UNDERSTANDING GENERALIZATION IN CLIENT-LEVEL DP-FL: A HIERARCHICAL STABILITY ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Differentially private federated learning (DP-FL) has been extensively studied for privacy and optimization, while its generalization remains less understood. Since data in FL are organized at the client and local-sample levels, they give rise to within-client and unseen-client generalization. We develop a hierarchical distributional-stability framework for client-level DP-FL based on sample- and client-replacement stability, revealing that the heterogeneity term explicitly enters the unseen-client generalization bound, whereas local dataset size explicitly governs the within-client generalization bound. Using total variation (TV) distance and KL divergence, we derive complementary generalization guarantees: the TV-based bound exhibits a tighter dependence on the client participation ratio, whereas the KL-based bound exhibits a tighter dependence on the training horizon. Furthermore, under suitable learning-rate decay, the KL-based bound can remain independent of the training horizon without requiring a finite cumulative step size. Combining these generalization guarantees with optimization analysis, we derive population excess-risk bounds that characterize the optimization–generalization trade-off induced by privacy, training dynamics, and heterogeneity. Experiments across multiple datasets empirically support the hierarchical decomposition and the main theoretical predictions.

1 INTRODUCTION

Federated learning (FL) enables a large number of distributed clients to collaboratively train a shared model without directly exchanging their raw data (McMahan et al., 2017). Despite this decentralized training paradigm, the model updates communicated between clients and the server may still reveal sensitive information through attacks such as membership inference and data reconstruction. Differentially private federated learning (DP-FL) has therefore been widely studied as a principled privacy-preserving framework, typically by clipping client updates or sample gradients and injecting calibrated random noise during training (Wei et al., 2020; Geyer et al., 2017; Andrew et al., 2021; McMahan et al., 2018; Shi et al., 2023). Although existing theoretical studies on DP-FL have mainly focused on privacy accounting and optimization convergence, its generalization behavior remains much less understood.

In FL, data follow a hierarchical generative process in which client distributions are drawn from a meta-distribution and local samples are then drawn from the corresponding client distributions (Yuan et al., 2022; Hu et al., 2023b; Wang et al., 2025). As illustrated in Figure 1, this hierarchy induces two levels of generalization: within-client generalization from observed local samples to unseen samples drawn from the same client distribution, and unseen-client generalization from participating clients to unseen client distributions. However, existing DP-FL generalization analyses often use a single overall generalization gap without explicitly distinguishing these two levels (He et al., 2021; Shi et al., 2025). Consequently, a single overall gap does not reveal whether generalization error arises primarily within observed clients or when extending to unseen clients.

Beyond this hierarchical generalization perspective, client-level DP introduces another fundamental challenge: protecting the entire contribution of a client rather than individual samples (Geyer et al., 2017; McMahan et al., 2018; Zhang et al., 2022a). However, existing privacy-based generalization

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

analyses for DP-FL largely retain the sample-level adjacency used in centralized DP analysis, extending privacy-to-generalization arguments to the federated setting (Shi et al., 2025; Wang et al., 2025). This formulation does not directly track the effect of replacing a client's entire contribution, leaving the joint influence of client heterogeneity and client participation on unseen-client generalization insufficiently characterized.

In this work, we develop a hierarchical distributional stability framework for heterogeneous client-level DP-FL. We analyze unseen-client and within-client generalization through client-replacement and sample-replacement stability, respectively, and derive complementary TV- and KL-based bounds that explicitly characterize how privacy, training dynamics, and client heterogeneity shape federated generalization. We further combine these generalization guarantees with optimization analysis to establish excess risk bounds and characterize the resulting optimization–generalization trade-offs. **To the best of our knowledge, we are the first to provide a hierarchical generalization analysis for client-level DP-FL.** Our key contributions are:

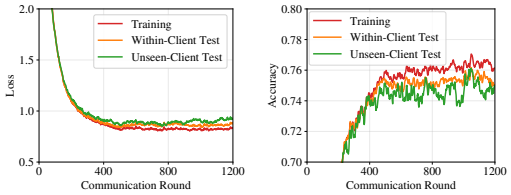

Figure 1: Hierarchical generalization in client-level DP-FL on FEMNIST. Performance generally degrades from training to within-client testing and further to unseen-client testing, illustrating generalization gaps at both levels.

- We develop a hierarchical distributional stability framework for heterogeneous client-level DP-FL that decomposes federated generalization into unseen-client and within-client generalization through client-replacement and sample-replacement stability, respectively. Our analysis reveals explicit dependence on client heterogeneity in client-replacement stability and the corresponding unseen-client generalization bound, while local dataset size enters sample-replacement stability and the corresponding within-client generalization bound.
- We derive complementary TV- and KL-based generalization bounds, with the TV-based bound exhibiting a tighter dependence on the client participation ratio and the KL-based bound exhibiting a tighter dependence on the training horizon. Furthermore, under suitable learning-rate decay, the KL-based bound can remain independent of the training horizon without requiring a finite cumulative step size.
- We combine the generalization guarantees with optimization analysis to derive population excess-risk bounds that characterize the optimization–generalization trade-off induced by privacy, training dynamics, and heterogeneity. Experiments across multiple datasets empirically support our hierarchical generalization framework and the main theoretical findings.

2 Related Work

Generalization in Federated Learning. Theoretical analyses of generalization in federated learning have been developed from several complementary perspectives. One line of work studies population or excess risk by combining statistical learning arguments with the optimization dynamics of federated algorithms, thereby characterizing the effects of local updates, client participation, and data heterogeneity (Chen et al., 2021; Sun et al., 2023). Another line develops PAC-Bayesian and information-theoretic analyses. Sefidgaran et al. (2024) derive PAC-Bayesian and rate-distortion bounds for federated learning, while Hu et al. (2023b) establish generalization guarantees under a two-level sampling framework and Wang et al. (2025) develop a conditional mutual information framework that separately characterizes different sources of federated generalization. Related Bayesian formulations have also been studied for personalized federated learning (Achituve et al., 2021; Zhang et al., 2022b). A further important direction is stability-based generalization analysis, which provides algorithm-dependent guarantees by relating the sensitivity of the learning procedure to perturbations of the training data. Lei et al. (2023) establish stability and generalization bounds for minibatch SGD and local SGD, Sun et al. (2024) characterize how data heterogeneity affects the stability-based generalization of FedAvg, FedProx, and SCAFFOLD, and Liu et al. (2025) extend this perspective to personalized federated learning. Despite these advances, existing stability-based

analyses largely characterize federated generalization through a single generalization gap, rather than explicitly accounting for the hierarchical structure of clients and local samples.

Generalization in Differentially Private Learning and DP-FL. The connection between differential privacy and generalization has been extensively studied in centralized learning. Early work shows that differential privacy can imply stability and generalization guarantees (Dwork et al., 2015; Bassily et al., 2016; Jung et al., 2021). Subsequent studies further characterize the generalization behavior of randomized and noisy learning algorithms from distributional and information-theoretic perspectives (Kulynych et al., 2022; Pensia et al., 2018; Wang et al., 2023; Issa et al., 2023). Beyond generalization guarantees derived from differential privacy, algorithm-dependent analyses track how perturbations propagate through iterative optimization. Li et al. (2020) study non-convex noisy gradient methods through Bayes stability, He et al. (2021) derive tighter generalization guarantees for iterative differentially private algorithms by tracking privacy-induced stability across iterations, Banerjee et al. (2022) and Dadi & Cevher (2025) further investigate stability and generalization in Gaussian-perturbed optimization. In comparison, existing DP-FL generalization results largely build on privacy-based generalization frameworks. He et al. (2021) apply a generic iterative-DP generalization framework to FL, while Shi et al. (2025) derive generalization guarantees from iterative differential privacy under sample-level adjacency. However, both analyses characterize generalization through a single overall gap, without distinguishing within-client generalization from generalization to unseen clients. Wang et al. (2025) further extend privacy-based analysis to hierarchical FL generalization by using sample-level DP to control conditional mutual information across two generalization levels. However, their privacy analysis is based on sample-level adjacency, which does not directly align with the client-level replacements underlying unseen-client generalization, and it does not explicitly characterize the effect of client heterogeneity. Overall, these privacy-based analyses do not explicitly disentangle how privacy, training dynamics, and client heterogeneity contribute to different generalization components. In contrast, we study hierarchical generalization in client-level DP-FL and provide a finer-grained characterization of these effects through distributional stability.

3 Problem Formulation

FL Setup. We consider a standard federated learning (FL) framework consisting of a central server and M clients. Each client $i \in [M]$ holds a local dataset $\mathcal{D}_i = \{z_{i,j}\}_{j=1}^{n_i}$, where $z_{i,j} = (x_{i,j}, y_{i,j}) \overset{\text{iid}}{\sim} \mathcal{P}_i$ and $\mathcal{P}_i \overset{\text{iid}}{\sim} \Pi$. Here, $\mathcal{P}_i$ denotes the local data distribution of client i and Π denotes a meta distribution over client data distributions. The federated dataset is denoted by $\mathcal{D} = (\mathcal{D}_1, \ldots, \mathcal{D}_M)$.

Loss and Risk Functions. Let $\phi(\theta; z)$ denote the per-sample loss for data point z with respect to model parameters θ. The local empirical risk $\hat{\Phi}_i(\theta)$ and global empirical risk $\hat{\Phi}(\theta)$ are defined as $\hat{\Phi}_i(\theta) = \frac{1}{|\mathcal{D}_i|}\sum_{z\in\mathcal{D}_i}\phi(\theta; z)$ and $\hat{\Phi}(\theta) = \frac{1}{M}\sum_{i=1}^{M}\hat{\Phi}_i(\theta)$, respectively. When the underlying federated dataset must be made explicit, we write these same empirical risks as $\hat{\Phi}_{i,\mathcal{D}}(\theta)$ and $\hat{\Phi}_{\mathcal{D}}(\theta)$. Correspondingly, the local population risk and participating-client population risk are defined as $\Phi_i(\theta) = \mathbb{E}_{z\sim\mathcal{P}_i}[\phi(\theta; z)]$ and $\Phi_{\text{part}}(\theta) = \frac{1}{M}\sum_{i=1}^{M}\Phi_i(\theta)$, respectively. Furthermore, the meta population risk over the client meta distribution Π is defined as $\Phi_\Pi(\theta) = \mathbb{E}_{\mathcal{P}\sim\Pi}\mathbb{E}_{z\sim\mathcal{P}}[\phi(\theta; z)]$. Here, $\Phi_{\text{part}}(\theta)$ characterizes the expected risk over the participating client distributions, while $\Phi_\Pi(\theta)$ characterizes the expected risk over a new client distribution drawn from Π.

Generalization Error. For a randomized FL algorithm with output $\theta_{\mathcal{D}}$, we define the unseen-client generalization error and within-client sample generalization error as $\text{Gen}_{\text{c}} = \mathbb{E}[\Phi_\Pi(\theta_{\mathcal{D}}) - \Phi_{\text{part}}(\theta_{\mathcal{D}})]$ and $\text{Gen}_{\text{s}} = \mathbb{E}[\Phi_{\text{part}}(\theta_{\mathcal{D}}) - \hat{\Phi}_{\mathcal{D}}(\theta_{\mathcal{D}})]$, respectively. Accordingly, the total generalization error can be decomposed as $\text{Gen}_{\text{total}} = \text{Gen}_{\text{c}} + \text{Gen}_{\text{s}} = \mathbb{E}[\Phi_\Pi(\theta_{\mathcal{D}}) - \hat{\Phi}_{\mathcal{D}}(\theta_{\mathcal{D}})]$. Here, Gen_{s} characterizes generalization from the observed local samples to new samples from the participating client distributions, while Gen_{c} characterizes generalization from the participating client distributions to an unseen client distribution drawn from Π.

Excess Risk. Let $\theta^\star \in \arg\min_\theta \Phi_\Pi(\theta)$ and $\hat{\theta}^\star \in \arg\min_\theta \hat{\Phi}(\theta)$ denote the optimal models with respect to the meta population risk and global empirical risk, respectively. For a randomized DP-FL algorithm that outputs $\theta^{T\tau}$ after T communication rounds and τ local steps per round, since

$\hat{\Phi}(\hat{\theta}^\star) \leq \hat{\Phi}(\theta^\star)$ and $\mathbb{E}_{\mathcal{D}}[\hat{\Phi}_{\mathcal{D}}(\theta^\star)] = \Phi_\Pi(\theta^\star)$, the expected population excess risk can be upper bounded by the total generalization error and the optimization error as

$$\mathbb{E}\left[\Phi_\Pi(\theta^{T\tau}) - \Phi_\Pi(\theta^\star)\right] \leq \mathrm{Gen_c} + \mathrm{Gen_s} + \mathbb{E}\left[\hat{\Phi}(\theta^{T\tau}) - \hat{\Phi}(\hat{\theta}^\star)\right], \tag{1}$$

where the last term is the optimization error.

Client-Level DP-FL. We consider central differential privacy with respect to the sequence of aggregated global models released during training, while individual client updates remain internal to a trusted aggregation functionality and are not observed by the adversary. Client-level DP aims to protect the entire contribution of an individual client. Two federated datasets $\mathcal{D}$ and $\mathcal{D}'$ are said to be client-level neighboring, denoted by $\mathcal{D} \simeq_c \mathcal{D}'$, if one can be obtained from the other by adding or removing the entire local dataset of a single client. A randomized FL algorithm $\mathcal{A}$ satisfies (ϵ, δ) differential privacy if, for any $\mathcal{D} \simeq_c \mathcal{D}'$ and any measurable set $\mathcal{S}$, $\Pr[\mathcal{A}(\mathcal{D}) \in \mathcal{S}] \leq e^\epsilon \Pr[\mathcal{A}(\mathcal{D}') \in \mathcal{S}] + \delta$.

At communication round t, the server samples a subset of clients $\mathcal{M}_t \subseteq [M]$ with $|\mathcal{M}_t| = m$ and broadcasts the current global model $\theta^{t\tau}$ to the selected clients. Each selected client initializes $\theta_i^{t\tau} = \theta^{t\tau}$ and performs τ local SGD steps. At local step $r \in \{0, \ldots, \tau - 1\}$, client i uniformly samples a mini-batch $\mathcal{B}_i^{t\tau+r} \subseteq \mathcal{D}_i$ of size b and updates $\theta_i^{t\tau+r+1} = \theta_i^{t\tau+r} - \frac{\eta_t}{b} \sum_{z \in \mathcal{B}_i^{t\tau+r}} \nabla_\theta \phi(\theta_i^{t\tau+r}; z)$. The local update is $\Delta_i^t \triangleq \theta_i^{t\tau+\tau} - \theta_i^{t\tau}$, and is clipped as $\bar{\Delta}_i^t = \Delta_i^t \min\left\{1, \frac{C}{\|\Delta_i^t\|_2}\right\}$. The trusted aggregation functionality then averages the clipped updates and adds Gaussian noise $\theta^{(t+1)\tau} = \theta^{t\tau} + \frac{1}{m}\sum_{i \in \mathcal{M}_t} \bar{\Delta}_i^t + \xi^t$, where $\xi^t \sim \mathcal{N}\left(0, \frac{\sigma^2 C^2}{m^2} I_d\right)$, σ is the noise multiplier, and d is the model dimension. The detailed algorithmic procedure is provided in Appendix B.

4 Stability-Based Generalization Analysis

We now analyze the hierarchical generalization behavior of client-level DP-FL through an algorithmic stability framework. To characterize the two levels of generalization, we introduce sample- and client-replacement perturbations.[1]

For the within-client sample generalization analysis, let $\mathcal{D}^{(u,v)} = (\mathcal{D}_1, \ldots, \mathcal{D}_u^{(v)}, \ldots, \mathcal{D}_M)$ denote the sample-replacement dataset, where $\mathcal{D}_u^{(v)} = \mathcal{D}_u \setminus \{z_{u,v}\} \cup \{z'_{u,v}\}$ and $z'_{u,v} \sim \mathcal{P}_u$. For the unseen-client generalization analysis, we independently draw $\mathcal{P}'_u \sim \Pi$ and $\mathcal{D}'_u \sim (\mathcal{P}'_u)^{|\mathcal{D}_u|}$, and define the client-replacement dataset as $\mathcal{D}^{(u)} = (\mathcal{D}_1, \ldots, \mathcal{D}'_u, \ldots, \mathcal{D}_M)$.

Definition 1 (Distributional Stability)**.** Let P_θ, $P_{\tilde{\theta}_\mathrm{s}}$, and $P_{\tilde{\theta}_\mathrm{c}}$ denote the output distributions of the randomized algorithm on $\mathcal{D}$, $\mathcal{D}^{(u,v)}$, and $\mathcal{D}^{(u)}$, respectively. The algorithm is Γ_s sample-replacement and Γ_c client-replacement distributionally stable if

$$\sup_z \left|\mathbb{E}_{\theta \sim P_\theta}[\phi(\theta; z)] - \mathbb{E}_{\tilde{\theta}_\mathrm{s} \sim P_{\tilde{\theta}_\mathrm{s}}}[\phi(\tilde{\theta}_\mathrm{s}; z)]\right| \leq \Gamma_\mathrm{s}, \quad \sup_z \left|\mathbb{E}_{\theta \sim P_\theta}[\phi(\theta; z)] - \mathbb{E}_{\tilde{\theta}_\mathrm{c} \sim P_{\tilde{\theta}_\mathrm{c}}}[\phi(\tilde{\theta}_\mathrm{c}; z)]\right| \leq \Gamma_\mathrm{c}. \tag{2}$$

Lemma 1 (Generalization from Distributional Stability)**.** *Suppose that a randomized learning algorithm $\mathcal{A}$ is Γ_s sample-replacement distributionally stable and Γ_c client-replacement distributionally stable. Then its expected generalization error satisfies*

$$\mathrm{Gen_{total}} \leq \Gamma_\mathrm{s} + \Gamma_\mathrm{c}. \tag{3}$$

The detailed proof is provided in Appendix C.

Therefore, to bound the total generalization error of DP-FL, it suffices to derive upper bounds on the sample-replacement and client-replacement distributional stability parameters Γ_s and Γ_c, respectively. We give some assumptions first.

Assumption 1. The per-sample loss $\phi(\cdot; z)$ is β-smooth with respect to the model parameter, i.e., $\|\nabla\phi(\theta; z) - \nabla\phi(\theta'; z)\|_2 \leq \beta \|\theta - \theta'\|_2, \quad \forall\theta, \theta', \forall z.$

[1]The replacement perturbations are used solely for stability-based generalization analysis, whereas client-level differential privacy is defined under add/remove adjacency.

Assumption 2. For both the original federated dataset $\mathcal{D}$ and any client-replacement dataset $\mathcal{D}^{(u)}$, the stochastic gradient $g_i(\theta, z) := \nabla_\theta \phi(\theta; z)$ is unbiased with uniformly bounded variance, i.e., $\mathbb{E}[g_i(\theta, z)] = \nabla \hat{\Phi}_{i,\mathcal{D}'}(\theta), \quad \mathbb{E}\left[\left\| g_i(\theta, z) - \nabla \hat{\Phi}_{i,\mathcal{D}'}(\theta) \right\|_2^2\right] \le \varsigma^2, \quad \forall i, \theta, \quad \mathcal{D}' \in \{\mathcal{D}, \mathcal{D}^{(u)}\}$.

Assumption 3. For both the original federated dataset $\mathcal{D}$ and any client-replacement dataset $\mathcal{D}^{(u)}$, the gradient dissimilarity is uniformly bounded, i.e., $\left\| \nabla \hat{\Phi}_{i,\mathcal{D}'}(\theta) - \nabla \hat{\Phi}_{\mathcal{D}'}(\theta) \right\|_2 \le G, \quad \forall i, \theta, \quad \mathcal{D}' \in \{\mathcal{D}, \mathcal{D}^{(u)}\}$.

Assumption 4. The per-sample loss is bounded, i.e., $0 \le \phi(\theta; z) \le B, \quad \forall \theta, \forall z$.

Assumption 5. The global empirical risk $\hat{\Phi}(\theta)$ satisfies a trajectory-wise expected μ-PŁ condition along the optimization path, i.e., $\mathbb{E}\left[\hat{\Phi}(\theta^{t\tau}) - \hat{\Phi}(\hat{\theta}^\star)\right] \le \frac{1}{2\mu}\mathbb{E}\left[\left\| \nabla \hat{\Phi}(\theta^{t\tau}) \right\|_2^2\right], \quad 0 \le t \le T$.

Notably, Assumptions 1–4 are standard in the stability-based generalization literature (Hardt et al., 2016; Sun et al., 2023; Liu et al., 2025), whereas Assumption 5 is introduced only in the excess-risk analysis to convert non-convex stationarity guarantees into empirical suboptimality bounds. We next characterize the hierarchical generalization of client-level DP-FL through two complementary distributional measures: total variation (TV) distance and KL divergence. For two probability distributions P and P', we define

$$\mathrm{TV}(P, P') = \sup_A |P(A) - P'(A)|, \qquad \mathrm{KL}(P \| P') = \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}P'}\right) \mathrm{d}P. \tag{4}$$

TV distance quantifies the maximal discrepancy between two distributions over all measurable events A, while KL divergence measures their relative divergence. Next, we establish stability and generalization bounds under these two distributional measures.

Theorem 1 (TV-Based Client-Level DP-FL)**.** *Under Assumptions 1–4, suppose that the learning rate is constant within each communication round. Then, for $\sigma > 0$, the generalization errors of client-level DP-FL satisfy*

$$\begin{aligned}
&\mathrm{Gen_c} \le \Gamma_\mathrm{c} \le B \min\left\{1, \frac{m}{M\sigma}\sqrt{\frac{2}{\pi}} \min\left\{T, \frac{\tau}{C}\left(\varsigma + \frac{MG}{M-1}\right)\sum_{t=0}^{T-1} \eta_t e^{\eta_t \beta \tau}\right\}\right\}, \\
&\mathrm{Gen_s} \le \Gamma_\mathrm{s} \le B \min\left\{1, \frac{m\tau b}{M|\mathcal{D}_{\min}|\sigma}\sqrt{\frac{2}{\pi}} \min\left\{T, \frac{\tau\sqrt{2\beta B}}{bC}\sum_{t=0}^{T-1} \eta_t e^{\eta_t \beta \tau}\right\}\right\}, \\
&\mathrm{Gen_{total}} \le \Gamma_\mathrm{c} + \Gamma_\mathrm{s},
\end{aligned} \tag{5}$$

where $|\mathcal{D}_{\min}|$ denotes the minimum local dataset size among all clients, and the detailed proof is provided in Appendix D.

Theorem 2 (KL-Based Client-Level DP-FL)**.** *Under Assumptions 1–4, suppose that the learning rate is constant within each communication round. Then, for $\sigma > 0$, the generalization errors of client-level DP-FL satisfy*

$$\begin{aligned}
&\mathrm{Gen_c} \le \Gamma_\mathrm{c} \le B \min\left\{1, \frac{1}{\sigma}\sqrt{\frac{m}{M}} \min\left\{\sqrt{T}, \frac{\tau}{C}\left(\varsigma + \frac{MG}{M-1}\right)\left(\sum_{t=0}^{T-1} \eta_t^2 e^{2\eta_t \beta \tau}\right)^{\frac{1}{2}}\right\}\right\}, \\
&\mathrm{Gen_s} \le \Gamma_\mathrm{s} \le B \min\left\{1, \frac{1}{\sigma}\sqrt{\frac{m\tau b}{M|\mathcal{D}_{\min}|}} \min\left\{\sqrt{T}, \frac{\tau\sqrt{2\beta B}}{bC}\left(\sum_{t=0}^{T-1} \eta_t^2 e^{2\eta_t \beta \tau}\right)^{\frac{1}{2}}\right\}\right\}, \\
&\mathrm{Gen_{total}} \le \Gamma_\mathrm{c} + \Gamma_\mathrm{s}.
\end{aligned} \tag{6}$$

The detailed proof is provided in Appendix E.

Remark 1 (Comparison of TV- and KL-based bounds)**.** *Unlike in centralized Gaussian DP-SGD, where the TV-based bound is tighter than the KL-based bound (Wang et al., 2023), the two bounds exhibit complementary dependencies on client participation and the training horizon in client-level*

DP-FL. The TV-based bound has a tighter dependence on the client participation ratio, scaling as $\frac{m}{M}$ rather than $\sqrt{\frac{m}{M}}$, whereas the KL-based bound has a tighter dependence on the training horizon, involving $\left(\sum_{t=0}^{T-1}\eta_t^2 e^{2\eta_t\beta\tau}\right)^{1/2}$ rather than $\sum_{t=0}^{T-1}\eta_t e^{\eta_t\beta\tau}$.

Remark 2 (Distinct dependencies across generalization levels)**.** *The unseen-client generalization bound (via client-replacement stability) depends on the client heterogeneity measure G, whereas the within-client generalization bound (via sample-replacement stability) depends on the local dataset size. Replacing a client changes its local data distribution, introducing dependence on client heterogeneity through G. In contrast, sample replacement affects only one sample's contribution to the local empirical distribution, whose weight is inversely proportional to the local dataset size.*

Corollary 1 (Effect of learning-rate decay)**.** *Under the conditions of Theorems 1 and 2, suppose that the round-wise learning rate satisfies $\eta_t = \frac{\eta_0}{(t+1)^a}$ for $a \geq 0$. Suppressing factors independent of T, the TV-based bound $\mathrm{Gen}_{\mathrm{total}}^{\mathrm{TV}}$ and the KL-based bound $\mathrm{Gen}_{\mathrm{total}}^{\mathrm{KL}}$ satisfy* [2]

$$\mathrm{Gen}_{\mathrm{total}}^{\mathrm{TV}} = \begin{cases} \mathcal{O}\left(T^{1-a}\right), & 0 \leq a < 1, \\ \mathcal{O}\left(\log T\right), & a = 1, \\ \mathcal{O}\left(1\right), & a > 1, \end{cases} \qquad \mathrm{Gen}_{\mathrm{total}}^{\mathrm{KL}} = \begin{cases} \mathcal{O}\left(T^{0.5-a}\right), & 0 \leq a < 0.5, \\ \mathcal{O}\left(\sqrt{\log T}\right), & a = 0.5, \\ \mathcal{O}\left(1\right), & a > 0.5. \end{cases} \tag{7}$$

Remark 3 (Horizon independence with non-summable learning rates)**.** *When $0.5 < a \leq 1$, we have $\sum_t \eta_t = \infty$ while $\sum_t \eta_t^2 < \infty$. Hence, the KL-based generalization bound remains $\mathcal{O}(1)$ even though the cumulative step size is unbounded. This shows that horizon independence arises from the square-summable accumulation in the KL-based stability analysis rather than from a finite total optimization movement. In contrast, the TV-based bound requires $a > 1$, for which $\sum_t \eta_t < \infty$.*

Remark 4 (Comparison of training-horizon dependence)**.** *Unlike stability-based FL generalization analyses that employ $\eta_t = \mathcal{O}(\frac{1}{t})$ yet obtain a horizon-dependent bound of order $\mathcal{O}\left(T^{\lambda}\right)$ for some $0 < \lambda < 1$ (Liu et al., 2025; Sun et al., 2024), our KL-based analysis achieves an $\mathcal{O}(1)$ bound under the less restrictive decay condition $\eta_t = \frac{\eta_0}{(t+1)^a}$ with $a > 0.5$. Moreover, existing horizon-independent generalization guarantees for noisy iterative learning rely on conditions such as dissipativity (Futami & Fujisawa, 2023), isoperimetric inequalities (Dadi & Cevher, 2025), or restrictions on batch size and the learning-rate-to-noise ratio (Li et al., 2020), whereas our result requires no such assumptions.*

We next consider a fixed client-level privacy budget (ϵ, δ). Existing analyses of client-level DP-FL (Hu et al., 2023a; Zhang et al., 2022a) show that the noise multiplier can be calibrated as

$$\sigma = \kappa \frac{m}{M} \frac{\sqrt{T \log(\frac{1}{\delta})}}{\epsilon}, \quad \kappa > 0. \tag{8}$$

Corollary 2 (Generalization under Privacy Calibration)**.** *Under the conditions of Theorems 1 and 2, suppose that the learning rate is constant and the noise multiplier is calibrated according to condition (8) to ensure client-level (ϵ, δ)-DP. Then, the total generalization error satisfies*

$$\mathrm{Gen}_{\mathrm{total}} \leq \frac{B\epsilon}{\kappa\sqrt{\log(1/\delta)}} \min\left\{ \sqrt{\frac{2T}{\pi}} \left(\Lambda_c + \frac{\tau b}{|\mathcal{D}_{\min}|} \Lambda_s \right), \sqrt{\frac{M}{m}} \left(\Lambda_c + \sqrt{\frac{\tau b}{|\mathcal{D}_{\min}|}} \Lambda_s \right) \right\}, \tag{9}$$

where $\Lambda_c = \min\left\{1, \frac{\eta\tau e^{\eta\beta\tau}}{C}\left(\varsigma + \frac{MG}{M-1}\right)\right\}, \Lambda_s = \min\left\{1, \frac{\eta\tau e^{\eta\beta\tau}\sqrt{2\beta B}}{bC}\right\}$. The detailed proof is provided in Appendix F.

Remark 5 (Comparison with privacy-based guarantees)**.** *Under a fixed privacy budget and a constant learning rate, our KL-based bound exhibits an $\mathcal{O}(1)$ dependence on T, matching the horizon dependence of existing privacy-derived guarantees when their privacy parameters are held fixed (He et al., 2021; Wang et al., 2025). Unlike these guarantees, our bound additionally makes explicit the joint dependence on federated factors such as client heterogeneity and the participation ratio.*

[2] Here we focus on the nontrivial branch obtained by relaxing the outer truncation; see Appendix A for a detailed discussion.

Corollary 3 (Excess risk of client-level DP-FL)**.** *Under Assumptions 1–5, Proposition 1 implies that the optimization error of client-level DP-FL with a constant learning rate η satisfying $\eta \leq \min\{\frac{1}{64\beta\tau}, \frac{1}{\mu\tau}\}$ is bounded by*

$$\mathbb{E}\left[\hat{\Phi}(\theta^{T\tau}) - \hat{\Phi}(\hat{\theta}^{\star})\right] \leq \mathcal{O}\left(q^T + \mathcal{F}(C) + \frac{m+1}{m}\left(G^2 + \frac{\varsigma^2}{\tau}\right) + \frac{\sigma^2 C^2}{m^2\tau}\right), \tag{10}$$

where $q = 1 - \frac{\mu\eta\tau}{2}$, and $\mathcal{F}(C)$ collects the clipping bias term. Combining this optimization bound with the specializations of Theorems 1 and 2 under a constant learning rate yields

$$\mathbb{E}\left[\Phi_{\Pi}(\theta^{T\tau}) - \Phi_{\Pi}(\theta^{\star})\right] \leq \mathcal{O}\Bigg(q^T + \mathcal{F}(C) + \frac{m+1}{m}\left(G^2 + \frac{\varsigma^2}{\tau}\right) + \frac{\sigma^2 C^2}{m^2\tau} + \min\left\{\frac{mT\tau^2(1+G)}{M\sigma C}, \frac{\sqrt{mT}\tau^{\frac{3}{2}}(1+G)}{\sqrt{M}\sigma C}\right\}\Bigg). \tag{11}$$

Corollary 3 reveals how privacy, training dynamics, and heterogeneity jointly shape the optimization–generalization trade-off in client-level DP-FL. Stronger Gaussian perturbation improves distributional stability but increases the optimization error. The clipping threshold C can have a non-monotonic effect through its influence on clipping bias, noise magnitude, and stability. Increasing client participation m can improve optimization while enlarging the stability-based generalization bound, whereas additional local steps τ balance optimization progress against local drift and instability. Greater client heterogeneity worsens both the optimization and generalization terms, thereby increasing the excess-risk bound.

5 Experiments

In this section, we evaluate our theoretical predictions on hierarchical generalization, training-horizon dependence, and optimization–generalization trade-offs across multiple datasets. Unless otherwise specified, all experiments use LeNet (LeCun et al., 1998), cross-entropy loss, a mini-batch size of 64, and SGD with a learning rate of 0.1. Our code is available at `https://anonymous.4open.science/r/ICLR_27`.

5.1 Details of Figure 1

We illustrate hierarchical generalization using TFF Federated EMNIST (FEMNIST) (Cohen et al., 2017), preserving its original writer-based partition. We randomly select 625 writers as clients, with 500 participating in training and 125 held out as unseen clients. We use $(M, m, T, \tau, C, \sigma) = (500, 50, 2000, 5, 0.2, 0.9)$. Participating clients train on their local training samples, and both participating and unseen clients are evaluated on their respective test samples.

5.2 Effects of Heterogeneity and Data Size on Hierarchical Generalization

In this section, we conduct two experiments on Fashion-MNIST (Xiao et al., 2017) to examine the effects of heterogeneity and local dataset size on hierarchical generalization. Unless otherwise specified, both experiments use $(M, m, T, \tau, C, \sigma) = (500, 50, 1200, 5, 0.2, 0.9)$, with 500 participating clients and 125 additional unseen clients.

We first employ Dirichlet data partitioning (Hsu et al., 2019) to examine the effect of heterogeneity strength. For each client, the class proportions are sampled from $\mathrm{Dir}(\alpha)$ while keeping the number of samples per client fixed. The same class proportions are used for the training and test data of each participating client, while unseen clients are independently generated from the same Dirichlet distribution. We vary α, where a smaller α corresponds to stronger heterogeneity. Then, we isolate the effect of local dataset size under an IID client partition. We vary the local dataset size as $|\mathcal{D}_i| \in \{80, 100, 120\}$. All participating clients have the same number of training samples within each setting, while the seen-client and unseen-client test partitions are held fixed across the three settings.

As shown in Figure 2, decreasing α substantially enlarges the unseen-client generalization gap (unseen-client test loss minus seen-client test loss), whereas the within-client generalization gap

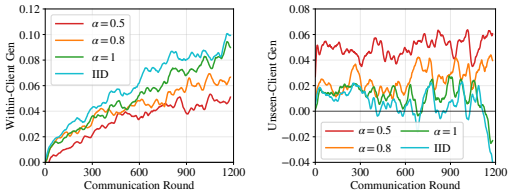

Figure 2: Effect of data heterogeneity on within-client generalization (left) and unseen-client generalization (right).

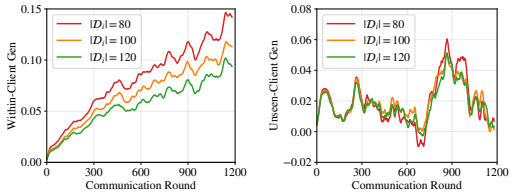

Figure 3: Effect of local dataset size on within-client generalization (left) and unseen-client generalization (right).

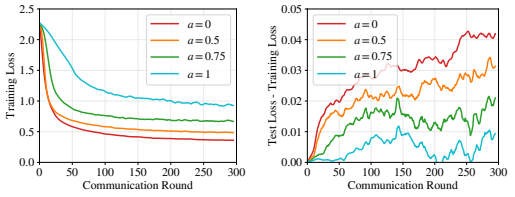

Figure 4: Effect of learning-rate decay on optimization and generalization in client-level DP-FL.

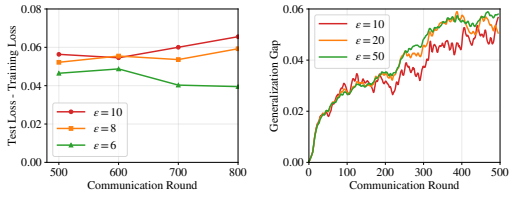

Figure 5: Generalization gaps across different training horizons (left) and gap trajectories under different privacy budgets ϵ (right).

(seen-client test loss minus training loss) does not exhibit a corresponding increase. Figure 3 shows that increasing the local dataset size leaves the unseen-client gap largely unchanged while reducing the within-client gap. These complementary observations empirically support the hierarchical decomposition: client heterogeneity primarily affects unseen-client generalization, whereas the amount of local data primarily governs within-client generalization.

5.3 Empirical Validation of Horizon-Independent Generalization

We next empirically examine the two mechanisms through which our analysis yields horizon-independent generalization. Both experiments are conducted on Fashion-MNIST (Xiao et al., 2017) with LeNet under an IID client partition, using $M = 500$, $m = 50$, $\tau = 5$, and $C = 0.2$.

Horizon Independence via Learning-Rate Decay. To examine Corollary 1, we fix the noise multiplier at $\sigma = 0.9$ and set $T = 300$. The round-wise learning rate follows $\eta_t = \frac{\eta_0}{(t+1)^a}$, where $a \in \{0, 0.5, 0.75, 1\}$. As shown in Figure 4, faster learning-rate decay progressively suppresses the accumulation of the generalization gap, at the cost of slower optimization. With $a = 0$, the gap continues to increase over training, whereas its growth becomes weaker for $a = 0.5$ and remains substantially more stable for $a = 0.75$ and $a = 1$. This behavior is qualitatively consistent with the $\mathcal{O}(1)$ KL-based bound predicted for $a > 0.5$. In particular, $a = 0.75$ lies in the regime $0.5 < a \leq 1$, where $\sum_t \eta_t = \infty$ but $\sum_t \eta_t^2 < \infty$, showing that the observed suppression does not rely on a summable learning-rate schedule.

Horizon Independence under a Fixed Privacy Budget. We further examine Corollary 2 under a constant learning rate. For each target horizon T, we independently train a model with the noise multiplier σ separately calibrated by an RDP accountant (Mironov, 2017; Wang et al., 2019) to satisfy the prescribed (ϵ, δ) privacy budget, where $\delta = \frac{1}{M} = 0.002$. Figure 5 (left) shows that, under multiple fixed privacy budgets, the terminal generalization gap remains within a comparable range across different training horizons. Figure 5 (right) further shows that the generalization gap generally increases with a larger privacy budget ϵ. These observations are consistent with Corollary 2.

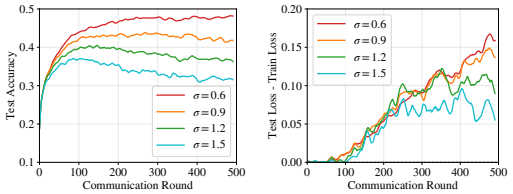

Figure 6: Test accuracy and generalization gap under different noise multipliers for client-level DP-FL.

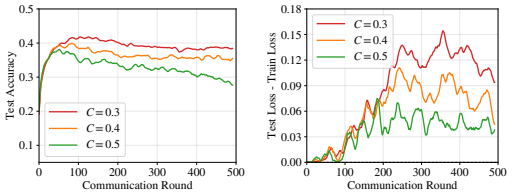

Figure 7: Test accuracy and generalization gap under different clipping thresholds for client-level DP-FL.

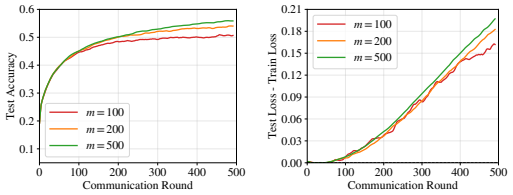

Figure 8: Test accuracy and generalization gap under different numbers of participating clients m for client-level DP-FL.

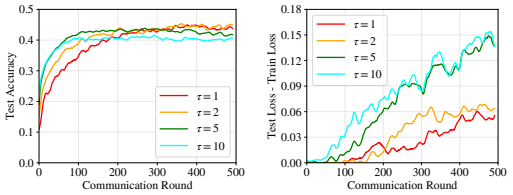

Figure 9: Test accuracy and generalization gap under different numbers of local steps τ for client-level DP-FL.

5.4 Empirical Evaluation of Influential Factors

We next evaluate the key factors identified by our theory on CIFAR-10 (Krizhevsky et al., 2009), focusing on the total generalization gap, empirically measured as the difference between the test loss and the training loss. Unless otherwise specified, we use an IID client partition with $(M, m, T, \tau, C, \sigma) = (500, 50, 500, 5, 0.2, 0.9)$. Each experiment varies one factor while keeping the others fixed. Overall, the observed trends are consistent with the parameter dependencies predicted by Theorem 1 and the optimization–generalization trade-offs characterized in Corollary 3.

Figures 6–9 illustrate the effects of key training and privacy factors. **(i) Communication rounds T and noise multiplier σ:** the generalization gap generally increases with T but decreases with a larger σ, while a larger σ reduces test accuracy, illustrating the trade-off between generalization and optimization. **(ii) Clipping threshold C:** increasing C generally reduces the generalization gap while decreasing test accuracy over the tested range, reflecting the competing effects of reduced clipping bias, improved stability, and the increased scale of Gaussian perturbation. **(iii) Client participation m:** increasing m generally enlarges the generalization gap while improving test accuracy; more participating clients strengthen the stability dependence on client participation but reduce aggregation noise and client-sampling variability. **(iv) Local steps τ:** increasing τ generally enlarges the generalization gap, while additional local computation can improve optimization, consistent with the stronger accumulation of perturbations caused by data replacement over more local updates.

6 Conclusion

In this work, we developed a distributional stability framework for client-level DP-FL that characterizes unseen-client and within-client generalization through client-replacement and sample-replacement stability, respectively. Our analysis identifies the respective effects of client heterogeneity and local dataset size on these two levels. We derived complementary TV- and KL-based generalization bounds, with the TV-based bound exhibiting a tighter dependence on the participation ratio and the KL-based bound exhibiting a tighter dependence on the training horizon. Furthermore, the KL-based bound can remain independent of the training horizon under suitable decaying, non-summable learning-rate schedules. Combining these guarantees with optimization analysis yields excess risk bounds capturing the trade-offs induced by privacy, training dynamics, and heterogeneity. Experiments across multiple datasets validate our main theoretical findings.

AI USE STATEMENT

Generative AI tools were used to aid and polish manuscript writing, support retrieval and discovery of relevant literature, and assist with code development and debugging. All AI-assisted code was reviewed and verified by the authors before use. All mathematical arguments, experimental results, and final claims were independently reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

REFERENCES

Idan Achituve, Aviv Shamsian, Aviv Navon, Gal Chechik, and Ethan Fetaya. Personalized federated learning with Gaussian processes. In *Advances in Neural Information Processing Systems*, 2021.

Galen Andrew, Om Thakkar, Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. In *Advances in Neural Information Processing Systems*, 2021.

Arindam Banerjee, Tiancong Chen, Xinyan Li, and Yingxue Zhou. Stability based generalization bounds for exponential family Langevin dynamics. In *International Conference on Machine Learning*, 2022.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, 2016.

Shuxiao Chen, Qinqing Zheng, Qi Long, and Weijie J Su. A theorem of the alternative for personalized federated learning. *arXiv preprint arXiv:2103.01901*, 2021.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. EMNIST: Extending MNIST to handwritten letters. In *International Joint Conference on Neural Networks*, 2017.

Leello Tadesse Dadi and Volkan Cevher. Generalization of noisy SGD in unbounded non-convex settings. In *International Conference on Machine Learning*, 2025.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems*, 2015.

Futoshi Futami and Masahiro Fujisawa. Time-independent information-theoretic generalization bounds for SGLD. In *Advances in Neural Information Processing Systems*, 2023.

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

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, 2016.

Fengxiang He, Bohan Wang, and Dacheng Tao. Tighter generalization bounds for iterative differentially private learning algorithms. In *Uncertainty in Artificial Intelligence*, 2021.

Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*, 2019.

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

Xiaolin Hu, Shaojie Li, and Yong Liu. Generalization bounds for federated learning: Fast rates, unparticipating clients and unbounded losses. In *International Conference on Learning Representations*, 2023b.

Ibrahim Issa, Amedeo Roberto Esposito, and Michael Gastpar. Generalization error bounds for noisy, iterative algorithms via maximal leakage. In *Annual Conference on Learning Theory*, 2023.

Christopher Jung, Katrina Ligett, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Moshe Shenfeld. A new analysis of differential privacy's generalization guarantees. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, 2021.

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

Bogdan Kulynych, Yao-Yuan Yang, Yaodong Yu, Jarosław Błasiok, and Preetum Nakkiran. What you see is what you get: Principled deep learning via distributional generalization. In *Advances in Neural Information Processing Systems*, 2022.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

Yunwen Lei, Tao Sun, and Mingrui Liu. Stability and generalization for minibatch SGD and local SGD. *arXiv preprint arXiv:2310.01139*, 2023.

Jian Li, Xuanyuan Luo, and Mingda Qiao. On generalization error bounds of noisy gradient methods for non-convex learning. In *International Conference on Learning Representations*, 2020.

Yingqi Liu, Qinglun Li, Jie Tan, Yifan Shi, Li Shen, and Xiaochun Cao. Understanding the stability-based generalization of personalized federated learning. In *International Conference on Learning Representations*, 2025.

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

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

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium*. IEEE, 2017.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Generalization error bounds for noisy, iterative algorithms. In *IEEE International Symposium on Information Theory*, 2018.

Milad Sefidgaran, Romain Chor, Abdellatif Zaidi, and Yijun Wan. Lessons from generalization error analysis of federated learning: You may communicate less often! In *International Conference on Machine Learning*, 2024.

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

Yifan Shi, Kang Wei, Li Shen, Yingqi Liu, Xueqian Wang, Bo Yuan, and Dacheng Tao. Towards the flatter landscape and better generalization in federated learning under client-level differential privacy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Yan Sun, Li Shen, and Dacheng Tao. Understanding how consistency works in federated learning via stage-wise relaxed initialization. In *Advances in Neural Information Processing Systems*, 2023.

Zhenyu Sun, Xiaochun Niu, and Ermin Wei. Understanding generalization of federated learning via stability: Heterogeneity matters. In *International Conference on Artificial Intelligence and Statistics*, 2024.

Hao Wang, Rui Gao, and Flavio P Calmon. Generalization bounds for noisy iterative algorithms using properties of additive noise channels. *Journal of Machine Learning Research*, 2023.

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled Rényi differential privacy and analytical moments accountant. In *International Conference on Artificial Intelligence and Statistics*, 2019.

Ziqiao Wang, Cheng Long, and Yongyi Mao. Generalization in federated learning: A conditional mutual information framework. In *International Conference on Machine Learning*, 2025.

Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 2020.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. 2017.

Honglin Yuan, Warren Morningstar, Lin Ning, and Karan Singhal. What do we mean by generalization in federated learning? In *International Conference on Learning Representations*, 2022.

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

Xu Zhang, Yinchuan Li, Wenpeng Li, Kaiyang Guo, and Yunfeng Shao. Personalized federated learning via variational Bayesian inference. In *International Conference on Machine Learning*, 2022b.

A Discussion on the nontrivial branch

Since Assumption 4 induces a trivial constant cap through the outer $\min\{1, \cdot\}$ in Theorems 1 and 2, when discussing the dependence on the training horizon in Corollary 1, we focus on the nontrivial branch of the stability-based upper bounds. Specifically, we relax the outer truncation using $\min\{1, x\} \leq x$ and study how the resulting upper bound accumulates with T. This allows us to distinguish horizon-independent stability of the relaxed bound from horizon independence caused solely by the bounded-loss truncation.

B Algorithmic Details of DP-FL

Algorithm 1 Client-Level DP-FL.

Input: Local datasets $\{\mathcal{D}_i\}_{i=1}^M$, communication rounds T, selected clients per round m, local steps τ, mini-batch size b, learning rate η_t, clipping threshold C, noise multiplier σ, initial global model θ^0.
Output: Final global model $\theta^{T\tau}$.
for communication round $t = 0, 1, \ldots, T-1$ **do**
  Server samples a client subset $\mathcal{M}_t \subseteq [M]$ with $|\mathcal{M}_t| = m$, and broadcasts $\theta^{t\tau}$ to all clients $i \in \mathcal{M}_t$.
  for each client $i \in \mathcal{M}_t$ in parallel **do**
    Initialize local model $\theta_i^{t\tau} \leftarrow \theta^{t\tau}$.
    for local step $r = 0, 1, \ldots, \tau - 1$ **do**
      Client i uniformly samples a mini-batch $\mathcal{B}_i^{t\tau+r} \subseteq \mathcal{D}_i$ with $|\mathcal{B}_i^{t\tau+r}| = b$, and updates $\theta_i^{t\tau+r+1} = \theta_i^{t\tau+r} - \frac{\eta_t}{b} \sum_{z \in \mathcal{B}_i^{t\tau+r}} \nabla_\theta \phi(\theta_i^{t\tau+r}; z)$.
    end for
    Compute the local update $\Delta_i^t \leftarrow \theta_i^{t\tau+\tau} - \theta^{t\tau}$.
    Clip the local update as $\bar{\Delta}_i^t \leftarrow \Delta_i^t \min\left\{1, \frac{C}{\|\Delta_i^t\|_2}\right\}$ and submit $\bar{\Delta}_i^t$ to the trusted aggregation functionality.
  end for
  The aggregation functionality draws $\xi^t \sim \mathcal{N}\left(0, \frac{\sigma^2 C^2}{m^2} I_d\right)$ and updates $\theta^{(t+1)\tau} = \theta^{t\tau} + \frac{1}{m} \sum_{i \in \mathcal{M}_t} \bar{\Delta}_i^t + \xi^t$.
end for

C Proof of Lemma 1

Proof. Let $\theta_{\mathcal{D}} \sim P_{\mathcal{A}(\mathcal{D})}$ denote the output of the randomized algorithm on the federated dataset $\mathcal{D}$. We first consider the unseen-client generalization error. Let $z'_u \sim \mathcal{P}'_u$ and $z_u \sim \mathcal{P}_u$ be independent samples. By the definitions of Φ_Π and Φ_{part},

$$\mathrm{Gen}_{\mathrm{c}} = \frac{1}{M} \sum_{u=1}^{M} \mathbb{E}\left[\phi(\theta_{\mathcal{D}}; z'_u) - \phi(\theta_{\mathcal{D}}; z_u)\right]. \tag{12}$$

Since $(\mathcal{P}_u, \mathcal{D}_u, z_u)$ and $(\mathcal{P}'_u, \mathcal{D}'_u, z'_u)$ are identically distributed,

$$\mathbb{E}\left[\phi(\theta_{\mathcal{D}}; z_u)\right] = \mathbb{E}\left[\phi(\theta_{\mathcal{D}^{(u)}}; z'_u)\right]. \tag{13}$$

Hence,

$$\mathrm{Gen}_{\mathrm{c}} = \frac{1}{M} \sum_{u=1}^{M} \mathbb{E}\left[\phi(\theta_{\mathcal{D}}; z'_u) - \phi(\theta_{\mathcal{D}^{(u)}}; z'_u)\right] \leq \Gamma_{\mathrm{c}}, \tag{14}$$

where the inequality follows from Γ_{c}-client-replacement distributional stability.

We next consider the within-client sample generalization error. Let $z'_{u,v} \sim \mathcal{P}_u$ be an independent sample. By the definitions of Φ_{part} and $\hat{\Phi}_{\mathcal{D}}$,

$$\mathrm{Gen}_{\mathrm{s}} = \frac{1}{M} \sum_{u=1}^{M} \frac{1}{|\mathcal{D}_u|} \sum_{v=1}^{|\mathcal{D}_u|} \mathbb{E}\left[\phi(\theta_{\mathcal{D}}; z'_{u,v}) - \phi(\theta_{\mathcal{D}}; z_{u,v})\right]. \tag{15}$$

Since $z_{u,v}$ and $z'_{u,v}$ are identically distributed,

$$\mathbb{E}\left[\phi(\theta_{\mathcal{D}}; z_{u,v})\right] = \mathbb{E}\left[\phi(\theta_{\mathcal{D}^{(u,v)}}; z'_{u,v})\right]. \tag{16}$$

Therefore,

$$\mathrm{Gen_s} = \frac{1}{M}\sum_{u=1}^{M}\frac{1}{|\mathcal{D}_u|}\sum_{v=1}^{|\mathcal{D}_u|}\mathbb{E}\left[\phi(\theta_{\mathcal{D}}; z'_{u,v}) - \phi(\theta_{\mathcal{D}^{(u,v)}}; z'_{u,v})\right] \le \Gamma_{\mathrm{s}}, \tag{17}$$

where the inequality follows from Γ_{s}-sample-replacement distributional stability.

Finally, since $\mathrm{Gen_{total}} = \mathrm{Gen_c} + \mathrm{Gen_s}$, we obtain

$$\mathrm{Gen_{total}} \le \Gamma_{\mathrm{c}} + \Gamma_{\mathrm{s}}. \tag{18}$$

□

D Proof of Theorem 1

We first establish the following lemma.

Lemma 2. *Under Assumptions 1 and 4, the per-sample gradient satisfies*

$$\|\nabla\phi(\theta; z)\|_2 \le \sqrt{2\beta B}, \quad \forall\theta, z. \tag{19}$$

Proof. By Assumption 1, the per-sample loss $\phi(\cdot; z)$ is β-smooth. Therefore, for any θ, θ' and z, the descent lemma gives

$$\phi(\theta'; z) \le \phi(\theta; z) + \langle\nabla\phi(\theta; z), \theta' - \theta\rangle + \frac{\beta}{2}\|\theta' - \theta\|_2^2. \tag{20}$$

Taking $\theta' = \theta - \frac{1}{\beta}\nabla\phi(\theta; z)$, we obtain

$$\phi\left(\theta - \frac{1}{\beta}\nabla\phi(\theta; z); z\right) \le \phi(\theta; z) - \frac{1}{2\beta}\|\nabla\phi(\theta; z)\|_2^2. \tag{21}$$

By Assumption 4, $\phi(\theta - \frac{1}{\beta}\nabla\phi(\theta; z); z) \ge 0$, and hence

$$\|\nabla\phi(\theta; z)\|_2^2 \le 2\beta\phi(\theta; z) \le 2\beta B. \tag{22}$$

Therefore,

$$\|\nabla\phi(\theta; z)\|_2 \le \sqrt{2\beta B}. \tag{23}$$

This completes the proof. □

Client-replacement stability. Let $\theta^{T\tau}$ and $\tilde{\theta}^{T\tau}$ denote the final global models produced by client-level DP-FL on $\mathcal{D}$ and $\mathcal{D}^{(u)}$, respectively, and denote their output distributions by $P_{\theta^{T\tau}}$ and $P_{\tilde{\theta}^{T\tau}}$. According to Assumption 4, for any test sample z, we have

$$\left|\mathbb{E}_{\theta\sim P_{\theta^{T\tau}}}[\phi(\theta; z)] - \mathbb{E}_{\tilde{\theta}\sim P_{\tilde{\theta}^{T\tau}}}[\phi(\tilde{\theta}; z)]\right| \le B\,\mathrm{TV}\left(P_{\theta^{T\tau}}, P_{\tilde{\theta}^{T\tau}}\right). \tag{24}$$

Define $\gamma_t^{\mathrm{c}} = \mathrm{TV}(P_{\theta^{t\tau}}, P_{\tilde{\theta}^{t\tau}})$. Since the two executions start from the same initialization, $\gamma_0^{\mathrm{c}} = 0$. Let K_t and $\tilde{K}_t$ denote the Markov transition kernels of round t under $\mathcal{D}$ and $\mathcal{D}^{(u)}$, respectively. By the triangle inequality and the contraction of total variation distance under a common Markov transition,

$$\gamma_{t+1}^{\mathrm{c}} \le \gamma_t^{\mathrm{c}} + \mathrm{TV}\left(P_{\tilde{\theta}^{t\tau}}K_t, P_{\tilde{\theta}^{t\tau}}\tilde{K}_t\right). \tag{25}$$

Hence, it suffices to bound the one-round transition distance when the two executions start from the same global model. Let $\mathcal{E}_t^{\mathrm{c}}$ denote the event that client u is selected in round t. Since $|\mathcal{M}_t| = m$, we have

$$\Pr(\mathcal{E}_t^{\mathrm{c}}) = \frac{m}{M}. \tag{26}$$

If $\mathcal{E}_t^{\mathrm{c}}$ does not occur, the two executions have identical one-round transition distributions. Therefore,

$$\mathrm{TV}\left(P_{\tilde{\theta}^{t\tau}}K_t, P_{\tilde{\theta}^{t\tau}}\tilde{K}_t\right) \le \frac{m}{M}\,\mathrm{TV}\left(P_{\theta^{(t+1)\tau}|\mathcal{E}_t^{\mathrm{c}}}, P_{\tilde{\theta}^{(t+1)\tau}|\mathcal{E}_t^{\mathrm{c}}}\right). \tag{27}$$

We next bound the conditional transition distance given $\mathcal{E}_t^{\mathrm{c}}$. Both local trajectories of client u are initialized from the same global model, i.e., $\theta_u^{t\tau} = \tilde{\theta}_u^{t\tau} = \theta^{t\tau}$. For $r = 0, \dots, \tau$, define

$$d_{t\tau+r}^{\mathrm{c}} = \mathbb{E}\left\|\theta_u^{t\tau+r} - \tilde{\theta}_u^{t\tau+r}\right\|_2. \tag{28}$$

We first bound the difference between the local empirical gradients of the replaced clients. Since $\mathcal{D}$ and $\mathcal{D}^{(u)}$ differ only in client u, we have

$$\nabla\hat{\Phi}_{\mathcal{D}}(\theta) - \nabla\hat{\Phi}_{\mathcal{D}^{(u)}}(\theta) = \frac{1}{M}\left[\nabla\hat{\Phi}_{u,\mathcal{D}}(\theta) - \nabla\hat{\Phi}_{u,\mathcal{D}^{(u)}}(\theta)\right]. \tag{29}$$

Therefore, by Assumption 3,

$$\begin{aligned}\left\|\nabla\hat{\Phi}_{u,\mathcal{D}}(\theta) - \nabla\hat{\Phi}_{u,\mathcal{D}^{(u)}}(\theta)\right\|_2 &\le \left\|\nabla\hat{\Phi}_{u,\mathcal{D}}(\theta) - \nabla\hat{\Phi}_{\mathcal{D}}(\theta)\right\|_2 + \left\|\nabla\hat{\Phi}_{\mathcal{D}}(\theta) - \nabla\hat{\Phi}_{\mathcal{D}^{(u)}}(\theta)\right\|_2 \\ &\quad + \left\|\nabla\hat{\Phi}_{\mathcal{D}^{(u)}}(\theta) - \nabla\hat{\Phi}_{u,\mathcal{D}^{(u)}}(\theta)\right\|_2 \\ &\le 2G + \frac{1}{M}\left\|\nabla\hat{\Phi}_{u,\mathcal{D}}(\theta) - \nabla\hat{\Phi}_{u,\mathcal{D}^{(u)}}(\theta)\right\|_2. \end{aligned} \tag{30}$$

Since $M > 1$, it follows that

$$\left\|\nabla\hat{\Phi}_{u,\mathcal{D}}(\theta) - \nabla\hat{\Phi}_{u,\mathcal{D}^{(u)}}(\theta)\right\|_2 \le \frac{2MG}{M-1}. \tag{31}$$

For a mini-batch $\mathcal{B}$, denote its averaged stochastic gradient by $\bar{g}_u(\theta;\mathcal{B}) = \frac{1}{b}\sum_{z\in\mathcal{B}} g_u(\theta, z)$. Conditioning on the history before local step r, Assumption 2 and Jensen's inequality give

$$\mathbb{E}\left\|\bar{g}_u(\theta;\mathcal{B}) - \nabla\hat{\Phi}_u(\theta)\right\|_2 \le \varsigma. \tag{32}$$

Moreover, Assumption 1 implies that each local empirical risk is β-smooth. Hence,

$$\mathbb{E}\left\|\bar{g}_{u,\mathcal{D}}(\theta_u^{t\tau+r};\mathcal{B}) - \bar{g}_{u,\mathcal{D}^{(u)}}(\tilde{\theta}_u^{t\tau+r};\widetilde{\mathcal{B}})\right\|_2 \le 2\varsigma + \beta d_{t\tau+r}^{\mathrm{c}} + \frac{2MG}{M-1}. \tag{33}$$

Therefore,

$$d_{t\tau+r+1}^{\mathrm{c}} \le (1+\eta_t\beta)d_{t\tau+r}^{\mathrm{c}} + 2\eta_t\left(\varsigma + \frac{MG}{M-1}\right). \tag{34}$$

Using $d_{t\tau}^{\mathrm{c}} = 0$ and iterating over $r = 0, \dots, \tau-1$, we obtain

$$d_{t\tau+\tau}^{\mathrm{c}} \le \frac{2}{\beta}\left(\varsigma + \frac{MG}{M-1}\right)\left[(1+\eta_t\beta)^\tau - 1\right]. \tag{35}$$

Since $\Delta_u^t = \theta_u^{t\tau+\tau} - \theta^{t\tau}$ and $\widetilde{\Delta}_u^t = \tilde{\theta}_u^{t\tau+\tau} - \theta^{t\tau}$, radial projection onto the radius-C Euclidean ball is non-expansive and both clipped updates have norm at most C, yielding

$$\mathbb{E}\left\|\bar{\Delta}_u^t - \widetilde{\bar{\Delta}}_u^t\right\|_2 \le \min\left\{2C, \frac{2}{\beta}\left(\varsigma + \frac{MG}{M-1}\right)\left[(1+\eta_t\beta)^\tau - 1\right]\right\}. \tag{36}$$

Conditioning on $\mathcal{E}_t^{\mathrm{c}}$ and the local randomness before the Gaussian perturbation, the two global models follow Gaussian distributions with the same covariance $\frac{\sigma^2C^2}{m^2}I_d$, and their mean difference satisfies $\|\mu_t - \mu_t'\|_2 = \frac{1}{m}\|\bar{\Delta}_u^t - \widetilde{\bar{\Delta}}_u^t\|_2$. Let $\Psi(\cdot)$ denote the cumulative distribution function of the standard normal distribution. Therefore, using the TV distance formula for two Gaussian distributions with the same covariance, the concavity of $2\Psi(x) - 1$ for $x \ge 0$, and Jensen's inequality,

$$\mathrm{TV}\left(P_{\theta^{(t+1)\tau}|\mathcal{E}_t^{\mathrm{c}}}, P_{\tilde{\theta}^{(t+1)\tau}|\mathcal{E}_t^{\mathrm{c}}}\right) \le 2\Psi\left(\frac{1}{2\sigma C}\min\left\{2C, \frac{2}{\beta}\left(\varsigma + \frac{MG}{M-1}\right)\left[(1+\eta_t\beta)^\tau - 1\right]\right\}\right) - 1. \tag{37}$$

Consequently,

$$\gamma_{t+1}^{\mathrm{c}} \le \gamma_t^{\mathrm{c}} + \frac{m}{M}\left[2\Psi\left(\frac{1}{2\sigma C}\min\left\{2C, \frac{2}{\beta}\left(\varsigma + \frac{MG}{M-1}\right)\left[(1+\eta_t\beta)^\tau - 1\right]\right\}\right) - 1\right]. \tag{38}$$

Since $\gamma_0^{\mathrm{c}} = 0$, unrolling over $t = 0, \ldots, T-1$ gives

$$\gamma_T^{\mathrm{c}} \le \frac{m}{M}\sum_{t=0}^{T-1}\left[2\Psi\left(\frac{1}{2\sigma C}\min\left\{2C, \frac{2}{\beta}\left(\varsigma + \frac{MG}{M-1}\right)\left[(1+\eta_t\beta)^\tau - 1\right]\right\}\right) - 1\right]. \tag{39}$$

Using $2\Psi(x) - 1 \le \sqrt{\frac{2}{\pi}}x$ and $(1+\eta_t\beta)^\tau - 1 \le \eta_t\beta\tau e^{\eta_t\beta\tau}$, we obtain

$$\begin{aligned}\gamma_T^{\mathrm{c}} &\le \frac{m}{M\sigma}\sqrt{\frac{2}{\pi}}\sum_{t=0}^{T-1}\min\left\{1, \frac{\tau\eta_t e^{\eta_t\beta\tau}}{C}\left(\varsigma + \frac{MG}{M-1}\right)\right\} \\ &\le \frac{m}{M\sigma}\sqrt{\frac{2}{\pi}}\min\left\{T, \frac{\tau}{C}\left(\varsigma + \frac{MG}{M-1}\right)\sum_{t=0}^{T-1}\eta_t e^{\eta_t\beta\tau}\right\}.\end{aligned} \tag{40}$$

Therefore,

$$\Gamma_{\mathrm{c}} \le B\min\left\{1, \frac{m}{M\sigma}\sqrt{\frac{2}{\pi}}\min\left\{T, \frac{\tau}{C}\left(\varsigma + \frac{MG}{M-1}\right)\sum_{t=0}^{T-1}\eta_t e^{\eta_t\beta\tau}\right\}\right\}. \tag{41}$$

Sample-replacement stability. Let $\theta^{T\tau}$ and $\tilde{\theta}^{T\tau}$ denote the final global models produced by client-level DP-FL on $\mathcal{D}$ and $\mathcal{D}^{(u,v)}$, respectively, and define $\gamma_t^{\mathrm{s}} = \mathrm{TV}(P_{\theta^{t\tau}}, P_{\tilde{\theta}^{t\tau}})$. As before,

$$\gamma_{t+1}^{\mathrm{s}} \le \gamma_t^{\mathrm{s}} + \mathrm{TV}\left(P_{\tilde{\theta}^{t\tau}}K_t, P_{\tilde{\theta}^{t\tau}}\tilde{K}_t\right). \tag{42}$$

Let $\mathcal{E}_t^{\mathrm{s}}$ denote the event that client u is selected and the replacement sample is included in at least one of the τ local mini-batches. Then

$$\Pr(\mathcal{E}_t^{\mathrm{s}}) \le \frac{m}{M}\min\left\{1, \frac{\tau b}{|\mathcal{D}_u|}\right\} \le \frac{m\tau b}{M|\mathcal{D}_u|}. \tag{43}$$

If $\mathcal{E}_t^{\mathrm{s}}$ does not occur, the two executions have identical one-round transition distributions. Hence,

$$\mathrm{TV}\left(P_{\tilde{\theta}^{t\tau}}K_t, P_{\tilde{\theta}^{t\tau}}\tilde{K}_t\right) \le \frac{m\tau b}{M|\mathcal{D}_u|}\,\mathrm{TV}\left(P_{\theta^{(t+1)\tau}|\mathcal{E}_t^{\mathrm{s}}}, P_{\tilde{\theta}^{(t+1)\tau}|\mathcal{E}_t^{\mathrm{s}}}\right). \tag{44}$$

Conditioning on $\mathcal{E}_t^{\mathrm{s}}$, define $d_{t\tau+r}^{\mathrm{s}} = \mathbb{E}\|\theta_u^{t\tau+r} - \tilde{\theta}_u^{t\tau+r}\|_2$ with $d_{t\tau}^{\mathrm{s}} = 0$. Let $I_{t,r}$ denote the indicator that the replacement sample is selected at local step r. If $I_{t,r} = 0$, the two coupled mini-batches contain exactly the same samples. By Assumption 1,

$$d_{t\tau+r+1}^{\mathrm{s}} \le (1+\eta_t\beta)d_{t\tau+r}^{\mathrm{s}}. \tag{45}$$

If $I_{t,r} = 1$, the two mini-batches can be written as $\mathcal{B} = \mathcal{S} \cup \{z_{u,v}\}$ and $\widetilde{\mathcal{B}} = \mathcal{S} \cup \{z'_{u,v}\}$, where $|\mathcal{S}| = b - 1$. By Assumption 1 and Lemma 2,

$$\begin{aligned}&\left\|g_u(\theta_u^{t\tau+r}, z_{u,v}) - g_u(\tilde{\theta}_u^{t\tau+r}, z'_{u,v})\right\|_2 \\ &\le \left\|g_u(\theta_u^{t\tau+r}, z_{u,v}) - g_u(\tilde{\theta}_u^{t\tau+r}, z_{u,v})\right\|_2 + \left\|g_u(\tilde{\theta}_u^{t\tau+r}, z_{u,v}) - g_u(\tilde{\theta}_u^{t\tau+r}, z'_{u,v})\right\|_2 \\ &\le \beta\left\|\theta_u^{t\tau+r} - \tilde{\theta}_u^{t\tau+r}\right\|_2 + 2\sqrt{2\beta B}.\end{aligned} \tag{46}$$

Therefore, in both cases,

$$d_{t\tau+r+1}^{\mathrm{s}} \le (1+\eta_t\beta)d_{t\tau+r}^{\mathrm{s}} + \frac{2\eta_t\sqrt{2\beta B}}{b}. \tag{47}$$

Since $d^{\mathrm{s}}_{t\tau} = 0$, iterating over $r = 0, \ldots, \tau - 1$ yields

$$d^{\mathrm{s}}_{t\tau+\tau} \le \frac{2\sqrt{2\beta B}}{\beta b} \left[(1+\eta_t\beta)^\tau - 1\right]. \tag{48}$$

Consequently, after clipping,

$$\mathbb{E}\left\| \bar{\Delta}^t_u - \tilde{\bar{\Delta}}^t_u \right\|_2 \le \min\left\{2C, \frac{2\sqrt{2\beta B}}{\beta b}\left[(1+\eta_t\beta)^\tau - 1\right]\right\}. \tag{49}$$

As in the client-replacement case, conditioning on the local randomness before Gaussian perturbation and using the Gaussian TV formula give

$$\mathrm{TV}\left(P_{\theta^{(t+1)\tau}|\mathcal{E}^{\mathrm{s}}_t}, P_{\tilde{\theta}^{(t+1)\tau}|\mathcal{E}^{\mathrm{s}}_t}\right) \le 2\Psi\left(\frac{1}{2\sigma C}\min\left\{2C, \frac{2\sqrt{2\beta B}}{\beta b}\left[(1+\eta_t\beta)^\tau - 1\right]\right\}\right) - 1. \tag{50}$$

Therefore,

$$\gamma^{\mathrm{s}}_{t+1} \le \gamma^{\mathrm{s}}_t + \frac{m\tau b}{M|\mathcal{D}_u|}\left[2\Psi\left(\frac{1}{2\sigma C}\min\left\{2C, \frac{2\sqrt{2\beta B}}{\beta b}\left[(1+\eta_t\beta)^\tau - 1\right]\right\}\right) - 1\right]. \tag{51}$$

Since $\gamma^{\mathrm{s}}_0 = 0$, unrolling the recursion over $t = 0, \ldots, T-1$ gives

$$\gamma^{\mathrm{s}}_T \le \frac{m\tau b}{M|\mathcal{D}_u|}\sum_{t=0}^{T-1}\left[2\Psi\left(\frac{1}{2\sigma C}\min\left\{2C, \frac{2\sqrt{2\beta B}}{\beta b}\left[(1+\eta_t\beta)^\tau - 1\right]\right\}\right) - 1\right]. \tag{52}$$

Using $2\Psi(x) - 1 \le \sqrt{\frac{2}{\pi}}x$, $(1+\eta_t\beta)^\tau - 1 \le \eta_t\beta\tau e^{\eta_t\beta\tau}$, and $|\mathcal{D}_u| \ge |\mathcal{D}_{\min}|$, we obtain

$$\begin{aligned}
\gamma^{\mathrm{s}}_T &\le \frac{m\tau b}{M|\mathcal{D}_{\min}|\sigma}\sqrt{\frac{2}{\pi}}\sum_{t=0}^{T-1}\min\left\{1, \frac{\tau\eta_t e^{\eta_t\beta\tau}\sqrt{2\beta B}}{bC}\right\} \\
&\le \frac{m\tau b}{M|\mathcal{D}_{\min}|\sigma}\sqrt{\frac{2}{\pi}}\min\left\{T, \frac{\tau\sqrt{2\beta B}}{bC}\sum_{t=0}^{T-1}\eta_t e^{\eta_t\beta\tau}\right\}.
\end{aligned} \tag{53}$$

Therefore,

$$\Gamma_{\mathrm{s}} \le B\min\left\{1, \frac{m\tau b}{M|\mathcal{D}_{\min}|\sigma}\sqrt{\frac{2}{\pi}}\min\left\{T, \frac{\tau\sqrt{2\beta B}}{bC}\sum_{t=0}^{T-1}\eta_t e^{\eta_t\beta\tau}\right\}\right\}. \tag{54}$$

Finally, by Lemma 1,

$$\begin{aligned}
\mathrm{Gen}_{\mathrm{total}} &\le \Gamma_{\mathrm{c}} + \Gamma_{\mathrm{s}} \\
&\le B\min\left\{1, \frac{m}{M\sigma}\sqrt{\frac{2}{\pi}}\min\left\{T, \frac{\tau}{C}\left(\varsigma + \frac{MG}{M-1}\right)\sum_{t=0}^{T-1}\eta_t e^{\eta_t\beta\tau}\right\}\right\} \\
&\quad + B\min\left\{1, \frac{m\tau b}{M|\mathcal{D}_{\min}|\sigma}\sqrt{\frac{2}{\pi}}\min\left\{T, \frac{\tau\sqrt{2\beta B}}{bC}\sum_{t=0}^{T-1}\eta_t e^{\eta_t\beta\tau}\right\}\right\}.
\end{aligned} \tag{55}$$

This completes the proof.

E Proof of Theorem 2

Client-replacement stability. Let $\theta^{T\tau}$ and $\tilde{\theta}^{T\tau}$ denote the final global models produced by client-level DP-FL on $\mathcal{D}$ and $\mathcal{D}^{(u)}$, respectively, and denote their output distributions by $P_{\theta^{T\tau}}$ and $P_{\tilde{\theta}^{T\tau}}$. According to Assumption 4 and Pinsker's inequality, for any test sample z, we have

$$\left|\mathbb{E}_{\theta\sim P_{\theta^{T\tau}}}[\phi(\theta; z)] - \mathbb{E}_{\tilde{\theta}\sim P_{\tilde{\theta}^{T\tau}}}[\phi(\tilde{\theta}; z)]\right| \le B\,\mathrm{TV}\left(P_{\theta^{T\tau}}, P_{\tilde{\theta}^{T\tau}}\right) \le B\sqrt{\frac{1}{2}\,\mathrm{KL}\left(P_{\theta^{T\tau}} \| P_{\tilde{\theta}^{T\tau}}\right)}. \tag{56}$$

Let K_t and $\tilde{K}_t$ denote the Markov transition kernels of round t under $\mathcal{D}$ and $\mathcal{D}^{(u)}$, respectively. By the chain rule for KL divergence and data processing, for each round t,

$$\mathrm{KL}\left(P_{\theta^{(t+1)\tau}} \| P_{\tilde{\theta}^{(t+1)\tau}}\right) \le \mathrm{KL}\left(P_{\theta^{t\tau}} \| P_{\tilde{\theta}^{t\tau}}\right) + \mathbb{E}_{\theta \sim P_{\theta^{t\tau}}}\left[\mathrm{KL}\left(K_t(\theta,\cdot) \| \tilde{K}_t(\theta,\cdot)\right)\right]. \tag{57}$$

Hence, it suffices to bound the one-round transition divergence when the two executions start from the same global model. Let $\mathcal{E}_t^{\mathrm{c}}$ denote the event that client u is selected in round t. Since $|\mathcal{M}_t| = m$, we have

$$\Pr(\mathcal{E}_t^{\mathrm{c}}) = \frac{m}{M}. \tag{58}$$

If $\mathcal{E}_t^{\mathrm{c}}$ does not occur, the two executions have identical one-round transition distributions. Therefore, by the joint convexity of KL divergence,

$$\mathrm{KL}\left(K_t(\theta,\cdot) \| \tilde{K}_t(\theta,\cdot)\right) \le \frac{m}{M}\,\mathrm{KL}\left(P_{\theta^{(t+1)\tau}|\mathcal{E}_t^{\mathrm{c}}} \| P_{\tilde{\theta}^{(t+1)\tau}|\mathcal{E}_t^{\mathrm{c}}}\right). \tag{59}$$

We next bound the conditional transition divergence given $\mathcal{E}_t^{\mathrm{c}}$. Both local trajectories of client u are initialized from the same global model, i.e., $\theta_u^{t\tau} = \tilde{\theta}_u^{t\tau} = \theta^{t\tau}$. As in the proof of Theorem 1, for any θ, Assumption 3 gives

$$\left\|\nabla\hat{\Phi}_{u,\mathcal{D}}(\theta) - \nabla\hat{\Phi}_{u,\mathcal{D}^{(u)}}(\theta)\right\|_2 \le \frac{2MG}{M-1}. \tag{60}$$

For a mini-batch $\mathcal{B}$, denote its averaged stochastic gradient by $\bar{g}_u(\theta;\mathcal{B}) = \frac{1}{b}\sum_{z\in\mathcal{B}} g_u(\theta,z)$. Conditioning on the history before local step r, Assumption 2 and Jensen's inequality give

$$\left(\mathbb{E}\left\|\bar{g}_u(\theta;\mathcal{B}) - \nabla\hat{\Phi}_u(\theta)\right\|_2^2\right)^{\frac{1}{2}} \le \varsigma. \tag{61}$$

Moreover, Assumption 1 implies that each local empirical risk is β-smooth. By Minkowski's inequality,

$$\left(\mathbb{E}\left\|\bar{g}_{u,\mathcal{D}}(\theta_u^{t\tau+r};\mathcal{B}) - \bar{g}_{u,\mathcal{D}^{(u)}}(\tilde{\theta}_u^{t\tau+r};\widetilde{\mathcal{B}})\right\|_2^2\right)^{\frac{1}{2}} \le 2\varsigma + \beta\left(\mathbb{E}\left\|\theta_u^{t\tau+r} - \tilde{\theta}_u^{t\tau+r}\right\|_2^2\right)^{\frac{1}{2}} + \frac{2MG}{M-1}. \tag{62}$$

Therefore,

$$\left(\mathbb{E}\left\|\theta_u^{t\tau+r+1} - \tilde{\theta}_u^{t\tau+r+1}\right\|_2^2\right)^{\frac{1}{2}} \le (1+\eta_t\beta)\left(\mathbb{E}\left\|\theta_u^{t\tau+r} - \tilde{\theta}_u^{t\tau+r}\right\|_2^2\right)^{\frac{1}{2}} + 2\eta_t\left(\varsigma + \frac{MG}{M-1}\right). \tag{63}$$

Using $\theta_u^{t\tau} = \tilde{\theta}_u^{t\tau}$ and iterating over $r = 0, \ldots, \tau-1$, we obtain

$$\left(\mathbb{E}\left\|\theta_u^{t\tau+\tau} - \tilde{\theta}_u^{t\tau+\tau}\right\|_2^2\right)^{\frac{1}{2}} \le \frac{2}{\beta}\left(\varsigma + \frac{MG}{M-1}\right)\left[(1+\eta_t\beta)^\tau - 1\right]. \tag{64}$$

Since $\Delta_u^t = \theta_u^{t\tau+\tau} - \theta^{t\tau}$ and $\widetilde{\Delta}_u^t = \tilde{\theta}_u^{t\tau+\tau} - \theta^{t\tau}$, radial projection onto the radius-C Euclidean ball is non-expansive and both clipped updates have norm at most C, yielding

$$\mathbb{E}\left\|\bar{\Delta}_u^t - \widetilde{\bar{\Delta}}_u^t\right\|_2^2 \le 4\min\left\{C^2, \frac{1}{\beta^2}\left(\varsigma + \frac{MG}{M-1}\right)^2\left[(1+\eta_t\beta)^\tau - 1\right]^2\right\}. \tag{65}$$

Conditioning on $\mathcal{E}_t^{\mathrm{c}}$ and the local randomness before the Gaussian perturbation, the two global models follow Gaussian distributions with the same covariance $\frac{\sigma^2C^2}{m^2}I_d$, and their mean difference satisfies $\|\mu_t - \mu_t'\|_2 = \frac{1}{m}\|\bar{\Delta}_u^t - \widetilde{\bar{\Delta}}_u^t\|_2$. Therefore, using the KL divergence formula for two Gaussian distributions with the same covariance and the joint convexity of KL divergence,

$$\begin{aligned}\mathrm{KL}\left(P_{\theta^{(t+1)\tau}|\mathcal{E}_t^{\mathrm{c}}} \| P_{\tilde{\theta}^{(t+1)\tau}|\mathcal{E}_t^{\mathrm{c}}}\right) &\le \frac{1}{2\sigma^2C^2}\mathbb{E}\left\|\bar{\Delta}_u^t - \widetilde{\bar{\Delta}}_u^t\right\|_2^2 \\ &\le \frac{2}{\sigma^2C^2}\min\left\{C^2, \frac{1}{\beta^2}\left(\varsigma + \frac{MG}{M-1}\right)^2\left[(1+\eta_t\beta)^\tau - 1\right]^2\right\}. \end{aligned} \tag{66}$$

Consequently,

$$\mathrm{KL}\left(K_t(\theta,\cdot)\|\tilde{K}_t(\theta,\cdot)\right) \le \frac{2m}{M\sigma^2C^2}\min\left\{C^2, \frac{1}{\beta^2}\left(\varsigma+\frac{MG}{M-1}\right)^2\left[(1+\eta_t\beta)^\tau-1\right]^2\right\}. \quad (67)$$

Since the two executions start from the same initialization, unrolling the KL recursion over $t = 0,\dots,T-1$ gives

$$\mathrm{KL}\left(P_{\theta^{T\tau}}\|P_{\tilde{\theta}^{T\tau}}\right) \le \frac{2m}{M\sigma^2C^2}\sum_{t=0}^{T-1}\min\left\{C^2, \frac{1}{\beta^2}\left(\varsigma+\frac{MG}{M-1}\right)^2\left[(1+\eta_t\beta)^\tau-1\right]^2\right\}. \quad (68)$$

Using $(1+\eta_t\beta)^\tau - 1 \le \eta_t\beta\tau e^{\eta_t\beta\tau}$, we obtain

$$\begin{aligned}\mathrm{KL}\left(P_{\theta^{T\tau}}\|P_{\tilde{\theta}^{T\tau}}\right) &\le \frac{2m}{M\sigma^2}\sum_{t=0}^{T-1}\min\left\{1, \frac{\tau^2\eta_t^2e^{2\eta_t\beta\tau}}{C^2}\left(\varsigma+\frac{MG}{M-1}\right)^2\right\}\\ &\le \frac{2m}{M\sigma^2}\min\left\{T, \frac{\tau^2}{C^2}\left(\varsigma+\frac{MG}{M-1}\right)^2\sum_{t=0}^{T-1}\eta_t^2e^{2\eta_t\beta\tau}\right\}. \quad (69)\end{aligned}$$

Therefore, by Pinsker's inequality,

$$\Gamma_{\mathrm{c}} \le B\min\left\{1, \frac{1}{\sigma}\sqrt{\frac{m}{M}}\min\left\{\sqrt{T}, \frac{\tau}{C}\left(\varsigma+\frac{MG}{M-1}\right)\left(\sum_{t=0}^{T-1}\eta_t^2e^{2\eta_t\beta\tau}\right)^{\frac{1}{2}}\right\}\right\}. \quad (70)$$

Sample-replacement stability. Let $\theta^{T\tau}$ and $\tilde{\theta}^{T\tau}$ denote the final global models produced by client-level DP-FL on $\mathcal{D}$ and $\mathcal{D}^{(u,v)}$, respectively. According to Assumption 4 and Pinsker's inequality, for any test sample z, we have

$$\left|\mathbb{E}_{\theta\sim P_{\theta^{T\tau}}}[\phi(\theta;z)] - \mathbb{E}_{\tilde{\theta}\sim P_{\tilde{\theta}^{T\tau}}}[\phi(\tilde{\theta};z)]\right| \le B\sqrt{\frac{1}{2}\mathrm{KL}\left(P_{\theta^{T\tau}}\|P_{\tilde{\theta}^{T\tau}}\right)}. \quad (71)$$

Let K_t and $\tilde{K}_t$ denote the Markov transition kernels of round t under $\mathcal{D}$ and $\mathcal{D}^{(u,v)}$, respectively. As in the client-replacement case, by the chain rule for KL divergence and data processing,

$$\mathrm{KL}\left(P_{\theta^{(t+1)\tau}}\|P_{\tilde{\theta}^{(t+1)\tau}}\right) \le \mathrm{KL}\left(P_{\theta^{t\tau}}\|P_{\tilde{\theta}^{t\tau}}\right) + \mathbb{E}_{\theta\sim P_{\theta^{t\tau}}}\left[\mathrm{KL}\left(K_t(\theta,\cdot)\|\tilde{K}_t(\theta,\cdot)\right)\right]. \quad (72)$$

Let $\mathcal{E}_t^{\mathrm{s}}$ denote the event that client u is selected and the replacement sample is included in at least one of the τ local mini-batches. Then

$$\Pr(\mathcal{E}_t^{\mathrm{s}}) \le \frac{m}{M}\min\left\{1, \frac{\tau b}{|\mathcal{D}_u|}\right\} \le \frac{m\tau b}{M|\mathcal{D}_u|}. \quad (73)$$

If $\mathcal{E}_t^{\mathrm{s}}$ does not occur, the two executions have identical one-round transition distributions. Hence, by the joint convexity of KL divergence,

$$\mathrm{KL}\left(K_t(\theta,\cdot)\|\tilde{K}_t(\theta,\cdot)\right) \le \frac{m\tau b}{M|\mathcal{D}_u|}\mathrm{KL}\left(P_{\theta^{(t+1)\tau}|\mathcal{E}_t^{\mathrm{s}}}\|P_{\tilde{\theta}^{(t+1)\tau}|\mathcal{E}_t^{\mathrm{s}}}\right). \quad (74)$$

Conditioning on $\mathcal{E}_t^{\mathrm{s}}$, let $I_{t,r}$ denote the indicator that the replacement sample is selected at local step r. If $I_{t,r} = 0$, the two coupled mini-batches contain exactly the same samples. By Assumption 1,

$$\left\|\bar{g}_{u,\mathcal{D}}(\theta_u^{t\tau+r};\mathcal{B}) - \bar{g}_{u,\mathcal{D}^{(u,v)}}(\tilde{\theta}_u^{t\tau+r};\widetilde{\mathcal{B}})\right\|_2 \le \beta\left\|\theta_u^{t\tau+r} - \tilde{\theta}_u^{t\tau+r}\right\|_2. \quad (75)$$

If $I_{t,r} = 1$, the two mini-batches can be written as $\mathcal{B} = \mathcal{S}\cup\{z_{u,v}\}$ and $\widetilde{\mathcal{B}} = \mathcal{S}\cup\{z'_{u,v}\}$, where $|\mathcal{S}| = b-1$. By Assumption 1 and Lemma 2,

$$\left\|\bar{g}_{u,\mathcal{D}}(\theta_u^{t\tau+r};\mathcal{B}) - \bar{g}_{u,\mathcal{D}^{(u,v)}}(\tilde{\theta}_u^{t\tau+r};\widetilde{\mathcal{B}})\right\|_2 \le \beta\left\|\theta_u^{t\tau+r} - \tilde{\theta}_u^{t\tau+r}\right\|_2 + \frac{2\sqrt{2\beta B}}{b}. \quad (76)$$

Therefore, in both cases, by Minkowski's inequality,

$$\left(\mathbb{E}\left\|\theta_u^{t\tau+r+1}-\tilde{\theta}_u^{t\tau+r+1}\right\|_2^2\right)^{\frac{1}{2}} \le (1+\eta_t\beta)\left(\mathbb{E}\left\|\theta_u^{t\tau+r}-\tilde{\theta}_u^{t\tau+r}\right\|_2^2\right)^{\frac{1}{2}} + \frac{2\eta_t\sqrt{2\beta B}}{b}. \tag{77}$$

Since $\theta_u^{t\tau}=\tilde{\theta}_u^{t\tau}$, iterating over $r=0,\dots,\tau-1$ yields

$$\left(\mathbb{E}\left\|\theta_u^{t\tau+\tau}-\tilde{\theta}_u^{t\tau+\tau}\right\|_2^2\right)^{\frac{1}{2}} \le \frac{2\sqrt{2\beta B}}{\beta b}\left[(1+\eta_t\beta)^\tau-1\right]. \tag{78}$$

Consequently, after clipping,

$$\mathbb{E}\left\|\bar{\Delta}_u^t-\tilde{\bar{\Delta}}_u^t\right\|_2^2 \le 4\min\left\{C^2, \frac{2\beta B}{\beta^2 b^2}\left[(1+\eta_t\beta)^\tau-1\right]^2\right\}. \tag{79}$$

As in the client-replacement case, conditioning on the local randomness before the Gaussian perturbation, using the Gaussian KL formula and the joint convexity of KL divergence gives

$$\begin{aligned}\mathrm{KL}\left(P_{\theta^{(t+1)\tau}|\mathcal{E}_t^s}\|P_{\tilde{\theta}^{(t+1)\tau}|\mathcal{E}_t^s}\right) &\le \frac{1}{2\sigma^2C^2}\mathbb{E}\left\|\bar{\Delta}_u^t-\tilde{\bar{\Delta}}_u^t\right\|_2^2 \\ &\le \frac{2}{\sigma^2C^2}\min\left\{C^2, \frac{2\beta B}{\beta^2b^2}\left[(1+\eta_t\beta)^\tau-1\right]^2\right\}.\end{aligned} \tag{80}$$

Therefore,

$$\mathrm{KL}\left(K_t(\theta,\cdot)\|\tilde{K}_t(\theta,\cdot)\right) \le \frac{2m\tau b}{M|\mathcal{D}_u|\sigma^2C^2}\min\left\{C^2, \frac{2\beta B}{\beta^2b^2}\left[(1+\eta_t\beta)^\tau-1\right]^2\right\}. \tag{81}$$

Since the two executions start from the same initialization, unrolling the KL recursion over $t=0,\dots,T-1$ and using $|\mathcal{D}_u|\ge|\mathcal{D}_{\min}|$ give

$$\mathrm{KL}\left(P_{\theta^{T\tau}}\|P_{\tilde{\theta}^{T\tau}}\right) \le \frac{2m\tau b}{M|\mathcal{D}_{\min}|\sigma^2C^2}\sum_{t=0}^{T-1}\min\left\{C^2, \frac{2\beta B}{\beta^2b^2}\left[(1+\eta_t\beta)^\tau-1\right]^2\right\}. \tag{82}$$

Using $(1+\eta_t\beta)^\tau-1\le\eta_t\beta\tau e^{\eta_t\beta\tau}$, we obtain

$$\begin{aligned}\mathrm{KL}\left(P_{\theta^{T\tau}}\|P_{\tilde{\theta}^{T\tau}}\right) &\le \frac{2m\tau b}{M|\mathcal{D}_{\min}|\sigma^2}\sum_{t=0}^{T-1}\min\left\{1, \frac{2\beta B\tau^2\eta_t^2e^{2\eta_t\beta\tau}}{b^2C^2}\right\} \\ &\le \frac{2m\tau b}{M|\mathcal{D}_{\min}|\sigma^2}\min\left\{T, \frac{2\beta B\tau^2}{b^2C^2}\sum_{t=0}^{T-1}\eta_t^2e^{2\eta_t\beta\tau}\right\}.\end{aligned} \tag{83}$$

Therefore, by Pinsker's inequality,

$$\Gamma_{\mathrm{s}} \le B\min\left\{1, \frac{1}{\sigma}\sqrt{\frac{m\tau b}{M|\mathcal{D}_{\min}|}}\min\left\{\sqrt{T}, \frac{\tau\sqrt{2\beta B}}{bC}\left(\sum_{t=0}^{T-1}\eta_t^2e^{2\eta_t\beta\tau}\right)^{\frac{1}{2}}\right\}\right\}. \tag{84}$$

Finally, by Lemma 1,

$$\begin{aligned}\mathrm{Gen}_{\mathrm{total}} &\le \Gamma_{\mathrm{c}}+\Gamma_{\mathrm{s}} \\ &\le B\min\left\{1, \frac{1}{\sigma}\sqrt{\frac{m}{M}}\min\left\{\sqrt{T}, \frac{\tau}{C}\left(\varsigma+\frac{MG}{M-1}\right)\left(\sum_{t=0}^{T-1}\eta_t^2e^{2\eta_t\beta\tau}\right)^{\frac{1}{2}}\right\}\right\} \\ &\quad + B\min\left\{1, \frac{1}{\sigma}\sqrt{\frac{m\tau b}{M|\mathcal{D}_{\min}|}}\min\left\{\sqrt{T}, \frac{\tau\sqrt{2\beta B}}{bC}\left(\sum_{t=0}^{T-1}\eta_t^2e^{2\eta_t\beta\tau}\right)^{\frac{1}{2}}\right\}\right\}.\end{aligned} \tag{85}$$

This completes the proof.

F Proof of Corollary 2

According to privacy calibration,

$$\sigma \geq \kappa \frac{m}{M} \frac{\sqrt{T \log(\frac{1}{\delta})}}{\epsilon}, \tag{86}$$

and hence

$$\frac{m}{M\sigma} \leq \frac{\epsilon}{\kappa \sqrt{T \log(\frac{1}{\delta})}}. \tag{87}$$

For a constant learning rate $\eta_t = \eta$, we have

$$\sum_{t=0}^{T-1} \eta_t e^{\eta_t \beta \tau} = T\eta e^{\eta \beta \tau}, \qquad \left(\sum_{t=0}^{T-1} \eta_t^2 e^{2\eta_t \beta \tau} \right)^{\frac{1}{2}} = \sqrt{T} \eta e^{\eta \beta \tau}. \tag{88}$$

Using equation 87 and relaxing the outer truncations via $\min\{1, x\} \leq x$ for $x \geq 0$, Theorem 1 gives the TV-based bound

$$\begin{aligned} \mathrm{Gen}_{\mathrm{total}} \leq \frac{B\epsilon}{\kappa \sqrt{\log(\frac{1}{\delta})}} \sqrt{\frac{2T}{\pi}} \Bigg[& \min\left\{1, \frac{\eta \tau e^{\eta \beta \tau}}{C} \left(\varsigma + \frac{MG}{M-1} \right) \right\} \\ & + \frac{\tau b}{|\mathcal{D}_{\min}|} \min\left\{1, \frac{\eta \tau e^{\eta \beta \tau} \sqrt{2\beta B}}{bC} \right\} \Bigg]. \end{aligned} \tag{89}$$

Similarly, Theorem 2 gives the KL-based bound

$$\begin{aligned} \mathrm{Gen}_{\mathrm{total}} \leq \frac{B\epsilon}{\kappa \sqrt{\log(\frac{1}{\delta})}} \sqrt{\frac{M}{m}} \Bigg[& \min\left\{1, \frac{\eta \tau e^{\eta \beta \tau}}{C} \left(\varsigma + \frac{MG}{M-1} \right) \right\} \\ & + \sqrt{\frac{\tau b}{|\mathcal{D}_{\min}|}} \min\left\{1, \frac{\eta \tau e^{\eta \beta \tau} \sqrt{2\beta B}}{bC} \right\} \Bigg]. \end{aligned} \tag{90}$$

Since both bounds apply to the same total generalization error, we obtain

$$\begin{aligned} \mathrm{Gen}_{\mathrm{total}} \leq \frac{B\epsilon}{\kappa \sqrt{\log(\frac{1}{\delta})}} \min \Bigg\{ & \\ & \sqrt{\frac{2T}{\pi}} \left[\min\left\{1, \frac{\eta \tau e^{\eta \beta \tau}}{C} \left(\varsigma + \frac{MG}{M-1} \right) \right\} + \frac{\tau b}{|\mathcal{D}_{\min}|} \min\left\{1, \frac{\eta \tau e^{\eta \beta \tau} \sqrt{2\beta B}}{bC} \right\} \right], \\ & \sqrt{\frac{M}{m}} \left[\min\left\{1, \frac{\eta \tau e^{\eta \beta \tau}}{C} \left(\varsigma + \frac{MG}{M-1} \right) \right\} + \sqrt{\frac{\tau b}{|\mathcal{D}_{\min}|}} \min\left\{1, \frac{\eta \tau e^{\eta \beta \tau} \sqrt{2\beta B}}{bC} \right\} \right] \Bigg\}. \end{aligned} \tag{91}$$

This completes the proof.

G Optimization Analysis of Client-Level DP-FL

We provide the optimization analysis of client-level DP-FL under Assumptions 1–5.

Proposition 1 (Optimization error of client-level DP-FL)**.** *Under Assumptions 1–5, suppose that the learning rate satisfies $\eta\tau \leq \min\left\{\frac{1}{64\beta}, \frac{1}{\mu}\right\}$. Then, the output of client-level DP-FL satisfies*

$$\begin{aligned} & \mathbb{E}\left[\hat{\Phi}(\theta^{T\tau}) - \hat{\Phi}(\hat{\theta}^\star)\right] \leq \left(1 - \frac{\mu\eta\tau}{2}\right)^T \left[\hat{\Phi}(\theta^0) - \hat{\Phi}(\hat{\theta}^\star)\right] \\ & + \mathcal{O}\left(\frac{1}{\mu} \left[\max\left\{0, \sqrt{2\beta B} - \frac{C}{\eta\tau}\right\}^2 + \left(1 + \frac{1}{m}\right) G^2 + \frac{1}{\tau}\left(1 + \frac{1}{m}\right) \frac{\varsigma^2}{b} + \frac{\beta d \sigma^2 C^2}{m^2 \eta \tau} \right] \right). \end{aligned} \tag{92}$$

Proof. We first bound the local-model drift before client-level clipping. Since $\theta_i^{t\tau} = \theta^{t\tau}$, iterating the local SGD update for r steps gives

$$\theta_i^{t\tau+r} - \theta^{t\tau} = -\frac{\eta}{b}\sum_{s=0}^{r-1}\sum_{z\in\mathcal{B}_i^{t\tau+s}} g_i(\theta_i^{t\tau+s}, z). \tag{93}$$

Conditioning on the history before local step s, Assumption 2 gives

$$\mathbb{E}\left[\left\|\frac{1}{b}\sum_{z\in\mathcal{B}_i^{t\tau+s}} g_i(\theta_i^{t\tau+s}, z) - \nabla\hat{\Phi}_i(\theta_i^{t\tau+s})\right\|_2^2\right] \le \frac{\varsigma^2}{b}. \tag{94}$$

Moreover, Assumptions 1 and 3 imply

$$\left\|\nabla\hat{\Phi}_i(\theta_i^{t\tau+s})\right\|_2^2 \le 4\left(\left\|\nabla\hat{\Phi}(\theta^{t\tau})\right\|_2^2 + G^2 + \beta^2\left\|\theta_i^{t\tau+s} - \theta^{t\tau}\right\|_2^2\right). \tag{95}$$

Conditioning on the history before each local step, we decompose the mini-batch gradient into its conditional mean and a centered stochastic perturbation. Using $\|\sum_{s=0}^{r-1} a_s\|_2^2 \le r\sum_{s=0}^{r-1}\|a_s\|_2^2$ and the martingale-difference property of the centered perturbations, we obtain

$$\mathbb{E}\left[\left\|\theta_i^{t\tau+r} - \theta^{t\tau}\right\|_2^2\right] \le 8\eta^2 r\sum_{s=0}^{r-1}\mathbb{E}\left[\left\|\nabla\hat{\Phi}(\theta^{t\tau})\right\|_2^2 + G^2 + \beta^2\left\|\theta_i^{t\tau+s} - \theta^{t\tau}\right\|_2^2\right] + 2\eta^2 r\frac{\varsigma^2}{b}. \tag{96}$$

Using $r \le \tau$, the preceding inequality further gives

$$\mathbb{E}\left[\left\|\theta_i^{t\tau+r} - \theta^{t\tau}\right\|_2^2\right] \le 8\eta^2\tau^2\mathbb{E}\left[\left\|\nabla\hat{\Phi}(\theta^{t\tau})\right\|_2^2 + G^2\right] + 8\beta^2\eta^2\tau\sum_{s=0}^{r-1}\mathbb{E}\left[\left\|\theta_i^{t\tau+s} - \theta^{t\tau}\right\|_2^2\right] + 2\eta^2\tau\frac{\varsigma^2}{b}. \tag{97}$$

Taking the maximum over $0 \le r \le \tau$ and using $\sum_{s=0}^{r-1}\mathbb{E}[\|\theta_i^{t\tau+s} - \theta^{t\tau}\|_2^2] \le \tau\max_{0\le r\le\tau}\mathbb{E}[\|\theta_i^{t\tau+r} - \theta^{t\tau}\|_2^2]$, we obtain

$$\left(1 - 8\beta^2\eta^2\tau^2\right)\max_{0\le r\le\tau}\mathbb{E}\left[\left\|\theta_i^{t\tau+r} - \theta^{t\tau}\right\|_2^2\right] \le 8\eta^2\tau^2\mathbb{E}\left[\left\|\nabla\hat{\Phi}(\theta^{t\tau})\right\|_2^2 + G^2\right] + 2\eta^2\tau\frac{\varsigma^2}{b}. \tag{98}$$

Since $\beta\eta\tau \le 1/64$, we have $1 - 8\beta^2\eta^2\tau^2 = \Omega(1)$. Therefore,

$$\mathbb{E}\left[\left\|\theta_i^{t\tau+r} - \theta^{t\tau}\right\|_2^2\right] \le \mathcal{O}\left(\eta^2\tau^2\mathbb{E}\left[\left\|\nabla\hat{\Phi}(\theta^{t\tau})\right\|_2^2 + G^2\right] + \eta^2\tau\frac{\varsigma^2}{b}\right). \tag{99}$$

We next characterize the error introduced by client-level clipping. By Lemma 2, $\|g_i(\theta, z)\|_2 \le \sqrt{2\beta B}$. Therefore, the norm of the unclipped local update satisfies

$$\|\Delta_i^t\|_2 = \left\|-\frac{\eta}{b}\sum_{r=0}^{\tau-1}\sum_{z\in\mathcal{B}_i^{t\tau+r}} g_i(\theta_i^{t\tau+r}, z)\right\|_2 \le \eta\tau\sqrt{2\beta B}. \tag{100}$$

Since $\bar{\Delta}_i^t = \Delta_i^t\min\left\{1, \frac{C}{\|\Delta_i^t\|_2}\right\}$, we have

$$\left\|\bar{\Delta}_i^t - \Delta_i^t\right\|_2 \le \max\left\{0, \eta\tau\sqrt{2\beta B} - C\right\}. \tag{101}$$

Hence, the averaged clipping error satisfies

$$\left\|\frac{1}{m\eta\tau}\sum_{i\in\mathcal{M}_t}\left(\bar{\Delta}_i^t - \Delta_i^t\right)\right\|_2^2 \le \max\left\{0, \sqrt{2\beta B} - \frac{C}{\eta\tau}\right\}^2. \tag{102}$$

In particular, the clipping error vanishes whenever $C \ge \eta\tau\sqrt{2\beta B}$.

We now analyze one communication round. By Algorithm 1, the global update can be written as

$$\theta^{(t+1)\tau} - \theta^{t\tau} = -\frac{\eta}{mb} \sum_{i\in\mathcal{M}_t} \sum_{r=0}^{\tau-1} \sum_{z\in\mathcal{B}_i^{t\tau+r}} g_i(\theta_i^{t\tau+r}, z) + \frac{1}{m} \sum_{i\in\mathcal{M}_t} \left(\bar{\Delta}_i^t - \Delta_i^t\right) + \xi^t. \tag{103}$$

The averaged local direction can be decomposed into the global gradient $\nabla\hat{\Phi}(\theta^{t\tau})$, client-sampling error, local-model drift, centered mini-batch stochastic perturbation, and client-level clipping error. We bound these terms separately.

Since $\mathcal{M}_t$ is sampled uniformly without replacement, Assumption 3 gives

$$\mathbb{E}\left[\left\|\frac{1}{m}\sum_{i\in\mathcal{M}_t}\nabla\hat{\Phi}_i(\theta^{t\tau}) - \nabla\hat{\Phi}(\theta^{t\tau})\right\|_2^2\right] \le \frac{G^2}{m}. \tag{104}$$

By Assumption 1, Jensen's inequality, and equation 99, the error caused by evaluating local gradients along the local trajectories satisfies

$$\mathbb{E}\left[\left\|\frac{1}{m\tau}\sum_{i\in\mathcal{M}_t}\sum_{r=0}^{\tau-1}\left[\nabla\hat{\Phi}_i(\theta_i^{t\tau+r}) - \nabla\hat{\Phi}_i(\theta^{t\tau})\right]\right\|_2^2\right] \le \mathcal{O}\left(\beta^2\eta^2\tau^2\mathbb{E}\left[\left\|\nabla\hat{\Phi}(\theta^{t\tau})\right\|_2^2 + G^2\right] + \beta^2\eta^2\tau\frac{\varsigma^2}{b}\right). \tag{105}$$

Moreover, by conditional independence across clients and the martingale-difference property across local steps, the averaged centered stochastic perturbation satisfies

$$\mathbb{E}\left[\left\|\frac{1}{m\tau}\sum_{i\in\mathcal{M}_t}\sum_{r=0}^{\tau-1}\left(\frac{1}{b}\sum_{z\in\mathcal{B}_i^{t\tau+r}} g_i(\theta_i^{t\tau+r}, z) - \nabla\hat{\Phi}_i(\theta_i^{t\tau+r})\right)\right\|_2^2\right] \le \frac{\varsigma^2}{mb\tau}. \tag{106}$$

Finally, since $\xi^t \sim \mathcal{N}\left(0, \frac{\sigma^2C^2}{m^2}I_d\right)$,

$$\mathbb{E}\left[\|\xi^t\|_2^2\right] = \frac{d\sigma^2C^2}{m^2}. \tag{107}$$

By Assumption 1, the global empirical risk $\hat{\Phi}$ is also β-smooth. Therefore, the descent lemma gives

$$\hat{\Phi}(\theta^{(t+1)\tau}) \le \hat{\Phi}(\theta^{t\tau}) + \left\langle\nabla\hat{\Phi}(\theta^{t\tau}), \theta^{(t+1)\tau} - \theta^{t\tau}\right\rangle + \frac{\beta}{2}\left\|\theta^{(t+1)\tau} - \theta^{t\tau}\right\|_2^2. \tag{108}$$

After substituting equation 103, the client-sampling and centered stochastic terms vanish in the inner product after taking expectation. The Gaussian perturbation also has zero conditional mean and is independent of the update before noise injection. Applying Young's inequality to the local-model drift and clipping error, using equation 102-equation 107, and absorbing the terms proportional to $\mathbb{E}[\|\nabla\hat{\Phi}(\theta^{t\tau})\|_2^2]$ under $\beta\eta\tau \le 1/64$, we obtain

$$\begin{aligned}\mathbb{E}\left[\hat{\Phi}(\theta^{(t+1)\tau})\right] \le \mathbb{E}\left[\hat{\Phi}(\theta^{t\tau})\right] - \frac{\eta\tau}{4}\mathbb{E}\left[\left\|\nabla\hat{\Phi}(\theta^{t\tau})\right\|_2^2\right] + \mathcal{O}\left(\eta\tau\max\left\{0, \sqrt{2\beta B} - \frac{C}{\eta\tau}\right\}^2\right.\\ \left.+\beta\eta^2\tau^2\left(\frac{G^2}{m} + \frac{\varsigma^2}{mb\tau}\right) + \beta^2\eta^3\tau^3\left(G^2 + \frac{\varsigma^2}{b\tau}\right) + \frac{\beta d\sigma^2C^2}{m^2}\right).\end{aligned} \tag{109}$$

Since $\beta\eta\tau \le 1/64$, we have

$$\beta\eta^2\tau^2 = \eta\tau(\beta\eta\tau) \le \frac{\eta\tau}{64}, \qquad \beta^2\eta^3\tau^3 = \eta\tau(\beta\eta\tau)^2 \le \frac{\eta\tau}{64^2}. \tag{110}$$

Therefore, equation 109 can be relaxed as

$$\begin{aligned}\mathbb{E}\left[\hat{\Phi}(\theta^{(t+1)\tau})\right] \le \mathbb{E}\left[\hat{\Phi}(\theta^{t\tau})\right] - \frac{\eta\tau}{4}\mathbb{E}\left[\left\|\nabla\hat{\Phi}(\theta^{t\tau})\right\|_2^2\right] + \mathcal{O}\left(\eta\tau\left[\max\left\{0, \sqrt{2\beta B} - \frac{C}{\eta\tau}\right\}^2\right.\right.\\ \left.\left.+\left(1+\frac{1}{m}\right)G^2 + \frac{1}{\tau}\left(1+\frac{1}{m}\right)\frac{\varsigma^2}{b}\right] + \frac{\beta d\sigma^2C^2}{m^2}\right).\end{aligned} \tag{111}$$

By Assumption 5,

$$\mathbb{E}\left[\left\|\nabla\hat{\Phi}(\theta^{t\tau})\right\|_2^2\right] \geq 2\mu\mathbb{E}\left[\hat{\Phi}(\theta^{t\tau}) - \hat{\Phi}(\hat{\theta}^\star)\right]. \tag{112}$$

Substituting this inequality into equation 111 yields

$$\begin{aligned}\mathbb{E}\left[\hat{\Phi}(\theta^{(t+1)\tau}) - \hat{\Phi}(\hat{\theta}^\star)\right] &\leq \left(1 - \frac{\mu\eta\tau}{2}\right)\mathbb{E}\left[\hat{\Phi}(\theta^{t\tau}) - \hat{\Phi}(\hat{\theta}^\star)\right] \\ &+ \mathcal{O}\left(\eta\tau\left[\max\left\{0, \sqrt{2\beta B} - \frac{C}{\eta\tau}\right\}^2 + \left(1 + \frac{1}{m}\right)G^2 + \frac{1}{\tau}\left(1 + \frac{1}{m}\right)\frac{\varsigma^2}{b}\right] + \frac{\beta d\sigma^2 C^2}{m^2}\right).\end{aligned} \tag{113}$$

Since $\mu\eta\tau \leq 1$, unrolling equation 113 over $t = 0, \ldots, T-1$ and using

$$\sum_{t=0}^{T-1}\left(1 - \frac{\mu\eta\tau}{2}\right)^t \leq \frac{2}{\mu\eta\tau} \tag{114}$$

give

$$\begin{aligned}\mathbb{E}\left[\hat{\Phi}(\theta^{T\tau}) - \hat{\Phi}(\hat{\theta}^\star)\right] &\leq \left(1 - \frac{\mu\eta\tau}{2}\right)^T\left[\hat{\Phi}(\theta^0) - \hat{\Phi}(\hat{\theta}^\star)\right] \\ &+ \mathcal{O}\left(\frac{1}{\mu}\left[\max\left\{0, \sqrt{2\beta B} - \frac{C}{\eta\tau}\right\}^2 + \left(1 + \frac{1}{m}\right)G^2 + \frac{1}{\tau}\left(1 + \frac{1}{m}\right)\frac{\varsigma^2}{b} + \frac{\beta d\sigma^2 C^2}{m^2\eta\tau}\right]\right).\end{aligned} \tag{115}$$

This completes the proof. □