paper is the first work to demonstrate convergent privacy for the general
292 FL-DP paradigms. We comprehensively study and illustrate the fine-grained privacy level for
293 Noisy-FedAvg and Noisy-FedProx methods based on f -DP analysis, an information-theoretic
294 lossless DP definition. Moreover, we conduct comprehensive analysis with existing work on other DP
295 frameworks and highlight the long-term cognitive bias of the privacy lower bound. Our analysis fills
296 the theoretical gap in the convergent privacy of FL-DP while further providing a reliable theoretical
297 guarantee for its privacy protection performance. Moreover, We conduct a series of experiments to
298 verify the boundedness of global sensitivity and its influence on different variables, further validating
299 that our theoretical analysis aligns more closely with practical scenarios.

References

- Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.
- Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*.
- Jason Altschuler and Kunal Talwar. Privacy of noisy stochastic gradient descent: More iterations without more privacy loss. *Advances in Neural Information Processing Systems*, 35:3788–3800, 2022.
- Jason M Altschuler, Jinho Bok, and Kunal Talwar. On the privacy of noisy stochastic gradient descent for convex optimization. *SIAM Journal on Computing*, 53(4):969–1001, 2024.
- Pathum Chamikara Mahawaga Arachchige, Peter Bertok, Ibrahim Khalil, Dongxi Liu, Seyit Camtepe, and Mohammed Atiquzzaman. Local differential privacy for deep learning. *IEEE Internet of Things Journal*, 7(7):5827–5842, 2019.
- Nicola Bastianello, Changxin Liu, and Karl H Johansson. Enhancing privacy in federated learning through local training. *arXiv preprint arXiv:2403.17572*, 2024.
- Jinho Bok, Weijie Su, and Jason M Altschuler. Shifted interpolation for differential privacy. *arXiv preprint arXiv:2403.00278*, 2024.
- Mingzhe Chen, Nir Shlezinger, H Vincent Poor, Yonina C Eldar, and Shuguang Cui. Communication-efficient federated learning. *Proceedings of the National Academy of Sciences*, 118(17):e2024789118, 2021.
- Shuzhen Chen, Dongxiao Yu, Yifei Zou, Jiguo Yu, and Xiuzhen Cheng. Decentralized wireless federated learning with differential privacy. *IEEE Transactions on Industrial Informatics*, 18(9):6273–6282, 2022.
- Anda Cheng, Peisong Wang, Xi Sheryl Zhang, and Jian Cheng. Differentially private federated learning with local regularization and sparsification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10122–10131, 2022.
- Rishav Chourasia, Jiayuan Ye, and Reza Shokri. Differential privacy dynamics of langevin diffusion and noisy gradient descent. *Advances in Neural Information Processing Systems*, 34:14771–14781, 2021.
- Rong Dai, Li Shen, Fengxiang He, Xinmei Tian, and Dacheng Tao. Displf: Towards communication-efficient personalized federated learning via decentralized sparse training. In *International conference on machine learning*, pages 4587–4604. PMLR, 2022.
- Rudrajit Das, Anish Acharya, Abolfazl Hashemi, Sujay Sanghavi, Inderjit S Dhillon, and Ufuk Topcu. Faster non-convex federated learning via global and local momentum. In *Uncertainty in Artificial Intelligence*, pages 496–506. PMLR, 2022.
- Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):3–37, 2022.
- Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12. Springer, 2006.
- Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Advances in Cryptology-EUROCRYPT 2006: 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28-June 1, 2006. Proceedings 25*, pages 486–503. Springer, 2006a.
- 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, 2006b.
- 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.
- Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 650–669. IEEE, 2015.

348 Takumi Fukami, Tomoya Murata, Kenta Niwa, and Iifan Tyou. Dp-norm: Differential privacy primal-dual
349 algorithm for decentralized federated learning. *IEEE Transactions on Information Forensics and Security*,
350 2024.

351 Yuanyuan Gao, Lei Zhang, Lulu Wang, Kim-Kwang Raymond Choo, and Rui Zhang. Privacy-preserving and
352 reliable decentralized federated learning. *IEEE Transactions on Services Computing*, 16(4):2879–2891, 2023.

353 Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients-how easy is it
354 to break privacy in federated learning? *Advances in neural information processing systems*, 33:16937–16947,
355 2020.

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

358 Antonious Girgis, Deepesh Data, Suhas Diggavi, Peter Kairouz, and Ananda Theertha Suresh. Shuffled model of
359 differential privacy in federated learning. In *International Conference on Artificial Intelligence and Statistics*,
360 pages 2521–2529. PMLR, 2021.

361 Eduard Gorbunov, Filip Hanzely, and Peter Richtárik. Local sgd: Unified theory and new efficient methods. In
362 *International Conference on Artificial Intelligence and Statistics*, pages 3556–3564. PMLR, 2021.

363 Michał Grudzień, Grigory Malinovsky, and Peter Richtárik. Can 5th generation local training methods support
364 client sampling? yes! In *International Conference on Artificial Intelligence and Statistics*, pages 1055–1092.
365 PMLR, 2023.

366 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In
367 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

368 Zaobo He, Lintao Wang, and Zhipeng Cai. Clustered federated learning with adaptive local differential privacy
369 on heterogeneous iot data. *IEEE Internet of Things Journal*, 2023.

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

372 Rui Hu, Yuanxiong Guo, and Yanmin Gong. Federated learning with sparsified model perturbation: Improving
373 accuracy under client-level differential privacy. *IEEE Transactions on Mobile Computing*, 2023.

374 Divyansh Jhunjhunwala, Pranay Sharma, Aushim Nagarkatti, and Gauri Joshi. Fedvarp: Tackling the variance
375 due to partial client participation in federated learning. In *Uncertainty in Artificial Intelligence*, pages 906–916.
376 PMLR, 2022.

377 Sanxiu Jiao, Lecai Cai, Xinjie Wang, Kui Cheng, and Xiang Gao. A differential privacy federated learning
378 scheme based on adaptive gaussian noise. *CMES-Computer Modeling in Engineering & Sciences*, 138(2),
379 2024.

380 Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In
381 *International conference on machine learning*, pages 1376–1385. PMLR, 2015.

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

385 Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha
386 Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International conference on*
387 *machine learning*, pages 5132–5143. PMLR, 2020.

388 Prashant Khanduri, Pranay Sharma, Haibo Yang, Mingyi Hong, Jia Liu, Ketan Rajawat, and Pramod Varshney.
389 Stem: A stochastic two-sided momentum algorithm achieving near-optimal sample and communication
390 complexities for federated learning. *Advances in Neural Information Processing Systems*, 34:6050–6061,
391 2021.

392 Muah Kim, Onur Günlü, and Rafael F Schaefer. Federated learning with local differential privacy: Trade-offs
393 between privacy, utility, and communication. In *ICASSP 2021-2021 IEEE International Conference on*
394 *Acoustics, Speech and Signal Processing (ICASSP)*, pages 2650–2654. IEEE, 2021.

395 Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

396 Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document
397 recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

398 Bo Li, Mikkel N Schmidt, Tommy S Alstrøm, and Sebastian U Stich. On the effectiveness of partial variance
399 reduction in federated learning with heterogeneous data. In *Proceedings of the IEEE/CVF Conference on*
400 *Computer Vision and Pattern Recognition*, pages 3964–3973, 2023.

401 Tao Lin, Sebastian Urban Stich, Kumar Kshitij Patel, and Martin Jaggi. Don’t use large mini-batches, use local
402 sgd. In *Proceedings of the 8th International Conference on Learning Representations*, 2019.

403 Jie Ling, Junchang Zheng, and Jiahui Chen. Efficient federated learning privacy preservation method with
404 heterogeneous differential privacy. *Computers & Security*, 139:103715, 2024.

405 Wei Liu, Li Chen, Yunfei Chen, and Wenyi Zhang. Accelerating federated learning via momentum gradient
406 descent. *IEEE Transactions on Parallel and Distributed Systems*, 31(8):1754–1766, 2020.

407 Andrew Lowy, Ali Ghafelebashi, and Meisam Razaviyayn. Private non-convex federated learning without a
408 trusted server. In *International Conference on Artificial Intelligence and Statistics*, pages 5749–5786. PMLR,
409 2023.

410 Grigory Malinovsky, Kai Yi, and Peter Richtárik. Variance reduced proxskip: Algorithm, theory and application
411 to federated learning. *Advances in Neural Information Processing Systems*, 35:15176–15189, 2022.

412 Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-
413 efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages
414 1273–1282. PMLR, 2017.

415 Matias Mendieta, Taojiannan Yang, Pu Wang, Minwoo Lee, Zhengming Ding, and Chen Chen. Local learning
416 matters: Rethinking data heterogeneity in federated learning. In *Proceedings of the IEEE/CVF Conference on*
417 *Computer Vision and Pattern Recognition*, pages 8397–8406, 2022.

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

420 Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtárik. Proxskip: Yes! local
421 gradient steps provably lead to communication acceleration! finally! In *International Conference on Machine*
422 *Learning*, pages 15750–15769. PMLR, 2022.

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

426 Maxence Noble, Aurélien Bellet, and Aymeric Dieuleveut. Differentially private federated learning on heteroge-
427 neous data. In *International Conference on Artificial Intelligence and Statistics*, pages 10110–10145. PMLR,
428 2022.

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

432 Nuria Rodríguez-Barroso, Goran Stipich, Daniel Jiménez-López, José Antonio Ruiz-Millán, Eugenio Martínez-
433 Cámara, Gerardo González-Seco, M Victoria Luzón, Miguel Angel Veganzones, and Francisco Herrera.
434 Federated learning and differential privacy: Software tools analysis, the sherpa. ai fl framework and method-
435 ological guidelines for preserving data privacy. *Information Fusion*, 64:270–292, 2020.

436 Lu Shi, Jiangang Shu, Weizhe Zhang, and Yang Liu. Hfl-dp: Hierarchical federated learning with differential
437 privacy. In *2021 IEEE Global Communications Conference (GLOBECOM)*, pages 1–7. IEEE, 2021.

438 Yifan Shi, Yingqi Liu, Kang Wei, Li Shen, Xueqian Wang, and Dacheng Tao. Make landscape flatter in
439 differentially private federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision*
440 *and Pattern Recognition*, pages 24552–24562, 2023.

441 Nir Shlezinger, Mingzhe Chen, Yonina C Eldar, H Vincent Poor, and Shuguang Cui. Uveqfed: Universal vector
442 quantization for federated learning. *IEEE Transactions on Signal Processing*, 69:500–514, 2020.

443 Sebastian Urban Stich. Local sgd converges fast and communicates little. In *ICLR 2019-International Conference*
444 *on Learning Representations*, 2019.

445 Yan Sun, Li Shen, Shixiang Chen, Liang Ding, and Dacheng Tao. Dynamic regularized sharpness aware
446 minimization in federated learning: Approaching global consistency and smooth landscape. In *International*
447 *Conference on Machine Learning*, pages 32991–33013. PMLR, 2023a.

448 Yan Sun, Li Shen, Tiansheng Huang, Liang Ding, and Dacheng Tao. Fedsspeed: Larger local interval, less
449 communication round, and higher generalization accuracy. *arXiv preprint arXiv:2302.10429*, 2023b.

450 Yan Sun, Li Shen, Hao Sun, Liang Ding, and Dacheng Tao. Efficient federated learning via local adaptive
451 amended optimizer with linear speedup. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45
452 (12):14453–14464, 2023c.

453 Yan Sun, Li Shen, and Dacheng Tao. Understanding how consistency works in federated learning via stage-wise
454 relaxed initialization. *Advances in Neural Information Processing Systems*, 36, 2024.

455 Han Wang, Siddhartha Marella, and James Anderson. Fedadmm: A federated primal-dual algorithm allowing
456 partial participation. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 287–294. IEEE,
457 2022.

458 Yansheng Wang, Yongxin Tong, and Dingyuan Shi. Federated latent dirichlet allocation: A local differential
459 privacy based framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages
460 6283–6290, 2020.

461 Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American*
462 *Statistical Association*, 105(489):375–389, 2010.

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

466 Kang Wei, Jun Li, Ming Ding, Chuan Ma, Hang Su, Bo Zhang, and H Vincent Poor. User-level privacy-preserving
467 federated learning: Analysis and performance optimization. *IEEE Transactions on Mobile Computing*, 21(9):
468 3388–3401, 2021.

469 Kang Wei, Jun Li, Chuan Ma, Ming Ding, Wen Chen, Jun Wu, Meixia Tao, and H Vincent Poor. Personalized
470 federated learning with differential privacy and convergence guarantee. *IEEE Transactions on Information*
471 *Forensics and Security*, 2023.

472 Thorsten Wittkopp and Alexander Acker. Decentralized federated learning preserves model and data privacy. In
473 *International Conference on Service-Oriented Computing*, pages 176–187. Springer, 2020.

474 Blake Woodworth, Kumar Kshitij Patel, Sebastian Stich, Zhen Dai, Brian Bullins, Brendan McMahan, Ohad
475 Shamir, and Nathan Srebro. Is local sgd better than minibatch sgd? In *International Conference on Machine*
476 *Learning*, pages 10334–10343. PMLR, 2020.

477 Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan, Vladimir Braverman, and Zhanxing Zhu. On the noisy
478 gradient descent that generalizes as sgd. In *International Conference on Machine Learning*, pages 10367–
479 10376. PMLR, 2020.

480 Xiang Wu, Yongting Zhang, Minyu Shi, Pei Li, Ruirui Li, and Neal N Xiong. An adaptive federated learning
481 scheme with differential privacy preserving. *Future Generation Computer Systems*, 127:362–372, 2022.

482 Jian Xu, Xinyi Tong, and Shao-Lun Huang. Personalized federated learning with feature alignment and classifier
483 collaboration. In *The Eleventh International Conference on Learning Representations*.

484 Ge Yang, Shaowei Wang, and Haijie Wang. Federated learning with personalized local differential privacy. In
485 *2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS)*, pages 484–489.
486 IEEE, 2021.

487 Xinyu Yang and Weisan Wu. A federated learning differential privacy algorithm for non-gaussian heterogeneous
488 data. *Scientific Reports*, 13(1):5819, 2023.

489 Xiyuan Yang, Wenke Huang, and Mang Ye. Dynamic personalized federated learning with adaptive differential
490 privacy. *Advances in Neural Information Processing Systems*, 36:72181–72192, 2023.

491 Xin Yao, Tianchi Huang, Chenglei Wu, Ruixiao Zhang, and Lifeng Sun. Towards faster and better federated
492 learning: A feature fusion approach. In *2019 IEEE International Conference on Image Processing (ICIP)*,
493 pages 175–179. IEEE, 2019.

494 Jiayuan Ye and Reza Shokri. Differentially private learning needs hidden state (or much faster convergence).
495 *Advances in Neural Information Processing Systems*, 35:703–715, 2022.

496 Lin Zhang, Yong Luo, Yan Bai, Bo Du, and Ling-Yu Duan. Federated learning for non-iid data via unified feature
497 learning and optimization objective alignment. In *Proceedings of the IEEE/CVF international conference on*
498 *computer vision*, pages 4420–4428, 2021a.

- 499 Meng Zhang, Ermin Wei, and Randall Berry. Faithful edge federated learning: Scalability and privacy. *IEEE*
500 *Journal on Selected Areas in Communications*, 39(12):3790–3804, 2021b.
- 501 Xinwei Zhang, Mingyi Hong, Sairaj Dhople, Wotao Yin, and Yang Liu. Fedpd: A federated learning framework
502 with adaptivity to non-iid data. *IEEE Transactions on Signal Processing*, 69:6055–6070, 2021c.
- 503 Yaling Zhang and Dongtai Tang. A differential privacy federated learning framework for accelerating con-
504 vergence. In *2022 18th International Conference on Computational Intelligence and Security (CIS)*, pages
505 122–126. IEEE, 2022.
- 506 Jingwen Zhao, Yunfang Chen, and Wei Zhang. Differential privacy preservation in deep learning: Challenges,
507 opportunities and solutions. *IEEE Access*, 7:48901–48911, 2019.
- 508 Qinqing Zheng, Shuxiao Chen, Qi Long, and Weijie Su. Federated f-differential privacy. In *International*
509 *conference on artificial intelligence and statistics*, pages 2251–2259. PMLR, 2021.
- 510 Linghui Zhu, Xinyi Liu, Yiming Li, Xue Yang, Shu-Tao Xia, and Rongxing Lu. A fine-grained differentially
511 private federated learning against leakage from gradients. *IEEE Internet of Things Journal*, 9(13):11500–
512 11512, 2021.

Limitations and Broader Impacts. Our paper provides the first convergence-privacy analysis for the FL framework. The current analysis primarily includes the impact of multi-step updates on local nodes and the effect of multi-clients aggregation on the privacy bounds. A limitation of this paper is the inability to directly extend the privacy analysis to stability analysis. Stability analysis of convergence has always been a crucial theoretical objective in non-convex optimization. Although the trade-off function constructed by f -DP incorporates certain iterative properties of stability terms, it currently cannot directly derive convergence bounds for stability. Moreover, the theoretical analysis in this paper provides a crucial theoretical basis for privacy preservation, demonstrating that privacy can still be maintained under finite noise and infinitely long learning processes. This implies that many online methods can ensure privacy through cumulative noise accumulation, which may provide valuable guidance for privacy preservation in future engineering applications.

A General FL-DP Framework

FL framework usually allows local clients to train several iterations and then aggregates these optimized local models for global consistency guarantees. Though indirect access to the dataset significantly mitigates the risk of data leakage, vanilla gradients or parameters communicated to the server still bring privacy concerns, i.e. indirect leakage. Thus, DP techniques are introduced by adding isotropic noises on local parameters before communication, to further enhance privacy protection.

Algorithm 1 General FL-DP Framework

Input: initial parameters w_0 , round T , interval K

Output: global parameters w_T

```

1: for  $t = 0, 1, 2, \dots, T - 1$  do
2:   activate local clients and communications
3:   for client  $i \in \mathcal{I}$  in parallel do
4:     set the initialization  $w_{i,0,t} = w_t$ 
5:     for  $k = 0, 1, 2, \dots, K - 1$  do
6:        $w_{i,k+1,t} = L\text{-update}(w_{i,k,t})$ 
7:     end for
8:     generate a noise  $n_i \sim \mathcal{N}(0, \sigma^2 I_d)$ 
9:     communicate  $w_{i,K,t} + n_i$  to the server
10:  end for
11:   $w_{t+1} = G\text{-update}(\{w_{i,K,t} + n_i\})$ 
12: end for

```

In our analysis, we consider the FL-DP framework with the classical normal client-level noises, as shown in Algorithm 1. At the beginning of each communication round t , the server activates local clients and communicates necessary variables. Then local clients begin the training in parallel. We describe this process as a total of $K > 1$ steps of $L\text{-update}$ function updates. Depending on algorithm designs, the specific form of local update functions varies. After training, the local clients enhance local privacy by adding noise perturbations to the uploaded model parameters. Our analysis primarily considers the properties of the isotropic Gaussian noise distribution, i.e. $n_i \sim \mathcal{N}(0, \sigma^2 I_d)$. Then the global server aggregates the noisy parameters to generate the global model w_{t+1} via the $G\text{-update}$ function. Repeat this for T rounds and return w_T as output.

B Preliminary Properties of f -DP

In this section, we mainly supplement some basic properties of f -DP, all of which are lemmas proposed by Dong et al. [2022]. Specifically, Lemmas 1 and 2 are employed in our theoretical analysis, whereas Lemmas 3 and 4 facilitate a lossless translation of our results into other standard DP frameworks for comparative purposes.

Lemma 1 (Post-processing) *If a randomized mechanism $\mathcal{M}$ is f -DP, any post processing mechanism based on $\mathcal{M}$ is still f -DP, i.e. $T(P'; Q') \geq T(P; Q)$ for any post-processing mapping which leads to $P \rightarrow P'$ and $Q \rightarrow Q'$.*

Intuitively, post-processing mappings bring some changes in the original distributions. However, such changes can not allow the updated distributions to be much easier to discern. This lemma also widely exists in other DP relaxations and stands as one of the foundational elements in current privacy analyses. In f -DP, this lemma also clearly demonstrates that the difficulty of hypothesis testing problems can not be simplified with the addition of known information, which still preserves the original distinguishability.

Lemma 2 (Composition) *We have a series of mechanisms $\mathcal{M}_i$ and a joint serial composition mechanism $\mathcal{M}$. Let each private mechanism $\mathcal{M}_i(\cdot, y_1, \dots, y_{i-1})$ be f_i -DP for all $y_1 \in Y_1, \dots, y_{i-1} \in Y_{i-1}$. Then the n -fold composed mechanism $\mathcal{M} : X \rightarrow Y_1 \times \dots \times Y_n$ is $f_1 \otimes \dots \otimes f_n$ -DP, where $\otimes$ denotes the joint distribution. For instance, if $f = T(P; Q)$ and $g = T(P'; Q')$, then $f \otimes g = T(P \times P'; Q \times Q')$.*

The composition in the f -DP framework is *closed* and *tight*. This is also one of the advantages of privacy representation in f -DP. Correspondingly, the advanced composition theorem for (ϵ, δ) -DP can not admit the optimal parameters to exactly capture the privacy in the composition process [Dwork et al., 2015]. However, the trade-off function has an exact probabilistic interpretation and can precisely measure the composition.

Lemma 3 (GDP $\rightarrow (\epsilon, \delta)$ -DP) *A μ -GDP mechanism with a trade-off function $T_G(\mu)$ is also $(\epsilon, \delta(\epsilon))$ -DP for all $\epsilon \geq 0$ where*

$$\delta(\epsilon) = \Phi\left(-\frac{\epsilon}{\mu} + \frac{\mu}{2}\right) - e^\epsilon \Phi\left(-\frac{\epsilon}{\mu} - \frac{\mu}{2}\right). \quad (20)$$

Lemma 4 (GDP $\rightarrow$ RDP) *A μ -GDP mechanism with a trade-off function $T_G(\mu)$ is also $(\zeta, \frac{1}{2}\mu^2\zeta)$ -RDP for any $\zeta > 1$.*

We state the transition and conversion calculations from f -DP (we specifically consider the GDP) to other DP relaxations, e.g. for the (ϵ, δ) -DP and RDP. These lemmas can effectively compare our theoretical results with existing ones. Our comparison primarily aims to demonstrate that the convergent privacy obtained in our analysis would directly derive bounded privacy budgets in other DP relaxations. Moreover, we will illustrate how the convergent f -DP further addresses conclusions that current FL-DP work cannot cover theoretically, which provides solid support for understanding its reliability of privacy protection.

C Proof of Main Theorems

C.1 Proofs of Theorem 1

We consider the general updates on the adjacent datasets $\mathcal{C}$ and $\mathcal{C}'$ on round t as follows:

$$\begin{aligned} w_{t+1} &= \phi(w_t) + \frac{1}{m} \sum_{i \in \mathcal{I}} n_{i,t}, \\ w'_{t+1} &= \phi'(w'_t) + \frac{1}{m} \sum_{i \in \mathcal{I}} n'_{i,t}, \end{aligned} \quad (21)$$

where w_0 is the initial state. $n_{i,t}$ and $n'_{i,t}$ are two noises generated from the normal distribution $\mathcal{N}(0, \sigma^2 I_d)$. To construct the interpolated sequence, we introduce the concentration coefficients λ_t to provide a convex combination of the updates above, which is,

$$\tilde{w}_{t+1} = \lambda_{t+1} \phi(w_t) + (1 - \lambda_{t+1}) \phi'(\tilde{w}_t) + \frac{1}{m} \sum_{i \in \mathcal{I}} n_{i,t}, \quad (22)$$

for $t = t_0, t_0 + 1, \dots, T - 1$. Furthermore, we set $\lambda_T = 1$ to let $\tilde{w}_T = \phi(w_{T-1}) + \frac{1}{m} \sum_{i \in \mathcal{I}} n_{i,T-1} = w_T$, and we add the definition of $\tilde{w}_{t_0} = w'_{t_0}$ as the interpolation beginning. t_0 determines the length of the interpolation sequence.

Lemma 5 *According to the expansion of trade-off functions, for the general updates in Eq.(22), we have the following recurrence relation:*

$$T(\tilde{w}_{t+1}; w'_{t+1}) \geq T(\tilde{w}_t; w'_t) \otimes T_G\left(\frac{\sqrt{m}}{\sigma} \lambda_{t+1} \|\phi(w_t) - \phi'(\tilde{w}_t)\|\right). \quad (23)$$

587 **Proof.** Based on the post-processing and compositions, let z and z' be the corresponding noises
 588 above, for any constant $\lambda \in [0, 1]$, we have (subscripts are temporarily omitted):

$$\begin{aligned}
 & T(\lambda\phi(w) + (1-\lambda)\phi'(\tilde{w}) + z; \phi'(w') + z') \\
 &= T(\phi'(\tilde{w}) + \lambda(\phi(w) - \phi'(\tilde{w})) + z; \phi'(w') + z') \\
 &\geq T((\phi'(\tilde{w}), \lambda(\phi(w) - \phi'(\tilde{w})) + z); (\phi'(w'), z')) \\
 &\geq T(\phi'(\tilde{w}); \phi'(w')) \otimes T(\lambda(\phi(w) - \phi'(\tilde{w})) + z; z') \\
 &\geq T(\tilde{w}; w') \otimes T(\lambda(\phi(w) - \phi'(\tilde{w})) + z; z'),
 \end{aligned}$$

589 where z and z' are two Gaussian noises that can be considered to be sampled from $\mathcal{N}(0, \frac{\sigma^2}{m} I_d)$ (av-
 590 erage of m isotropic Gaussian noises). Therefore, the distinguishability between the first term and
 591 the second term does not exceed the mean shift of the distribution, which is $\|\frac{\sqrt{m}}{\sigma}\lambda(\phi(w) - \phi'(\tilde{w}))\|$.
 592 By taking $w = w_t$ and $\lambda = \lambda_{t+1}$, the proofs are completed.
 593

594 According to the above lemma, by expanding it from $t = t_0$ to $T - 1$ and the factor $T(\tilde{w}_{t_0}; w'_{t_0}) =$
 595 $T_G(0)$, we can prove the formulation in Eq. (10).

596 C.2 Proofs of Theorem 2

597 Lemma 5 provides the general recursive relationship on the global states along the communication
 598 round t . To obtain the lower bound of the trade-off function, we only need to solve for the gaps
 599 $\|\phi(w) - \phi'(\tilde{w})\|$. It is worth noting that the local update process here involves dual replacement of
 600 both the dataset (ϕ and ϕ') and the initial state (w and $\tilde{w}$). Therefore, we measure their maximum
 601 discrepancy by assessing their respective distances to the intermediate variable constructed by the
 602 cross-items:

$$\|\phi(w) - \phi'(\tilde{w})\| \leq \underbrace{\|\phi(w) - \phi'(w)\|}_{\text{Data Sensitivity}} + \underbrace{\|\phi'(w) - \phi'(\tilde{w})\|}_{\text{Model Sensitivity}}. \quad (24)$$

603 The first term measures the disparity in training on different datasets and the second term measures
 604 the gap in training from different initial models. One of our contributions is to provide their general
 605 gaps. In our paper, we expand the update function $\phi(x)$ by considering the multiple local iterations
 606 and federated cross-device settings. By simply setting the local interval to 1 and the number of clients
 607 to 1, our results can easily reproduce the original conclusion in [Bok et al., 2024]. Furthermore, our
 608 comprehensive considerations have led to a new understanding of the impact of local updates on
 609 privacy.

610 $\phi(w_t)$ and $\phi'(w_t)$ begin from w_t . $\phi'(w_t)$ and $\phi'(\tilde{w}_t)$ adopt the data samples $\varepsilon' \in \mathcal{C}'$. We naturally use
 611 $w_{i,k,t}$ and $\tilde{w}_{i,k,t}$ to represent individual states in $\phi(w_t)$ and $\phi'(\tilde{w}_t)$, respectively. **To avoid ambiguity,**
 612 **we define the states in $\phi'(w_t)$ as $\hat{w}_{i,k,t}$.** When $i \neq i^*$, since $\varepsilon = \varepsilon'$, then $w_{i,k,t}$ only differs from
 613 $\hat{w}_{i,k,t}$ on i^* -th client.

614 on the Noisy-FedAvg Method:

615 **Lemma 6 (Data Sensitivity.)** The *data sensitivity* caused by gradient descent steps can be bounded
 616 as:

$$\|\phi(w_t) - \phi'(w_t)\| \leq \frac{2V}{m} \sum_{k=0}^{K-1} \eta_{k,t}, \quad (25)$$

617 where $\eta_{k,t}$ is the learning rate at the k -th iteration of t -th communication round.

618 **Proof.** By directly expanding the update functions ϕ and ϕ' at w_t , we have:

$$\begin{aligned}
 & \|\phi(w_t) - \phi'(w_t)\| \\
 &= \|w_t - \frac{1}{m} \sum_{i \in \mathcal{I}} \sum_{k=0}^{K-1} \eta_{k,t} \nabla f_i(w_{i,k,t}, \varepsilon) - w_t + \frac{1}{m} \sum_{i \in \mathcal{I}} \sum_{k=0}^{K-1} \eta_{k,t} \nabla f_i(\hat{w}_{i,k,t}, \varepsilon')\| \\
 &\leq \frac{1}{m} \sum_{i \in \mathcal{I}} \sum_{k=0}^{K-1} \eta_{k,t} \|\nabla f_i(w_{i,k,t}, \varepsilon) - \nabla f_i(\hat{w}_{i,k,t}, \varepsilon')\|
 \end{aligned}$$

$$= \frac{1}{m} \sum_{k=0}^{K-1} \eta_{k,t} \|\nabla f_{i^*}(w_{i^*,k,t}, \varepsilon) - \nabla f_{i^*}(\hat{w}_{i^*,k,t}, \varepsilon')\| \leq \frac{2V}{m} \sum_{k=0}^{K-1} \eta_{k,t}.$$

619 The last equation adopts $\varepsilon = \varepsilon'$ when $i \neq i^*$. This completes the proofs.

620

621 **Lemma 7 (Model Sensitivity.)** *The model sensitivity caused by gradient descent steps can be*
 622 *bounded as:*

$$\|\phi'(w_t) - \phi'(\tilde{w}_t)\| \leq (1 + \eta(K, t)L) \|w_t - \tilde{w}_t\|, \quad (26)$$

623 where $\eta(K, t) = \eta_{0,t} + \sum_{k=1}^{K-1} \eta_{k,t} \prod_{j=0}^{k-1} (1 + \eta_{j,t}L)$ is a constant related the selection of learning
 624 rates.

625 **Proof.** We first learn an individual case. On the t -th round, we assume the initial states of two
 626 sequences are w_t and $\tilde{w}_t$. Each is performed by the update function ϕ' for local K steps. For each
 627 step, we have:

$$\begin{aligned} & \|\hat{w}_{i,k+1,t} - \tilde{w}_{i,k+1,t}\| \\ & \leq \|\hat{w}_{i,k,t} - \tilde{w}_{i,k,t}\| + \eta_{k,t} \|\nabla f_i(\hat{w}_{i,k,t}, \varepsilon') - \nabla f_i(\tilde{w}_{i,k,t}, \varepsilon')\| \\ & \leq (1 + \eta_{k,t}L) \|\hat{w}_{i,k,t} - \tilde{w}_{i,k,t}\|. \end{aligned}$$

628 This implies each gap when $k \geq 1$ can be upper bounded by:

$$\|\hat{w}_{i,k,t} - \tilde{w}_{i,k,t}\| \leq (1 + \eta_{k-1,t}L) \|\hat{w}_{i,k-1,t} - \tilde{w}_{i,k-1,t}\| \leq \dots \leq \prod_{j=0}^{k-1} (1 + \eta_{j,t}L) \|w_t - \tilde{w}_t\|.$$

629 Then we consider the recursive formulation of the stability gaps along the iterations k . We can
 630 directly apply Eq.(22) to obtain the relationship for the differences updated from different initial
 631 states on the same dataset. By directly expanding the update function ϕ' at w_t and $\tilde{w}_t$, we have:

$$\begin{aligned} & \|\phi'(w_t) - \phi'(\tilde{w}_t)\| \\ & = \|w_t - \frac{1}{m} \sum_{i \in \mathcal{I}} \sum_{k=0}^{K-1} \eta_{k,t} \nabla f_i(\hat{w}_{i,k,t}, \varepsilon') - \tilde{w}_t + \frac{1}{m} \sum_{i \in \mathcal{I}} \sum_{k=0}^{K-1} \eta_{k,t} \nabla f_i(\tilde{w}_{i,k,t}, \varepsilon')\| \\ & \leq \|w_t - \tilde{w}_t\| + \left\| \frac{1}{m} \sum_{i \in \mathcal{I}} \sum_{k=0}^{K-1} \eta_{k,t} (\nabla f_i(\hat{w}_{i,k,t}, \varepsilon') - \nabla f_i(\tilde{w}_{i,k,t}, \varepsilon')) \right\| \\ & \leq \|w_t - \tilde{w}_t\| + \frac{L}{m} \sum_{i \in \mathcal{I}} \sum_{k=0}^{K-1} \eta_{k,t} \|\hat{w}_{i,k,t} - \tilde{w}_{i,k,t}\| \\ & \leq \left[1 + \left(\eta_{0,t} + \sum_{k=1}^{K-1} \eta_{k,t} \prod_{j=0}^{k-1} (1 + \eta_{j,t}L) \right) L \right] \|w_t - \tilde{w}_t\|. \end{aligned}$$

632 This completes the proofs.

633

634 We have successfully quantified the specific form of the problem as above. By solving for a series
 635 of reasonable values of the auxiliary variable λ to minimize the above problem, we obtain the tight
 636 lower bound on privacy. Before that, let's discuss the learning rate to simplify this expression. Both
 637 $\eta(K, t)$ and $\sum \eta_{k,t}$ terms are highly related to the selections of learning rates. Typically, this choice
 638 is determined by the optimization process. Whether it's generalization or privacy analysis, both are
 639 based on the assumption that the optimization can converge properly. Therefore, we selected several
 640 different learning rate designs based on various combination methods to complete the subsequent
 641 analysis. Due to the unique two-stage learning perspective of federated learning, current methods
 642 for designing the learning rate generally choose between a constant rate or a rate that decreases
 643 with local rounds or iterations. Therefore, we discuss them separately including constant learning
 644 rate, cyclically decaying learning rate, stage-wise decaying learning rate, and continuously decaying
 645 learning rate. We provide a simple comparison in Figure 3.

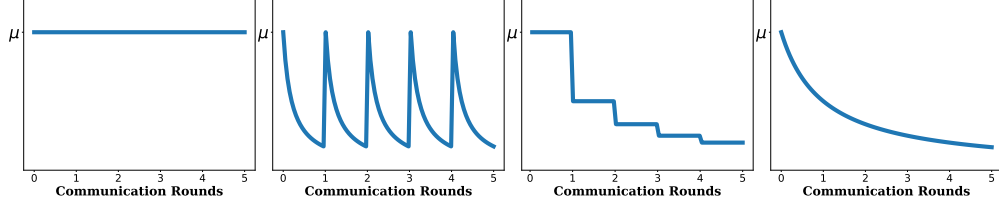


Figure 3: Four general setups of learning rate adopted in the federated learning community. From left to right, they are: *Constant learning rates*, *Cyclically decaying learning rates*, *Stage-wise decaying learning rate*, and *Continuously decaying learning rate*.

Constant learning rates This is currently the simplest case. We consider the learning rate to always be a constant, i.e. $\eta_{k,t} = \mu$. Then we have that the accumulation term $\sum_{k=0}^{K-1} \eta_{k,t} = \mu K$. For the $\eta(K, t)$ term, we have:

$$\eta(K, t) = \eta_{0,t} + \sum_{k=1}^{K-1} \eta_{k,t} \prod_{j=0}^{k-1} (1 + \eta_{j,t} L) = \mu \sum_{k=0}^{K-1} (1 + \mu L)^k = \frac{1}{L} ((1 + \mu L)^K - 1).$$

When K is selected, both of them can be considered as a constant related to K . The choice of μ also requires careful consideration. Although it is a constant, its selection is typically related to m , K , and T based on the optimization process. We will discuss this point in the final theorems.

Cyclically decaying learning rates Some works treat this learning process as an aggregation process of several local training processes, i.e. each local client learns from a better initial state (knowledge learned from other clients). And since the client pool is very large, most clients will exit after obtaining the model they desire. This setting is often used in “cross-device” scenarios [Kairouz et al., 2021]. Thus, local learning can be considered as an independent learning process. In this case, the learning rate is designed to decay in an inversely proportional function to achieve optimal local accuracy, i.e. $\eta_{k,t} = \frac{\mu}{k+1}$, and is restored to a larger initial value at the start of each round, i.e. $\eta_{0,t} = \mu$. Then we have the accumulation term:

$$\ln(K+1) = \int_{k=0}^K \frac{1}{k+1} dk \leq \sum_{k=0}^{K-1} \frac{1}{k+1} \leq 1 + \int_0^{K-1} \frac{1}{k+1} dk = 1 + \ln(K). \quad (27)$$

When K is large, this term is dominated by $\mathcal{O}(\ln(K))$. Based on the fact that K is very large in federated learning, we further approximate this term to $c \ln(K+1)$ where c is a scaled constant. It is easy to check that there must exist $1 \leq c < 1.543$ for any $K \geq 1$. Thus we have the accumulation term as $\sum_{k=0}^{K-1} \eta_{k,t} = c\mu \ln(K+1)$. For the $\eta(K, t)$ term, we have its upper bound:

$$\begin{aligned} \eta(K, t) &= \mu + \sum_{k=1}^{K-1} \frac{\mu}{k+1} \prod_{j=0}^{k-1} \left(1 + \frac{\mu L}{j+1}\right) \leq \mu + \sum_{k=1}^{K-1} \frac{\mu}{k+1} \prod_{j=0}^{k-1} \exp\left(\frac{\mu L}{j+1}\right) \\ &= \mu + \sum_{k=1}^{K-1} \frac{\mu}{k+1} \left[\exp\left(\sum_{j=0}^{k-1} \frac{1}{j+1}\right) \right]^{\mu L} = \sum_{k=0}^{K-1} \frac{\mu}{k+1} [\exp(c \ln(k+1))]^{\mu L} \\ &= \mu \sum_{k=0}^{K-1} (k+1)^{c\mu L-1} \leq \mu \int_{k=0}^K (k+1)^{c\mu L-1} dk = \frac{1}{cL} ((1+K)^{c\mu L} - 1). \end{aligned}$$

The first inequality adopts $1+x \leq e^x$ and the last adopts the concavity. Actually, we still can learn its general lower bound by a scaling constant. By adopting a scaling b , we can have $1+x \geq e^{bx}$, which is equal to $b \leq \frac{\ln(x+1)}{x}$. It is also easy to check $0.693 < b < 1$ when $0 < x \leq 1$. Thus we have:

$$\eta(K, t) = \mu + \sum_{k=1}^{K-1} \frac{\mu}{k+1} \prod_{j=0}^{k-1} \left(1 + \frac{\mu L}{j+1}\right) \geq \mu + \sum_{k=1}^{K-1} \frac{\mu}{k+1} \prod_{j=0}^{k-1} \exp\left(\frac{\mu b L}{j+1}\right)$$

$$\begin{aligned}
&= \mu + \sum_{k=1}^{K-1} \frac{\mu}{k+1} \left[\exp \left(\sum_{j=0}^{k-1} \frac{1}{j+1} \right) \right]^{\mu b L} = \sum_{k=0}^{K-1} \frac{\mu}{k+1} [\exp(c \ln(k+1))]^{\mu b L} \\
&= \mu \sum_{k=0}^{K-1} (k+1)^{c \mu b L - 1} \geq \mu \int_{k=-1}^{K-1} (k+1)^{c \mu b L - 1} dk = \frac{1}{c b L} K^{c \mu b L}.
\end{aligned}$$

668 The last inequality also adopts concavity. Through this simple scaling, we learn the general bounds
669 for the learning rate function $\eta(K, t)$ as:

$$\frac{1}{c b L} K^{c \mu b L} \leq \eta(K, t) \leq \frac{1}{c L} ((1 + K)^{c \mu L} - 1), \quad (28)$$

670 where $1 \leq c < 1.543$, $0.693 < b < 1$ and $\mu \leq \frac{1}{L}$ (this condition is almost universally satisfied in
671 current optimization theories). Although we cannot precisely find the tight bound of this function
672 $\eta(K, t)$, we can still treat it as a form based on constants to complete the subsequent analysis, i.e. it
673 could be approximated as a larger upper bound $\frac{1}{L} ((1 + K)^{c \mu L} - 1)$. More importantly, we have
674 determined that this learning rate function still diverges as K increases.

675 **Stage-wise decaying learning rates** This is one of the most common selections of learning rate
676 in the current federated community, which is commonly applied in “cross-silo” scenarios [Kairouz
677 et al., 2021]. When the client pool is not very large, clients who participate in the training often aim
678 to establish long-term cooperation to continuously improve their models. Therefore, each client will
679 contribute to the entire training process over a long period. From a learning perspective, local training
680 is more like exploring the path to a local optimum rather than actually achieving the local optimum.
681 Therefore, each local training will adopt a constant learning rate and perform several update steps, i.e.
682 $\eta_{k,t} = \eta_t$. At each communication round, the learning rate decays once and continues to the next
683 stage, i.e. $\eta_t = \frac{\mu}{t+1}$. Based on the analysis of the constant learning rate, the accumulation term is
684 $\sum_{k=0}^{K-1} \eta_{k,t} = \frac{\mu K}{t+1}$. For the $\eta(K, t)$ term, we have:

$$\begin{aligned}
\eta(K, t) &= \frac{\mu}{t+1} + \sum_{k=1}^{K-1} \frac{\mu}{t+1} \prod_{j=0}^{k-1} \left(1 + \frac{\mu L}{t+1} \right) \\
&= \frac{\mu L}{t+1} \sum_{k=0}^{K-1} \left(1 + \frac{\mu L}{t+1} \right)^k = \frac{1}{L} \left(\left(1 + \frac{\mu L}{t+1} \right)^K - 1 \right).
\end{aligned}$$

685 It can be seen that the analysis of this function is more challenging because the learning rate function
686 $\eta(K, t)$ is decided by t , which introduces complexity to the subsequent analysis. We will explain this
687 in detail in the subsequent discussion.

688 **Continuously decaying learning rates** This is a common selection of learning rate in the federated
689 community, involving dual learning rate decay along both local training and global training. This
690 can almost be applied to all methods to adapt to the final training, including both the cross-silo and
691 cross-device cases. At the same time, its analysis is also more challenging because the learning rate
692 is coupled with communication rounds and local iterations, yielding new upper and lower bounds.
693 We consider the general case $\eta_{k,t} = \frac{\mu}{tK+k+1}$. Therefore, the accumulation term can be bounded as:

$$\begin{aligned}
\sum_{k=0}^{K-1} \frac{1}{tK+k+1} &> \int_{k=0}^K \frac{1}{tK+k+1} dk = \ln \left(\frac{tK+K+1}{tK+1} \right), \\
\sum_{k=0}^{K-1} \frac{1}{tK+k+1} &< \frac{1}{tK+1} + \int_{k=0}^{K-1} \frac{1}{tK+k+1} dk = \frac{1}{tK+1} + \ln \left(\frac{tK+K}{tK+1} \right).
\end{aligned}$$

694 Similarly, when K is large enough, this term is dominated by $\mathcal{O}(\ln(\frac{t+1}{t}))$. For simplicity in the
695 subsequent proof, we follow the process above and let it be $z \ln(\frac{t+2}{t+1})$ to include the term at $t = 0$.
696 It is also easy to check that $z > 1$ is a constant for any $K > 1$. And z is also a constant. It means

697 we can always select the lower bound as its representation. Therefore, for the learning rate function
 698 $\eta(K, t)$, we have:

$$\begin{aligned}
 \eta(K, t) &= \frac{\mu}{tK+1} + \sum_{k=1}^{K-1} \frac{\mu}{tK+k+1} \prod_{j=0}^{k-1} \left(1 + \frac{\mu L}{tK+j+1}\right) \\
 &\leq \frac{\mu}{tK+1} + \sum_{k=1}^{K-1} \frac{\mu}{tK+k+1} \left[\exp \left(\sum_{j=0}^{k-1} \frac{1}{tK+j+1} \right) \right]^{\mu L} \\
 &= \frac{\mu}{tK+1} + \sum_{k=1}^{K-1} \frac{\mu}{tK+k+1} \left[\exp \left(z \ln \left(\frac{tK+k+1}{tK+1} \right) \right) \right]^{\mu L} \\
 &= \frac{\mu}{(tK+1)^{z\mu L}} \sum_{k=0}^{K-1} (tK+k+1)^{z\mu L-1} \\
 &\leq \frac{\mu}{(tK+1)^{z\mu L}} \int_{k=0}^K (tK+k+1)^{z\mu L-1} dk = \frac{1}{zL} \left(\left(\frac{tK+K+1}{tK+1} \right)^{z\mu L} - 1 \right).
 \end{aligned}$$

699 Similarly, we introduce the coefficient b to provide the lower bound as:

$$\begin{aligned}
 \eta(K, t) &= \frac{\mu}{tK+1} + \sum_{k=1}^{K-1} \frac{\mu}{tK+k+1} \prod_{j=0}^{k-1} \left(1 + \frac{\mu L}{tK+j+1}\right) \\
 &\geq \frac{\mu}{tK+1} + \sum_{k=1}^{K-1} \frac{\mu}{tK+k+1} \left[\exp \left(\sum_{j=0}^{k-1} \frac{1}{tK+j+1} \right) \right]^{\mu b L} \\
 &= \frac{\mu}{tK+1} + \sum_{k=1}^{K-1} \frac{\mu}{tK+k+1} \left[\exp \left(z \ln \left(\frac{tK+k+1}{tK+1} \right) \right) \right]^{\mu b L} \\
 &= \frac{\mu}{(tK+1)^{z\mu b L}} \sum_{k=0}^{K-1} (tK+k+1)^{z\mu b L-1} \\
 &\geq \frac{\mu}{(tK+1)^{z\mu b L}} \int_{k=-1}^{K-1} (tK+k+1)^{z\mu b L-1} dk = \frac{1}{zbL} \left(\left(\frac{tK+K}{tK+1} \right)^{z\mu b L} - \left(\frac{tK}{tK+1} \right)^{z\mu b L} \right) \\
 &> \frac{1}{zbL} \left(\left(\frac{tK+K}{tK+1} \right)^{z\mu b L} - 1 \right).
 \end{aligned}$$

700 Through the sample scaling, we learn the general bounds for the learning rate function $\eta(K, t)$ as:

$$\frac{1}{zbL} \left(\left(\frac{tK+K}{tK+1} \right)^{z\mu b L} - 1 \right) < \eta(K, t) \leq \frac{1}{zL} \left(\left(\frac{tK+K+1}{tK+1} \right)^{z\mu L} - 1 \right), \quad (29)$$

701 where $1 < z, 0.693 < b < 1$ and $\mu \leq \frac{1}{L}$. Obviously, when K is large enough, the learning rate term
 702 is still dominated by $\mathcal{O} \left(\left(\frac{t+2}{t+1} \right)^{z\mu L} - 1 \right)$. Therefore, to learn the general cases, we can consider the
 703 specific form of the learning rate function based on the constant scaling as $\frac{1}{L} \left(\left(\frac{t+2}{t+1} \right)^{z\mu L} - 1 \right)$. As
 704 t increases, this function will approach zero.

705

706 **on the Noisy-FedProx Method:**

707 In this part, we will address the differential privacy analysis of a noisy version of another classical
 708 federated learning optimization method, i.e. the Noisy-FedProx method. The vanilla FedProx
 709 method is an optimization algorithm designed for cross-silo federated learning, particularly to address
 710 the challenges caused by data heterogeneity across different clients. Unlike traditional federated
 711 learning algorithms like FedAvg, which can struggle with variations in data distribution, it introduces
 712 a proximal term to the objective function. This helps to stabilize the training process and improve
 713 convergence. Specifically, it adopts the consistency as the penalized term to correct the local objective:

$$\min_w f_i(w) + \frac{\alpha}{2} \|w - w_t\|^2. \quad (30)$$

714 The proximal term is a very common regularization term in federated learning and has been widely
 715 used in both federated learning and personalized federated learning approaches. It introduces an
 716 additional penalty to the local objective, ensuring that local updates are optimized towards the
 717 local optimal solution while being subject to an extra global constraint, i.e. each local update does
 718 not stray too far from the initialization point. In fact, there are many optimization methods that
 719 apply such regularization terms. For example, various federated primal-dual methods based on
 720 the ADMM approach construct local Lagrangian functions, and in personalized federated learning,
 721 local privatization regularization terms are introduced to differentiate from the vanilla consistency
 722 objective. The analysis of the above methods is fundamentally based on a correct understanding of
 723 the advantages and significance of the proximal term in stability error. In this paper, to achieve a
 724 cross-comparison while maintaining generality, we consider the optimization process of local training
 725 as total K -step updates:

$$\phi(w_t) = w_t - \frac{1}{m} \sum_{i \in \mathcal{I}} \sum_{k=0}^{K-1} \eta_{k,t} (\nabla f_i(w_{i,k,t}, \varepsilon) + \alpha(w_{i,k,t} - w_t)). \quad (31)$$

726 Here, we also employ the proofs mentioned in the previous section, and our study of the difference
 727 term is based on both data sensitivity and model sensitivity perspectives. We provide these two main
 728 lemmas as follows.

729

730 **Lemma 8 (Data Sensitivity.)** *The local data sensitivity of the Noisy-FedProx method at t -th com-*
 731 *munication round can be upper bounded as:*

$$\|\phi(w_t) - \phi'(w_t)\| \leq \frac{2V}{m\alpha}. \quad (32)$$

732 **Proof.** *We first consider a single step in Eq.(31) as:*

$$w_{i,k+1,t} = w_{i,k,t} - \eta_{k,t} (\nabla f_i(w_{i,k,t}, \varepsilon) + \alpha(w_{i,k,t} - w_t)).$$

733 *The proximal term brings more opportunities to enhance the analysis of local updates. We can split*
 734 *the proximal term and subtract the w_t term on both sides, resulting in a recursive formula for the*
 735 *cumulative update term:*

$$w_{i,k+1,t} - w_t = (1 - \eta_{k,t}\alpha)(w_{i,k,t} - w_t) - \eta_{k,t} \nabla f_i(w_{i,k,t}, \varepsilon).$$

736 *The above equation indicates that a reduction factor $1 - \eta_{k,t}\alpha < 1$ can limit the scale of local*
 737 *updates. This is a very good property, allowing us to shift the analysis of the data sensitivity to their*
 738 *relationship of local updates. According to the above, we can upper bound the gaps between $\{w_{i,k,t}\}$*
 739 *and $\{\hat{w}_{i,k,t}\}$ sequences as:*

$$\begin{aligned} & \| (w_{i,k+1,t} - w_t) - (\hat{w}_{i,k+1,t} - w_t) \| \\ &= \| (1 - \eta_{k,t}\alpha) [(w_{i,k,t} - w_t) - (\hat{w}_{i,k,t} - w_t)] - \eta_{k,t} (\nabla f_i(w_{i,k,t}, \varepsilon) - \nabla f_i(\hat{w}_{i,k,t}, \varepsilon')) \| \\ &\leq (1 - \eta_{k,t}\alpha) \| (w_{i,k,t} - w_t) - (\hat{w}_{i,k,t} - w_t) \| + \eta_{k,t} \| \nabla f_i(w_{i,k,t}, \varepsilon) - \nabla f_i(\hat{w}_{i,k,t}, \varepsilon') \| \\ &\leq (1 - \eta_{k,t}\alpha) \| (w_{i,k,t} - w_t) - (\hat{w}_{i,k,t} - w_t) \| + 2\eta_{k,t}V. \end{aligned}$$

740 *Different from proofs in Lemma 6, the term $1 - \eta_{k,t}\alpha$ can further decrease the stability gap during*
 741 *accumulation. By summing from $k = 0$ to $K - 1$, we can obtain:*

$$\| (w_{i,K,t} - w_t) - (\hat{w}_{i,K,t} - w_t) \|\leq$$

$$\begin{aligned}
&\leq \prod_{k=0}^{K-1} (1 - \eta_{k,t}\alpha) \| (w_{i,0,t} - w_t) - (\hat{w}_{i,0,t} - w_t) \| + \sum_{k=0}^{K-1} \left(\prod_{j=k+1}^{K-1} (1 - \eta_{j,t}\alpha) \right) 2\eta_{k,t}V \\
&= 2V \sum_{k=0}^{K-1} \left(\prod_{j=k+1}^{K-1} (1 - \eta_{j,t}\alpha) \right) \eta_{k,t}.
\end{aligned}$$

742 Here, we provide a simple proof using a constant learning rate to demonstrate that its upper bound
743 can be independent of K . By considering $\eta_{k,t} = \mu$, we have:

$$\sum_{k=0}^{K-1} \left(\prod_{j=k+1}^{K-1} (1 - \eta_{j,t}\alpha) \right) \eta_{k,t} = \sum_{k=0}^{K-1} \left(\prod_{j=k+1}^{K-1} (1 - \mu\alpha) \right) \mu = \frac{1 - (1 - \mu\alpha)^K}{\alpha} < \frac{1}{\alpha}.$$

744 In fact, when the learning rate decays with k , it can still be easily proven to have a constant upper
745 bound. Therefore, in the subsequent proofs, we directly use the form of this constant upper bound
746 as the result of data sensitivity in the *Noisy-FedProx* method. Based on the definition of $\phi(w)$, we
747 have:

$$\begin{aligned}
\|\phi(w_t) - \phi'(w_t)\| &= \|(\phi(w_t) - w_t) - (\phi'(w_t) - w_t)\| = \left\| \frac{1}{m} \sum_{i \in \mathcal{I}} [(w_{i,K,t} - w_t) - (\hat{w}_{i,K,t} - w_t)] \right\| \\
&= \frac{1}{m} \| (w_{i^*,K,t} - w_t) - (\hat{w}_{i^*,K,t} - w_t) \| < \frac{2V}{m\alpha}.
\end{aligned}$$

748 This completes the proofs.

749

750 **Lemma 9 (Model Sensitivity.)** The local model sensitivity of the *Noisy-FedProx* method at t -th
751 communication round can be upper bounded as:

$$\|\phi'(w_t) - \phi'(\tilde{w}_t)\| \leq \frac{\alpha}{\alpha_L} \|w_t - \tilde{w}_t\|. \quad (33)$$

752 **Proof.** We also adopt the splitting above. Since both sequences are trained on the same dataset, the
753 gradient difference can be measured by the parameter difference. Therefore, we directly consider the
754 form of the parameter difference:

$$\begin{aligned}
&\|\hat{w}_{i,k+1,t} - \tilde{w}_{i,k+1,t}\| \\
&= \|(1 - \eta_{k,t}\alpha)(\hat{w}_{i,k,t} - \tilde{w}_{i,k,t}) - \eta_{k,t}(\nabla f_i(\hat{w}_{i,k,t}, \varepsilon') - \nabla f_i(\tilde{w}_{i,k,t}, \varepsilon')) - \eta_{k,t}\alpha(w_t - \tilde{w}_t)\| \\
&\leq (1 - \eta_{k,t}\alpha)\|\hat{w}_{i,k,t} - \tilde{w}_{i,k,t}\| + \eta_{k,t}L\|\hat{w}_{i,k,t} - \tilde{w}_{i,k,t}\| + \eta_{k,t}\alpha\|w_t - \tilde{w}_t\| \\
&= (1 - \eta_{k,t}\alpha_L)\|\hat{w}_{i,k,t} - \tilde{w}_{i,k,t}\| + \eta_{k,t}\alpha\|w_t - \tilde{w}_t\|,
\end{aligned}$$

755 where $\alpha_L = \alpha - L$ is a constant. Here, we consider $\alpha > L$. When $\alpha \leq L$, its upper bound can
756 not be guaranteed to be reduced. When $\alpha > L$, it can restore the property of decayed stability. By
757 summing from $k = 0$ to $K - 1$, we can obtain:

$$\begin{aligned}
&\|\hat{w}_{i,K,t} - \tilde{w}_{i,K,t}\| \\
&\leq \prod_{k=0}^{K-1} (1 - \eta_{k,t}\alpha_L) \|\hat{w}_{i,0,t} - \tilde{w}_{i,0,t}\| + \sum_{k=0}^{K-1} \left(\prod_{j=k+1}^{K-1} (1 - \eta_{j,t}\alpha_L) \right) \eta_{k,t}\alpha \|w_t - \tilde{w}_t\| \\
&= \left[\prod_{k=0}^{K-1} (1 - \eta_{k,t}\alpha_L) + \sum_{k=0}^{K-1} \left(\prod_{j=k+1}^{K-1} (1 - \eta_{j,t}\alpha_L) \right) \eta_{k,t}\alpha \right] \|w_t - \tilde{w}_t\|.
\end{aligned}$$

758 Similarly, we learn the upper bound from a simple constant learning rate. By select $\eta_{k,t} = \mu$, we
759 have:

$$\prod_{k=0}^{K-1} (1 - \eta_{k,t}\alpha_L) + \sum_{k=0}^{K-1} \left(\prod_{j=k+1}^{K-1} (1 - \eta_{j,t}\alpha_L) \right) \eta_{k,t}\alpha$$

$$\begin{aligned}
&= \prod_{k=0}^{K-1} (1 - \mu\alpha_L) + \sum_{k=0}^{K-1} \left(\prod_{j=k+1}^{K-1} (1 - \mu\alpha_L) \right) \mu\alpha \\
&= (1 - \mu\alpha_L)^K + \alpha \frac{1 - (1 - \mu\alpha_L)^K}{\alpha_L} \\
&= \frac{\alpha}{\alpha_L} - \frac{L(1 - \mu\alpha_L)^K}{\alpha_L} < \frac{\alpha}{\alpha_L}.
\end{aligned}$$

760 The same, it can also be checked that the general upper bound of the stability gaps is a constant
761 even if the learning rate is selected to be decayed along iteration k . Therefore, in the subsequent
762 proofs, we directly use the form of this constant upper bound as the result of model sensitivity in the
763 *Noisy-FedProx* method. Based on the definition of $\phi(w)$, we have:

$$\|\phi'(w_t) - \phi'(\tilde{w}_t)\| = \left\| \frac{1}{m} \sum_{i \in \mathcal{I}} (\hat{w}_{i,K,t} - \tilde{w}_{i,K,t}) \right\| \leq \frac{1}{m} \sum_{i \in \mathcal{I}} \|\hat{w}_{i,K,t} - \tilde{w}_{i,K,t}\| \leq \frac{\alpha}{\alpha_L} \|w_t - \tilde{w}_t\|.$$

764 This completes the proofs.

765 C.3 Solution of Eq. (13)

766 According to the recurrence relation in Lemma 5, we can confine the privacy amplification process to
767 a finite number of steps with the aid of an interpolation sequence, yielding to the convergent bound.
768 Therefore, we have:

$$\begin{aligned}
&T(w_T; w'_T) = T(\tilde{w}_T; w'_T) \\
&\geq T(\tilde{w}_{T-1}; w'_{T-1}) \otimes T_G \left(\frac{\sqrt{m}}{\sigma} \lambda_T \|\phi(w_{T-1}) - \phi'(\tilde{w}_{T-1})\| \right) \\
&\geq T(\tilde{w}_{t_0}; w'_{t_0}) \otimes \cdots \otimes T_G \left(\frac{\sqrt{m}}{\sigma} \lambda_T \|\phi(w_{T-1}) - \phi'(\tilde{w}_{T-1})\| \right) \\
&= T(w'_{t_0}; w'_{t_0}) \otimes T_G \left(\frac{\sqrt{m}}{\sigma} \sqrt{\sum_{t=t_0}^{T-1} \lambda_{t+1}^2 \|\phi(w_t) - \phi'(\tilde{w}_t)\|^2} \right) \\
&\geq T_G \left(\frac{\sqrt{m}}{\sigma} \sqrt{\sum_{t=t_0}^{T-1} \lambda_{t+1}^2 (\rho_t \|w_t - \tilde{w}_t\| + \gamma_t)^2} \right).
\end{aligned}$$

769 Although the above form appears promising, an inappropriate selection of the key parameters will
770 still cause divergence due to the recurrence term coefficient $1 + \eta(K, t)L > 1$, leading it to approach
771 infinity as t increases. For instance, small t_0 will result in a significantly increased λ and the bound
772 will be closed to the stability gap $\|w_T - w'_T\|$, and large t_0 will result in a long accumulation of the
773 stability gaps, which is also unsatisfied. At the same time, it is also crucial to choose appropriate
774 λ to ensure that the stability accumulation can be reasonably diluted. Therefore, we also need
775 to thoroughly investigate how significant the stability gap caused by the interpolation points is.
776 According to Eq.(21) and (22), we have:

$$\|w_{t+1} - \tilde{w}_{t+1}\| \leq (1 - \lambda_{t+1}) (\rho_t \|w_t - \tilde{w}_t\| + \gamma_t).$$

777 The above relationship further constrains the stability of the interpolation sequence. It is worth noting
778 that the upper bound of the final step is independent of the choice of λ . At the same time, since all
779 terms are positive, given a group of specific λ , taking the upper bound at each possible t will result in
780 the maximum error accumulation. This is also the worst-case privacy we have constructed. Therefore,
781 solving the worst privacy could be considered as solving the following problem:

$$\begin{aligned}
&\underbrace{\min_{\{\lambda_{t+1}\}, t_0} \max_{\{\|w_t - \tilde{w}_t\|\}} \sum_{t=t_0}^{T-1} \lambda_{t+1}^2 (\rho_t \|w_t - \tilde{w}_t\| + \gamma_t)^2}_{\text{worst privacy}} \\
&\quad \underbrace{\hspace{10em}}_{\text{tight privacy lower bound}} \\
&\text{s.t. } \|w_{t+1} - \tilde{w}_{t+1}\| \leq (1 - \lambda_{t+1}) (\rho_t \|w_t - \tilde{w}_t\| + \gamma_t).
\end{aligned} \tag{34}$$

Based on the above analysis, this problem can be directly transformed into a privacy minimization problem when the interpolation sequence reaches the maximum stability error. Therefore, we just need to solve the following problem:

$$\begin{aligned} \min_{\{\lambda_{t+1}\}, t_0} \quad & \sum_{t=t_0}^{T-1} \lambda_{t+1}^2 (\gamma_t \|w_t - \tilde{w}_t\| + \gamma_t)^2, \\ \text{s.t.} \quad & \|w_{t+1} - \tilde{w}_{t+1}\| = (1 - \lambda_{t+1}) (\rho_t \|w_t - \tilde{w}_t\| + \gamma_t). \end{aligned} \quad (35)$$

It is important to note that this upper bound condition is usually loose because the probability that the interpolation terms simultaneously reach their maximum deviation is very low. This is merely the theoretical worst-case privacy scenario.

Then we solve the minimization problem. By considering the worst stability conditions, we can provide the relationship between the gaps and coefficients λ_{t+1} as:

$$\|w_{t+1} - \tilde{w}_{t+1}\| = \rho_t \|w_t - \tilde{w}_t\| + \gamma_t - \lambda_{t+1} (\rho_t \|w_t - \tilde{w}_t\| + \gamma_t).$$

Expanding it from $t = t_0$ to T , we have:

$$0 = \|w_T - \tilde{w}_T\| = \left(\prod_{t=t_0}^{T-1} \rho_t \right) \|w_{t_0} - \tilde{w}_{t_0}\| + \sum_{t=t_0}^{T-1} \left(\prod_{j=t+1}^{T-1} \rho_j \right) [\gamma_t - \lambda_{t+1} (\rho_t \|w_t - \tilde{w}_t\| + \gamma_t)].$$

Due to the term $\lambda_{t+1} (\rho_t \|w_t - \tilde{w}_t\| + \gamma_t)$ being part of the analytical form of the minimization objective, we preserve the integrity of this algebraic form and only split it from the perspectives of coefficients λ_t , ρ_t and γ_t . According to the definition $\tilde{w}_{t_0} = w'_{t_0}$, then we have:

$$\sum_{t=t_0}^{T-1} \left(\prod_{j=t+1}^{T-1} \rho_j \right) \lambda_{t+1} (\rho_t \|w_t - \tilde{w}_t\| + \gamma_t) = \left(\prod_{t=t_0}^{T-1} \rho_t \right) \|w_{t_0} - w'_{t_0}\| + \sum_{t=t_0}^{T-1} \left(\prod_{j=t+1}^{T-1} \rho_j \right) \gamma_t. \quad (36)$$

The above equation presents the summation of the term $\lambda_{t+1} (\rho_t \|w_t - \tilde{w}_t\| + \gamma_t)$ accompanied by a scaling coefficient $\left(\prod_{j=t+1}^{T-1} \rho_j \right) > 1$. It naturally transforms the summation form into an initial stability gap and a constant term achieved through a combination of learning rates. To solve it, we can directly adopt the Cauchy-Schwarz inequality to separate the terms and construct a constant term based on the form of the scaling coefficient to find its achievable lower bound:

$$\begin{aligned} & \sum_{t=t_0}^{T-1} \lambda_{t+1}^2 (\rho_t \|w_t - \tilde{w}_t\| + \gamma_t)^2 \\ & \geq \left(\sum_{t=t_0}^{T-1} \left(\prod_{j=t+1}^{T-1} \rho_j \right) \lambda_{t+1} (\rho_t \|w_t - \tilde{w}_t\| + \gamma_t) \right)^2 \left(\sum_{t=t_0}^{T-1} \left(\prod_{j=t+1}^{T-1} \rho_j \right)^2 \right)^{-1} \\ & = \left(\left(\prod_{t=t_0}^{T-1} \rho_t \right) \|w_{t_0} - w'_{t_0}\| + \sum_{t=t_0}^{T-1} \left(\prod_{j=t+1}^{T-1} \rho_j \right) \gamma_t \right)^2 \left(\sum_{t=t_0}^{T-1} \left(\prod_{j=t+1}^{T-1} \rho_j \right)^2 \right)^{-1}. \end{aligned}$$

Although the original problem requires solving the λ_{t+1} , here we can know one possible minimum form of the problem no longer includes this parameter. In fact, this parameter has been transformed into the optimality condition of the Cauchy-Schwarz inequality.

Therefore, we only need to optimize it w.r.t the parameter t_0 . Unfortunately, this part highly correlates with the stability gaps $\|w_{t_0} - w'_{t_0}\|$. Current research progress indicates that in non-convex optimization, this term diverges as the number of training rounds t increases. This makes it difficult for us to accurately quantify its specific impact on the privacy bound. If t_0 is very small, it means that the introduced stability gap will also be very small. However, consequently, the coefficients of the ρ_t and γ_t terms will increase due to the accumulation over $T - t_0$ rounds. To detail this, we have to make certain compromises. Because t_0 is an integer belonging to $[0, T - 1]$, we denote its optimal selection by t^* (it certainly exists when T is given). Therefore, the privacy lower bound under other

810 choices of t_0 will certainly be more relaxed, i.e. $\text{Privacy}_{t_0} \leq \text{Privacy}_{t^*}$ (privacy is weak at other
 811 selection of t_0). This allows us to look for other asymptotic solutions instead of finding the optimal
 812 solution. Although we cannot ultimately achieve the form of the optimal solution, we can still provide
 813 a stable privacy lower bound. To eliminate the impact of stability error, we directly choose $t_0 = 0$,
 814 yielding the following bound:

$$\begin{aligned}
 \mathcal{H}_* \leq \mathcal{H}_0 &= \left(\left(\prod_{t=t_0}^{T-1} \rho_t \right) \|w_{t_0} - w'_{t_0}\| + \sum_{t=t_0}^{T-1} \left(\prod_{j=t+1}^{T-1} \rho_j \right) \gamma_t \right)^2 \left(\sum_{t=t_0}^{T-1} \left(\prod_{j=t+1}^{T-1} \rho_j \right)^2 \right)^{-1} \Big|_{t_0=0} \\
 &= \left(\sum_{t=0}^{T-1} \left(\prod_{j=t+1}^{T-1} \rho_j \right) \gamma_t \right)^2 \left(\sum_{t=0}^{T-1} \left(\prod_{j=t+1}^{T-1} \rho_j \right)^2 \right)^{-1}.
 \end{aligned}$$

815 By substituting the values of ρ_t and γ_t under different cases, then we can prove the main theorems in
 816 this paper.

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 have outlined the core contributions of this paper.

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 discussed the relaxed components in the theoretical analysis and provided auxiliary arguments to ensure the correctness of the conclusions.

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 a comprehensive proof for the main theorems.

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 provide all the necessary information for reproducibility.

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 all the necessary information for reproducibility.

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, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide all the necessary information for reproducibility.

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: [Yes]

Justification: We report the standard deviation in our evaluations.

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: Our experimental scale is relatively small, and they are easy to reproduce.

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: Fully compliant.

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: [NA]

Justification: The core contribution of our paper lies in machine learning theory.

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

1026 generate deepfakes for disinformation. On the other hand, it is not needed to point out
 1027 that a generic algorithm for optimizing neural networks could enable people to train
 1028 models that generate Deepfakes faster.

- 1029 • The authors should consider possible harms that could arise when the technology is
 1030 being used as intended and functioning correctly, harms that could arise when the
 1031 technology is being used as intended but gives incorrect results, and harms following
 1032 from (intentional or unintentional) misuse of the technology.
- 1033 • If there are negative societal impacts, the authors could also discuss possible mitigation
 1034 strategies (e.g., gated release of models, providing defenses in addition to attacks,
 1035 mechanisms for monitoring misuse, mechanisms to monitor how a system learns from
 1036 feedback over time, improving the efficiency and accessibility of ML).

1037 11. Safeguards

1038 Question: Does the paper describe safeguards that have been put in place for responsible
 1039 release of data or models that have a high risk for misuse (e.g., pretrained language models,
 1040 image generators, or scraped datasets)?

1041 Answer: [NA]

1042 Justification: Our paper poses no such risks.

1043 Guidelines:

- 1044 • The answer NA means that the paper poses no such risks.
- 1045 • Released models that have a high risk for misuse or dual-use should be released with
 1046 necessary safeguards to allow for controlled use of the model, for example by requiring
 1047 that users adhere to usage guidelines or restrictions to access the model or implementing
 1048 safety filters.
- 1049 • Datasets that have been scraped from the Internet could pose safety risks. The authors
 1050 should describe how they avoided releasing unsafe images.
- 1051 • We recognize that providing effective safeguards is challenging, and many papers do
 1052 not require this, but we encourage authors to take this into account and make a best
 1053 faith effort.

1054 12. Licenses for existing assets

1055 Question: Are the creators or original owners of assets (e.g., code, data, models), used in
 1056 the paper, properly credited and are the license and terms of use explicitly mentioned and
 1057 properly respected?

1058 Answer: [Yes]

1059 Justification: We cite all related papers and show our respects to those studies.

1060 Guidelines:

- 1061 • The answer NA means that the paper does not use existing assets.
- 1062 • The authors should cite the original paper that produced the code package or dataset.
- 1063 • The authors should state which version of the asset is used and, if possible, include a
 1064 URL.
- 1065 • The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- 1066 • For scraped data from a particular source (e.g., website), the copyright and terms of
 1067 service of that source should be provided.
- 1068 • If assets are released, the license, copyright information, and terms of use in the
 1069 package should be provided. For popular datasets, paperswithcode.com/datasets
 1070 has curated licenses for some datasets. Their licensing guide can help determine the
 1071 license of a dataset.
- 1072 • For existing datasets that are re-packaged, both the original license and the license of
 1073 the derived asset (if it has changed) should be provided.
- 1074 • If this information is not available online, the authors are encouraged to reach out to
 1075 the asset's creators.

1076 13. New assets

1077 Question: Are new assets introduced in the paper well documented and is the documentation
 1078 provided alongside the assets?

1079 Answer: [NA]

1080 Justification: The paper does not release new assets.

1081 Guidelines:

- 1082 • The answer NA means that the paper does not release new assets.
- 1083 • Researchers should communicate the details of the dataset/code/model as part of their
- 1084 submissions via structured templates. This includes details about training, license,
- 1085 limitations, etc.
- 1086 • The paper should discuss whether and how consent was obtained from people whose
- 1087 asset is used.
- 1088 • At submission time, remember to anonymize your assets (if applicable). You can either
- 1089 create an anonymized URL or include an anonymized zip file.

1090 **14. Crowdsourcing and research with human subjects**

1091 Question: For crowdsourcing experiments and research with human subjects, does the paper

1092 include the full text of instructions given to participants and screenshots, if applicable, as

1093 well as details about compensation (if any)?

1094 Answer: [NA]

1095 Justification: The paper does not involve crowdsourcing nor research with human subjects.

1096 Guidelines:

- 1097 • The answer NA means that the paper does not involve crowdsourcing nor research with
- 1098 human subjects.
- 1099 • Including this information in the supplemental material is fine, but if the main contribu-
- 1100 tion of the paper involves human subjects, then as much detail as possible should be
- 1101 included in the main paper.
- 1102 • According to the NeurIPS Code of Ethics, workers involved in data collection, curation,
- 1103 or other labor should be paid at least the minimum wage in the country of the data
- 1104 collector.

1105 **15. Institutional review board (IRB) approvals or equivalent for research with human**

1106 **subjects**

1107 Question: Does the paper describe potential risks incurred by study participants, whether

1108 such risks were disclosed to the subjects, and whether Institutional Review Board (IRB)

1109 approvals (or an equivalent approval/review based on the requirements of your country or

1110 institution) were obtained?

1111 Answer: [NA]

1112 Justification: The paper does not involve crowdsourcing nor research with human subjects.

1113 Guidelines:

- 1114 • The answer NA means that the paper does not involve crowdsourcing nor research with
- 1115 human subjects.
- 1116 • Depending on the country in which research is conducted, IRB approval (or equivalent)
- 1117 may be required for any human subjects research. If you obtained IRB approval, you
- 1118 should clearly state this in the paper.
- 1119 • We recognize that the procedures for this may vary significantly between institutions
- 1120 and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the
- 1121 guidelines for their institution.
- 1122 • For initial submissions, do not include any information that would break anonymity (if
- 1123 applicable), such as the institution conducting the review.

1124 **16. Declaration of LLM usage**

1125 Question: Does the paper describe the usage of LLMs if it is an important, original, or

1126 non-standard component of the core methods in this research? Note that if the LLM is used

1127 only for writing, editing, or formatting purposes and does not impact the core methodology,

1128 scientific rigorousness, or originality of the research, declaration is not required.

1129 Answer: [NA]

1130 Justification: The core method development in this research does not involve LLMs as any
1131 important, original, or non-standard components.
1132 Guidelines:
1133 • The answer NA means that the core method development in this research does not
1134 involve LLMs as any important, original, or non-standard components.
1135 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)
1136 for what should or should not be described.