UNVEILING THE NON-MONOTONIC EFFECT OF PRIVACY ON GENERALIZATION UNDER BYZANTINE ROBUSTNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Prior work has established a fundamental trilemma between Byzantine robustness, local differential privacy (LDP), and optimization error in distributed learning. We show that the trilemma does not extend to generalization error as such, but instead depends on privacy regimes. In the high-noise (strong privacy) regime, we prove that increasing privacy reduces the generalization error, i.e., there is no tension between robustness and privacy. In the low-noise (weaker privacy) regime, however, there is a tension between robustness and privacy, i.e., increasing privacy indeed degrades generalization. We characterize this non-monotonic behavior rigorously by providing matching lower and upper bounds on the algorithmic stability of Byzantine-robust distributed learning subject to LDP. Our characterization leverages a novel connection between the strength of Byzantine attacks and their ability to perform membership inference attacks. Through numerical experiments, we further show that this behavior is not confined to the theoretical worst case, but also manifests in real learning problems under existing Byzantine attack.

1 INTRODUCTION

Learning high-quality models often requires access to large amounts of data and substantial computational resources, making distributed learning systems such as Federated Learning (FL) (Kairouz et al., 2021) a promising approach to leverage data and computation across many participants. FL is typically implemented via distributed gradient descent algorithms coordinating local updates across participants. However, sharing model updates is known to leak information about the underlying training data, which can reduce participants' willingness to contribute and raises challenges for complying with data protection regulations. Local Differential Privacy (LDP, or local DP) is a natural privacy framework for distributed learning under the honest-but-curious threat model, where the server follows the protocol but may attempt to infer sensitive information from participant updates (Duchi et al., 2013; Kairouz et al., 2021). This assumption is practically motivated, as the server is often a third party that cannot be fully trusted. LDP typically enforces privacy by having each participant locally add noise to its updates before sending them. In addition to privacy concerns, real-world distributed systems face a range of unpredictable malfunctions, from benign hardware failures to adversarial misbehavior (e.g., poisoning attacks), broadly categorized under the umbrella of Byzantine failures (Guerraoui et al., 2024). This motivates a second line of defense known as Byzantine robustness, which aims to ensure reliable learning through robust aggregation rules that mitigate the influence of corrupted or malicious updates. While both LDP and Byzantine robustness are well-studied in isolation, their interplay remains largely underexplored, even though federated learning naturally gives rise to settings where both must be considered simultaneously.

Related work. A first line of work has investigated the impact of LDP alone, establishing a tight privacy-utility trade-off, where LDP constraints inflate the optimization of the empirical risk by a variance penalty (Wang et al., 2020; Girgis et al., 2021; Noble et al., 2022). LDP has also been studied in federated stochastic convex optimization (Duchi et al., 2013; Reshef & Levy, 2024), where the focus is on excess population risk guarantees, leading to similar qualitative insights.

Focusing on Byzantine robustness alone, recent work has extensively studied (f, κ)-robust aggregation as a principled approach to achieve optimal optimization guarantees in the presence of Byzantine failures (Allouah et al., 2023a). These results show that, while robust aggregation can tolerate a fraction of Byzantine participants, it introduces an unavoidable bias that prevents the optimization

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.

error from vanishing when data are heterogeneous across participants. This limitation extends to the statistical learning setting (Boudou et al., 2025), where the generalization error (the discrepancy between training and population risks) is further amplified by the constraints imposed by robust aggregation. However, their analysis does not consider formal privacy constraints.

Recently, a fundamental *trilemma* has been demonstrated for distributed algorithms that aim to be both Byzantine robust and satisfy LDP (Allouah et al., 2023b). Specifically, this work shows that the interaction between Byzantine robustness mechanisms and LDP constraints induces an additional optimization error term that does not appear when considering either requirement in isolation. This unavoidable term reveals an intrinsic tension between privacy and robustness. While this trilemma characterizes the optimization dynamics in terms of empirical risk, the true measure of a model's performance is its population risk, i.e., its behavior on unseen data. This risk is typically decomposed into the sum of optimization and generalization errors (Hardt et al., 2016), see Section 2. This naturally raises the question: *does this trilemma extend to generalization?* While it is intuitive that injecting privacy noise prior to aggregation can hinder the server's ability to distinguish between honest and Byzantine updates, thereby increasing optimization error, its effect on generalization is less clear. Indeed, such noise is often associated with an implicit regularization effect that can help control and even improve generalization (Wang et al., 2016; Neu et al., 2021). This creates a fundamental tension: *does LDP noise merely exacerbate the impact of Byzantine robustness, or can it also mitigate them?* The interplay between these two opposing effects remains poorly understood and is the focus of our work.

Contributions. We investigate this question and show that the privacy-robustness-optimization trilemma does not directly extend to generalization. By tightly analyzing the algorithmic stability of the algorithm introduced by Allouah et al. (2023b) (i.e., `Safe-D(S)DG`, detailed in Algorithm 1), which achieves optimal optimization error in the considered setting, we uncover a fundamental non-monotonic relationship between privacy guarantees and generalization error under Byzantine robustness. Crucially, this behavior is not driven solely by the raw noise magnitude, but is characterized by the ability of Byzantine participants to reliably detect the presence of individual data points, i.e., the performance of a *Membership Inference Attack* (MIA). Indeed, we show that membership information of honest participants' data is crucial to Byzantine participants' power to attack the learning process. Specifically, we identify a threshold corresponding to a qualitative change in the privacy regime, where an optimal MIA achieves a non-trivial probability of success. This transition delineates two distinct regimes, each exhibiting fundamentally different generalization behavior. **(i) Weak privacy regime (low noise):** In this regime, the injected noise interacts adversely with Byzantine-robust aggregation, leading to a degradation of generalization performance. Quantitatively, the generalization error increases with the noise level and with the fraction of Byzantine participants, causing an initial deterioration. **(ii) Strong privacy regime (high noise):** Beyond the MIA threshold, the attack becomes ineffective and the system enters a regime consistent with classical differential privacy behavior. In this regime, increasing the noise level improves stability, and the generalization error decreases, eventually vanishing as privacy strengthens.

While the MIA construction characterizes the theoretical worst-case limits, our numerical experiments further show that this non-monotonic behavior extends beyond this regime and manifests in practical learning problems under existing Byzantine attacks.

Overall, ur results refine the characterization of the privacy-robustness-utility trade-off in distributed learning.

2 PRELIMINARIES

Problem setting. We consider a distributed setup with n participants coordinated by a central server. We assume this system is subject to Byzantine failures, and requires to guarantee LDP.

We consider the Byzantine failures threat model (Lamport et al., 1982; Guerraoui et al., 2024), in which at most f ($< \frac{n}{2}$) participants can act arbitrarily. Their identities are unknown to the server and they may deviate from the prescribed algorithm, collude, and send adversarial updates while having access to all information exchanged between honest participants and the server. We note $\mathcal{H} \subseteq [n] := \{1, \dots, n\}$ the set of *honest* participants, with $|\mathcal{H}| = n - f$. Each honest participant $i \in \mathcal{H}$ holds a local dataset $\mathcal{D}_i = \{z^{(i,1)}, \dots, z^{(i,m)}\}$ composed of m i.i.d. data points (or samples) from an input space $\mathcal{Z}$ drawn from a distribution p_i.

We denote by $\mathcal{S} = \cup_{i\in\mathcal{H}}\mathcal{D}_i$ the overall dataset. Given a parameter vector $\theta \in \Theta \subset \mathbb{R}^d$ representing the model, a data point $z \in \mathcal{Z}$ incurs a loss defined by a real-valued function $\ell(\theta; z)$. The goal is to minimize the *population* risk over the honest participants

$$R_{\mathcal{H}}(\theta) = \tfrac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \mathbb{E}_{z\sim p_i}[\ell(\theta, z)].$$

Since only a finite number of samples are available from each distribution, we approximate this objective by minimizing the corresponding *empirical* risk over the (unknown) set of honest participants

$$\widehat{R}_{\mathcal{H}}(\theta) = \tfrac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \widehat{R}_i(\theta) = \tfrac{1}{n-f}\sum_{i\in\mathcal{H}} \tfrac{1}{m}\sum_{z\in\mathcal{D}_i} \ell(\theta, z).$$

The expected excess population risk of a learning algorithm's output $\mathcal{A} : \mathcal{Z}^{(n-f)m} \to \Theta, \mathcal{S} \mapsto \mathcal{A}(\mathcal{S})$ can be decomposed into generalization and optimization errors (e.g., Hardt et al., 2016) as follows

$$\mathbb{E}\big[R_{\mathcal{H}}(\mathcal{A}(\mathcal{S})) - \inf_{\theta\in\Theta} R_{\mathcal{H}}(\theta)\big] \leq \mathbb{E}\big[R_{\mathcal{H}}(\mathcal{A}(\mathcal{S})) - \widehat{R}_{\mathcal{H}}(\mathcal{A}(\mathcal{S}))\big] + \mathbb{E}\big[\widehat{R}_{\mathcal{H}}(\mathcal{A}(\mathcal{S})) - \inf_{\theta\in\Theta} \widehat{R}_{\mathcal{H}}(\theta)\big]. \quad (1)$$

Prior work on robust distributed learning has focused primarily on the optimization error, i.e., the second term in the bound (1). In contrast, we study the generalization error corresponding to the first term.

Robust distributed optimization. Robust distributed optimization algorithms aim to achieve low empirical risk. They are typically adaptations of standard first-order iterative optimization algorithms like gradient descent (GD) or stochastic gradient descent (SGD). These algorithms are made resilient by replacing the server-side averaging operator by a robust aggregation rule $F : (\mathbb{R}^d)^n \to \mathbb{R}^d$ (Guerraoui et al., 2024). Formally, given a learning rate γ, a parameter θ_t at iteration $t \in \{0, \ldots, T-1\}$ is updated as follows

$$\theta_{t+1} = G_\gamma^F(\theta_t) := \theta_t - \gamma F(g_t^{(1)}, \ldots, g_t^{(n)}),$$

where $g_t^{(i)}$ denotes the update sent by participant i at step t. For an honest participant $i \in \mathcal{H}$, we have $g_t^{(i)} = \nabla\widehat{R}_i(\theta_t)$ under GD, or $g_t^{(i)} = \nabla\ell(\theta_t; z_t^{(i)})$ under SGD, with $z_t^{(i)}$ uniformly sampled from the local dataset $\mathcal{D}_i$. For a misbehaving participant $i \notin \mathcal{H}$, $g_t^{(i)}$ is an arbitrary vector in $\mathbb{R}^d$.

Recent advances in robust distributed optimization have identified key properties sufficient for a robust aggregation rule to ensure the strong notion of (f, ρ)-resilience with respect to the empirical risk (Allouah et al., 2023a;b). The following definition captures a wide range of aggregation rules and has been shown to yield tight resilience guarantees across a variety of practical distributed learning settings.

Definition 2.1. *Let $\|\cdot\|_{\mathrm{sp}}$ denote the spectral norm for p.s.d. matrices (i.e., the largest eigenvalue). Let $n \geq 1$, $0 \leq f < n/2$ and $\kappa \geq 0$. An aggregation rule F is (f, κ)-robust if for any $g_1, \ldots, g_n \in \mathbb{R}^d$, any set $S \subset [n]$ of size $n - f$, with $\overline{g}_S = \frac{1}{|S|}\sum_{i\in S} g_i$ and $\Sigma_S = \frac{1}{|S|}\sum_{i\in S}(g_i - \overline{g}_S)(g_i - \overline{g}_S)^\intercal$,*

$$\|F(g_1, \ldots, g_n) - \overline{g}_S\|_2^2 \leq \kappa\|\Sigma_S\|_{\mathrm{sp}}.$$

f and κ are referred to as the robustness parameter and robustness coefficient of F, respectively.

An aggregation rule F that is $(f, \mathcal{O}(f/n))$-robust achieves optimal (f, ρ)-resilience. Specifically, iterative methods using such F attain an optimization error that matches the information-theoretic lower bound for first-order optimization under f Byzantine participants (Allouah et al., 2023b). An example is the *smallest maximum eigenvalue averaging* (SMEA) from Allouah et al. (2023b).

Definition 2.2. *Let $f, n \in \mathbb{N}$, $f < n/2$ and $g_1, \ldots, g_n \in \mathbb{R}^d$, the* SMEA *outputs*

$$\overline{g}_{S^*} = \frac{1}{|S^*|}\sum_{i\in S_t^*} g_i \quad \textit{with} \quad S^* \in \operatorname*{argmin}_{S\subseteq\{1,\ldots,n\}, |S|=n-f} \lambda_{\max}\left(\frac{1}{|S|}\sum_{i\in S}(g_i - \overline{g}_S)(g_i - \overline{g}_S)^\intercal\right).$$

Note that the use of the spectral norm in Definition 2.1 is motivated by our setting. Under LDP constraints, which are usually achieved by adding isotropic noise to gradients, the empirical covariance matrix of honest participants' gradients trace typically scales with the dimension, which is not the case when using the spectral norm, hence the usefulness of this matrix norm in high dimension. This highlights the focus on SMEA, which was specifically designed for our setting (Allouah et al., 2023b). The recent Covariance Bound Agnostic-Filter (CAF) from Allouah et al. (2025) is a computationally efficient variant of SMEA.

Differential privacy. A primary objective for honest participants is to protect the privacy of their datasets. To provide a rigorous, quantifiable guarantee, we leverage differential privacy (DP) (Dwork et al., 2006). DP ensures that an algorithm's output is indistinguishable whether or not any single individual's data is included in the input dataset. We adopt the standard definition of (ε, δ)-DP.

Definition 2.3. *Let $\varepsilon > 0$, $0 \leq \delta \leq 1$. A randomized algorithm $\mathcal{A}$ is (ε, δ)-DP if for all neighboring honest datasets S and S' (differing in a single entry) and for all measurable subset of output O,*

$$\mathbb{P}(\mathcal{A}(S) \in O) \leq e^{\varepsilon}\mathbb{P}(\mathcal{A}(S') \in O) + \delta$$

In the distributed setting, let $V_i(S)$ denote the (random) transcript of communications between the server and honest participant i generated by $\mathcal{A}(S)$. $\mathcal{A}$ is (ε, δ)-LDP if $\forall i \in \mathcal{H}$, V_i is (ε, δ)-DP.

To summarize, our threat model considers two types of adversaries. First, we model the central server as honest-but-curious: it follows the prescribed protocol (i.e., it is honest) but may simultaneously attempt to infer sensitive, private information about individual participants' datasets from model updates it observes (i.e., it is curious). Second, our system operates under the assumption of Byzantine failures. Note that this second adversary is strictly more powerful: it observes as much information as the server (since it can eavesdrop on its communications with honest participants), but also controls f Byzantine participants that can deviate from the protocol arbitrarily.

The `Safe-D(S)GD` algorithm. Owing to its strong empirical performance and baseline optimality guarantees, `Safe-D(S)GD` (Allouah et al., 2023b) constitutes a natural choice of algorithm for our study. We consider the variant without momentum (Algorithm 1). In the deterministic (GD) setting, momentum is not required to achieve optimal error rates, making `Safe-DGD` theoretically optimal. While momentum is necessary in the stochastic regime to control the noise floor induced by Byzantine failures in the optimization error guarantee (Karimireddy et al., 2021), omitting it here simplifies our analysis of algorithmic stability, which is already technically involved, and helps isolate the effects of privacy and Byzantine robustness. This simplification is further justified by prior work showing that momentum can negatively impact stability in distributed optimization (Attia & Koren, 2021; Ramezani-Kebrya et al., 2024).

Algorithm 1 `Safe-D(S)GD`

Initialization: θ_0, aggregation F, noise σ_{DP}, number of steps T, learning rate γ
for $t = 0$ **to** $T - 1$ **do**
  Server broadcasts θ_t to all participants
  for all honest participant $i \in \mathcal{H}$ **do**
    Compute $g_t^{(i)} = \frac{1}{m}\sum_{j=1}^{m} \nabla \ell(\theta_t; z^{(i,j)})$
    or $g_t^{(i)} = \nabla \ell(\theta_t; z^{(i,J_t^{(i)})})$ where $z^{(i,J_t^{(i)})}$ from $\mathcal{D}_i$ is sampled uniformly at random
    Add noise $\widetilde{g}_t^{(i)} = g_t^{(i)} + \xi_t^{(i)}; \xi_t^{(i)} \sim \mathcal{N}(0, \sigma_{\text{DP}}^2 I_d)$
    Send $\widetilde{g}_t^{(i)}$ to the server
  Server aggregates $R_t = F(\widetilde{g}_t^{(1)}, \ldots, \widetilde{g}_t^{(n)})$ # f participants may send an arbitrary vector
  Server updates model $\theta_{t+1} = \Pi_{\Theta}(\theta_t - \gamma_t R_t)$
return θ_T or $\hat{\theta}$ sampled uniformly from $\{\theta_0, \ldots, \theta_T\}$

The privacy analysis of `Safe-D(S)GD` follows the same, classic line of reasoning as in Allouah et al. (2023b, Theorem 4.1), namely the composition of the subsampled Gaussian mechanism. We state the theorem for `Safe-DGD` for simplicity, though `Safe-DSGD` yields similar guarantees.

Theorem 2.1. *If $\sigma_{\text{DP}} = \frac{2C\sqrt{2T\ln(1/\delta)}}{m\varepsilon}$, then* `Safe-DGD` *is (ε, δ)-LDP.*

3 Non-monotonic LDP effect for Byzantine-robust generalization

Our study leverages algorithmic stability to understand the generalization error of `Safe-DGD`. We do so via uniform stability, which measures the worst-case sensitivity of a learning algorithm to changes in its training data. This approach was introduced in Vapnik & Chervonenkis (1974); Devroye & Wagner (1979); Rogers & Wagner (1978), and later popularized by Bousquet & Elisseeff (2002);

Hardt et al. (2016); Wang et al. (2016). We revisit this framework to study generalization in private and robust distributed learning, where changes occur only in the honest participants' data.

Definition 3.1. *Consider a distributed setting with up to f Byzantine participants (out of n). A distributed algorithm $\mathcal{A}$ is ε-uniformly stable if, for all honest datasets $\mathcal{S}, \mathcal{S}' \in \mathcal{Z}^{(n-f)m}$ that are neighboring—i.e., differing in at most one sample—we have,*

$$\sup_{z \in \mathcal{Z}} \mathbb{E}_{\mathcal{A}}[\ell(\mathcal{A}(\mathcal{S}); z) - \ell(\mathcal{A}(\mathcal{S}'); z)] \leq \varepsilon. \tag{2}$$

By standard arguments, stability upper-bounds generalization error, and has been shown to tightly characterize generalization under Byzantine robustness (Boudou et al., 2025). The question then becomes: to which extent can Byzantine participants degrade the algorithmic stability of `Safe-DGD`?

Key intuition: MIA limits Byzantine adaptation. To maximally increase the stability gap in (2) (and thus generalization error), Byzantine participants must adapt their updates depending on whether training is performed on $\mathcal{S}$ or $\mathcal{S}'$. Since $\mathcal{S}$ and $\mathcal{S}'$ differ by only a single sample replacement, such adaptation requires inferring which differing sample is present from the communicated noisy gradients. The LDP guarantee limits precisely this inference, by controlling the performance of a *Membership Inference Attack* (MIA). Thus, LDP restricts Byzantine participants' ability to adapt their updates and amplify the algorithm's sensitivity to sample replacements.

Before stating our main results, we formalize the hypothesis-testing framework for MIAs that underlies our approach. Using this framework, we then derive matching lower and upper bounds on the stability of `Safe-DGD` under Byzantine failures and LDP constraint, thereby proving a non-monotonic effect of LDP on the generalization error under Byzantine robustness.

Denote $\mu_t = \mathrm{Law}(V_{1:t}(\mathcal{S}))$ and $\mu'_t = \mathrm{Law}(V_{1:t}(\mathcal{S}'))$ the distributions of the transcript of communication between the server and honest participants up to step $t \in [T]$, as observed by the Byzantine participants when running $\mathcal{A}$ on datasets $\mathcal{S}$ and $\mathcal{S}'$, respectively. Consistent with the LDP threat model, Byzantine participants know $(\mathcal{S}, \mathcal{S}')$, representing the hypothetical neighboring dataset in the stability analysis, the algorithm and the privacy mechanism, and thus have full knowledge of the candidate transcript distributions μ_t and μ'_t. Their uncertainty lies solely in determining whether the observed transcript of honest noisy gradients originates from training on $\mathcal{S}$ (hypothesis H_0) or $\mathcal{S}'$ (hypothesis H_1). Under this theoretical test, or MIA, LDP bounds the distinguishability of these two hypotheses, and therefore limit the success of any MIA (Dong et al., 2022). By standard hypothesis-testing bounds (Tsybakov, 2008, Lemma 2.1), the maximum success probability of any such test over the t-step transcript is bounded by the total variation distance between μ_t and μ'_t. That is, for any test,

$$\mathbb{P}(t\text{-step MIA success}) := 1 - \mathbb{P}(\text{reject } H_0 \mid H_0) - \mathbb{P}(\text{accept } H_0 \mid H_1) \leq d_{\mathrm{TV}}(\mu_t, \mu'_t) := \mathfrak{P}_t. \tag{3}$$

Moreover, there exist an optimal test for which equality holds in (3) (Dong et al., 2022, Neyman-Pearson Lemma as stated in Theorem A.1). Under (ε, δ)-LDP via the Gaussian mechanism, the optimal MIA success probability satisfies $\mathfrak{P}_T \in \Theta\big(\min\big\{1, \varepsilon/\sqrt{\ln(1/\delta)}\big\}\big)$; see (Dong et al., 2022). For completeness, we provide a proof in Lemma A.1.

Lower bounds. Any Byzantine attack instantiation provides a lower bound on the uniform stability as, by definition, the stability (or any upper bound) has to hold under any attack. To explicit how the MIA information limit translates into algorithmic instability, we describe, for intuition, the lower-bound construction of Theorem 3.1 as a two-phase Byzantine attack. Byzantine participants dedicate an initial fraction of the training steps pT, for a fixed $p \in (0, 1)$, to perform an optimal MIA with success probability $\mathfrak{P}_{pT} \in \Theta\left(\mathfrak{P}_T\right)$. They then use the remaining $(1-p)T$ steps to execute an attack conditioned on the MIA event. This construction allows us to lower bound the resulting instability as the product of two terms: the probability $\mathfrak{P}_{pT}$ of a successful MIA, and the instability that Byzantine participants can induce over the remaining steps, without the LDP constraint (i.e., conditioned on the MIA success), but still under a robust aggregation rule. The argument can be formalized to yield the bound stated in Theorem 3.1, with the proof deferred to Appendix C.3.

Theorem 3.1. *Consider the setting in Section 2 with full batch and the* SMEA *aggregation. Suppose there exist $\alpha_{\min}, \mu, \nu > 0$ such that $\frac{f}{n-f} \in [\alpha_{\min}, \frac{1}{2})$ and $\frac{d}{n-f} \in [\mu, \nu]$. Let $\mu_{\min}, N_{\min}, c_{\min}$ be problem-dependent constants defined in Assumption C.3, and assume $\mu \geq \mu_{\min}$, $n - f - 1 \geq N_{\min}$, $T \in \mathcal{O}\big(e^{c_{\min}\left(\sqrt{d}+\sqrt{n-f-1}\right)^2}\big)$. Then there exists $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}$ such that $\forall z \in \mathcal{Z}, \ell(\cdot; z)$ is C-Lipschitz, L-smooth and convex, and the uniform stability of $\mathcal{A}$ is lower bounded by*

$$\Omega\left(\mathfrak{P}_T \times \gamma C^2 T \left(\tfrac{1}{(n-f)m} + \tfrac{f}{n-2f} + \sqrt{\tfrac{f}{n-2f}}\tfrac{\sigma_{\mathrm{DP}}}{C}\left(1 + \sqrt{\tfrac{d}{n-f}}\right)\right)\right). \tag{4}$$

If we moreover assume $\frac{n}{3} \leq f < \frac{n}{2}$, the uniform stability of $\mathcal{A}$ is lower bounded by

$$\Omega\left(\mathfrak{P}_T \times \gamma C^2 T\left(\tfrac{1}{(n-f)m} + \sqrt{\tfrac{f}{n-2f}} + \sqrt{\tfrac{f}{n-2f}}\tfrac{\sigma_{\mathrm{DP}}}{C}\left(1+\sqrt{\tfrac{d}{n-f}}\right)\right)\right). \tag{5}$$

For both cases, if we assume the loss function bounded by ℓ_∞[1] and $T \in \mathcal{O}\big(\mathfrak{P}_T^2 e^{c_{\min}\left(\sqrt{d}+\sqrt{n-f-1}\right)^2}\big)$, then there exists $\sigma_{threshold} > 0$ such that for $\sigma_{DP} \geq \sigma_{threshold}$, the uniform stability of $\mathcal{A}$ is lower bounded by the minimum between (4) *(respectively* (5)*) and $\Omega\left(\ell_\infty \mathfrak{P}_T\right)$.*

The assumptions of Theorem 3.1, while technical, are fairly mild. They are either almost always satisfied in practice (e.g., the upper bound on T) or assumed to enable bounds on expectations and concentration of random matrices (e.g., $\mu_{\min}$ and $N_{\min}$). For details, see Appendix C.3.2. The two ranges on $\frac{f}{n}$ from Theorem 3.2 are inherited from Boudou et al. (2025). The term $\sqrt{\frac{f}{n-2f}}\frac{\sigma_{\mathrm{DP}}}{C}\big(1+\sqrt{\frac{d}{n-f}}\big)$, which is common to these two ranges, arises as the product of privacy- and robustness-driven terms and vanishes when $f = 0$. Intuitively, conditioned on a successful MIA, Byzantine participants can leverage the spectral dispersion of the honest noisy gradients to bias the update in a data-dependent direction. This, however, is capped by the spectral filter (SMEA) at the level of the honest spectral dispersion, i.e., $\mathbb{E}\|\Sigma_{\mathcal{H}}\|_{\mathrm{sp}} \in \Theta\big(\sigma_{\mathrm{DP}}\big(1+\sqrt{\frac{d}{n-f}}\big)\big)$. Notably, the MIA success probability $\mathfrak{P}_T \in \Theta\big(\min\big(1, \frac{\varepsilon}{\sqrt{\ln(1/\delta)}}\big)\big)$ (Lemma A.1) scales as $1/\sigma_{\mathrm{DP}}$ and saturates at 1. As privacy noise grows, $\mathfrak{P}_T$ decreases while the other term increases; combined with the boundedness of the loss, this interplay foreshadows the non-monotonic behavior formalized in the following paragraphs.

Upper bounds. We now upper-bound the stability of `Safe-DGD` (proof deferred to Appendix C.2).

Theorem 3.2. *Consider the setting described in Section 2. Let $\sigma_{\mathrm{DP}} = \frac{2C\sqrt{2T\ln(1/\delta)}}{m\varepsilon}$ and $\gamma \leq \frac{1}{L}$. Suppose $\forall z \in \mathcal{Z}$, $\ell(\cdot; z)$ C-Lipschitz and L-smooth. Then $\mathcal{A}$ is (ε, δ)-*LDP *and, up to numerical constants explicitly provided in Appendix C.2,*

(i) If $\ell(\cdot; z)$ is convex $\forall z \in \mathcal{Z}$, then the uniform stability of $\mathcal{A}$ is upper bounded by

$$2\gamma C^2 T\left(\tfrac{1}{(n-f)m} + \sqrt{\kappa} + \sqrt{\kappa}\tfrac{\sigma_{\mathrm{DP}}}{C}\sqrt{1+\tfrac{d}{n-f}}\right). \tag{6}$$

(ii) If $\ell(\cdot; z)$ is μ-strongly convex $\forall z \in \mathcal{Z}$, then the uniform stability of $\mathcal{A}$ is upper bounded by

$$\tfrac{2C^2}{\mu}\left(\tfrac{1}{(n-f)m} + \sqrt{\kappa} + \sqrt{\kappa}\tfrac{\sigma_{\mathrm{DP}}}{C}\sqrt{1+\tfrac{d}{n-f}}\right). \tag{7}$$

If we further assume the loss function to be bounded by ℓ_∞, then the uniform stability of $\mathcal{A}$ is upper bounded by the minimum between (6)*, respectively* (7)*, and $2\ell_\infty \mathfrak{P}_T$.*

When there is no Byzantine participant ($f = 0$, hence $\kappa = 0$), both eqs. (6) and (7) reduce to the classical stability bounds established in Hardt et al. (2016). When $\sigma_{\mathrm{DP}} = 0$, we recover the bound from Boudou et al. (2025). However, when both $f, \sigma_{\mathrm{DP}} > 0$, the analysis reveals an additive degradation term arising from the interaction between privacy and robustness. In the convex setting, this degradation accumulates over the T iterations, whereas in the strongly convex case, it appears as a one-time term. The bound scales linearly with the maximal sensitivity of the gradient update ($2\gamma C$), but also with the privacy noise magnitude. Thus, stronger privacy (more noise) degrades stability, in contrast with standard DP-to-stability arguments (Wang et al., 2016). We discuss this interesting phenomenon that was not observed when only considering DP in the following tightness analysis below.

Consequence: non-monotonic effect of privacy on generalization. First assume the loss function is bounded by ℓ_∞. Due to the last parts of Theorems 3.1 and 3.2, there exist $\sigma_{\text{threshold}} > 0$ such that for $\sigma_{\mathrm{DP}} \geq \sigma_{\text{threshold}}$, the bounds (4), (5), (6), (7) are of the order $\Theta(\ell_\infty \mathfrak{P}_T)$. That is, in this strong privacy regime, the guarantee eventually saturates and is solely characterized by the LDP constraint. In this regime, as the privacy strengthens (alternatively σ_{DP} increases), $\mathfrak{P}_T$ decreases. This explains why the stability improves with the noise magnitude in the strong privacy regime.

[1]This can be ensured by projecting parameters onto a compact Θ after the update Line 10 in Algorithm 1.

Alternatively, assume the weak privacy regime $\mathfrak{P}_T \in \Omega(1)$. For SMEA, $\sqrt{\kappa} \in \mathcal{O}\big(\sqrt{\frac{f}{n-f}}(1 + \frac{f}{n-2f})\big)$ (Allouah et al., 2023b, Proposition 5.1). Plugging this into our upper bound (6) establishes tightness with respect to (6) and (5), and shows that (6) matches (4) under the mild condition $\frac{\sigma_{\text{DP}}}{C}\big(1 + \sqrt{\frac{d}{n-f}}\big) \in \Omega(1)$. The tightness of our result highlights an unavoidable tension between LDP and Byzantine robustness in the stability of the studied algorithm. This explains why the stability—and consequently the generalization error (Boudou et al., 2025, Lemma 4.1)—worsens with the noise magnitude in the weak privacy regime.

Overall, for small σ_{DP} (large ε, weak privacy regime), the tight bounds of Theorems 3.1 and 3.2 linearly grows with σ_{DP}. Eventually, σ_{DP} becomes large enough (small ε, strong privacy regime) and the loss becomes saturated while Byzantine participants fail to consistently succeed their MIA, so as the bounds decrease with σ_{DP}. This transition highlights a surprising non-monotonic dependence of generalization on the level of privacy noise, as summarized by the following corollary.

Corollary 3.2.1. *Assume the setting of Theorem 3.1 with bounded loss. Let* $\sigma_{\text{DP}} = \frac{2C\sqrt{2T\ln(1/\delta)}}{m\varepsilon}$. *In the weak privacy regime* $\frac{\varepsilon}{\sqrt{\ln(1/\delta)}} \in \Omega(1)$, *or equivalently* $\sigma_{\text{DP}} \in \mathcal{O}(\sigma_{\text{base}})$ *with* $\sigma_{\text{base}} := \frac{C\sqrt{T}}{m}$, *the stability scales as*

$$\sigma_{\text{DP}} \times \Theta\Big(\gamma C T \sqrt{\tfrac{f}{n-2f}}\Big(1 + \sqrt{\tfrac{d}{n-f}}\Big)\Big), \tag{8}$$

and, in the strong privacy regime, it scales as

$$\tfrac{1}{\sigma_{\text{DP}}} \times \mathcal{O}\Big(\tfrac{\ell_\infty C\sqrt{T}}{m}\Big), \tag{9}$$

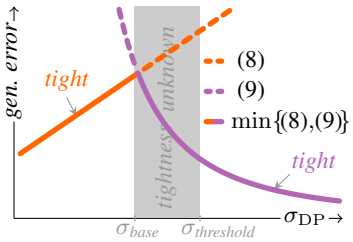


Figure 1: Generalization error as a function of σ_{DP}.

which is asymptotically tight in σ_{DP}. *In other words, there exists* $\varepsilon_{\text{threshold}} \in (0, 1)$ *such that, for* $\varepsilon \leq \varepsilon_{\text{threshold}}$, *the lower bound scale as* $\Theta(\ell_\infty \frac{\varepsilon}{\sqrt{\ln(1/\delta)}})$ *(cf. Theorems 3.1 and 3.2 and Lemma A.1). Generalization exhibits a similar behavior, following from (Boudou et al., 2025, Lemma 4.1).*

Corollary 3.2.1 shows the non-monotonic effect of LDP on generalization error under Byzantine robustness. Having established this phenomenon in the theoretical worst case, we next investigate whether it persists in practical learning scenarios under existing Byzantine attack.

4 Empirical Evaluations

In this section, we show that the non-monotonic behavior predicted by our worst-case stability analysis is not merely a theoretical artifact, but also arises in real learning problems under existing Byzantine attacks. All simulations are implemented via the ByzFL library (González et al., 2025); additional experimental details and evaluations appear in Appendix D.

Empirical evaluation settings. For each of the three datasets considered, we consider a common evaluation protocol. While a single training run is inexpensive, estimating expected generalization requires repetitions over datasets, seeds, and noise multipliers, which is the main bottleneck. This motivates the use of a fixed setting with $n = 7$ participants[2] and $m = 200$ local samples per honest participant. The model test loss is evaluated on a held-out test set of $N_{\text{test}} = 10000$ samples.

We consider three classification tasks: (i) synthetic data following Li et al. (2020); Noble et al. (2022) with logistic regression; (ii) MNIST (Lecun et al., 1998) with logistic regression; and (iii) CIFAR-10 (Krizhevsky, 2009) with a convolutional network.

All experiments use deterministic initialization and `Safe-DGD` with learning rate γ, T iterations, SMEA aggregation, and Gaussian noise $\sigma_{\text{DP}} = \sigma \times \sigma_{\text{base}}$, calibrated via a multiplier σ, which serves as a proxy for the privacy level and scales the base noise $\sigma_{\text{base}} := \frac{2C\sqrt{2T}}{m}$. We evaluate $\sigma \in \{0.125, 0.25, 0.5, 1, 2, 4, 8, 16, 32\}$, reporting both individual realizations and Monte Carlo estimates over datasets and seeds. Byzantine participants ($f = 1$ unless otherwise stated) use a

[2]The computational cost of SMEA scales poorly with n, hence the small number of participants.

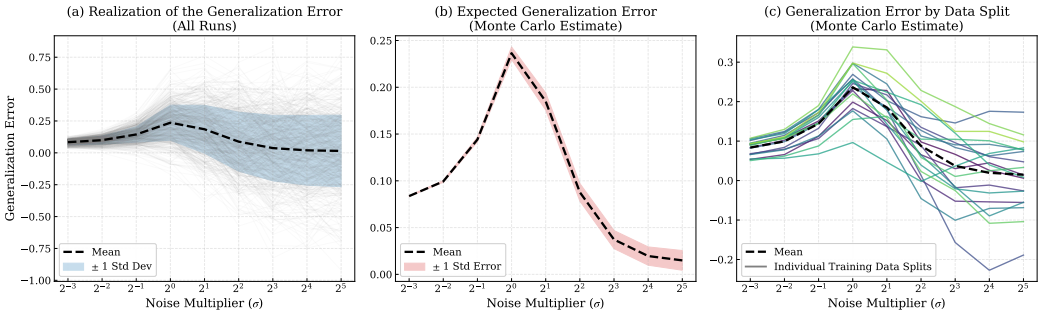


Figure 2: **Synthetic dataset.** Impact of LDP on generalization error under Byzantine failures. The Byzantine participant ($f = 1$) executes the Opt-IPM attack, the server uses SMEA as defense, with $\gamma = 0.4$, $T = 600$, and $d = 50$. **(a)** Realizations of the generalization error (as a random variable) demonstrating high variance across all independent runs (variance estimated on all runs). **(b)** The Monte Carlo estimate of the expected generalization error, where reported variance represents the mean estimation standard error. **(c)** Expected generalization error stratified by individual training dataset, isolating the variance attributable to data samples versus differential privacy noise.

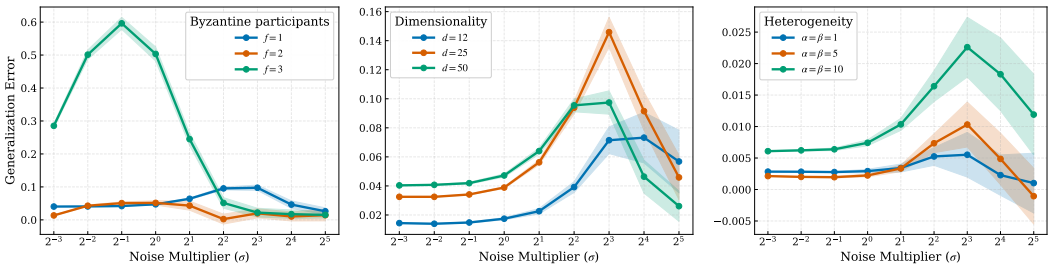


Figure 3: **Synthetic dataset.** Absolute expected generalization error under variations of parameters. **(a)** Left: varying $f \in \{1, 2, 3\}$ for fixed $d = 50$. **(b)** Middle: varying $d \in \{12, 25, 50\}$ for fixed $f = 1$. **(c)** Right: varying heterogeneity for fixed $f = 1$, where α, β are defined in Li et al. (2020) (the larger α and β, the more heterogenous). All evaluations share identical baseline parameters $T = 200$ and $\gamma = 0.2$.

state-of-the-art Opt-IPM attack (Xie et al., 2020), relying on line search at each iteration to compute the most damaging perturbation bypassing SMEA.

For (i)–(ii), the objective is convex and smooth. To be consistent with our theory, we enforce Lipschitzness and smoothness by L_2-normalizing features ($\|x\|_2 \leq 2$) (Ye & Shokri, 2022, Proposition F.1), and bounded loss by projecting parameters onto a ball of radius 100 after updates, enabling calibrated privacy noise. For CIFAR-10, gradients are clipped at 3.0 and parameters layerwise projected onto a ball of radius 10.

Below, we present the results, along with additional experiment-specific details.

Results and discussion. Our empirical results in Figures 2 to 5 show consistent and interpretable trends across settings. For logistic regression on synthetic data (Figures 2 and 3) and MNIST (Figure 4), which satisfy the assumptions of our theory, the observed dynamics closely match our predictions. On CIFAR-10 (Figure 5), where we train a CNN and thus operate outside the convex setting covered by our analysis, the behavior remains qualitatively consistent and exhibits meaningful structure despite more challenging interpretation, suggesting promising directions for further investigation.

Across Figures 2 and 4, the dependence on the noise multiplier aligns with our bounds: it grows linearly in the weak privacy regime. Figure 3 further highlights other dynamics: in the weak privacy regime, higher dimension consistently leads to worse generalization, in line with the predicted ordering, although the dependence departs from the square-root behavior suggested by the bound. In the strong privacy regime, the behavior becomes less regular, suggesting the presence of multiple

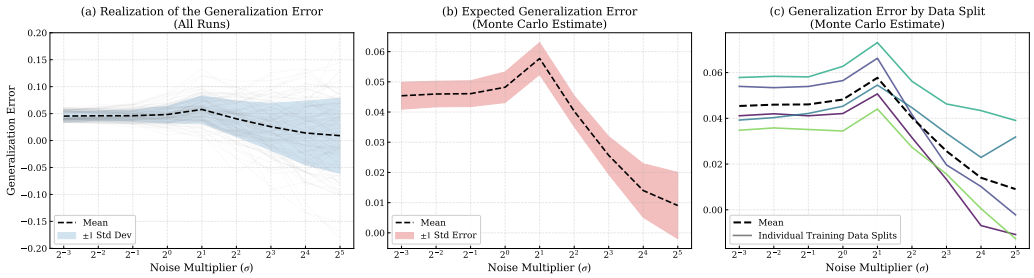


Figure 4: **MNIST**. Analogous to Figure 2 with $\gamma = 0.4$, $T = 400$, $f = 1$.

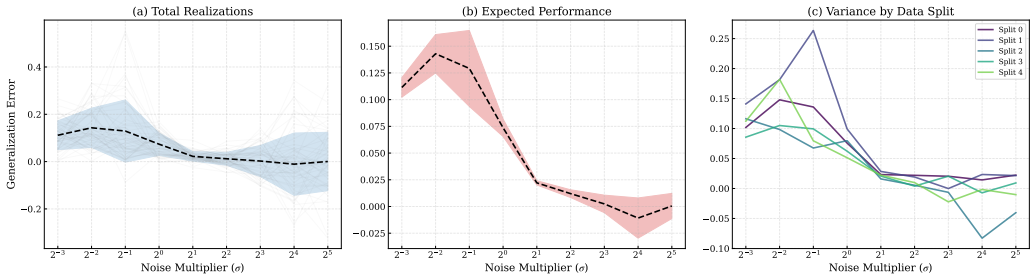


Figure 5: **CIFAR-10**. Analogous to Figure 2 with $\gamma = 0.1$, $T = 400$, $f = 1$. The convolutional network architecture consists of three convolutional layers (16, 32, 64 filters), ReLU activations, max-pooling, and a 1024-dim feature map followed by fully connected layers (128, 64, 10).

interacting effects. The dependence on the Byzantine fraction exhibits richer structure. The results are consistent with the change in regime in our bounds, which distinguish between $\frac{f}{n} \leq \frac{1}{3}$ (4) ($f = 1, 2$) and $\frac{1}{3} \leq \frac{f}{n} < \frac{1}{2}$ (5) ($f = 3$). For each fixed f, the behavior remains non-monotonic, as predicted, while the regime transition depends on f. Varying $\frac{f}{n-f}$, the dynamic is consistent with our bounds (inherited from Boudou et al., 2025), including a pronounced increase in attack severity at $f = 3$. Finally, the larger the heterogeneity, the less the model generalize.

Overall, these results support the qualitative trends of our analysis while revealing additional structure not captured by the bounds. As predicted, we observe an initial increase in generalization error under small noise (weak privacy), followed by a regime switch where the error decreases as noise increases (strong privacy). We leave several directions future work, including alternative optimizers, varying number of participants, different aggregation rules, and relaxing assumptions beyond convexity.

5 Conclusion

Prior optimization-focused theory overlooks the joint impact of LDP and Byzantine robustness on generalization, leaving an incomplete picture. We close this gap and uncover a surprising non-monotonic, two-phase behavior. Through tight stability bounds, we show that privacy noise actively degrades the generalization error in weak privacy regime, yet drives it to zero as privacy strengthens. Our theory offers several predictive insights and provides a qualitative and quantitative characterization of the transition between the two regimes. We moreover corroborate these findings in empirical evaluations, validating the usefulness of our bounds as predictive of the generalization error dynamics in real learning problems. While these findings offer a theoretical counterpoint to the optimization trilemma, practical privacy budgets often lie in the weak privacy regime. Thus, achieving privacy, robustness, and generalization simultaneously in practical distributed learning is fundamentally constrained. Moving forward, this necessitates the exploration of ways to systematically relax the threat models, whether by integrating cryptographic primitives to bound misbehaviors (Farhadkhani et al., 2022; Boudou et al., 2025), or by relaxing LDP as in Allouah et al. (2025), where a shared randomness seed among honest participants is assumed, enabling correlated noise mechanisms that yield improved optimization guarantees.

AI USE STATEMENT

In this work, we have not used generative AI tools for drafting paper sections, research ideation or execution, retrieval and discovery, generating synthetic datasets, conceiving the ideas or proving mathematical claims. Additionally, we used generative AI tools to aid and polish the writing of the manuscript. We have reviewed all AI-assisted work; all suggestions and edits were manually examined, revised, and verified for accuracy and tone by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Youssef Allouah, Sadegh Farhadkhani, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, and John Stephan. Fixing by mixing: A recipe for optimal byzantine ml under heterogeneity. In *AISTATS*, 2023a.

Youssef Allouah, Rachid Guerraoui, Nirupam Gupta, Rafael Pinot, and John Stephan. On the privacy-robustness-utility trilemma in distributed learning. In *ICML*, 2023b.

Youssef Allouah, Rachid Guerraoui, and John Stephan. Towards trustworthy federated learning with untrusted participants. In *ICML*, 2025.

Amit Attia and Tomer Koren. Algorithmic instabilities of accelerated gradient descent. *NeurIPS*, 2021.

F. Bach. *Learning Theory from First Principles*. Adaptive Computation and Machine Learning series. MIT Press, 2024.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Thomas Boudou, Batiste Le Bars, Nirupam Gupta, and Aurélien Bellet. Byzantine failures harm the generalization of robust distributed learning algorithms more than data poisoning. *arXiv preprint arXiv:2506.18020v2*, 2025.

Olivier Bousquet and André Elisseeff. Stability and generalization. *JMLR*, 2:499–526, 2002.

W. G. Cochran. The distribution of quadratic forms in a normal system, with applications to the analysis of covariance. *Mathematical Proceedings of the Cambridge Philosophical Society*, 30(2): 178–191, 1934.

Luc Devroye and T. Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, IT-25:601 – 604, 10 1979.

Jinshuo Dong, Aaron Roth, and Weijie J Su. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(1):3–37, 2022.

John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy, data processing inequalities, and statistical minimax rates. *arXiv preprint arXiv:1302.3203*, 2013.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, pp. 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

Sadegh Farhadkhani, Rachid Guerraoui, Oscar Villemaud, et al. An equivalence between data poisoning and byzantine gradient attacks. In *ICML*, 2022.

Antonious Girgis, Deepesh Data, Suhas Diggavi, Peter Kairouz, and Ananda Theertha Suresh. Shuffled model of differential privacy in federated learning. In *AISTATS*, 2021.

Marc González, Rachid Guerraoui, Rafael Pinot, Geovani Rizk, John Stephan, and François Taïani. Byzfl: Research framework for robust federated learning. *arXiv preprint arXiv:2505.24802*, 2025.

Rachid Guerraoui, Nirupam Gupta, and Rafael Pinot. Byzantine machine learning: A primer. *ACM Computing Surveys*, 56(7), 2024.

Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *ICML*, 2016.

Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konecný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2): 1–210, 2021.

Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Learning from history for byzantine robust optimization. In *ICML*, 2021.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Leslie Lamport, Robert Shostak, and Marshall Pease. The byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, 4(3):382–401, 1982.

B. Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28, 10 2000.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *MLSys*, 2020.

Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley Series in Probability and Statistics. 1982.

Gergely Neu, Gintare Karolina Dziugaite, Mahdi Haghifam, and Daniel M. Roy. Information-theoretic generalization bounds for stochastic gradient descent. In *COLT*, 2021.

Maxence Noble, Aurélien Bellet, and Aymeric Dieuleveut. Differentially private federated learning on heterogeneous data. In *AISTATS*, 2022.

Ali Ramezani-Kebrya, Kimon Antonakopoulos, Volkan Cevher, Ashish Khisti, and Ben Liang. On the generalization of stochastic gradient descent with momentum. *JMLR*, 25(22):1–56, 2024.

Roie Reshef and Kfir Yehuda Levy. Private and federated stochastic convex optimization: Efficient strategies for centralized systems. In *ICML*, 2024.

William Rogers and T. Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, 6, 05 1978.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. 2013.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2008.

Vladimir Vapnik and Alexey Chervonenkis. Theory of pattern recognition. *Nauka*, 1974.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. ISBN 978-1-108-41519-4.

Di Wang, Marco Gaboardi, Adam Smith, and Jinhui Xu. Empirical risk minimization in the non-interactive local model of differential privacy. *JMLR*, 21(200):1–39, 2020.

Yu-Xiang Wang, Jing Lei, and Stephen E Fienberg. Learning with differential privacy: Stability, learnability and the sufficiency and necessity of erm principle. *JMLR*, 17(183):1–40, 2016.

Cong Xie, Oluwasanmi Koyejo, and Indranil Gupta. Fall of empires: Breaking byzantine-tolerant sgd by inner product manipulation. In *UAI*, 2020.

Jiayuan Ye and Reza Shokri. Differentially private learning needs hidden state (or much faster convergence). In *NeurIPS*, 2022.

A MIA, DP AND STABILITY

We establish a formal connection between MIAs, differential privacy, and algorithmic stability.

Lemma A.1. *Consider the setting in Section 2 and the MIA framework described in Section 3. Let $\mathcal{A}$ be an algorithm that guarantees (ε, δ)-LDP via the Gaussian mechanism. For any $\varepsilon > 0$, the worst-case MIA success probability $\mathfrak{P}_T$ satisfies*

$$\mathfrak{P}_T = \Theta\left(\min\left\{1, \frac{\varepsilon}{\sqrt{\ln(1/\delta)}}\right\}\right). \tag{10}$$

Proof. As explainded in Section 3, by The Neyman-Pearson Lemma (Dong et al., 2022, Theorem A.1), $\mathfrak{P}_T = d_{\mathrm{TV}}(\mu_T, \mu'_T)$, where $\mu_t = \mathrm{Law}(V_{1:t}(\mathcal{S}))$ and $\mu'_t = \mathrm{Law}(V_{1:t}(\mathcal{S}'))$ denote the distributions of the transcript of communication between the server and honest participants up to step $t \in [T]$, as observed by the Byzantine participants when running $\mathcal{A}$ on datasets $\mathcal{S}$ and $\mathcal{S}'$, respectively.

That is, the transcripts $V_{1:T}(\mathcal{S})$ and $V_{1:T}(\mathcal{S}')$ are determined by the concatenated noisy gradient sequences $\tilde{G}_{\mathcal{S}} := (\tilde{g}_t^{(i)} = g_t^{(i)} + \xi_t^{(i)})_{i\in\mathcal{H}, t\in[T]}$ and $\tilde{G}_{\mathcal{S}'} := (\tilde{g}_t'^{(i)} = g_t'^{(i)} + \xi_t^{(i)})_{i\in\mathcal{H}, t\in[T]}$, respectively, with $\xi_t^{(i)} \sim \mathcal{N}(0, \sigma_{\mathrm{DP}}^2 I_d)$ independent noise vectors for each honest participants $i \in \mathcal{H}$ and each iteration $t \in [T]$. Indeed, as the server's broadcasted parameters are deterministic aggregations of the received gradients, no additional randomness is introduced. Thus, the statistical distance between the transcripts is the distance between the joint distributions of the independent noisy gradients.

Let $G_{\mathcal{S}}$ and $G_{\mathcal{S}'}$ denote the un-noised concatenated gradient sequences for $\mathcal{S}$ and $\mathcal{S}'$ respectively. Given the per-step gradient L_2-sensitivity $\Delta = 2C/m$, the T-step trajectory has an aggregate sensitivity of $\Delta_{\text{traj}} = \Delta\sqrt{T}$. Indeed, denoting the neighboring-dataset relation by $\mathcal{S} \sim \mathcal{S}'$, we have

$$\Delta_{\text{traj}} := \sup_{\mathcal{S}\sim\mathcal{S}'} \|G_{\mathcal{S}} - G_{\mathcal{S}'}\|_2 = \sup_{\mathcal{S}\sim\mathcal{S}'} \sqrt{\sum_{t=1}^{T}\sum_{i\in\mathcal{H}} \|g_t^{(i)} - g_t'^{(i)}\|_2^2} = \sqrt{\sum_{t=1}^{T} \Delta^2} = \Delta\sqrt{T}.$$

The Gaussian mechanism achieves (ε, δ)-DP by adding $\mathcal{N}(0, \sigma_{\mathrm{DP}}^2 I_d)$ noise at each step, where $\sigma_{\mathrm{DP}} = \frac{2C\sqrt{2T\ln(1/\delta)}}{m\varepsilon} \in \Theta\left(\frac{\Delta_{\text{traj}}\sqrt{\ln(1/\delta)}}{\varepsilon}\right)$. Consequently, by rotational invariance of the isotropic Gaussian distribution, we have

$$\mathfrak{P}_T = d_{\mathrm{TV}}(\mathcal{N}(0, \sigma_{\mathrm{DP}}^2), \mathcal{N}(\Delta_{\text{traj}}, \sigma_{\mathrm{DP}}^2)) = 2\Phi(x) - 1, \quad \text{where} \quad x := \frac{\Delta_{\text{traj}}}{2\sigma_{\mathrm{DP}}} \in \Theta\left(\frac{\varepsilon}{\sqrt{\ln(1/\delta)}}\right).$$

We now bound $2\Phi(x) - 1$ via its integral definition. For any $x > 0$,

$$2\Phi(x) - 1 = \sqrt{\frac{2}{\pi}} \int_0^x e^{-\frac{t^2}{2}}\, dt.$$

For $x \in (0, 1]$, the integrand is bounded by $e^{-1/2} \le e^{-t^2/2} \le 1$. Integrating this inequality yields

$$\sqrt{\frac{2}{\pi e}}x \le 2\Phi(x) - 1 \le \sqrt{\frac{2}{\pi}}x, \quad \text{proving} \quad \mathfrak{P}_T \in \Theta(x).$$

For $x > 1$, since Φ is strictly increasing, we have

$$2\Phi(1) - 1 \le 2\Phi(x) - 1 < 1 \quad \text{proving} \quad \mathfrak{P}_T \in \Theta(1). \qquad \square$$

B REGULARITY ASSUMPTIONS AND EXPANSIVITY RESULTS

To analyze the stability of iterative optimization algorithms, we typically rely on regularity assumptions on the loss function. We begin by stating standard regularity assumptions (Hardt et al., 2016) used throughout the paper, along with classical results that follow from them.

Definition B.1 (Regularity assumptions)**.** *Let $\ell : \Theta \to \mathbb{R}$ differentiable, $\Theta \subset \mathbb{R}^d$.*

(i) *ℓ is $C-$Lipschitz continuous (or $C-$Lipschitz) if there exists $C > 0$ such that,*

$$\forall u, v \in \Theta, \quad |\ell(u) - \ell(v)| \leq C\|u - v\|_2.$$

This property is equivalent to the norm of the gradient of ℓ being uniformly bounded by C.

(ii) *ℓ is $L-$Lipschitz smooth (or $L-$smooth) if its gradient is L-Lipschitz. This is equivalent to the following smoothness inequality:*

$$\forall u, v \in \Theta, \quad \ell(u) \leq \ell(v) + \langle \nabla\ell(v), u - v\rangle + \frac{L}{2}\|u - v\|_2^2. \tag{11}$$

(iii) *ℓ is convex if and only if, $\forall u, v \in \Theta$, $\langle \nabla\ell(u) - \nabla\ell(v), u - v\rangle \geq 0$.*

Moreover it is $\mu-$strongly convex if there exists $\mu > 0$ such that,

$$\forall u, v \in \Theta, \quad \ell(u) \geq \ell(v) + \langle \nabla\ell(v), u - v\rangle + \frac{\mu}{2}\|u - v\|_2^2.$$

We note that for a strongly convex function to have bounded gradients, it must be defined on a convex compact set, which then implies boundedness of both the loss and the gradient. To ensure this, we can either penalize the problem or restrict the parameter domain and apply an Euclidean projection at every step. Throughout the paper, we will tacitly assume this when referring to strongly convex functions, which does not limit the applicability of our analysis since the projection does not increase the distance between projected points.

Notably, convex and $L-$smooth functions satisfy an important property known as co-coercivity (see Proposition 5.4 in Bach, 2024).

Lemma B.1. *Let $\ell : \Theta \to \mathbb{R}$ differentiable, $\Theta \subset \mathbb{R}^d$. If ℓ is a convex and $L-$smooth function, then it satisfies the co-coercivity inequality,*

$$\forall u, v \in \Theta, \quad \langle \nabla\ell(u) - \nabla\ell(v), u - v\rangle \geq \frac{1}{L}\|\nabla\ell(u) - \nabla\ell(v)\|_2^2. \tag{12}$$

We focus on robust variants of distributed gradient descent and stochastic gradient descent. Below we recall properties of the one-step update. Let $\mathcal{A} \in \{\mathrm{GD}, \mathrm{SGD}\}$, $\ell : \Theta \to \mathbb{R}$ differentiable and $\gamma > 0$, we denote:

$$G_\gamma^{\mathcal{A}}(\theta) := \theta - \gamma\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} g^{(i)}.$$

Here, for each $i \in \mathcal{H}$, $g^{(i)}$ denotes the exact gradient $\nabla\ell(\theta)$ in the case of GD, or an unbiased estimate thereof—effectively computed with a sample from the local dataset—in the case of SGD. We recall the expansiveness definition for an update rule from Hardt et al. (2016, Definition 2.3).

Definition B.2. *An update rule $G : \Theta \to \Theta$ is said to be $\eta-$expansive if for all $\theta, \omega \in \Theta$,*

$$\|G(\theta) - G(\omega)\|_2 \leq \eta\|\theta - \omega\|_2.$$

A straightforward extension of Hardt et al. (2016, Lemma 3.6) to the distributed setting yields:

Lemma B.2. *Let $\mathcal{A} \in \{\mathrm{GD}, \mathrm{SGD}\}$ and $\gamma > 0$. Assume $\forall z \in \mathcal{Z}, \ell(\cdot, z) : \Theta \to \mathbb{R}$ is L-smooth, Then $G_\gamma^{\mathcal{A}}$ is $\eta_{G_\gamma^{\mathcal{A}}}-$expansive, with the following expansive coefficient,*

(i) *$\eta_{G_\gamma^{\mathcal{A}}} = (1 + \gamma L)$.*

(ii) *Assume in addition that ℓ is convex. Then, for any $\gamma \leq \frac{2}{L}$, $\eta_{G_\gamma^{\mathcal{A}}} = 1$.*

(iii) *Assume in addition that ℓ is μ-strongly convex. Then, for any $\gamma \leq \frac{2}{\mu+L}$, $\eta_{G_\gamma^{\mathcal{A}}} = (1 - \gamma\frac{L\mu}{\mu+L})$. For $\gamma \leq \frac{1}{L}$, since we have $L \geq \mu$, we can simplify the contractive constant to $\eta_{G_\gamma^{\mathcal{A}}} = (1 - \gamma\mu)$.*

$\eta_{G_\gamma^{\mathcal{A}}}$ is referred to as the expansivity coefficient of $\mathcal{A}$ with learning rate γ.

C Deferred proofs from section 3

In what follows, we first provide supporting lemmas, to then prove Theorems 3.1 and 3.2.

C.1 Supporting lemmas

For clarity and simplicity, we chose to emphasize our results under the broad bounded gradient assumption (i.e. Lipschitz-continuity of the loss function) in the main text. In this section, we present results derived under refined assumptions—namely, bounded heterogeneity and bounded variance—in place of the more general bounded gradient condition. These assumptions are more commonly used in the optimization literature (as opposed to the generalization literature).

We formally define these assumptions below.

Assumption C.1 (Bounded heterogeneity)**.** *There exists $G < \infty$ such that for all $\theta \in \Theta$: $\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \|\nabla \widehat{R}_i(\theta) - \nabla \widehat{R}_{\mathcal{H}}(\theta)\|_2^2 \leq G^2$.*

Assumption C.2 (Bounded variance)**.** *There exists $\sigma < \infty$ such that for each honest participant $i \in \mathcal{H}$ and all $\theta \in \Theta$: $\frac{1}{m}\sum_{x\in\mathcal{D}_i} \|\nabla \ell(\theta; x) - \nabla \widehat{R}_i(\theta)\|_2^2 \leq \sigma^2$.*

As shown in the following, our analysis techniques can yield tighter bounds when assuming bounded heterogeneity and bounded variance. However, we note that these bounds do not provide additional conceptual insights within the scope of our discussion. Indeed, within the uniform stability framework, if we assume only bounded gradients and nothing more, there exist distributed learning settings where G is a constant multiple of the Lipschitz constant C. For instance, consider a scenario where half of the honest participants produce gradients with norm C, and the other half produce gradients with norm $-C$; in this case, G would be C.

We will need an adaptation of the Lemma D.1 from Allouah et al. (2023b) that use a concentration argument on the empirical covariance matrix of Gaussian random variables:

Lemma C.1. *Consider the setting described in Section 2. Suppose Assumptions C.1, C.2. Then, for every $t \in \{1, \ldots, T-1\}$ we have*

$$\mathbb{E}_{\mathcal{A}}[\lambda_{\max}(\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}(\widetilde{g}_t^{(i)} - \overline{\widetilde{g}}_t)(\widetilde{g}_t^{(i)} - \overline{\widetilde{g}}_t)^T)] \leq 2\sigma^2 + 72\sigma_{\mathrm{DP}}^2(1 + \frac{d}{n-f}) + G^2$$
$$\leq 3C^2 + 72\sigma_{\mathrm{DP}}^2(1 + \frac{d}{n-f})$$

We note that, in the absence of momentum, the upper bound remains time-independent.

Proof. The proof leverages the ability to evaluate the expectation of the controllable noise—specifically, the differential privacy noise and the noise arising from random sample selection—while conditioning on the uncontrollable noise introduced by adversarial participants. The bound being independent of the latter noise, we establish the result by taking the total expectation.

Let $t \in \{0, \ldots, T-1\}$, we recall:

$$\overline{\widetilde{g}}_t = \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \widetilde{g}_t^{(i)} = \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \nabla \ell(\theta_t, x^{(i, J_t^{(i)})}) + \xi_t^{(i)}$$

$$\text{with } \xi_t^{(i)} \sim \mathcal{N}(0, \sigma_{\mathrm{DP}}^2 I_d) \text{ and } J_t^{(i)} \sim \mathcal{U}(\{1, \ldots, m\})$$

and an alternative formulation of the maximal eigenvalue

$$\Delta_t = \lambda_{\max}(\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}(\widetilde{g}_t^{(i)} - \overline{\widetilde{g}}_t)(\widetilde{g}_t^{(i)} - \overline{\widetilde{g}}_t)^T) = \sup_{\|v\|_2\leq 1} \frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}} \langle v, \widetilde{g}_t^{(i)} - \overline{\widetilde{g}}_t\rangle^2$$

We decompose

$$\widetilde{g}_t^{(i)} - \overline{\widetilde{g}}_t = \widetilde{g}_t^{(i)} - \nabla \widehat{R}_i(\theta_t) + \nabla \widehat{R}_i(\theta_t) - \nabla \widehat{R}_{\mathcal{H}}(\theta_t) + \nabla \widehat{R}_{\mathcal{H}}(\theta_t) - \overline{\widetilde{g}}_t$$

so that we have

$$\langle v, \widetilde{g}_t^{(i)} - \overline{\widetilde{g}}_t\rangle^2 = \langle v, \nabla\widehat{R}_i(\theta_t) - \nabla\widehat{R}_{\mathcal{H}}(\theta_t)\rangle^2 + \langle v, \widetilde{g}_t^{(i)} - \nabla\widehat{R}_i(\theta_t) + \nabla\widehat{R}_{\mathcal{H}}(\theta_t) - \overline{\widetilde{g}}_t\rangle^2 + 2\langle v, \widetilde{g}_t^{(i)} - \nabla\widehat{R}_i(\theta_t) + \nabla\widehat{R}_{\mathcal{H}}(\theta_t) - \overline{\widetilde{g}}_t\rangle\langle v, \nabla\widehat{R}_i(\theta_t) - \nabla\widehat{R}_{\mathcal{H}}(\theta_t)\rangle$$

Upon averaging, taking the supremum over the unit ball, and then taking total expectations, we get:

$$\mathbb{E}\left[\sup_{\|v\|_2\le 1}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\langle v, \widetilde{g}_t^{(i)} - \overline{\widetilde{g}}_t\rangle^2\right] = \mathbb{E}\left[\sup_{\|v\|_2\le 1}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\langle v, \nabla\widehat{R}_i(\theta_t) - \nabla\widehat{R}_{\mathcal{H}}(\theta_t)\rangle^2\right] + \mathbb{E}\left[\sup_{\|v\|_2\le 1}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\langle v, \widetilde{g}_t^{(i)} - \nabla\widehat{R}_i(\theta_t) + \nabla\widehat{R}_{\mathcal{H}}(\theta_t) - \overline{\widetilde{g}}_t\rangle^2\right] + 2\mathbb{E}\left[\sup_{\|v\|_2\le 1}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\langle v, \widetilde{g}_t^{(i)} - \nabla\widehat{R}_i(\theta_t) + \nabla\widehat{R}_{\mathcal{H}}(\theta_t) - \overline{\widetilde{g}}_t\rangle\langle v, \nabla\widehat{R}_i(\theta_t) - \nabla\widehat{R}_{\mathcal{H}}(\theta_t)\rangle\right] \quad (13)$$

We show that the last term in eq. (13) is non-positive. In fact, with $M = N + N^T$,

$$N = \sum_{i\in\mathcal{H}}(\widetilde{g}_t^{(i)} - \nabla\widehat{R}_i(\theta_t) + \nabla\widehat{R}_{\mathcal{H}}(\theta_t) - \overline{\widetilde{g}}_t)(\nabla\widehat{R}_i(\theta_t) - \nabla\widehat{R}_{\mathcal{H}}(\theta_t))^T$$

and using Lemma D.4 from Allouah et al. (2023b):

$$\mathbb{E}\left[\sup_{\|v\|_2\le 1} 2\sum_{i\in\mathcal{H}}\langle v, \widetilde{g}_t^{(i)} - \nabla\widehat{R}_i(\theta_t) + \nabla\widehat{R}_{\mathcal{H}}(\theta_t) - \overline{\widetilde{g}}_t\rangle\langle v, \nabla\widehat{R}_i(\theta_t) - \nabla\widehat{R}_{\mathcal{H}}(\theta_t)\rangle\right] = \mathbb{E}\left[\sup_{\|v\|_2\le 1}\langle v, Mv\rangle\right] \le 9^d \sup_{\|v\|_2\le 1}\mathbb{E}\left[\langle v, Mv\rangle\right]$$

$$= 9^d \sup_{\|v\|_2\le 1}\mathbb{E}\left[2\sum_{i\in\mathcal{H}}\langle v, \underbrace{\mathbb{E}[\widetilde{g}_t^{(i)} - \nabla\widehat{R}_i(\theta_t) + \nabla\widehat{R}_{\mathcal{H}}(\theta_t) - \overline{\widetilde{g}}_t|\theta_t]}_{=0}\rangle\langle v, \nabla\widehat{R}_i(\theta_t) - \nabla\widehat{R}_{\mathcal{H}}(\theta_t)\rangle\right]$$

We now bound the second term in eq. (13):

$$\mathbb{E}\left[\sup_{\|v\|_2\le 1}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\langle v, \widetilde{g}_t^{(i)} - \nabla\widehat{R}_i(\theta_t) + \nabla\widehat{R}_{\mathcal{H}}(\theta_t) - \overline{\widetilde{g}}_t\rangle^2\right] \underset{2ab\le a^2+b^2}{\le} 2\mathbb{E}\left[\sup_{\|v\|_2\le 1}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\langle v, g_t^{(i)} - \nabla\widehat{R}_i(\theta_t) + \nabla\widehat{R}_{\mathcal{H}}(\theta_t) - \overline{g}_t\rangle^2 + \langle v, \xi_t^{(i)} - \overline{\xi}_t\rangle^2\right] \underset{\substack{\text{bias-variance}\\ \text{decomposition}}}{\le} 2\mathbb{E}\left[\sup_{\|v\|_2\le 1}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\langle v, g_t^{(i)} - \nabla\widehat{R}_i(\theta_t)\rangle^2 + \langle v, \xi_t^{(i)}\rangle^2\right]$$

$$\le 2\underbrace{\mathbb{E}\left[\sup_{\|v\|_2\le 1}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\langle v, \xi_t^{(i)}\rangle^2\right]}_{\le 36\sigma_{\mathrm{DP}}^2(1+\frac{d}{n-f})\text{ Lemma D.6 from Allouah et al. (2023b)}} + 2\mathbb{E}\left[\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\underbrace{\mathbb{E}_{J_t^{(i)}}[\|g_t^{(i)} - \nabla\widehat{R}_i(\theta_t)\|_2^2]}_{\le\sigma^2}\right]$$

And finally we bound the first term in eq. (13) with assumption C.1:

$$\mathbb{E}\left[\sup_{\|v\|_2\le 1}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\langle v, \nabla\widehat{R}_i(\theta_t) - \nabla\widehat{R}_{\mathcal{H}}(\theta_t)\rangle^2\right] \le \sup_{\theta\in\Theta,\|v\|_2\le 1}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\langle v, \nabla\widehat{R}_i(\theta_t) - \nabla\widehat{R}_{\mathcal{H}}(\theta_t)\rangle^2 \le \sup_{\theta\in\Theta}\frac{1}{|\mathcal{H}|}\sum_{i\in\mathcal{H}}\|\nabla\widehat{R}_i(\theta_t) - \nabla\widehat{R}_{\mathcal{H}}(\theta_t)\|_2^2 \le G^2$$

□

C.2 Proof of Theorem 3.2

Theorem 3.2. *Consider the setting described in Section 2. Let $\sigma_{\mathrm{DP}} > 0$ and $\gamma \leq \frac{1}{L}$. Suppose $\forall z \in \mathcal{Z}$, $\ell(\cdot; z)$ C-Lipschitz and L-smooth. Then $\mathcal{A}$ is $(\varepsilon, \delta) - \mathrm{DP}$ and,*

(i) If $\ell(\cdot; z)$ is convex $\forall z \in \mathcal{Z}$, then the uniform stability of $\mathcal{A}$ is upper bounded by

$$2\gamma C^2 T \left(\frac{1}{(n-f)m} + \sqrt{3\kappa} + 2\sqrt{18}\sqrt{\kappa}\frac{\sigma_{\mathrm{DP}}}{C}\sqrt{1 + \frac{d}{n-f}} \right). \tag{14}$$

(ii) If $\ell(\cdot; z)$ is μ-strongly convex $\forall z \in \mathcal{Z}$, then the uniform stability of $\mathcal{A}$ is upper bounded by

$$\frac{2C^2}{\mu} \left(\frac{1}{(n-f)m} + \sqrt{3\kappa} + 2\sqrt{18}\sqrt{\kappa}\frac{\sigma_{\mathrm{DP}}}{C}\sqrt{1 + \frac{d}{n-f}} \right). \tag{15}$$

If we further assume the loss function to be bounded by ℓ_∞, then the uniform stability of $\mathcal{A}$ is upper bounded by the minimum between (6)*, respectively* (7)*, and $2\ell_\infty \mathfrak{P}_T$.*

Proof. The proof follows exactly all the arguments of Boudou et al. (2025, Theorem 3.1), the only difference is the use of Lemma C.1.

For the upper bound, it is due to the following. Let $\mu_S := \mathrm{Law}(\mathcal{A}(\mathcal{S}))$ and $\mu_{S'} := \mathrm{Law}(\mathcal{A}(\mathcal{S}'))$, then

$$\begin{aligned} |\mathbb{E}\left[\ell(\mathcal{A}(\mathcal{S}), z) - \ell(\mathcal{A}(\mathcal{S}'), z)\right]| &= \left| \int_{\mathbb{R}^d} \ell(w, z) d\mu_S(w) - \int_{\mathbb{R}^d} \ell(w, z) d\mu_{S'}(w) \right| \\ &\leq \int_{\mathbb{R}^d} |\ell(w, z)| |d\mu_S(w) - d\mu_{S'}(w)| \\ &\leq 2\ell_\infty d_{\mathrm{TV}}(\mu_S, \mu_{S'}) = 2\ell_\infty \mathfrak{P}_T. \end{aligned}$$

□

C.3 Proof of Theorem 3.1

In this appendix, we present the proof of Theorem 3.1 first by focusing on the case $d = 1$, as the extension to $d > 1$ is more technical yet follows an identical proof structure, making the $d = 1$ case sufficient to convey the fundamental insight and construction. We then provide the proof of Theorem 3.1 fot the case $d > 1$.

C.3.1 Proof for $d = 1$.

We first establish a variance concentration inequality, which serves as a technical prerequisite To our main stability lower bound.

Lemma C.2. *Let $1 \leq f < \frac{n}{2}$. Let $\mathcal{H} = \{1, \ldots, n-f\}$, where for each $i \in \mathcal{H}$, $g^{(i)} \sim \mathcal{N}(0, \sigma_{\mathrm{DP}}^2)$. Let $\mathcal{B} = \{n-f+1, \ldots, n\}$, where for each $i \in \mathcal{B}$, $g^{(i)} = \beta\sigma_{\mathrm{DP}}$, with $\beta := c\sqrt{\frac{(n-f-1)(n-f)}{f(n-2f)}}$. Let $c, c' \in (0, 1)$ be strictly positive constants satisfying $c^2 + c'^2 \leq \frac{f}{n-f}$. Let $\tilde{S}$ be a fixed subset of size $n-f$ consisting of $n-2f$ indices from $\mathcal{H}$ and all f indices from $\mathcal{B}$. Let $s := \frac{c'^2}{2}\mathbb{E}\left[\mathrm{Var}(\{g^{(i)}\}_{i\in\mathcal{H}})\right]$. Then, the following bounds hold.*

(a) Expectation bound: $\mathbb{E}\left[\mathrm{Var}(\{g^{(i)}\}_{i\in\tilde{S}})\right] \leq (1 - c'^2)\mathbb{E}\left[\mathrm{Var}(\{g^{(i)}\}_{i\in\mathcal{H}})\right]$.

(b) Lower tail bound: $\mathbb{P}\left(\mathrm{Var}(\{g^{(i)}\}_{i\in\mathcal{H}}) - \mathbb{E}\left[\mathrm{Var}(\{g^{(i)}\}_{i\in\mathcal{H}})\right] \leq -s\right) \leq e^{-\frac{c'^4}{16}(n-f-1)}$.

(c) Upper tail bound: Let $c_1 > 0$ be the absolute constant from the Hanson-Wright inequality,

$$\mathbb{P}\left(\mathrm{Var}(\{g^{(i)}\}_{i\in\tilde{S}}) - \mathbb{E}\left[\mathrm{Var}(\{g^{(i)}\}_{i\in\tilde{S}})\right] > s\right) \leq e^{-\frac{c_1 c'^2}{4}(n-f-1)\min\left\{\frac{c'^2(n-f-1)}{4(n-2f)}, 1\right\}} + e^{-\frac{c'^4(n-f)(n-f-1)}{128fc^2}}.$$

Proof. (a) We use the following variance identity to compute the expectation

$$\mathbb{E}\left[\mathrm{Var}(\{g^{(i)}\}_{i\in\tilde{S}})\right] = \mathbb{E}\left[\frac{1}{n-f}\sum_{i\in\tilde{S}}\left(g^{(i)}-\bar{g}_{\tilde{S}}\right)^2\right]$$
$$= \frac{1}{n-f}\mathbb{E}\left[\sum_{i\in\tilde{S}} {g^{(i)}}^2\right] - \mathbb{E}\left[\bar{g}_{\tilde{S}}^2\right] = \frac{1}{n-f}\mathbb{E}\left[\sum_{i\in\tilde{S}} {g^{(i)}}^2\right] - \mathrm{Var}(\bar{g}_{\tilde{S}}) - \mathbb{E}\left[\bar{g}_{\tilde{S}}\right]^2.$$

We compute the three terms separately. For the first one,

$$\mathbb{E}\left[\sum_{i\in\tilde{S}} {g^{(i)}}^2\right] = (n-2f)\sigma_{\mathrm{DP}}^2 + f\beta^2\sigma_{\mathrm{DP}}^2 = \sigma_{\mathrm{DP}}^2\left(n-2f+f\beta^2\right).$$

Moreover, $\mathbb{E}\left[\bar{g}_{\tilde{S}}\right] = \frac{f}{n-f}\beta\sigma_{\mathrm{DP}}$, and by independence,

$$\mathrm{Var}(\bar{g}_{\tilde{S}}) = \frac{1}{(n-f)^2}\left((n-2f)\,\mathrm{Var}(\mathcal{N}(0,\sigma_{\mathrm{DP}}^2)) + \mathrm{Var}(f\beta\sigma_{\mathrm{DP}})\right) = \frac{n-2f}{(n-f)^2}\sigma_{\mathrm{DP}}^2.$$

Rearranging the terms, with $\frac{n-2f}{n-f} - \frac{n-2f}{(n-f)^2} = \frac{(n-2f)(n-f-1)}{(n-f)^2}$ and $\frac{f}{n-f} - \frac{f^2}{(n-f)^2} = \frac{f(n-2f)}{(n-f)^2}$,

$$\mathbb{E}\left[\mathrm{Var}(\{g^{(i)}\}_{i\in\tilde{S}})\right] = \sigma_{\mathrm{DP}}^2\frac{n-f-1}{n-f}\left[\frac{(n-2f)}{n-f} + \beta^2\frac{f(n-2f)}{(n-f)(n-f-1)}\right]$$
$$= \sigma_{\mathrm{DP}}^2\frac{n-f-1}{n-f}\left[\frac{n-2f}{n-f} + c^2\right] = \sigma_{\mathrm{DP}}^2\frac{n-f-1}{n-f}\left[1-\left(\frac{f}{n-f}-c^2\right)\right]$$
$$\leq (1-c'^2)\sigma_{\mathrm{DP}}^2\frac{n-f-1}{n-f},$$

where the final inequality holds by assumption $\frac{f}{n-f} - c^2 \geq c'^2$.

(b) Let $g^{(\mathcal{H})} := (g^{(i)})_{i\in\mathcal{H}}$. By Cochran's Theorem (Cochran, 1934),

$$Y := \frac{1}{\sigma_{\mathrm{DP}}^2}\sum_{i\in\mathcal{H}}\left(g^{(i)}-\bar{g}\right)^2 = \frac{1}{\sigma_{\mathrm{DP}}^2}{g^{(\mathcal{H})}}^\intercal P_{n-f} g^{(\mathcal{H})} = \sum_{i=1}^{n-f-1} {Y^{(i)}}^2 \sim \chi^2_{n-f-1},$$

where $\forall i \in \{1,\dots,n-f-1\}, Y^{(i)} \sim \mathcal{N}(0,1)$, χ^2_{n-f-1} is the chi-squared distribution with $n-f-1$ degrees of freedom, $P_{n-f} = I_{n-f} - \frac{1}{n-f}1_{n-f}1_{n-f}^\intercal$ is the orthogonal projection matrix onto the $n-f-1$-dimensional subspace of vectors whose components sum to zero, I_{n-f} is the identity matrix of size $(n-f)\times(n-f)$, and 1_{n-f} the vector of size $n-f$ with every entry equal to 1. From Laurent & Massart (2000, Lemma 1), for any $u \geq 0$,

$$\mathbb{P}\left(Y \leq n-f-1-u\right) \leq e^{-\frac{u^2}{4(n-f-1)}}.$$

Because $\mathrm{Var}(\{g^{(i)}\}_{i\in\mathcal{H}}) = \frac{\sigma_{\mathrm{DP}}^2}{n-f}Y$, substituting $u = \frac{c'^2}{2}(n-f-1)$ equates to evaluating the tail at $s = \frac{c'^2}{2}\frac{n-f-1}{n-f}\sigma_{\mathrm{DP}}^2$, which yields

$$\mathbb{P}\left(\mathrm{Var}(\{g^{(i)}\}_{i\in\mathcal{H}}) - \mathbb{E}\left[\mathrm{Var}(\{g^{(i)}\}_{i\in\mathcal{H}})\right] \leq -s\right) = \mathbb{P}\left(Y \leq n-f-1-u\right) \leq e^{-\frac{c'^4}{16}(n-f-1)}.$$

(c) Let $g := \begin{pmatrix} h := (g^{(i)})_{i\in\mathcal{H}\cap\tilde{S}} \\ b := \beta\sigma_{\mathrm{DP}}1_f \end{pmatrix}$, we have

$$\mathrm{Var}(\{g^{(i)}\}_{i\in\tilde{S}}) = \frac{1}{n-f}g^\intercal P_{n-f} g = \underbrace{\frac{1}{n-f}h^\intercal\left(I_{n-2f} - \frac{1}{n-f}1_{n-2f}1_{n-2f}^\intercal\right)h}_{V_1}$$
$$\underbrace{-\frac{2}{(n-f)^2}h^\intercal 1_{n-2f}1_f^\intercal b}_{V_2} + \underbrace{\frac{1}{n-f}b^\intercal\left(I_f - \frac{1}{n-f}1_f1_f^\intercal\right)b}_{V_3}$$

$$= V_1 - \frac{2f\beta\sigma_{\mathrm{DP}}}{(n-f)^2}\sum_{i\in\mathcal{H}\cap\tilde{S}} g^{(i)} + \frac{f(n-2f)}{(n-f)^2}\beta^2{\sigma_{\mathrm{DP}}}^2.$$

Concentration of V_1 is bounded using the Hanson-Wright inequality (Rudelson & Vershynin, 2013, Theorem 1.1). With $M = I_{n-2f} - \frac{1}{n-f}1_{n-2f}1_{n-2f}^{\intercal}$, we have $\|M\|_{\mathrm{F}}^2 \leq n-2f$ and $\|M\|_{\mathrm{sp}} = 1$. Thus, for a constant $c_1 > 0$,

$$p_1(s) := \mathbb{P}\left(V_1 > \mathbb{E}\left[V_1\right] + s\right) \leq e^{-c_1 \min\left\{\frac{s^2(n-f)^2}{\sigma_{\mathrm{DP}}^4(n-2f)}, \frac{s(n-f)}{\sigma_{\mathrm{DP}}^2}\right\}}.$$

For V_2, the sum is normally distributed,

$$\frac{2f\beta\sigma_{\mathrm{DP}}}{(n-f)^2}\sum_{i\in\mathcal{H}\cap\tilde{S}} g^{(i)} \sim \mathcal{N}\left(0, \left(\frac{2f\beta\sigma_{\mathrm{DP}}}{(n-f)^2}\right)^2 (n-2f)\sigma_{\mathrm{DP}}^2\right)$$

$$p_2(s) := \mathbb{P}\left(V_2 < -s\right) \leq e^{-\frac{s^2(n-f)^4}{8f^2\beta^2(n-2f)\sigma_{\mathrm{DP}}^4}} = e^{-\frac{s^2(n-f)^3}{8fc^2(n-f-1)\sigma_{\mathrm{DP}}^4}}.$$

Finally, V_3 is perfectly concentrated as the variables in $\mathcal{B}$ are constants.

Evaluating the tails at $\frac{s}{2} = \frac{c'^2}{4}\frac{n-f-1}{n-f}\sigma_{\mathrm{DP}}^2$ and union bounding yields $\mathbb{P}\left(E_3^{\varsigma}\right) \leq p_1(\frac{s}{2}) + p_2(\frac{s}{2})$, evaluating to the stated bound. □

Armed with this concentration result, we now establish lower bounds on algorithmic stability.

Theorem 3.1. *Consider the setting in Section 2 with full batch and the* SMEA *aggregation rule. Suppose* $d = 1$, $T \in \mathcal{O}\big(e^{c_{\min}(n-f-1)}\big)$, *where* $c_{\min} > 0$, *and* $\sigma_{\mathrm{DP}} = \frac{2C\sqrt{2T\ln(1/\delta)}}{m\varepsilon}$. *Let* $\mathfrak{P}_T \in \Theta\big(\min\big\{1, \frac{\varepsilon}{\sqrt{\ln(1/\delta)}}\big\}\big)$ *as described in Section 3. Then there exists* $\ell \in \Theta \times \mathcal{Z} \to \mathbb{R}$ *such that* $\forall z \in \mathcal{Z}, \ell(\cdot; z)$ *is* C*-Lipschitz,* L*-smooth and convex, and the uniform stability of* $\mathcal{A}$ *is lower bounded by*

$$\Omega\left(\mathfrak{P}_T \times \gamma C^2 T\left(\frac{1}{(n-f)m} + \frac{f}{n-2f} + \frac{\sigma_{\mathrm{DP}}}{C}\sqrt{\frac{f}{n-2f}}\right)\right). \tag{16}$$

If we moreover assume $\frac{n}{3} < f < \frac{n}{2}$, *the uniform stability of* $\mathcal{A}$ *is lower bounded by*

$$\Omega\left(\mathfrak{P}_T \times \gamma C^2 T\left(\frac{1}{(n-f)m} + \sqrt{\frac{f}{n-2f}} + \sqrt{\frac{f}{n-2f}}\frac{\sigma_{\mathrm{DP}}}{C}\right)\right). \tag{17}$$

For both cases, if we assume the loss function bounded by ℓ_∞[3] *and* $T \in \mathcal{O}\big(\mathfrak{P}_T^2 e^{c_{\min}\left(\sqrt{d}+\sqrt{n-f-1}\right)^2}\big)$, *then there exists* $\sigma_{threshold} > 0$ *such that for* $\sigma_{DP} \geq \sigma_{threshold}$, *the uniform stability of* $\mathcal{A}$ *is lower bounded by the minimum between* (4) *(respectively* (5)*) and* $\Omega\left(\ell_\infty \mathfrak{P}_T\right)$.

Proof. Let $d = 1$, $n > 0$, $\mathcal{H} = \{1, \ldots, n-f\}$ and let $S, S' \in \mathcal{Z}^{(n-f)m}$ neighboring datasets where every sample equal 0, except for one sample in S' from the participant 1, equal to C. Let $\ell : (\theta, z) \to \ell(\theta; z) = \theta z$.[4] For $i \in \mathcal{H} := \{1, \ldots, n-f\}$, $0 \leq t \leq T$, let $\xi_t^{(i)} \sim \mathcal{N}(0, \sigma_{\mathrm{DP}}^2)$ be independent random gaussian noise.

Byzantine participants trigger. We analyze two attacker capabilities. First, if Byzantine participants have access to the random seeds of honest participants, they can circumvent the LDP guarantee. Indeed, this knowledge allows them to perfectly reconstruct the added noise, subtract it, and deterministically identify, at the first iteration, whether the first participant's true gradient was $\frac{C}{m}$ or 0. In the alternative scenario, where attackers only observe the final noisy gradient (lacking the seed), they cannot perform this deterministic inference. We must then rely on a statistical test to bound their success probability to distinguish the two cases. We explain a possible (optimal) test in what follows.

Let $p_{\text{test}} \in (0, 1)$. We define the behavior of Byzantine participants based on a trigger condition. This trigger is determined by performing a gaussian mean hypothesis test on the first $n_{\text{test}} := p_{\text{test}}T$

[3] This can be ensured by projecting parameters onto a compact set Θ after the update (Line 10 in Algorithm 1).

[4] We only require the gradient to be zero when $z = 0$ and C for the perturbed data.

sample, collected during an initial fraction of the total T iterations. The test evaluates whether the sample mean corresponds to hypothesis H_0 (mean 0, equivalently the dataset is S) or to hypothesis H_1 (mean $\frac{C}{m}$, equivalently the dataset is S'), assuming a known variance σ^2_{DP}. This test processes the statistic ζ which evaluates at $\frac{C}{m} + \bar{\xi}$ under H_1, and evaluates at $\bar{\xi}$ under H_0, where $\bar{\xi} = \frac{1}{n_{\text{test}}} \sum_{t=1}^{n_{\text{test}}} \xi_t^{(1)}$. Overall, the Byzantine participants undergo a membership inference attack to know whether the gradients from participant 1 equals $g_t^{(1)} = \xi_t^{(1)}$ or $g_t^{\prime(1)} = \frac{C}{m} + \xi_t^{(1)}$, i.e whether the dataset is S or S'. The decision is made depending on a trigger value c_{trigger},

$$\text{reject } H_0 \text{ if } \zeta > c_{\text{trigger}}, \text{ else accept } H_0.$$

We denote $Z \sim \mathcal{N}(0,1)$ a standard gaussian random variable, and $z_\alpha \in \mathbb{R}$ the value such that $\mathbb{P}(Z > z_\alpha) = \alpha$, $\alpha \in (0,1)$. Conditioned on H_0, $\zeta | H_0 \sim \mathcal{N}(0, \frac{\sigma^2_{\text{DP}}}{n_{\text{test}}})$, hence setting $c_{\text{trigger}} = z_\alpha \frac{\sigma_{\text{DP}}}{\sqrt{n_{\text{test}}}}$,

$$\mathbb{P}\left(\text{reject } H_0 | H_0\right) = \mathbb{P}\left(\zeta > c_{\text{trigger}} | H_0\right) = \mathbb{P}\left(Z > z_\alpha\right) = \alpha.$$

Let denote E_1 the event that the Byzantine participants successfully infer the underlying datasets from the gradient statistics (successful MIA). Under the optimal hypothesis test, we have

$$\begin{aligned} E_1 &= \left\{\bar{\xi} \text{ does not trigger } H_0 \text{ rejection}\right\} \cap \left\{\text{shifting } \bar{\xi} \text{ by} \frac{C}{m} \text{ triggers } H_0 \text{ rejection}\right\} \\ &= \left\{\bar{\xi} \leq c_{\text{trigger}}\right\} \cap \left\{\bar{\xi} + \frac{C}{m} > c_{\text{trigger}}\right\} \\ &= \left\{c_{\text{trigger}} - \frac{C}{m} < \bar{\xi} \leq c_{\text{trigger}}\right\}, \end{aligned}$$

and,

$$\mathbb{P}\left(E_1\right) = \mathbb{P}\left(z_\alpha - \frac{C\sqrt{n_{\text{test}}}}{m\sigma_{\text{DP}}} < Z \leq z_\alpha\right) = \mathbb{P}\left(z_\alpha - \frac{\varepsilon\sqrt{p_{\text{test}}}}{2\sqrt{2\ln(1/\delta)}} < Z \leq z_\alpha\right).$$

This means that $\mathbb{P}\left(E_1\right) \in \Omega(1)$ if and only if there exist a constant k such that $\varepsilon \geq k\sqrt{\ln(1/\delta)}$.

The event E_1 relates to the best false negative rate $\beta_\alpha = \inf_{\mathbb{P}(\text{reject } H_0 | H_0) \leq \alpha} \mathbb{P}(\text{accept } H_0 \mid H_1) = \mathbb{P}(Z - \frac{C\sqrt{n_{\text{test}}}}{m\,\sigma_{\text{DP}}} \leq z_\alpha) = \Phi(z_\alpha + \frac{\varepsilon\sqrt{p_{\text{test}}}}{2\sqrt{2\ln(1/\delta)}})$ (the Byzantine participants infer S when S') for a fixed false positive rate $\alpha = \mathbb{P}(\text{reject } H_0 \mid H_0) = \mathbb{P}(Z > z_\alpha) = 1 - \Phi(z_\alpha)$ (the Byzantine participants infer S' when S) through $\mathbb{P}(E_1) = 1 - \alpha - \beta_\alpha = \Phi(z_\alpha) - \Phi(z_\alpha + \frac{\varepsilon\sqrt{p_{\text{test}}}}{2\sqrt{2\ln(1/\delta)}})$. That is, $\mathbb{P}(E_1) = 1 - \mathbb{P}(\text{reject } H_0 \mid H_0) - \mathbb{P}(\text{accept } H_0 \mid H_1)$ quantifies the Byzantine participants maximum inference capability, representing the probability of MIA success beyond random chance (or MIA advantage).

Byzantine participants behavior. We assign the following behavior (Algorithm 2) to the Byzantine participants $i \in \{n - f + 1, n\}$, for $t \geq n_{\text{test}} + 1$.

Algorithm 2 Byzantine Participant Behavior

```
# β (defined below) depends on f and n
if ζ > c_trigger then
    return β σ_DP
else
    return −β σ_DP
```

That is, depending on the input ζ, which itself depends only on the observed noisy gradients, the Byzantine participants bias the trajectory toward either positive or negative values. In what follows, we decompose our analyze depending on whether the Byzantine participants have a successful MIA (i.e., infer whether the dataset is $\mathcal{S}$ or $\mathcal{S}'$) in the two optimization trajectories (E_1), or not ($E_1^\complement$).

(i) Under E_1 and the trajectory produced by S' (i.e., with $g_t^{(1)\prime} = \frac{1}{m}\sum_{j=1}^{m} z^{(1,j)\prime} + \xi_t^{(1)} = \frac{C}{m} + \xi_t^{(1)}$), the Byzantine participants favor the positive noises. This can be established by computations

analogous to those carried out in Boudou et al. (2025, Theorem 3.2). Following this reasoning, we set the Byzantine values consistently with Algorithm 2 to be

$$\forall i \in \{n-f+1,\dots,n\}, t \in \{n_{\text{test}}+1,\dots,T\}, g_t^{(i)} := \beta\sigma_{\text{DP}},$$

where

$$\beta := c\sqrt{\frac{(n-f-1)(n-f)}{f(n-2f)}},$$

for constants $c, c' \in (0,1)$ satisfying $c^2 + c'^2 \leq \frac{f}{n-f}$. We then show that, with high probability,

$$\text{for } t \in \{n_{\text{test}}+1,\dots,T\}, \text{Var}(\{g_t^{(i)}\}_{i\in S_t}) < \text{Var}(\{g_t^{(i)}\}_{i\in\mathcal{H}}),$$

where $S_t' = \{n-f+1,\dots,n\} \cup \text{Select}_{n-2f}(\mathcal{H})$. Here, $\text{Select}_{n-2f}(\mathcal{H})$ denotes the subset of $n-2f$ honest participants whose gradients, when combined with the Byzantine gradients, minimize the variance of S_t'. Consequently, with high probability, $\text{SMEA}(g_t^{(1)},\dots,g_t^{(n)}) = \frac{1}{n-f}\sum_{i\in S_t'} g_t^{(i)}$. This is made possible because of the following concentration arguments.

Let $\tilde{S}$ be a subset of size $n-f$ containing the f Byzantine participants and $n-2f$ honest participants. Because there are $n-f$ honest participants in total, and $n-f \geq n-2f+1$ (assuming $f \geq 1$), we can specifically construct $\tilde{S}$ such that $1 \notin \tilde{S}$. By definition of SMEA, S_t' minimizes the variance of such subsets, ensuring that $\text{Var}(\{g_t'^{(i)}\}_{i\in S_t'}) \leq \text{Var}(\{g_t'^{(i)}\}_{i\in\tilde{S}})$. Furthermore, since $1 \notin \tilde{S}$, we have $g_t'^{(i)} = g_t^{(i)} = \xi_t^{(i)}$ for $i \in \tilde{S} \setminus \{n-f+1,\dots,n\}$. Because $\xi_t^{(i)}$ are independent, centered, and symmetric Gaussians, and $g_t'^{(i)} = \beta\sigma_{\text{DP}} = -g_t^{(i)}$ for $i \in \{n-f+1,\dots,n\}$, the variables $\text{Var}(\{g_t'^{(i)}\}_{i\in\tilde{S}})$ and $\text{Var}(\{g_t^{(i)}\}_{i\in\tilde{S}})$ follow the same probability distribution. Thus, their expectations and tail bounds are identical.

Let $s := \frac{c'^2}{2}\mathbb{E}\left[\text{Var}(\{g_t^{(i)}\}_{i\in\mathcal{H}})\right] = \frac{c'^2}{2}\frac{n-f-1}{n-f}\sigma_{\text{DP}}^2$. Note that,

$$\mathbb{E}\big[\text{Var}(\{g_t'^{(i)}\}_{i\in\mathcal{H}})\big] = \frac{n-f-1}{n-f}\sigma_{\text{DP}}^2 + \frac{n-f-1}{(n-f)^2}\frac{C^2}{m^2} > \mathbb{E}\big[\text{Var}(\{g_t^{(i)}\}_{i\in\mathcal{H}})\big].$$

For any $t > n_{\text{test}}$,

(a) We show $\mathbb{E}\left[\text{Var}(\{g_t^{(i)}\}_{i\in\tilde{S}})\right] \leq (1-c'^2)\mathbb{E}\left[\text{Var}(\{g_t^{(i)}\}_{i\in\mathcal{H}})\right]$.

(b) We control the lower tail probability of the event

$$E_{2,t}'^{\mathsf{c}} := \big\{\text{Var}(\{g_t'^{(i)}\}_{i\in\mathcal{H}}) - \mathbb{E}\left[\text{Var}(\{g_t^{(i)}\}_{i\in\mathcal{H}})\right] \leq -s\big\}.$$

We prove that the probability of the variance collapsing below a threshold is lower under the dataset shift. Because the $n-f$ participants are drawn identically from $\mathcal{N}(0,\sigma_{\text{DP}}^2)$, the scaled variance follows a central chi-squared distribution: $\frac{(n-f-1)\,\text{Var}(\{g_t^{(i)}\}_{i\in\mathcal{H}})}{\sigma_{\text{DP}}^2} \sim \chi^2_{n-f-1}$. Under the shifted trajectory, one honest element is deterministically shifted by $\frac{C}{m}$. The shifted sample variance $\text{Var}(\{g_t'^{(i)}\}_{i\in\mathcal{H}})$ therefore follows a non-central chi-squared distribution with a positive non-centrality parameter, $\frac{(n-f-1)\,\text{Var}(\{g_t'^{(i)}\}_{i\in\mathcal{H}})}{\sigma_{\text{DP}}^2} \sim \chi^2_{n-f-1}(\frac{C^2}{m^2})$.

By Muirhead (1982, Corollary 1.3.5), the cumulative distribution function of a non-central chi-squared variable evaluates to a Poisson mixture of central chi-squared CDFs. Let F_d denote the CDF of a central χ_d^2 variable. For any threshold $x > 0$,

$$\mathbb{P}\left(\frac{(n-f-1)\,\text{Var}(\{g_t'^{(i)}\}_{i\in\mathcal{H}})}{\sigma_{\text{DP}}^2} \leq x\right) = \sum_{j=0}^{\infty} e^{-\lambda/2}\frac{(\lambda/2)^j}{j!}F_{n-f-1+2j}(x).$$

A central χ_d^2 random variable can be defined as the sum of d independent squared standard normal variables. Increasing the degrees of freedom adds positive random variables to the

sum, which deterministically shifts the probability mass toward larger values. Consequently, the CDF is strictly decreasing in d: for any $j \geq 0$, we have $F_{n-f-1+2j}(x) \leq F_{n-f-1}(x)$. Applying this inequality term-by-term to the Poisson mixture yields the step-by-step stochastic dominance bound

$$\mathbb{P}\left(\frac{(n-f-1)\operatorname{Var}(\{g_t'^{(i)}\}_{i\in\mathcal{H}})}{\sigma_{\mathrm{DP}}^2} \leq x\right) \leq F_{n-f-1}(x)\sum_{j=0}^{\infty} e^{-\lambda/2}\frac{(\lambda/2)^j}{j!}$$
$$= \mathbb{P}\left(\frac{(n-f-1)\operatorname{Var}(\{g_t^{(i)}\}_{i\in\mathcal{H}})}{\sigma_{\mathrm{DP}}^2} \leq x\right).$$

By defining $x := \frac{(n-f-1)(\mathbb{E}[\operatorname{Var}(\{g_t^{(i)}\}_{i\in\mathcal{H}})]-s)}{\sigma_{\mathrm{DP}}^2}$, the relation guarantees that the shifted tail failure probability is bounded by the unshifted tail failure probability. Therefore, we control the event bounds via $\mathbb{P}(E_{2,t}'^{\mathsf{c}}) \leq \mathbb{P}(E_{2,t}^{\mathsf{c}})$, where the unshifted failure event is defined as

$$E_{2,t}^{\mathsf{c}} := \left\{\operatorname{Var}(\{g_t^{(i)}\}_{i\in\mathcal{H}}) - \mathbb{E}\left[\operatorname{Var}(\{g_t^{(i)}\}_{i\in\mathcal{H}})\right] \leq -s\right\}. \tag{18}$$

(c) We control the upper tail probability the events

$$E_{3,t}^{\mathsf{c}} := \left\{\operatorname{Var}(\{g_t^{(i)}\}_{i\in\tilde{S}}) - \mathbb{E}\left[\operatorname{Var}(\{g_t^{(i)}\}_{i\in\tilde{S}})\right] > s\right\},$$
$$E_{3,t}'^{\mathsf{c}} := \left\{\operatorname{Var}(\{g_t'^{(i)}\}_{i\in\tilde{S}}) - \mathbb{E}\left[\operatorname{Var}(\{g_t'^{(i)}\}_{i\in\tilde{S}})\right] > s\right\},$$

that have the same distribution.

Indeed, on the event $E_{2,t}' \cap E_{3,t}'$ we have

$$\operatorname{Var}(\{g_t'^{(i)}\}_{i\in S_t'}) \leq \operatorname{Var}(\{g_t'^{(i)}\}_{i\in\tilde{S}}) \leq \mathbb{E}\left[\operatorname{Var}(\{g_t'^{(i)}\}_{i\in\tilde{S}})\right] + s$$
$$= \mathbb{E}\left[\operatorname{Var}(\{g_t^{(i)}\}_{i\in\tilde{S}})\right] + s \leq (1-\frac{c'^2}{2})\mathbb{E}\left[\operatorname{Var}(\{g_t^{(i)}\}_{i\in\mathcal{H}})\right]$$
$$= \mathbb{E}\left[\operatorname{Var}(\{g_t^{(i)}\}_{i\in\mathcal{H}})\right] - s < \operatorname{Var}(\{g_t'^{(i)}\}_{i\in\mathcal{H}}).$$

Indeed, for any single-iteration, which are independent, *(a)*, *(b)*, and *(c)* are ensured by Lemma C.2.

(ii) Under E_1 and the trajectory produced by S (i.e., with $g_t^{(1)} = \frac{1}{m}\sum_{j=1}^m z^{(1,j)} + \xi_t^{(1)} = \xi_t^{(1)}$), we do an analogous analysis as in *(i)*, in the opposite direction, and define analogously $E_{2,t}$ and $E_{3,t}$ with identical concentration properties. That is, due to Algorithm 2, the Byzantine participants successfully steer the optimization trajectory in the positive direction by favoring the negative noises.

(iii) Otherwise, in the complementary event E_1^{c}, the influence of Byzantine participants is considered negligible, and any potential contribution to instability from this case is disregarded.

Consequence for the lower bound. We are now able to control the probability of the successful attack event (MIA and tail succes),

$$E := E_1 \cap E_2, \text{ where } E_2 := \bigcap_{t=n_{\text{test}}+1}^{T} (E_{2,t} \cap E_{3,t} \cap E_{2,t}' \cap E_{3,t}')$$

Indeed, for $t, s \in \{n_{\text{test}}+1, \ldots, T\}$, $t \neq s$, E_1 is independent of $E_{2,t}$, $E_{3,t}$, $E_{2,t}'$ and $E_{3,t}'$, and $E_{2,t} \cap E_{3,t} \cap E_{2,t}' \cap E_{3,t}'$ is independent of $E_{2,s} \cap E_{3,s} \cap E_{2,s}' \cap E_{3,s}'$. As a result, each iteration are independent, and for $t > n_{\text{test}}$, we evaluate the joint success over the final $T - n_{\text{test}}$ steps. By applying a union bound on the failure events per iteration, and leveraging the bounds $\mathbb{P}(E_{2,t}'^{\mathsf{c}}) \leq \mathbb{P}(E_{2,t}^{\mathsf{c}})$, $\mathbb{P}(E_{3,t}'^{\mathsf{c}}) = \mathbb{P}(E_{3,t}^{\mathsf{c}})$ and Lemma C.2, we have,

$$\mathbb{P}(E) = \mathbb{P}(E_1) \prod_{t\in\{n_{\text{test}}+1,\ldots,T\}} \mathbb{P}\left(E_{2,t} \cap E_{3,t} \cap E_{2,t}' \cap E_{3,t}'\right)$$

$$\geq \mathbb{P}(E_1) \prod_{t\in\{n_{\text{test}}+1,\dots,T\}} \left(1 - \mathbb{P}\left(E_{2,t}^{\mathsf{c}}\right) - \mathbb{P}\left(E_{3,t}^{\mathsf{c}}\right) - \mathbb{P}\left(E'_{2,t}\right) - \mathbb{P}\left(E'_{3,t}\right)\right)$$

$$\geq \mathbb{P}(E_1) \left(1 - 2\left(e^{-\frac{c'^4(n-f-1)}{16}} + e^{-\frac{c_1 c'^2(n-f-1)}{4}\min\{\frac{c'^2(n-f-1)}{4(n-2f)},1\}} + e^{-\frac{c'^4(n-f)(n-f-1)}{128c^2 f}}\right)\right)^{T-n_{\text{test}}}, \tag{19}$$

Consequently, $\mathbb{P}(E_2) \in \Omega(1)$ if and only if

$$T \in \mathcal{O}\Big(e^{c_{\min}(n-f-1)}\Big),$$

where $c_{\min} > 0$ is a strictly positive constant defined as

$$\min\left\{\frac{c'^4}{16}, \frac{c_1 c'^2}{4}\min\left\{\frac{c'^2(n-f-1)}{4(n-2f)}, 1\right\}, \frac{c'^4(n-f)}{128c^2 f}\right\} \geq \frac{c'^4}{16}\min\left\{1, c_1, \frac{1}{8c^2}\right\} := c_{\min},$$

where we used $n > 2f$ and $f \geq 1$ (otherwise SMEA is the standard average aggregation) to bound $\frac{n-f-1}{n-2f} \geq 1$ and $\frac{n-f}{f} > 1$. This condition is ensured by our mild assumption on T.

We decompose the expected divergence of the trajectories into three terms to match our derivation,[5]

$$\mathbb{E}\left[\theta_T - \theta'_T\right] = \mathbb{E}\left[(\theta_T - \theta'_T)\mathbf{1}_E\right] + \mathbb{E}\left[(\theta_T - \theta'_T)\mathbf{1}_{E_1^{\mathsf{c}}\cap E_2}\right] + \mathbb{E}\left[(\theta_T - \theta'_T)\mathbf{1}_{E_2^{\mathsf{c}}}\right]$$

Applying the reverse triangle inequality yields

$$\left|\mathbb{E}\left[\theta_T - \theta'_T\right]\right| \geq \left|\mathbb{E}\left[\theta_T - \theta'_T \mid E\right]\mathbb{P}(E) + \mathbb{E}\left[(\theta_T - \theta'_T)\mathbf{1}_{E_1^{\mathsf{c}}\cap E_2}\right]\right| - \left|\mathbb{E}\left[(\theta_T - \theta'_T)\mathbf{1}_{E_2^{\mathsf{c}}}\right]\right|. \tag{20}$$

We next show that (1) the first term consists of two nonnegative contributions, and hence $\left|\mathbb{E}\left[\theta_T - \theta'_T \mid E\right]\mathbb{P}(E) + \mathbb{E}\left[(\theta_T - \theta'_T)\mathbf{1}_{E_1^{\mathsf{c}}\cap E_2}\right]\right| \geq \mathbb{E}\left[\theta_T - \theta'_T \mid E\right]\mathbb{P}(E)$; and (2) the last term can be neglected compared to the first term.

Phase 1 ($t \leq n_{test}$). For the first n_{test} iterations, the Byzantine participants send sufficiently large vectors, forcing SMEA to exclusively select the $n - f$ honest gradients. Because the datasets S and S' differ only at participant 1 by C/m, the parameters diverge deterministically,

$$\theta'_{n_{\text{test}}} = -\frac{\gamma}{n-f}\sum_{t=1}^{n_{\text{test}}}\left(\frac{C}{m} + \sum_{i=1}^{n-f}\xi_t^{(i)}\right) = \theta_{n_{\text{test}}} - \frac{\gamma C n_{\text{test}}}{(n-f)m}.$$

Thus, the difference $\theta_{n_{\text{test}}} - \theta'_{n_{\text{test}}} = \frac{\gamma C n_{\text{test}}}{(n-f)m}$ is a strictly positive deterministic constant. Because Phase 1 is independent of Phase 2, it holds for all subsequent conditional expectations.

Phase 2 ($t > n_{test}$) under MIA and tail success event ($E = E_1 \cap E_2$). Conditioned on E, the MIA successfully distinguishes the datasets (E_1), and SMEA selects the Byzantine vectors at each steps (E_2).

Let $B = \{n - f + 1, \dots, n\}$. and S_t and S'_t denote the subsets of indices chosen by SMEA on S and S' respectively, for $t \in \{n_{\text{text}} + 1, \dots, T\}$. The final parameters are

$$\theta_T = \theta_{n_{\text{test}}} - \frac{\gamma}{n-f}\sum_{t=n_{\text{test}}+1}^{T}\left(-f\beta\sigma_{\text{DP}} + \sum_{i\in S_t\setminus B}\xi_t^{(i)}\right),$$

$$\theta'_T = \theta'_{n_{\text{test}}} - \frac{\gamma}{n-f}\sum_{t=n_{\text{test}}+1}^{T}\left(f\beta\sigma_{\text{DP}} + \sum_{i\in S'_t\setminus B}g_t'^{(i)}\right).$$

Let $\Delta = \frac{\gamma}{n-f}\sum_{t=n_{\text{test}}+1}^{T}\left(\sum_{i\in S'_t\setminus B}g_t'^{(i)} - \sum_{i\in S_t\setminus B}\xi_t^{(i)}\right)$, subtracting the two trajectories yields

$$\mathbb{E}\left[\theta_T - \theta'_T \mid E\right] = \frac{\gamma C n_{\text{test}}}{(n-f)m} + \frac{2\gamma f\beta\sigma_{\text{DP}}(T - n_{\text{test}})}{n-f} + \mathbb{E}\left[\Delta \mid E\right]. \tag{21}$$

Moreover, we show that $\mathbb{E}\left[\Delta \mid E\right] \geq 0$ with probability 1, which therefore adds to the other positive terms in (21), and hence $\mathbb{E}\left[\theta_T - \theta'_T \mid E\right] \geq \frac{\gamma C n_{\text{test}}}{(n-f)m} + \frac{2\gamma f\beta\sigma_{\text{DP}}(T-n_{\text{test}})}{n-f} > 0$. Indeed, this is due

[5]Note that, by definition, $\mathbb{E}\left[(\theta_T - \theta'_T)\mathbf{1}_E\right] = \mathbb{E}\left[(\theta_T - \theta'_T) \mid E\right]\mathbb{P}(E)$.

to (1) $g_t'^{(i)} \geq \xi_t^{(i)}$ for all honest participants (with equality for $i \neq 1$ and $g_t'^{(1)} = \xi_t^{(1)} + C/m$), (2) under E and the trajectory produced by S' the SMEA favor the honest positive noises due to the Byzantine successful attack, and (3) under E and the trajectory produced by S the SMEA favor the honest negative noises due to the Byzantine successful attack.

To further justify (2) and (3), we recall that under E and the trajectory produced by S', the Byzantine participants fool SMEA and $\forall i \in B, t \in \{n_{\text{test}}+1, \ldots, T\}, g_t^{(i)} = \beta\sigma_{\text{DP}}$ are selected, i.e., $S_t' = B \cup \text{Select}_{n-2f}(\mathcal{H})$. Here, $\text{Select}_{n-2f}(\mathcal{H})$ denotes the subset of $n-2f$ honest participants whose gradients, when combined with the Byzantine gradients, minimize the variance of S_t'. Since the f Byzantine values are all equal to $\beta\sigma_{\text{DP}} > 0$, minimizing the variance of S_t' amounts to selecting the $n-2f$ honest gradients with the highest values. In other words, the selection favors honest gradients with positive value. Specifically, let $\sigma_t' : \mathcal{H} \to \mathcal{H}$ denote the permutation that sorts the honest gradients $(g_t'^{(i)})_{i\in\mathcal{H}}$ in non-increasing order (highest values first). Then, $\text{Select}_{n-2f}(\mathcal{H}) = S_t' \setminus B = \{\sigma_t'(1), \ldots, \sigma_t'(n-2f)\}$. Analogously, we define σ_t for S_t that sorts $(\xi_t^{(i)})_{i\in\mathcal{H}}$ in a non-decreasing order (lowest values first), where the negative Byzantine values $-\beta\sigma_{\text{DP}}$ force SMEA to favor the honest gradients with the lowest values, yielding $S_t \setminus B = \{\sigma_t(1), \ldots, \sigma_t(n-2f)\}$. Consequently,

$$\sum_{i\in S_t'\setminus B} g_t'^{(i)} = \sum_{i=1}^{n-2f} g_t'^{(\sigma'(i))} \geq \sum_{i=1}^{n-2f} \xi_t^{(\sigma'(i))} \geq \sum_{i=1}^{n-2f} \xi_t^{(\sigma(i))} = \sum_{i\in S_t\setminus B} \xi_t^{(i)},$$

where the first inequality holds because of (1), and the second inequality holds because the sum of the $n-2f$ smallest elements of ξ_t (sorted by σ_t) is by definition smaller than any sum $n-2f$ elements of ξ_t. Thus, $\mathbb{E}[\Delta \mid E] \geq 0$, which adds to the positive terms in (21).

Phase 2 ($t > n_{test}$) under MIA failure but tail success event ($E_1^{\text{c}} \cap E_2$). Conditioned on E_1^{c}, the MIA fails to distinguish the datasets. Consequently, the Byzantine participants observe identical trigger conditions and execute the same strategy for both S and S'. That is, either $\zeta \leq c_{\text{trigger}} - \frac{C}{m}$, and the Byzantine participants send $-\beta\sigma_{\text{DP}}$ for both trajectories produced by S and S' or $\zeta > c_{\text{trigger}}$ and the Byzantine participants send $\beta\sigma_{\text{DP}}$ for both trajectories produced by S and S'. The processed vectors by SMEA at each iteration are identical in every component except the first honest participant, where $g_t'^{(1)} = g_t^{(1)} + \frac{C}{m}$.

Conditioned further on E_2, SMEA selects the Byzantine participants on both trajectories. Let H_t and H_t' denote the $n-2f$ honest participants selected for S and S' at t. By definition, let $b \in \{\beta\sigma_{\text{DP}}, -\beta\sigma_{\text{DP}}\}$,

$$\begin{aligned}\text{SMEA}(g_t') - \text{SMEA}(g_t) &= \frac{1}{n-f}\left(\sum_{i\in B} b + \sum_{i\in H_t'} g_t'^{(i)}\right) - \frac{1}{n-f}\left(\sum_{i\in B} b + \sum_{i\in H_t} g_t^{(i)}\right) \\ &= \frac{1}{n-f}\left(\sum_{i\in H_t'} g_t'^{(i)} - \sum_{i\in H_t} g_t^{(i)}\right).\end{aligned}$$

Because the Byzantine vectors are identical on both trajectories, they cancel. To minimize variance, SMEA selects the $n-2f$ honest elements closest to b. Whether $b = \beta\sigma_{\text{DP}}$ (forcing the selection of the $n-2f$ largest elements) or $b = -\beta\sigma_{\text{DP}}$ (forcing the selection of the $n-2f$ smallest elements), the underlying inputs satisfy $g_t'^{(i)} \geq g_t^{(i)}$ for all $i \in \mathcal{H}$. Therefore,

$$\sum_{i\in H_t'} g_t'^{(i)} \geq \sum_{i\in H_t} g_t^{(i)} \quad \text{and hence,} \quad \text{SMEA}(g_t') - \text{SMEA}(g_t) \geq 0.$$

That is,

$$\mathbb{E}\left[(\theta_T - \theta_T')\mathbf{1}_{E_1^{\text{c}}\cap E_2}\right] = \theta_{n_{\text{test}}} - \theta'_{n_{\text{test}}} + \gamma \sum_{t=n_{\text{test}}+1}^{T} \mathbb{E}\left[\big(\text{SMEA}(g_t') - \text{SMEA}(g_t)\big)\mathbf{1}_{E_1^{\text{c}}\cap E_2}\right] \geq 0. \tag{22}$$

Phase 2 ($t > n_{test}$) under tail failure event (E_2^{c}). The goal here is to formally prove that we can neglect the instability from the event E_2^{c} when accounting for our stability lower bound. Since this

union event is difficult to analyze directly, we instead adopt a worst-case bound and exploit the fact that $\mathbb{P}(E_2^{\mathsf{c}}) \le 6(T - n_{\text{test}})e^{-c_{\min}(n-f)}$ is small to show that

$$\left|\mathbb{E}\left[(\theta_T - \theta'_T)\mathbf{1}_{E_2^{\mathsf{c}}}\right]\right| \le \frac{1}{4}\mathbb{E}[(\theta_T - \theta'_T)\mathbf{1}_{E_1 \cap E_2}]. \tag{23}$$

For any $t > n_{\text{test}}$,

$$\left|\text{SMEA}(g'_t) - \text{SMEA}(g_t)\right| \le \frac{2f\beta\sigma_{\text{DP}}}{n-f} + \frac{1}{n-f}\sum_{i\in\mathcal{H}}\left(|g_t'^{(i)}| + |g_t^{(i)}|\right).$$

Let $N_t := \frac{1}{n-f}\sum_{i\in\mathcal{H}}(|\xi_t^{(i)} + C/m| + |\xi_t^{(i)}|) \le \frac{2}{n-f}\sum_{i\in\mathcal{H}}|\xi_t^{(i)}| + \frac{C}{(n-f)m}$. Over T steps,

$$\begin{aligned}\left|\mathbb{E}\left[(\theta_T - \theta'_T)\mathbf{1}_{E_2^{\mathsf{c}}}\right]\right| &\le \theta_{n_{\text{test}}} - \theta'_{n_{\text{test}}} + \gamma\sum_{t=n_{\text{test}}+1}^{T}\mathbb{E}\left[\left(\frac{2f\beta\sigma_{\text{DP}}}{n-f} + N_t\right)\mathbf{1}_{E_2^{\mathsf{c}}}\right] \\ &= \theta_{n_{\text{test}}} - \theta'_{n_{\text{test}}} + \gamma(T - n_{\text{test}})\frac{2f\beta\sigma_{\text{DP}}}{n-f}\mathbb{P}(E_2^{\mathsf{c}}) + \gamma\sum_{t=n_{\text{test}}+1}^{T}\mathbb{E}\left[N_t\mathbf{1}_{E_2^{\mathsf{c}}}\right].\end{aligned}$$

We apply the Cauchy-Schwarz inequality, $\mathbb{E}[N_t\mathbf{1}_{E_2^{\mathsf{c}}}] \le \sqrt{\mathbb{E}[N_t^2]}\sqrt{\mathbb{P}(E_2^{\mathsf{c}})}$. By Jensen's inequality on the sum, $\mathbb{E}[N_t^2] \le 2(4\sigma_{\text{DP}}^2 + \frac{C^2}{(n-f)^2m^2}) := V_N^2$. Thus,

$$\left|\mathbb{E}\left[(\theta_T - \theta'_T)\mathbf{1}_{E_2^{\mathsf{c}}}\right]\right| \le \theta_{n_{\text{test}}} - \theta'_{n_{\text{test}}} + \gamma(T - n_{\text{test}})\left(\frac{2f\beta\sigma_{\text{DP}}}{n-f}\mathbb{P}(E_2^{\mathsf{c}}) + V_N\sqrt{\mathbb{P}(E_2^{\mathsf{c}})}\right). \tag{24}$$

By substituting (21) and (24) in (23), it is sufficient to show that

$$\frac{2f\beta\sigma_{\text{DP}}}{n-f}\mathbb{P}(E_2^{\mathsf{c}}) + V_N\sqrt{\mathbb{P}(E_2^{\mathsf{c}})} \le \frac{1}{4}\frac{2f\beta\sigma_{\text{DP}}}{n-f}\mathbb{P}(E_1)\big(1 - \mathbb{P}(E_2^{\mathsf{c}})\big), \tag{25}$$

which is a quadratic inequality for $\sqrt{\mathbb{P}(E_2^{\mathsf{c}})} \in [0,1]$. We show there exists $p_2^{\max} > 0$ such that the inequality holds for all $\sqrt{\mathbb{P}(E_2^{\mathsf{c}})} \le p_2^{\max}$, and that it is ensured by the mild assumption on T.

Indeed, let $x = \sqrt{\mathbb{P}(E_2^{\mathsf{c}})} \in [0,1]$, $V_N = \sigma_{\text{DP}}\sqrt{8 + \frac{2C^2}{(n-f)^2m^2\sigma_{\text{DP}}^2}} \le 3\sigma_{\text{DP}}$,[6]

$$A := \frac{2f\beta\sigma_{\text{DP}}}{n-f}(1+\frac{1}{4}\mathbb{P}(E_1)) = 2c\sigma_{\text{DP}}\sqrt{\frac{f(n-f-1)}{(n-f)(n-2f)}}\left(1 + \frac{1}{4}\mathbb{P}(E_1)\right) \le \frac{5c\sigma_{\text{DP}}}{2}\sqrt{\frac{f(n-f-1)}{(n-f)(n-2f)}},$$

and

$$C := \frac{f\beta\sigma_{\text{DP}}}{2(n-f)}\mathbb{P}(E_1) = \frac{c}{2}\sigma_{\text{DP}}\sqrt{\frac{f(n-f-1)}{(n-f)(n-2f)}}\mathbb{P}(E_1).$$

We solve

$$Ax^2 + V_N x - C \le 0.$$

Because $A > 0$ and $-C < 0$, the parabola is negative at $x = 0$ and opens upwards, hence the inequality has a unique positive solution, equalt to

$$p_2^{\max} = \frac{-V_N + \sqrt{V_N^2 + 4AC}}{2A} = \frac{2C}{\sqrt{V_N^2 + 4AC} + V_N},$$

so that (25) is satisfied for all $x \le p_2^{\max}$. We now lower bound $p_2^{\max}$, using $\sqrt{\frac{f(n-f-1)}{(n-f)(n-2f)}} \ge \frac{1}{\sqrt{n}}$,

$$p_2^{\max} \ge \frac{C}{\sqrt{AC} + 3\sigma_{\text{DP}}} = \frac{\frac{C}{\sigma_{\text{DP}}}}{\sqrt{\frac{AC}{\sigma_{\text{DP}}^2}} + 3} = \frac{\frac{c\sqrt{f(n-f-1)}}{2\sqrt{(n-f)(n-2f)}}\mathbb{P}(E_1)}{\frac{c\sqrt{f(n-f-1)}}{2\sqrt{(n-f)(n-2f)}}\sqrt{5\mathbb{P}(E_1)} + 3} \ge \frac{\mathbb{P}(E_1)}{\sqrt{5} + \frac{6\sqrt{(n-f)(n-2f)}}{c\sqrt{f(n-f-1)}}}$$

[6]The condition $\frac{2C^2}{(n-f)^2m^2\sigma_{\text{DP}}^2} \le 1$ is equivalent to $\varepsilon \le (n-f)^2 4T\ln(1/\delta)$, which is a extremely mild condition and is typically satisfied in both theoretical and practical settings.

$$\geq \frac{\mathbb{P}(E_1)}{\sqrt{5}+\frac{6\sqrt{n}}{c}} \geq \frac{\mathbb{P}(E_1)}{(\sqrt{5}+\frac{6}{c})\sqrt{n}}.$$

Consequently, it is sufficient to have $\mathbb{P}(E_2^{\mathsf{c}}) \leq (\sqrt{5}+\frac{6}{c})^{-2}\frac{\mathbb{P}(E_1)^2}{n}$ for (25) to be satisfied. Recall that $\mathbb{P}(E_2^{\mathsf{c}}) \leq 6(T-n_{\text{test}})e^{-c_{\min}(n-f)}$, then a sufficient condition is

$$T \leq \frac{1}{6(\sqrt{5}+\frac{6}{c})^2(1-p_{\text{test}})}\mathbb{P}(E_1)^2 e^{c_{\min}(n-f)-\ln n} \in \mathcal{O}\left(\mathbb{P}(E_1)^2 e^{\frac{c_{\min}}{2}(n-f)}\right).$$

Final Lower Bound: Plugging (21), (22) and (23) in (20) yields

$$\begin{aligned}
\left|\mathbb{E}\left[\theta_T-\theta'_T\right]\right| &\geq \mathbb{E}\left[\theta_T-\theta'_T \mid E_1\cap E_2\right]\mathbb{P}(E_1\cap E_2) + 0 - \frac{1}{4}\mathbb{E}\left[\theta_T-\theta'_T \mid E_1\cap E_2\right]\mathbb{P}(E_1\cap E_2)\\
&= \frac{3}{4}\mathbb{E}\left[\theta_T-\theta'_T \mid E_1\cap E_2\right]\mathbb{P}(E_1)\mathbb{P}(E_2)\\
&= \frac{3}{4}\gamma T\left(\frac{p_{\text{test}}C}{(n-f)m}+\frac{2(1-p_{\text{test}})f\beta\sigma_{\text{DP}}}{n-f}\right)\mathbb{P}(E_1)\mathbb{P}(E_2)\\
&\in \Omega\left(\gamma T\left(\frac{C}{(n-f)m}+\sigma_{\text{DP}}\sqrt{\frac{f}{n-2f}}\right)\mathbb{P}(E_1)\right),
\end{aligned}$$

where we used (19), $n_{\text{test}} = p_{\text{test}}T$, and $\beta = c\sqrt{\frac{(n-f-1)(n-f)}{f(n-2f)}}$ in the last inequality. Recall that $\mathbb{P}(E_1)$ is governed by the hypothesis test signal-to-noise ratio $\frac{\varepsilon\sqrt{p_{\text{test}}}}{2\sqrt{2\ln(1/\delta)}}$, that is the privacy contraint.

We conclude the proof with the following derivation, enabled by the fact that the loss function is linear

$$\sup_{z\in[-C,C]}\left|\mathbb{E}\left[\ell(\theta_T,z)-\ell(\theta'_T,z)\right]\right| = \sup_{z\in[-C,C]} z\,\left|\mathbb{E}\left[\theta_T-\theta'_T\right]\right| = C\,\left|\mathbb{E}\left[\theta_T-\theta'_T\right]\right|.$$

Indeed, by assumption $T \in \mathcal{O}\big(\mathbb{P}(E_1)^2 e^{c_{\min}(n-f-1)}\big)$. Consequently, (1) If $\mathbb{P}(E_1) \in \Omega(1)$, then $T \in \mathcal{O}\big(e^{c_{\min}(n-f-1)}\big)$, which corresponds to the high-privacy regime. (2) If additionally the loss function is bounded by ℓ_∞. As we use linear loss functions, where bounded loss is equivalent to bounded domain, we ensure the bounded loss assumption by restricting the domain of the parameters to $[-D, D]$, with $D = \frac{\ell_\infty}{C}$. Consequently, $|\ell(\theta;z)| = |\theta z| \leq DC = \ell_\infty$. The only required modification of `Safe-DGD` is to add a projection onto the parameter compact set after the server update (Line 10 in Algorithm 1). Following our proof, there exists $\sigma_{\text{threshold}} > 0$ such that for $\sigma_{\text{DP}} \geq \sigma_{\text{threshold}}$, the parameter eventually hit the domain limit $\theta_T = D$ and $\theta'_T = -D$ so we prove the saturated lower bound.

Gluing lower bounds. Moreover, when $\sigma_{\text{DP}} = 0$, the uniform stability is lower bounded by $\Omega\left(\gamma C^2 T\left(\frac{1}{(n-f)m}+\sqrt{\frac{f}{n-2f}}\right)\right)$ when $\frac{n}{3} \leq f < \frac{n}{2}$, and by $\Omega\left(\gamma C^2 T\left(\frac{1}{(n-f)m}+\frac{f}{n-2f}\right)\right)$ when $f < \frac{n}{3}$ (Boudou et al., 2025, Theorem 3.2). We conclude the proof with the formula $\max\{a,b\} \geq \frac{a+b}{2} \in \Omega(a+b)$, i.e. when $\frac{n}{3} \leq f < \frac{n}{2}$ the uniform stability is lower bounded by

$$\Omega\left(\gamma C^2 T\left(\frac{1}{(n-f)m}+\sqrt{\frac{f}{n-2f}}\left(1+\frac{\sigma_{\text{DP}}}{C}\right)\right)\right).$$

The case $f < \frac{n}{3}$ is analogous. Note that $\sqrt{\kappa_{\text{SMEA}}}$ is of the same order as $\sqrt{\frac{f}{n-2f}}$ if we assume there exist a constant $\nu > 0$ such that $f \leq \frac{n}{2+\nu}$, i.e. the lower bound matches the upper bound. □

C.3.2 Proof for $d > 1$.

Assumption C.3. *We assume the the following.*

1. *There exists a constant $\alpha_{\min} > 0$ such that $\alpha := \frac{f}{n-f} \in [\alpha_{\min}, 1/2)$.*
2. *There exist constants $\mu, \nu > 0$ such that $\mu \leq \frac{d}{n-f} \leq \nu$.*

3. *$\mu > \mu_{\min}$ and $n - f - 1 \geq N_{\min}$, where $\mu_{\min}$ and $N_{\min}$ are constants that depends on assumption 1. & 2., and the problem spectral and finite-sample structure. That is,*

$$\mu_{\min} := \frac{4}{c_{\max}^2(1 - r_{\max})},$$

and

$$N_{\min} := \max\left\{ \left(\frac{3(\sqrt{\nu} + \sqrt{1-\alpha_{\min}} + \max\{C_1, C_2\})}{1 - \sqrt{1-\alpha_{\min}}} \right)^2, \frac{16\min(1,\nu)}{c_\delta \sqrt{\mu}\sqrt{2 + \frac{\alpha_{\min}\mu c^2}{\min(1,\nu)}}} \right\},$$

where $C_1, C_2 > 0$ are universal constants, and $c_{\max}, r_{\max} \in (0,1)$ are defined as

$$r_{\max} := \left(\frac{\sqrt{\nu} + \frac{1}{3} + \frac{2}{3}\sqrt{1-\alpha_{\min}}}{\sqrt{\nu} + \frac{2}{3} + \frac{1}{3}\sqrt{1-\alpha_{\min}}} \right)^2 \geq \frac{\mathbb{E}\|X_\perp\|_{\text{sp}}^2}{\mathbb{E}\|X\|_{\text{sp}}^2},$$

$$c_{\max}^2 := 1 - \frac{1-\alpha_{\min}}{\left(\sqrt{\nu} + 1 - \frac{1-\sqrt{1-\alpha_{\min}}}{3}\right)^2} \leq \frac{\mathbb{E}\left[(DZ)^\intercal P_{n-f} DZ\right]}{\mathbb{E}\|X\|_{\text{sp}}^2},$$

where $Z \sim \mathcal{N}(0, I_{n-f})$, $X \in \mathbb{R}^{d\times(n-f-1)}$, $X_\perp \in \mathbb{R}^{(d-1)\times(n-2f)}$ are random matrices with i.i.d. standard normal entries. $D \in \mathbb{R}^{(n-f)\times(n-f)}$ is a diagonal masking matrix with $n-2f$ ones and f zeros, and $P_{n-f} = I_{n-f} - \frac{1}{n-f}1_{n-f}1_{n-f}^\intercal$ is a centering projection matrix.

Assumption C.3, while technical are fairly mild. They are either practically always true (e.g., T) or assumed to bound expectation and concentration of random matrices (e.g., $\mu_{\min}$ and $N_{\min}$).

The quantities $\frac{\mathbb{E}\|X_\perp\|_{\text{sp}}^2}{\mathbb{E}\|X\|_{\text{sp}}^2}$ and $\frac{\mathbb{E}[(DZ)^\intercal P_{n-f} DZ]}{\mathbb{E}\|X\|_{\text{sp}}^2}$ have interpretation within the context of our proof. Indeed, $\frac{\mathbb{E}\|X_\perp\|_{\text{sp}}^2}{\mathbb{E}\|X\|_{\text{sp}}^2}$ is the ratio between the expected principal variance of the honest noise orthogonal to the attack and the expected principal variance of all honest noise. Intuitively, it measures how much of the honest noise aligns with the attack direction (which is arbitrary in our proof), dictating how easily the attacker can hide inside the honest variance. Finally, $\frac{\mathbb{E}[(DZ)^\intercal P_{n-f} DZ]}{\mathbb{E}\|X\|_{\text{sp}}^2}$ is the ratio between the expected variance of the honest noise along the attack axis and the expected principal variance of all the honest noise. Intuitively, this measures how much the orthogonal honest noise tilts the principal eigenvector away from the attack direction. We now turn to the proof of Theorem 3.1.

Theorem 3.1. *Consider the setting in Section 2 with full batch and the* SMEA *aggregation rule. Suppose Assumption C.3, $T \in \mathcal{O}\left(e^{c_{\min}\left(\sqrt{d}+\sqrt{n-f-1}\right)^2}\right)$ where $c_{\min}$ is a problem-dependent constant, and $\sigma_{\text{DP}} = \frac{2C\sqrt{2T\ln(1/\delta)}}{m\varepsilon}$. Let $\mathfrak{P}_T \in \Theta\left(\min\left\{1, \frac{\varepsilon}{\sqrt{\ln(1/\delta)}}\right\}\right)$ as described in Section 3. Then there exists $\ell \in \Theta \times \mathcal{Z} \to \mathbb{R}$ such that $\forall z \in \mathcal{Z}, \ell(\cdot; z)$ is C-Lipschitz, L-smooth and convex, and the uniform stability of $\mathcal{A}$ is lower bounded by*

$$\Omega\left(\mathfrak{P}_T \times \gamma C^2 T\left(\frac{1}{(n-f)m} + \frac{f}{n-2f} + \sqrt{\frac{f}{n-2f}}\frac{\sigma_{\text{DP}}}{C}\left(1 + \sqrt{\frac{d}{n-f}}\right)\right)\right). \tag{26}$$

If we moreover assume $\frac{n}{3} < f < \frac{n}{2}$, the uniform stability of $\mathcal{A}$ is lower bounded by

$$\Omega\left(\mathfrak{P}_T \times \gamma C^2 T\left(\frac{1}{(n-f)m} + \sqrt{\frac{f}{n-2f}} + \sqrt{\frac{f}{n-2f}}\frac{\sigma_{\text{DP}}}{C}\left(1 + \sqrt{\frac{d}{n-f}}\right)\right)\right). \tag{27}$$

For both cases, if we assume the loss function bounded by ℓ_∞[7] and $T \in \mathcal{O}\left(\mathfrak{P}_T^2 e^{c_{\min}\left(\sqrt{d}+\sqrt{n-f-1}\right)^2}\right)$, then there exists $\sigma_{threshold} > 0$ such that for $\sigma_{DP} \geq \sigma_{threshold}$, the uniform stability of $\mathcal{A}$ is lower bounded by the minimum between (4) *(respectively* (5)*) and $\Omega\left(\ell_\infty \mathfrak{P}_T\right)$.*

[7] This can be ensured by projecting parameters onto a compact set Θ after the update (Line 10 in Algorithm 1).

Proof. Let $d, n \geq 1$, $\mathcal{H} = \{1, \ldots, n-f\}$ and let $S, S' \in \mathcal{Z}^{(n-f)m}$ neighboring datasets where every sample equal 0, except for one sample in S' from the participant 1, equal to $-Ce_1$. Let $\ell : \mathbb{R}^d \times \overline{B}(0, C) \to \mathbb{R}, (\theta, z) \mapsto \ell(\theta; z) = \theta^\intercal z$. For $i \in \mathcal{H} := \{1, \ldots, n-f\}$, $0 \leq t \leq T$, let $\xi_t^{(i)} \sim \mathcal{N}(0, \sigma_{\text{DP}}^2 I_d)$ be independent random gaussian noise.

Byzantine participants trigger. By projecting the gradients onto the axis e_1, the MIA simplifies into a one dimenstional scalar hypothesis test based on the statistic ζ. That is, $\zeta = \bar{\xi} = \frac{1}{n_{\text{test}}} \sum_{t=1}^{n_{\text{test}}} {\xi_t^{(1)}}^\intercal e_1$ under H_0, and $\zeta = -\frac{C}{m} + \bar{\xi}$ under H_1, analogous to Theorem 3.1. Because the Gaussian noise is symmetric around zero, this yields the same MIA success probability $\mathbb{P}(E_1)$ as in Theorem 3.1.

Byzantine participants behavior under the SMEA **aggregation rule.** We assign the following behavior (Algorithm 3) to the Byzantine participants $i \in \{n-f+1, n\}$, for $t \geq n_{\text{test}} + 1$.

Algorithm 3 Byzantine Participant Behavior

```
# β (defined below) depends on f and n
if ζ ≥ −c_trigger then
    return β σ_DP e_1
else
    return −β σ_DP e_1
```

That is, depending on the input ζ statistic, which itself depends only on the observed noisy gradients, the Byzantine participants attempt to steer the optimization trajectory toward either positive or negative values (with respect to the hyperplane orthogonal to e_1).

In what follows, we decompose our analyze depending on whether the Byzantine participants undergo a successful MIA (i.e., infer whether the dataset is $\mathcal{S}$ or $\mathcal{S}'$ successfully) in the two coupled optimization trajectories (E_1), or not (E_1^{c}). In the complementary event E_1^{c}, the influence of Byzantine participants is considered negligible, and any potential contribution to instability from this case is disregarded. Consequently, we fix their behavior to sending an arbitrary vector, for example, 0. Under E_1, the attack Algorithm 3 output $\beta\, \sigma_{\text{DP}}\, e_1$ for the trajectory produced by S' (i.e., with $g_t^{(1)\prime} = \frac{1}{m} \sum_{j=1}^m z^{(1,j)\prime} + \xi_t^{(1)} = -\frac{C}{m} e_1 + \xi_t^{(1)}$), and output $-\beta\, \sigma_{\text{DP}}\, e_1$ for the trajectory produced by S (i.e., with $g_t^{(1)} = \frac{1}{m} \sum_{j=1}^m z^{(1,j)} + \xi_t^{(1)} = \xi_t^{(1)}$). That is, under E_1, the Byzantine participants attempt to steer the optimization trajectory in different directions depending on the dataset they have successfully inferred. The analysis is analogous for both trajectories. For the remaining of this paragraph, we work under E_1 and the trajectory induced by S.

Let $\Lambda := \mathbb{E}\|\Sigma_{\mathcal{H}}\|_{\text{sp}}$. Following Algorithm 3 after n_{test} steps, all Byzantine participants $i \in \mathcal{B} := \{n-f+1, \ldots, n\}$ send to the server a constant vectors in the direction e_1,

$$\text{for all } i \in \mathcal{B}, t \in \{n_{\text{test}} + 1, \ldots, T\}, \quad g_t^{(i)} := \beta \sigma_{\text{DP}} e_1,$$

where

$$\beta := c \frac{n-f}{\sqrt{f(n-2f)}} \frac{\sqrt{\Lambda}}{\sigma_{\text{DP}}}, \tag{28}$$

where $c \in (0, c_{\text{opt}})$ and c_{opt} are parameter-independent constants defined in step *(iv)* below. β is chosen so that SMEA consistently selects the Byzantine vectors, as explained below.

- **Note on Λ and the magnitude of $\frac{\sqrt{\Lambda}}{\sigma_{\text{DP}}}$.** By standard derivation, $\Sigma_{\mathcal{H}}$ is identically distributed to $\frac{\sigma_{\text{DP}}^2}{n-f} XX^\intercal$, where $X \in \mathbb{R}^{d \times (n-f-1)}$ is a random matrix with i.i.d. standard normal entries $\mathcal{N}(0,1)$.[8] Hence, $\lambda_{\max}(\Sigma_{\mathcal{H}}) = \frac{\sigma_{\text{DP}}^2}{n-f} s_{\max}(X)^2$, where $s_{\max}(X)$ is the maximum singular

[8] Indeed, denoting Ξ_t the matrix whose column are $\frac{\xi_t^{(i)}}{\sigma_{\text{DP}}}$ for $i \in \mathcal{H}$ and $P_{n-f} = I_{n-f} - \frac{1}{n-f} 1_{n-f} 1_{n-f}^\intercal$ the centering matrix which is an orthogonal projection matrix, $\Sigma_t = \frac{\sigma_{\text{DP}}^2}{n-f} \Xi_t P_{n-f} \Xi_t^\intercal$. P_{n-f} is diagonal-

value of X. Taking the expectation yields

$$\Lambda = \mathbb{E}[\lambda_{\max}(\Sigma_{\mathcal{H}})] = \frac{\sigma_{\text{DP}}^2}{n-f}\mathbb{E}[s_{\max}(X)^2]. \tag{29}$$

To relate Λ to the Marchenko-Pastur baseline $\sqrt{n-f-1}+\sqrt{d}$, we multiply and divide Equation (29) by this baseline, yielding

$$\frac{\sqrt{\Lambda}}{\sigma_{\text{DP}}} = c_\Lambda \sqrt{\frac{n-f-1}{n-f}}\left(1+\sqrt{\frac{d}{n-f-1}}\right), \quad \text{where} \quad c_\Lambda := \frac{\sqrt{\mathbb{E}[s_{\max}(X)^2]}}{\sqrt{n-f-1}+\sqrt{d}}. \tag{30}$$

To prove that c_Λ can be considered as a constant, we rely on results from high-dimensional probability (Vershynin, 2018, Chapter 7). By Vershynin (2018, Theorems 7.3.1) and standard Gaussian comparison arguments, there exists a constant $C > 0$ such that

$$\sqrt{d}+\sqrt{n-f-1}-C \le \mathbb{E}[s_{\max}(X)] \le \sqrt{d}+\sqrt{n-f-1}.$$

Furthermore, because the map $X \mapsto s_{\max}(X)$ is 1-Lipschitz,[9] the Gaussian Poincaré inequality guarantees $\text{Var}(s_{\max}(X)) \le 1$ (Boucheron et al., 2013, Section 3.7). Using the identity $\mathbb{E}[s_{\max}(X)^2] = \big(\mathbb{E}[s_{\max}(X)]\big)^2 + \text{Var}(s_{\max}(X))$ yields

$$(\sqrt{d}+\sqrt{n-f-1}-C)^2 \le \mathbb{E}[s_{\max}(X)^2] \le (\sqrt{d}+\sqrt{n-f-1})^2+1.$$

Taking the square root and dividing by the baseline dimension $(\sqrt{d}+\sqrt{n-f-1})$, we obtain

$$\max\{\frac{1}{2}, 1-\frac{C}{2}\} \le 1-\frac{C}{\sqrt{d}+\sqrt{n-f-1}} \le c_\Lambda \le \sqrt{1+\frac{1}{(\sqrt{d}+\sqrt{n-f-1})^2}} \le \sqrt{\frac{5}{4}}, \tag{31}$$

where the numerical lower bound is obtained through the following inequalities $\mathbb{E}s_{\max}(X)^2 \ge \frac{\mathbb{E}\|X\|_F^2}{\min(d,n-f-1)} = \frac{d(n-f-1)}{\min(d,n-f-1)} = \max(d, n-f-1)$, and hence $c_\Lambda \ge \frac{\max(d,n-f-1)}{\sqrt{n-f-1}+\sqrt{d}} \ge \frac{1}{2}$.

Let $\Sigma_S := \frac{1}{|S|}\sum_{i\in S}(g_t^{(i)}-\bar{g}_S)(g_t^{(i)}-\bar{g}_S)^\intercal$ denote the sample covariance matrix of a subset S. To ensure the aggregator selects the subset $S_t = H_t \cup \mathcal{B}$, where $H_t \subset \mathcal{H}$ is composed of $n-2f$ indices from $\mathcal{H}$, over the honest subset $\mathcal{H}$, we must guarantee with high probability that

$$\text{for } t \in \{n_{\text{test}}+1, \dots, T\}, \quad \lambda_{\max}(\Sigma_{S_t}) < \lambda_{\max}(\Sigma_{\mathcal{H}}).$$

That is $\text{SMEA}(g_t^{(1)}, \dots, g_t^{(n)}) = \bar{g}_{S_t} = \frac{1}{n-f}\sum_{i\in S_t} g_t^{(i)} = \frac{f}{n-f}\beta\,\sigma_{\text{DP}}\,e_1 + \frac{n-2f}{n-f}\bar{\xi}_{H_t}$. Indeed, all other subsets than S_t and $\mathcal{H}$ necessaryly have greater spectral norm.[10] We demonstrate this via a concentration argument, first bounding the expected spectra of both matrices.

To tightly bound $\lambda_{\max}(\Sigma_{S_t})$, we leverage the fact that, due to the subset $\mathcal{B}$, the dominant eigenvector of Σ_{S_t} consistently aligns with e_1, thereby reducing the dimensional dependence.[11] To do so, we decompose Σ_{S_t} into a 2×2 block matrix structure, where one diagonal block accounts for the

izable, $P_{n-f} = U\Lambda U^\intercal$, where U is an orthogonal matrix and $\Lambda := \text{diag}(1, \dots, 1, 0)$. Consequently, $\Sigma_t = \frac{\sigma_{\text{DP}}^2}{n-f}(\Xi_t U)\Lambda(\Xi_t U)^\intercal = \frac{\sigma_{\text{DP}}^2}{n-f}\sum_{j=1}^{n-f-1}(\Xi_t U)_j(\Xi_t U)_j^\intercal := \frac{\sigma_{\text{DP}}^2}{n-f}XX^\intercal$, with $X_j = (\Xi_t U)_j$.

[9] $|s_{\max}(X) - s_{\max}(Y)| = |\|X\|_{\text{sp}} - \|Y\|_{\text{sp}}| \le \|X-Y\|_{\text{sp}} \le \|X-Y\|_{\text{F}}$.

[10] For any intermediate subset S_k containing exactly $0 < k < f$ Byzantine vectors, the expected variance along the direction e_1 introduces a between-group mixture term proportional to $k(n-f-k)$. Because this forms a strictly concave quadratic function of k, its minimum over the closed interval $[0, f]$ is necessarily achieved at the boundaries ($k = 0$ or $k = f$). By tuning β to equalize these endpoints with a slight margin favoring $k = f$, we guarantee that SMEA selects either the purely honest subset $\mathcal{H}$ or the subset S_t over any intermediate mixture.

[11] This avoids the worst-case—loose in our setting—application of Weyl's inequality. The reasoning is as follows. By definition, S_t is partitioned into two disjoint groups: the honest subset H_t of size $n-2f$ (with

covariance along the direction e_1, and the other for the covariance in the $(d-1)$-dimensional orthogonal complement $e_1^\perp$. Geometrically, we demonstrate that the principal eigenvector of Σ_{S_t} is overwhelmingly aligned with the attack axis e_1.

Any unit vector $v \in \mathbb{R}^d$ can be uniquely decomposed as $v = xe_1 + yw$, where $w \in e_1^\perp$ is a unit vector and $x^2 + y^2 = 1$. Hence,

$$\begin{aligned} v^\intercal \Sigma_{S_t} v &= x^2(e_1^\intercal \Sigma_{S_t} e_1) + y^2(w^\intercal \Sigma_{S_t} w) + 2xy(e_1^\intercal \Sigma_{S_t} w) \\ &= x^2\Sigma_{11} + y^2(w^\intercal \Sigma_{22} w) + 2xyw^\intercal \Sigma_{12}, \end{aligned} \tag{32}$$

where we define the following blocks.

- $\Sigma_{11} := e_1^\intercal \Sigma_{S_t} e_1 \in \mathbb{R}$ is the variance along the attack axis e_1.
- $\Sigma_{22} := P_{e_1^\perp} \Sigma_{S_t} P_{e_1^\perp} \in \mathbb{R}^{(d-1)\times(d-1)}$ is the orthogonal covariance block, where $P_{e_1^\perp} := I_d - e_1 e_1^\intercal$ denote the orthogonal projection matrix zeroing out the first coordinate. Because the attack vector spans e_1, Σ_{22} is entirely independent of the Byzantine vectors.
- $\Sigma_{12} := P_{e_1^\perp} \Sigma_{S_t} e_1 \in \mathbb{R}^{d-1}$ is the cross-covariance vector. Note that we identify $(0 \quad \Sigma_{2,1} \quad \dots \quad \Sigma_{d,1})$ with $\Sigma_{12}^\intercal$, since $w \in e_1^\intercal$. We identify Σ_{22} analogously.

To bound the maximal eigenvalue, $\lambda_{\max}(\Sigma_{S_t}) = \sup_{v \in \mathbb{S}^{d-1}} v^\intercal \Sigma_{S_t} v$, we aim at upper bounding (32),

$$v^\intercal \Sigma_{S_t} v \le x^2\Sigma_{11} + y^2\lambda_{\max}(\Sigma_{22}) + 2|xy|\|\Sigma_{12}\|_2 = (|x| \quad |y|) \begin{pmatrix} \Sigma_{11} & \|\Sigma_{12}\|_2 \\ \|\Sigma_{12}\|_2 & \lambda_{\max}(\Sigma_{22}) \end{pmatrix} \begin{pmatrix} |x| \\ |y| \end{pmatrix},$$

where $(|x| \quad |y|)$ is a unit vector in $\mathbb{R}^2$. Consequently, this decomposition yields[12]

$$\begin{aligned} \lambda_{\max}(\Sigma_{S_t}) &\le \lambda_{\max}(\begin{pmatrix} \Sigma_{11} & \|\Sigma_{12}\|_2 \\ \|\Sigma_{12}\|_2 & \lambda_{\max}(\Sigma_{22}) \end{pmatrix}) \\ &= \frac{\Sigma_{11} + \lambda_{\max}(\Sigma_{22})}{2} + \frac{1}{2}\sqrt{(\Sigma_{11} - \lambda_{\max}(\Sigma_{22}))^2 + 4\|\Sigma_{12}\|_2^2}. \end{aligned} \tag{33}$$

(i) **Bounding Σ_{11}.** The variance along e_1 decomposes into

$$e_1^\intercal \Sigma_{S_t} e_1 = \underbrace{\frac{1}{n-f} h_{t,1}^\intercal P_{n-f} h_{t,1}}_{:=V_1} - \underbrace{\frac{2f\beta\sigma_{\text{DP}}}{(n-f)^2} \sum_{i \in \mathcal{H} \cap S_t} \xi_{t,1}^{(i)}}_{:=V_2} + \underbrace{\frac{f(n-2f)}{(n-f)^2}\beta^2\sigma_{\text{DP}}^2}_{:=V_3},$$

empirical mean $\bar{\xi}_{H_t}$ and covariance Σ_{H_t}) and the Byzantine subset $\mathcal{B}$ of size f (with empirical mean $\beta\sigma_{\text{DP}}e_1$ and covariance 0). Using the standard decomposition formula for the empirical covariance for the partition $S_t = H_t \cup \mathcal{B}$, we have

$$\begin{aligned} \Sigma_{S_t} &= \frac{|H_t|}{|S_t|}\Sigma_{H_t} + \frac{|\mathcal{B}|}{|S_t|}\Sigma_{\mathcal{B}} + \frac{|H_t||\mathcal{B}|}{|S_t|^2}(\bar{\xi}_{H_t} - \beta\sigma_{\text{DP}}e_1)(\bar{\xi}_{H_t} - \beta\sigma_{\text{DP}}e_1)^\intercal \\ &= \frac{n-2f}{n-f}\Sigma_{H_t} + \frac{f(n-2f)}{(n-f)^2}(\bar{\xi}_{H_t} - \beta\sigma_{\text{DP}}e_1)(\bar{\xi}_{H_t} - \beta\sigma_{\text{DP}}e_1)^\intercal. \end{aligned}$$

We have $\lambda_{\max}((\bar{\xi}_{H_t} - \beta\sigma_{\text{DP}}e_1)(\bar{\xi}_{H_t} - \beta\sigma_{\text{DP}}e_1)^\intercal) = \|\bar{\xi}_{H_t} - \beta\sigma_{\text{DP}}e_1\|_2^2$ and, by Weyl's inequality, the maximum eigenvalue of a sum of symmetric matrices is bounded by the sum of their maximum eigenvalues,

$$\lambda_{\max}(\Sigma_{S_t}) \le \text{UB}_{S_t} := \frac{n-2f}{n-f}\lambda_{\max}(\Sigma_{H_t}) + \frac{f(n-2f)}{(n-f)^2}\|\bar{\xi}_{H_t} - \beta\sigma_{\text{DP}}e_1\|_2^2.$$

UB_{S_t} does not leverage the fact that, because of the subset $\mathcal{B}$, the dominant eigenvector aligns with e_1, which introduces a dimensional dependence and makes the analysis loose. Note that general alternatives to Weyl's inequality for matrix perturbation exist, such as those based on the Davis–Kahan $\sin\Theta$ theorem (Theorem 4.1.15 Vershynin, 2018, , i.e., eigen-space perturbation theory). However, our technique directly leverages the problem structure, making the overall reasoning tighter and simpler.

[12]We might use the inequality $\sqrt{x^2+y^2} \le \min\{|x|+|y|, |x| + \frac{y^2}{2|x|}\}$ to the square root term to simplify the proof. Though, we place our analysis on the edge of what the Byzantine participants can do. That is, in the inequality above, it is difficult to know if $|x|$ and $|y|$ will have the same order or not. As we require to be as tight as possible, we apply Sylvester's criterion to avoid explicitly handling inequalities.

where $\xi_{t,1}^{(i)} = \xi_t^{(i)\intercal} e_1$, $P_{n-f} = I_{n-f} - \frac{1}{n-f} 1_{n-f} 1_{n-f}^\intercal$, and h_t is the vector containing the $n-2f$ honest noise from S_t, and filled with 0 for the remaining f dimension. By the definition of β, the deterministic Byzantine contribution simplifies to $V_3 = c^2\Lambda$. Taking the expectation yields.[13]

$$\mathbb{E}[\Sigma_{11}] = \sigma_0^2 + c^2\Lambda := U_{11}, \text{ where } \sigma_0^2 := \frac{(n-2f)(n-f-1)}{(n-f)^2}\sigma_{\text{DP}}^2. \tag{34}$$

(ii) **Bounding $\mathbb{E}\lambda_{\max}(\Sigma_{22})$ via Loewner partial order.** Let $u_i := P_{e_1^\perp} g_t^{(i)}$ denote the projected gradients and $\mu := \frac{1}{n-f}\sum_{i\in S_t} u_i$ their subset mean. We have, $\Sigma_{22} = M_{22} - \mu\mu^\intercal$, where $M_{22} := \frac{1}{n-f}\sum_{i\in S_t} u_i u_i^\intercal$. Because $\mu\mu^\intercal \succeq 0$, we have $\Sigma_{22} \preceq M_{22}$, and hence $\lambda_{\max}(\Sigma_{22}) \le \lambda_{\max}(M_{22})$. Crucially, because the f Byzantine vectors evaluate to zero in $e_1^\perp$, they contribute zero to the uncentered sum,

$$M_{22} = \frac{1}{n-f}\sum_{i\in H_t} \xi_{t,\perp}^{(i)} \xi_{t,\perp}^{(i)\intercal} := \frac{\sigma_{\text{DP}}^2}{n-f} X_\perp X_\perp^\intercal,$$

where $X_\perp \in \mathbb{R}^{(d-1)\times(n-2f)}$ consists of i.i.d. standard normal entries. Because $f \ge 1$, $X_\perp$ is formed by removing columns (discarding $f-1$ excess honest participants) and a row (projecting onto $e_1^\perp$) from the matrix $X \in \mathbb{R}^{d\times(n-f-1)}$ defined in (29). Thus, $s_{\max}(X_\perp) \le s_{\max}(X)$ deterministically. Furthermore, because X is composed by independent normal random variables, the extra row and columns almost surely project non-trivially onto the dominant singular vectors of $X_\perp$. Consequently, the spectral norm strictly expands with probability 1 (i.e., $\mathbb{P}(s_{\max}(X_\perp) < s_{\max}(X)) = 1$).[14] Taking the expectation yields $r_\Lambda := \frac{\mathbb{E}[s_{\max}(X_\perp)^2]}{\mathbb{E}[s_{\max}(X)^2]} < 1$, and hence

$$\mathbb{E}\lambda_{\max}(\Sigma_{22}) \le \mathbb{E}\lambda_{\max}(M_{22}) = r_\Lambda\Lambda := U_{22}. \tag{35}$$

(iii) **Bounding $\mathbb{E}\|\Sigma_{12}\|_2^2$.** Because the Byzantine vectors project to zero in $e_1^\perp$ and evaluate to $\beta\sigma_{\text{DP}}$ in e_1, the cross-covariance simplifies to

$$\Sigma_{12} = (1-\alpha)(\Sigma_{H_t})_{12} + \alpha(1-\alpha)(\bar{\xi}_{H_t,1} - \beta\sigma_{\text{DP}})\bar{\xi}_{H_t,\perp},$$

where $(\Sigma_{H_t})_{12} = \frac{1}{n-2f}\sum_{i\in H_t}(\xi_{i,\perp} - \bar{\xi}_{H_t,\perp})(\xi_{i,1} - \bar{\xi}_{H_t,1})$. To compute the expected squared norm $\mathbb{E}\|\Sigma_{12}\|_2^2$, we square the sum. By the independence of the Gaussian sample mean and sample covariance, the expected cross-term evaluates to zero (i.e., $\mathbb{E}[(\Sigma_{H_t})_{12}^\intercal \bar{\xi}_{H_t,\perp}] = \mathbb{E}[(\Sigma_{H_t})_{12}]^\intercal \mathbb{E}[\bar{\xi}_{H_t,\perp}] = 0$). This yields the decomposition,

$$\mathbb{E}\|\Sigma_{12}\|_2^2 = (1-\alpha)^2\mathbb{E}\|(\Sigma_{H_t})_{12}\|_2^2 + \alpha^2(1-\alpha)^2\mathbb{E}\left[(\bar{\xi}_{H_t,1} - \beta\sigma_{\text{DP}})^2\|\bar{\xi}_{H_t,\perp}\|_2^2\right].$$

Furthermore, $\bar{\xi}_{H_t,1}$ is independent of $\bar{\xi}_{H_t,\perp}$, we have

$$\mathbb{E}\left[(\bar{\xi}_{H_t,1} - \beta\sigma_{\text{DP}})^2\|\bar{\xi}_{H_t,\perp}\|_2^2\right] = \left(\frac{\sigma_{\text{DP}}^2}{n-2f} + \beta^2\sigma_{\text{DP}}^2\right) U_V^2.$$

Additionally,[15]

$$\mathbb{E}\|(\Sigma_{H_t})_{12}\|_2^2 = \frac{n-2f-1}{(n-2f)^2}(d-1)\sigma_{\text{DP}}^4 = \frac{n-2f-1}{n-2f}\sigma_{\text{DP}}^2 U_V^2 \text{ where } U_V^2 := \frac{d-1}{n-2f}\sigma_{\text{DP}}^2.$$

[13]This expectation is derived using the trace trick for quadratic forms. Because the honest variables are zero-mean, $\mathbb{E}[h_{t,1}^\intercal P_{n-f} h_{t,1}] = \text{Tr}(P_{n-f}\Sigma_h) = \text{Tr}(\Sigma_h) - \frac{1}{n-f}\text{Tr}(1_{n-f}1_{n-f}^\intercal\Sigma_h)$, where $\Sigma_h := \mathbb{E}[h_{t,1}h_{t,1}^\intercal]$. The vector $h_{t,1} \in \mathbb{R}^{n-f}$ contains $n-2f$ independent honest coordinates $\sim \mathcal{N}(0,\sigma_{\text{DP}}^2)$ and f zeros. Thus, Σ_h is diagonal, yielding $\text{Tr}(\Sigma_h) = (n-2f)\sigma_{\text{DP}}^2$. Moreover, $\text{Tr}(1_{n-f}1_{n-f}^\intercal\Sigma_h) = 1_{n-f}^\intercal\Sigma_h 1_{n-f} = (n-2f)\sigma_{\text{DP}}^2$. Consequently, $\frac{1}{n-f}\mathbb{E}[h_{t,1}^\intercal P_{n-f} h_{t,1}] = \frac{1}{n-f}\big(n-2f-\frac{n-2f}{n-f}\big)\sigma_{\text{DP}}^2 = \frac{(n-2f)(n-f-1)}{(n-f)^2}\sigma_{\text{DP}}^2$.

[14]Let $X_\perp$ be a submatrix obtained by deleting rows or columns from X. By definition, $\|X\|_{\text{sp}} = \sup_{\|v\|_2=1}\|Xv\|_2$. For any unit vector u in the domain of $X_\perp$, appending zeros to the deleted coordinates yields a unit vector $\tilde{u}$ in the domain of X such that $\|X\tilde{u}\|_2 = \|X_\perp u\|_2$. Taking the supremum over all such vectors guarantees $\|X_\perp\|_{\text{sp}} \le \|X\|_{\text{sp}}$ deterministically. Consequently, $\mathbb{E}[s_{\max}(X_\perp)^2] \le \mathbb{E}[s_{\max}(X)^2]$.

[15]We write $(\Sigma_{H_t})_{12} = \frac{\sigma_{\text{DP}}^2}{n-2f} X_\perp P_{n-2f} X_1$, with the projection matrix $P_{n-2f} := I_{n-2f} - \frac{1}{n-2f}1_{n-2f}1_{n-2f}^\intercal$, and $X_\perp \in \mathbb{R}^{(d-1)\times(n-2f)}$, $X_1 \in \mathbb{R}^{n-2f}$ denote an independent standard normal matrix and vector. Evaluating $\mathbb{E}\|X_\perp P_{n-2f}X_1\|_2^2 = \mathbb{E}[\text{Tr}(XP_{n-2f}\mathbb{E}[X_1X_1^\intercal \mid X_\perp]P_{n-2f}X^\intercal)] = \mathbb{E}[\text{Tr}(X_\perp P_{n-2f}^2 X_\perp^\intercal)] = \text{Tr}(P_{n-2f}\mathbb{E}[X_\perp^\intercal X_\perp]) = \text{Tr}(P_{n-2f}(d-1)I_{n-f}) = (n-2f-1)(d-1)$ yield the result.

Thus,

$$\begin{aligned}\mathbb{E}\|\Sigma_{12}\|_2^2 &= \frac{(n-2f)(n-2f-1)}{(n-f)^2}\sigma_{\text{DP}}^2 U_V^2 + \frac{\alpha^2(1-\alpha)^2}{n-2f}\sigma_{\text{DP}}^2 U_V^2 + \alpha(1-\alpha)U_V^2 c^2\Lambda \\ &= \frac{1-\alpha}{n-f}\left(n-2f-1+\alpha^2\right)\sigma_{\text{DP}}^2 U_V^2 + \alpha(1-\alpha)c^2 U_V^2\Lambda \\ &= (1-\alpha)U_V^2\left[\frac{n-2f-1+\alpha^2}{n-f}\sigma_{\text{DP}}^2 + \alpha c^2\Lambda\right] \\ &= (1-\alpha)U_V^2\left[(1-\alpha)\frac{n-f-1-\alpha}{n-f}\sigma_{\text{DP}}^2 + \alpha c^2\Lambda\right] \\ &\le (1-\alpha)U_V^2\left[(1-\alpha)\frac{n-f-1}{n-f}\sigma_{\text{DP}}^2 + \alpha c^2\Lambda\right] := U_{12}^2 \end{aligned} \tag{36}$$

(iv) **The margin via Sylvester's criterion (fixing c_{opt} and δ).** Let define the following matrices

$$M := \begin{pmatrix} \Sigma_{11} & \|\Sigma_{12}\|_2 \\ \|\Sigma_{12}\|_2 & \lambda_{\max}(\Sigma_{22}) \end{pmatrix}$$

and, where we used (34) (35) (36),

$$M_{\text{exp}} := \begin{pmatrix} U_{11} & U_{12} \\ U_{12} & U_{22} \end{pmatrix} = \begin{pmatrix} \sigma_0^2 + c^2\Lambda & U_{12} \\ U_{12} & r_\Lambda\Lambda \end{pmatrix}.$$

We have

$$\lambda_{\max}(\Sigma_{S_t}) \le \lambda_{\max}(M) = \lambda_{\max}(M_{\text{exp}} + (M - M_{\text{exp}})) \le \lambda_{\max}(M_{\text{exp}}) + \lambda_{\max}(\Delta M),$$

where we define $\Delta M = M - M_{\text{exp}}$. We require $\lambda_{\max}(M_{\text{exp}}) \le \Lambda - 2\delta$ for a chosen margin $\delta > 0$. This is equivalent to requiring the shifted matrix $(\Lambda - 2\delta)I - M_{\text{exp}}$ to be positive semi-definite. By Sylvester's Criterion, a 2×2 symmetric matrix is positive semi-definite if and only if its diagonal terms and its determinant are non-negative

$$\Lambda - U_{11} - 2\delta \ge 0, \tag{37}$$

$$(\Lambda - U_{11} - 2\delta)(\Lambda - U_{22} - 2\delta) \ge U_{12}^2. \tag{38}$$

Satisfying condition (37). Recall that $\Lambda = \frac{\sigma_{\text{DP}}^2}{n-f}\mathbb{E}[s_{\max}(X)^2]$ for a $d \times (n-f-1)$ standard Gaussian matrix X, and $\mathbb{E}[s_{\max}(X)^2] \ge \frac{\mathbb{E}\|X\|_F^2}{\min(d, n-f-1)} = \max(d, n-f-1)$. Hence,

$$(1-\alpha)\Lambda \ge \frac{(1-\alpha)\sigma_{\text{DP}}^2}{n-f}\max(d, n-f-1) = \sigma_0^2 \max(\frac{d}{n-f-1}, 1).$$

Consequently, (37) is possible with the following mild assumptions.

(A) *The Byzantine participant number scale with the network scale*. There exists ι such that $f \ge \iota n$, hence $\alpha \ge \frac{\iota}{1-\iota}$. Consequently, $\frac{\sigma_0^2}{\Lambda} \le (1 - \frac{\iota}{1-\iota}) := \rho_{\max} < 1$.

(B) *Overparametrization regime*. There exists $\eta > 0$ such that $d \ge (1+\eta)(n-f-1)$. Hence $\frac{\sigma_0^2}{\Lambda} \le \frac{(1-\alpha)(n-f-1)}{d} \le \frac{1}{1+\eta} := \rho_{\max} < 1$.

(C) We can use a more refined bound on $\frac{\sigma_0^2}{\Lambda} = (1-\alpha)\frac{n-f-1}{\mathbb{E}[s_{\max}(X)^2]}$. Assume there exists $\nu > 0$ such that $d \ge \nu(n-f-1)$. We prove that, by Vershynin (2018, Chapter 7), there exists a constant $C > 0$ such that $\mathbb{E}[s_{\max}(X)^2] \ge \left(\sqrt{d} + \sqrt{n-f-1} - C\right)^2$. Because of Assumption C.3, $\alpha \ge \alpha_{\min}$ and $\frac{C}{\sqrt{n-f-1}} \le \bar{\epsilon}$. Hence,

$$\frac{\sigma_0^2}{\Lambda} \le \frac{1-\alpha}{\left(\sqrt{\frac{d}{n-f-1}} + 1 - \frac{C}{\sqrt{n-f-1}}\right)^2} \le \frac{1-\alpha_{\min}}{\left(\sqrt{\nu} + 1 - \bar{\epsilon}\right)^2} := \rho_{\max} < 1.$$

Assuming $\rho_{\max} < 1$ exists, and define $c_{\max}^2 := 1 - \rho_{\max} > 0$. Because $c \in (0, c_{\max})$, and to satisfy the condition (37), we fix the margin δ as follows.

$$\delta := c_\delta \Lambda = \frac{\min\{c_{\max}^2 - c^2, 1 - r_\Lambda\}}{4} \Lambda. \tag{39}$$

Substituting this margin into condition (37) yields

$$\Lambda - U_{11} - 2\delta = (1 - c^2)\Lambda - \sigma_0^2 - 2\delta \geq (1 - \rho_{\max} - c^2)\Lambda - 2\delta = \frac{c_{\max}^2 - c^2}{2} \Lambda > 0.$$

Satisfying condition (38). We require

$$(\Lambda - U_{11} - 2\delta)(\Lambda - U_{22} - 2\delta) \geq U_{12}^2.$$

By factoring out Λ, and substituting (34), (35) (39), yields

$$(\Lambda - U_{11} - 2\delta)(\Lambda - U_{22} - 2\delta) \geq \frac{(c_{\max}^2 - c^2)(1 - r_\Lambda)}{4} \Lambda^2. \tag{40}$$

Conversely, we upper bound U_{12}^2 from (36). Because $\Lambda \geq \frac{d}{n-f}\sigma_{\text{DP}}^2$, hence $\sigma_{\text{DP}}^2 \leq \frac{n-f}{d}\Lambda$. Substituting this alongside $U_V^2 \leq \frac{1}{1-\alpha}\Lambda$ yields

$$U_{12}^2 = (1-\alpha)U_V^2 \left[(1-\alpha)\frac{n-f-1}{n-f}\sigma_{\text{DP}}^2 + \alpha c^2 \Lambda\right] \leq \Lambda^2 \left[\alpha c^2 + (1-\alpha)\frac{n-f}{d}\right]. \tag{41}$$

Combining the bounds from (40) and (41) gives the condition

$$\frac{(c_{\max}^2 - c^2)(1 - r_\Lambda)}{4} \geq \alpha c^2 + (1-\alpha)\frac{n-f}{d}.$$

Rearranging,

$$c^2 \left(\alpha + \frac{1 - r_\Lambda}{4}\right) \leq \frac{c_{\max}^2(1 - r_\Lambda)}{4} - (1-\alpha)\frac{n-f}{d}. \tag{42}$$

Consequently, by Vershynin (2018, Chapter 7), combined with standard Gaussian concentration, there exist absolute constants $C_1, C_2 > 0$ such that

$$\begin{aligned} r_\Lambda = \frac{\mathbb{E}[s_{\max}(X_\perp)^2]}{\mathbb{E}[s_{\max}(X)^2]} &\leq \left(\frac{\sqrt{d-1} + \sqrt{n-2f} + C_1}{\sqrt{d} + \sqrt{n-f-1} - C_2}\right)^2 \\ &= \left(\frac{\sqrt{\frac{d-1}{n-f-1}} + \sqrt{\frac{n-2f}{n-f-1}} + \frac{C_1}{\sqrt{n-f-1}}}{\sqrt{\frac{d}{n-f-1}} + 1 - \frac{C_2}{\sqrt{n-f-1}}}\right)^2. \end{aligned}$$

By Assumption C.3, $\sqrt{\frac{d}{n-f-1}} = \sqrt{\frac{d}{n-f}}\sqrt{1 + \frac{1}{n-f-1}} \leq \sqrt{\nu}(1 + \sqrt{\frac{1}{n-f-1}})$, and $\sqrt{\frac{n-2f}{n-f-1}} = \sqrt{\frac{n-2f}{n-f}}\sqrt{1 + \frac{1}{n-f-1}} \leq \sqrt{1-\alpha}(1 + \sqrt{\frac{1}{n-f-1}})$. We aggregate these terms and define the residual $\epsilon_n := \frac{\sqrt{\nu} + \sqrt{1-\alpha} + \max(C_1, C_2)}{\sqrt{n-f-1}}$. Again, by Assumption C.3, because we assume, $n - f - 1 \geq \left(\frac{3(\sqrt{\nu} + \sqrt{1-\alpha_{\min}} + \max\{C_1, C_2\})}{1 - \sqrt{1-\alpha_{\min}}}\right)^2$ we have $\epsilon_n \leq \bar{\epsilon} := \frac{1 - \sqrt{1-\alpha_{\min}}}{3}$. Consequently, because r_Λ decreases with respect to α and increases with respect to the residual ϵ_n, substituting their bound yield

$$\begin{aligned} r_\Lambda = \frac{\mathbb{E}[s_{\max}(X_\perp)^2]}{\mathbb{E}[s_{\max}(X)^2]} &< \left(\frac{\sqrt{\nu} + \sqrt{1-\alpha} + \epsilon_n}{\sqrt{\nu} + 1 - \epsilon_n}\right)^2 \\ &\leq \left(\frac{\sqrt{\nu} + \frac{1}{3} + \frac{2}{3}\sqrt{1-\alpha_{\min}}}{\sqrt{\nu} + \frac{2}{3} + \frac{1}{3}\sqrt{1-\alpha_{\min}}}\right)^2 := r_{\max} \in (0, 1). \end{aligned}$$

Concurrently, satisfying Sylvester's criterion requires

$$c^2\left(\alpha+\frac{1-r_\Lambda}{4}\right)\le\frac{c_{\max}^2(1-r_\Lambda)}{4}-(1-\alpha)\frac{n-f}{d}.$$

This condition is implied when, defining $c_{\text{opt}}^2:=\frac{c_{\max}^2(1-r_{\max})-4\mu^{-1}}{2+(1-r_{\max})}$,

$$c^2\le c_{\text{opt}}^2\le\frac{c_{\max}^2(1-r_\Lambda)-4(1-\alpha)\frac{n-f}{d}}{4\alpha+1-r_\Lambda}.$$

Note that, by Assumption C.3, $\mu>\frac{4}{c_{\max}^2(1-r_{\max})}$ and hence $c_{\text{opt}}^2\in(0,1)$.

(v) **Probabilistic consequence.** Recall that

$$\lambda_{\max}(\Sigma_{S_t})\le\lambda_{\max}(M_{\text{exp}})+\lambda_{\max}(\Delta M),\ \text{with}\ \Delta M=M-M_{\text{exp}}:=\begin{pmatrix}\Delta_{11}&\Delta_{12}\\ \Delta_{12}&\Delta_{22}\end{pmatrix}.$$

We prove $\lambda_{\max}(M_{\text{exp}})\le\Lambda-2\delta$ for $\delta=\frac{\min\{c_{\max}^2-c^2,1-r_\Lambda\}}{4}\Lambda$. We now control the following tail failure events for $t\in\{n_{\text{test}}+1,\dots,T\}$ with high probability

$$E_{2,t}^{\mathsf{c}}:=\{\lambda_{\max}(\Sigma_{\mathcal{H}})\le\Lambda-\delta\},$$
$$E_{3,t}^{\mathsf{c}}:=\{\lambda_{\max}(\Delta M)\ge\delta\}\subset\left\{|\Delta_{11}|\ge\frac{\delta}{4}\right\}\cup\left\{|\Delta_{12}|\ge\frac{\delta}{4}\right\}\cup\left\{|\Delta_{22}|\ge\frac{\delta}{4}\right\},$$

where we used $\lambda_{\max}(\Delta M)\le\|\Delta M\|_{\text{F}}\le\sqrt{\Delta_{11}^2+2\Delta_{12}^2+\Delta_{22}^2}\le|\Delta_{11}|+2|\Delta_{12}|+|\Delta_{22}|$.

High probability bound for $E_{2,t}^{\mathsf{c}}$. Recall that $\Sigma_{\mathcal{H}}$ is identically distributed to $\frac{\sigma_{\text{DP}}^2}{n-f}XX^{\intercal}$, where $X\in\mathbb{R}^{d\times(n-f-1)}$ is a random matrix with i.i.d. standard normal entries $\mathcal{N}(0,1)$. Hence, $\lambda_{\max}(\Sigma_{\mathcal{H}})$ behaves as the squared spectral norm of a Gaussian matrix, scaled by $\frac{\sigma_{\text{DP}}^2}{n-f}$. Because the spectral norm map $X\mapsto s_{\max}(X)$ is 1-Lipschitz continuous, Gaussian Lipschitz concentration (Boucheron et al., 2013, Theorem 5.6) guarantees sub-Gaussian tails for $s_{\max}(X)$. By definition, $\Lambda=\frac{\sigma_{\text{DP}}^2}{n-f}\mathbb{E}[s_{\max}(X)^2]$, hence $E_{2,t}^{\mathsf{c}}$ is equivalent to $s_{\max}(X)^2\le\mathbb{E}[s_{\max}(X)^2]-\delta\frac{n-f}{\sigma_{\text{DP}}^2}$. Moreover, by the Gaussian Poincaré inequality (Boucheron et al., 2013, Section 3.7), the variance of the 1-Lipschitz function is bounded, meaning $\mathbb{E}[s_{\max}(X)^2]\le(\mathbb{E}[s_{\max}(X)])^2+1$. Substituting this yields the implication

$$E_{2,t}^{\mathsf{c}}\implies\left\{s_{\max}(X)\le\mathbb{E}[s_{\max}(X)]-\frac{\delta\frac{n-f}{\sigma_{\text{DP}}^2}-1}{2\mathbb{E}[s_{\max}(X)]}\right\},$$

where we used the inequality $\sqrt{x-y}\le\sqrt{x}-\frac{y}{2\sqrt{x}}$ for $0<y<x$. Because the margin δ scales proportionally with Λ, there exists a constant $\tilde{c}_2\in(0,1)$ such that $\delta\frac{n-f}{\sigma_{\text{DP}}^2}-1\ge\tilde{c}_2\delta\frac{n-f}{\sigma_{\text{DP}}^2}$. Furthermore, applying Jensen's inequality, we have $\mathbb{E}[s_{\max}(X)]\le\sqrt{\mathbb{E}[s_{\max}(X)^2]}=\sqrt{\frac{n-f}{\sigma_{\text{DP}}^2}\Lambda}$. Consequently, the required deviation s from the mean is bounded by

$$s\ge\frac{\tilde{c}_2\delta\frac{n-f}{\sigma_{\text{DP}}^2}}{2\sqrt{\frac{n-f}{\sigma_{\text{DP}}^2}\Lambda}}=\frac{\tilde{c}_2}{2}\delta\sqrt{\frac{n-f}{\sigma_{\text{DP}}^2\Lambda}}.$$

By Gaussian Lipschitz concentration (Boucheron et al., 2013, Theorem 5.6), for any $s>0$ we have $\mathbb{P}(s_{\max}(X)\le\mathbb{E}[s_{\max}(X)]-s)\le e^{-s^2/2}$. That is, there exists a constant $c_2>0$ such that

$$\mathbb{P}\left(E_{2,t}^{\mathsf{c}}\right)\le\mathbb{P}(s_{\max}(X)\le\mathbb{E}[s_{\max}(X)]-s)\le\exp\left(-c_2\frac{\delta^2(n-f)}{\sigma_{\text{DP}}^2\Lambda}\right).$$

High probability bound for $E_{3,t}^{\mathsf{c}}$. Concurrently, we bound the entries of ΔM using concentration argument.

Bounding $\mathbb{P}(|\Delta_{11}| \geq \frac{\delta}{4})$. Recall that $\Delta_{11} = (V_1 - \mathbb{E}[V_1]) - V_2$. By the Hanson-Wright inequality (Rudelson & Vershynin, 2013, Theorem 1.1) applied to V_1, and standard Gaussian tail bounds for the term V_2, we have for universal constants $\tilde{c}_1, \tilde{c}_3 > 0$

$$\mathbb{P}\left(|V_1 - \mathbb{E}[V_1]| \geq \frac{\delta}{8}\right) \leq 2\exp\left(-\tilde{c}_1 \min\left\{\frac{\delta^2(n-f)^2}{64\sigma_{\text{DP}}^4(n-2f)}, \frac{\delta(n-f)}{8\sigma_{\text{DP}}^2}\right\}\right),$$
$$\mathbb{P}\left(|V_2| \geq \frac{\delta}{8}\right) \leq 2\exp\left(-\tilde{c}_3 \frac{\delta^2(n-f)^3}{128f(n-f-1)\beta^2\sigma_{\text{DP}}^4}\right).$$

Hence, by the union bound there exisits $\tilde{c}_4 > 0$ such that

$$\mathbb{P}\left(|\Delta_{11}| \geq \frac{\delta}{4}\right) \leq \mathbb{P}\left(|V_1 - \mathbb{E}[V_1]| \geq \frac{\delta}{8}\right) + \mathbb{P}\left(|V_2| \geq \frac{\delta}{8}\right)$$
$$\leq 4e^{-\tilde{c}_4 \min\left\{\frac{\delta^2(n-f)^2}{64\sigma_{\text{DP}}^4(n-2f)}, \frac{\delta(n-f)}{8\sigma_{\text{DP}}^2}, \frac{\delta^2(n-f)^3}{128f(n-f-1)\beta^2\sigma_{\text{DP}}^4}\right\}}.$$

Bounding $\mathbb{P}(|\Delta_{22}| \geq \frac{\delta}{4})$. We bound the absolute deviation of $\Delta_{22} := \lambda_{\max}(\Sigma_{22}) - U_{22}$. Recall that $\Sigma_{22} = M_{22} - \mu\mu^\intercal$, where $\mu := \frac{1}{n-f}\sum_{i\in H_t} \xi_{t,\perp}^{(i)} = \frac{\sigma_{\text{DP}}}{n-f} X_\perp 1_{n-2f}$, $M_{22} = \frac{\sigma_{\text{DP}}^2}{n-f} X_\perp X_\perp^\intercal$, and by definition, $U_{22} = \mathbb{E}[\lambda_{\max}(M_{22})]$. We recall that $X_\perp \in \mathbb{R}^{(d-1)\times(n-2f)}$ is the matrix with standard normal entries and that $1_{n-2f} \in \mathbb{R}^{n-2f}$ is the all-ones vector.

For the upper tail, because $\mu\mu^\intercal \succeq 0$, we have $\Sigma_{22} \preceq M_{22}$, which implies $\lambda_{\max}(\Sigma_{22}) \leq \lambda_{\max}(M_{22})$. Consequently, $\Delta_{22} \leq \lambda_{\max}(M_{22}) - U_{22}$. By a analogous derivation done for bounding $E_{3,t}^{\mathsf{c}}$ in high probability, there exists an absolute constant $\tilde{c}_5 > 0$ such that

$$\mathbb{P}\left(\Delta_{22} \geq \frac{\delta}{4}\right) \leq \mathbb{P}\left(\lambda_{\max}(M_{22}) - U_{22} \geq \frac{\delta}{4}\right) \leq \exp\left(-\tilde{c}_5 \frac{\delta^2(n-f)}{\sigma_{\text{DP}}^2\Lambda}\right).$$

For the lower tail, because $n - f > n - 2f$, we have $\frac{1}{(n-f)^2} < \frac{1}{(n-f)(n-2f)}$ and we define a surrogate matrix Σ_0 with the larger penalty to obtain the positive-semidefinite ordering

$$\Sigma_{22} \succeq M_{22} - \frac{\sigma_{\text{DP}}^2}{(n-f)(n-2f)} X_\perp 1_{n-2f} 1_{n-2f}^\intercal X_\perp^\intercal$$
$$= \frac{\sigma_{\text{DP}}^2}{n-f} X_\perp \left(I_{n-2f} - \frac{1}{n-2f} 1_{n-2f} 1_{n-2f}^\intercal\right) X_\perp^\intercal := \Sigma_0.$$

The matrix $I_{n-2f} - \frac{1}{n-2f} 1_{n-2f} 1_{n-2f}^\intercal$ is a projection matrix of rank $n - 2f - 1$. Therefore, by the rotational invariance of the Gaussian matrix $X_\perp$, the surrogate matrix Σ_0 is identically distributed to $\frac{\sigma_{\text{DP}}^2}{n-f} \tilde{X}_\perp \tilde{X}_\perp^\intercal$, where $\tilde{X}_\perp \in \mathbb{R}^{(d-1)\times(n-2f-1)}$ consists of i.i.d. standard normal entries. Because $\Sigma_{22} \succeq \Sigma_0$, we have $\lambda_{\max}(\Sigma_{22}) \geq \lambda_{\max}(\Sigma_0)$. Thus, the lower deviation of Σ_{22} is contained within the lower tail of Σ_0. Again, we apply the Gaussian Lipschitz concentration (Boucheron et al., 2013, Theorem 5.6) and analogous derivation as done for $E_{2,t}^{\mathsf{c}}$.

Consequently, combining the upper tail of M_{22} and the lower tail of Σ_0, there exists a constant $\tilde{c}_6 > 0$ such that

$$\mathbb{P}\left(|\Delta_{22}| \geq \frac{\delta}{4}\right) \leq 2\exp\left(-\tilde{c}_6 \frac{\delta^2(n-f)}{\sigma_{\text{DP}}^2\Lambda}\right).$$

Bounding $\mathbb{P}(|\Delta_{12}| \geq \frac{\delta}{4})$. We bound the absolute deviation $|\Delta_{12}| = |\|\Sigma_{12}\|_2 - U_{12}|$. To map this to concentration inequalities, we transition to the squared norm and prove that it is sufficient to control the deviation $|\|\Sigma_{12}\|_2^2 - \mathbb{E}\|\Sigma_{12}\|_2^2|$. From the expansion in (36), we have $U_{12}^2 = \mathbb{E}\|\Sigma_{12}\|_2^2 + \frac{\alpha(1-\alpha)^2}{n-f}\sigma_{\text{DP}}^2 U_V^2$. For readibility, define $G := \frac{\alpha(1-\alpha)^2}{n-f}\sigma_{\text{DP}}^2 U_V^2 = \frac{\alpha(1-\alpha)^2}{n-f}\frac{d-1}{n-2f}\sigma_{\text{DP}}^4$. Squaring the bounds of the absolute deviation event $\{|\|\Sigma_{12}\|_2 - U_{12}| \geq \delta/4\}$, we have the following.

– *The upper tail* $\|\Sigma_{12}\|_2 \geq U_{12} + \frac{\delta}{4}$. The squared upper tail, requires $\|\Sigma_{12}\|_2^2 \geq U_{12}^2 + \frac{\delta}{2}U_{12} + \frac{\delta^2}{16}$. Define $\tau := \frac{\delta U_{12}}{8}$, we have

$$\begin{aligned}\mathbb{P}\left(\|\Sigma_{12}\|_2 \geq U_{12} + \frac{\delta}{4}\right) &= \mathbb{P}\left(\|\Sigma_{12}\|_2^2 \geq U_{12}^2 + \frac{\delta}{2}U_{12} + \frac{\delta^2}{16}\right) \\ &= \mathbb{P}\left(\|\Sigma_{12}\|_2^2 \geq \mathbb{E}\|\Sigma_{12}\|_2^2 + \tau + \left(G + \frac{3\delta U_{12}}{8} + \frac{\delta^2}{16}\right)\right) \\ &\geq \mathbb{P}\left(\|\Sigma_{12}\|_2^2 \geq \mathbb{E}\|\Sigma_{12}\|_2^2 + \tau\right).\end{aligned}$$

– *The lower tail* $\|\Sigma_{12}\|_2 \leq U_{12} - \frac{\delta}{4}$. First, note that if $U_{12} < \frac{\delta}{4}$ the probability of the event is 0. We only need to control the regime where $U_{12} > \frac{\delta}{4}$, which is equivalent to $2\tau = \frac{\delta U_{12}}{4} \geq \frac{\delta^2}{16}$. Under this condition, the squared lower tail event write

$$\left\{\|\Sigma_{12}\|_2^2 \leq U_{12}^2 - \frac{\delta}{2}U_{12} + \frac{\delta^2}{16} = \mathbb{E}\|\Sigma_{12}\|_2^2 - \tau - (2\tau - \frac{\delta^2}{16}) - (\tau - G)\right\}$$

Hence, if $G \leq \tau$, which we prove below, we have

$$\begin{aligned}\mathbb{P}\left(\|\Sigma_{12}\|_2 \leq U_{12} - \frac{\delta}{4}\right) &= \mathbb{P}\left(\|\Sigma_{12}\|_2^2 \leq U_{12}^2 - \frac{\delta}{2}U_{12} + \frac{\delta^2}{16}\right) \\ &= \mathbb{P}\left(\|\Sigma_{12}\|_2^2 \leq \mathbb{E}\|\Sigma_{12}\|_2^2 - \tau - (2\tau - \frac{\delta^2}{16}) - (\tau - G)\right) \\ &\geq \mathbb{P}\left(\|\Sigma_{12}\|_2^2 \leq \mathbb{E}\|\Sigma_{12}\|_2^2 - \tau\right).\end{aligned}$$

– *Ensuring* $G \leq \tau$. Recall that $G^2 = \frac{\alpha^2(1-\alpha)^4}{(n-f)^2}\sigma_{\text{DP}}^4 U_V^4$, where $U_V^2 = \frac{d-1}{n-2f}\sigma_{\text{DP}}^2$, and

$$U_{12}^2 = (1-\alpha)U_V^2\left[(1-\alpha)\frac{n-f-1}{n-f}\sigma_{\text{DP}}^2 + \alpha c^2\Lambda\right].$$

Hence, the requirement $G \leq \tau = \frac{c_\delta \Lambda U_{12}}{8}$ is equivalent to

$$\frac{\alpha^2(1-\alpha)^3}{(n-f)^2}\sigma_{\text{DP}}^4 U_V^2 \leq \left(\frac{c_\delta}{8}\right)^2 \Lambda^2\left[(1-\alpha)\frac{n-f-1}{n-f}\sigma_{\text{DP}}^2 + \alpha c^2\Lambda\right].$$

Applying (30) and Assumption C.3, we have

$$\Lambda \geq c_\Lambda^2 \frac{d}{n-f}\sigma_{\text{DP}}^2\left(1 + \sqrt{\frac{n-f-1}{d}}\right)^2 \geq \frac{c_\Lambda^2\mu}{\min(1,\nu)}\sigma_{\text{DP}}^2.$$

Substituting this lower bound and U_V^2, the condition $G \leq \tau$ is implied by

$$\frac{\alpha^2(1-\alpha)^2}{n-f} \leq d\left(\frac{c_\delta c_\Lambda^2}{8\min(1,\nu)}\right)^2\left[(1-\alpha)\frac{n-f-1}{n-f} + \alpha\frac{\mu c^2 c_\Lambda^2}{\min(1,\nu)}\right].$$

The following is stronger inequality, as $d \geq \mu(n-f)$,

$$\mu\left[(1-\alpha)(n-f-1)(n-f) + \alpha\frac{\mu c^2 c_\Lambda^2}{\min(1,\nu)}(n-f)^2\right] \geq \left(\frac{8\min(1,\nu)}{c_\delta c_\Lambda^2}\right)^2 \alpha^2(1-\alpha)^2.$$

Because $(n-f)^2 \geq (n-f)(n-f-1) \geq (n-f-1)^2$, we also have the condition

$$\mu(n-f-1)^2\left[1 - \alpha + \alpha\frac{\mu c^2 c_\Lambda^2}{\min(1,\nu)}\right] \geq \left(\frac{8\min(1,\nu)}{c_\delta c_\Lambda^2}\right)^2 \alpha^2(1-\alpha)^2.$$

The following is stronger inequality

$$(n-f-1)^2\left[\frac{1}{2} + \frac{\alpha_{\min}\mu c^2 c_\Lambda^2}{\min(1,\nu)}\right] \geq \left(\frac{2\min(1,\nu)}{\sqrt{\mu}c_\delta c_\Lambda^2}\right)^2.$$

Hence

$$n - f - 1 \geq \frac{2\min(1,\nu)}{c_\delta c_\Lambda^2 \sqrt{\mu}\sqrt{\frac{1}{2} + \frac{\alpha_{\min}\mu c^2 c_\Lambda^2}{\min(1,\nu)}}}.$$

Using the fact that $\frac{1}{2} \leq c_\Lambda \leq \sqrt{\frac{5}{4}}$, we have

$$n - f - 1 \geq \frac{16\min(1,\nu)}{c_\delta \sqrt{\mu}\sqrt{2 + \frac{\alpha_{\min}\mu c^2}{\min(1,\nu)}}},$$

which is ensured by Assumption C.3.

Consequently, we prove $\mathbb{P}(|\Delta_{12}| \geq \frac{\delta}{4}) \geq \mathbb{P}(|\|\Sigma_{12}\|_2^2 - \mathbb{E}\|\Sigma_{12}\|_2^2| \geq \tau)$. We next control the right hand side.

Bounding $\mathbb{P}(|\|\Sigma_{12}\|_2^2 - \mathbb{E}\|\Sigma_{12}\|_2^2| \geq \tau)$. Recall the cross-covariance decomposition

$$\Sigma_{12} = (1-\alpha)(\Sigma_{H_t})_{12} + \alpha(1-\alpha)(\bar{\xi}_{H_t,1} - \beta\sigma_{\text{DP}})\bar{\xi}_{H_t,\perp}.$$

Both terms are sums of scaled products of independent standard Gaussian vectors and matrices. Indeed, for the first term, we use the centering projection $P_{n-2f} := I_{n-2f} - \frac{1}{n-2f}1_{n-2f}1_{n-2f}^\intercal$. Let $X_\perp \in \mathbb{R}^{(d-1)\times(n-2f)}$ and $X_1 \in \mathbb{R}^{n-2f}$ denote an independent standard normal matrix and vector. We can write

$$(\Sigma_{H_t})_{12} = \frac{\sigma_{\text{DP}}^2}{n-2f} X_\perp \left(I_{n-2f} - \frac{1}{n-2f}1_{n-2f}1_{n-2f}^\intercal\right) X_1 = \frac{\sigma_{\text{DP}}^2}{n-2f} X_\perp P_{n-2f} X_1.$$

For the second term, we have $\bar{\xi}_{H_t,1} \sim \mathcal{N}(0, \frac{\sigma_{\text{DP}}^2}{n-2f})$ and $\bar{\xi}_{H_t,\perp} \sim \mathcal{N}(0, \frac{\sigma_{\text{DP}}^2}{n-2f} I_{d-1})$. Reparameterizing with independent standard normal variable $z_1 \sim \mathcal{N}(0,1)$ and $\bar{\xi}_{H_t,\perp} = \frac{\sigma_{\text{DP}}}{n-2f} X_\perp 1_{n-2f}$, where $\frac{X_\perp 1_{n-2f}}{\sqrt{n-2f}} \sim \mathcal{N}(0, I_{d-1})$, we have

$$(\bar{\xi}_{H_t,1} - \beta\sigma_{\text{DP}}) = \frac{\sigma_{\text{DP}}^2}{(n-2f)^{\frac{3}{2}}}(z_1 - \beta\sqrt{n-2f}) X_\perp 1_{n-2f}.$$

Consequently, we have

$$\begin{aligned}
\Sigma_{12} &= \frac{(1-\alpha)\sigma_{\text{DP}}^2}{n-2f} X_\perp P_{n-2f} X_1 + \frac{\alpha(1-\alpha)\sigma_{\text{DP}}^2}{(n-2f)^{\frac{3}{2}}}\left(z_1 + \beta\sqrt{n-2f}\right) X_\perp 1_{n-2f}, \\
&= X_\perp \underbrace{\left[\frac{(1-\alpha)\sigma_{\text{DP}}^2}{n-2f} P_{n-2f} X_1 + \frac{\alpha(1-\alpha)\sigma_{\text{DP}}^2}{(n-2f)^{\frac{3}{2}}}(z_1 - \beta\sqrt{n-2f}) 1_{n-2f}\right]}_{:= v_{\text{cond}}}.
\end{aligned}$$

We condition on the σ-algebra $\mathcal{F}_1 := \sigma(X_1, z_1)$. Given $\mathcal{F}_1$, the vector $v_{\text{cond}} \in \mathbb{R}^{n-2f}$ is deterministic. Because $X_\perp$ consists of independent standard Gaussians, right-multiplying by v_{cond} guarantees $\Sigma_{12} \mid \mathcal{F}_1 \sim \mathcal{N}(0, \|v_{\text{cond}}\|_2^2 I_{d-1})$. The squared norm of such combination is $\|a(P_{n-2f}X_1) + b(1_{n-2f})\|_2^2 = a^2\|P_{n-2f}X_1\|_2^2 + b^2(n-2f) + 2abX_1^\intercal P_{n-2f}^\intercal 1_{n-2f} = a^2\|P_{n-2f}X_1\|_2^2 + b^2(n-2f)$.[16] Consequently, the conditional variance simplifies to

$$\sigma_{\text{cond}}^2 := \|v_{\text{cond}}\|_2^2 = \left(\frac{(1-\alpha)\sigma_{\text{DP}}^2}{n-2f}\right)^2 \|P_{n-2f}X_1\|_2^2 + \left(\frac{\alpha(1-\alpha)\sigma_{\text{DP}}^2}{n-2f}\right)^2 (z_1 - \beta\sigma_{\text{DP}})^2.$$

As a result, we have the following distributional equality $\|\Sigma_{12}\|_2^2 \mid \mathcal{F}_1 \overset{d}{=} \sigma_{\text{cond}}^2 W$, where $W \sim \chi_{d-1}^2$. Crucially, the standard χ_{d-1}^2 distribution of W originates from $X_\perp$ and is hence independent of $\mathcal{F}_1$. By the law of total probability, integrating out $\mathcal{F}_1$ proves that the unconditional distribution of $\|\Sigma_{12}\|_2^2$ is equivalent to the product of the two independent

[16] $P_{n-2f}^\intercal 1_{n-2f} = 1_{n-2f} - \frac{1}{n-2f}1_{n-2f}1_{n-2f}^\intercal 1_{n-2f} = 0$.

random variables, $\|\Sigma_{12}\|_2^2 \overset{d}{=} \sigma_{\text{cond}}^2 W$. Thus, $\mathbb{E}\|\Sigma_{12}\|_2^2 = \mathbb{E}[\sigma_{\text{cond}}^2 W] = \mathbb{E}[\sigma_{\text{cond}}^2]\mathbb{E}[W]$. We bound the product deviation using the identity

$$\begin{aligned}\|\Sigma_{12}\|_2^2 - \mathbb{E}\|\Sigma_{12}\|_2^2 &= \sigma_{\text{cond}}^2 W - \mathbb{E}[\sigma_{\text{cond}}^2]\mathbb{E}[W] \\ &= (\sigma_{\text{cond}}^2 - \mathbb{E}[\sigma_{\text{cond}}^2])(W - \mathbb{E}[W]) \\ &\quad + (\sigma_{\text{cond}}^2 - \mathbb{E}[\sigma_{\text{cond}}^2])\mathbb{E}[W] + \mathbb{E}[\sigma_{\text{cond}}^2](W - \mathbb{E}[W]).\end{aligned}$$

We define the intersection of independent high-probability events

$$\mathcal{E} := \{|\sigma_{\text{cond}}^2 - \mathbb{E}[\sigma_{\text{cond}}^2]| \le t_A\} \cap \{|W - \mathbb{E}[W]| \le t_B\}.$$

Define $t_A = \min\{\frac{\tau}{3\mathbb{E}[W]}, \sqrt{\frac{\tau}{3}}\}$ and $t_B = \min\{\frac{\tau}{3\mathbb{E}[\sigma_{\text{cond}}^2]}, \sqrt{\frac{\tau}{3}}\}$. By the triangle inequality, conditioned on $\mathcal{E}$, we have

$$|\|\Sigma_{12}\|_2^2 - \mathbb{E}\|\Sigma_{12}\|_2^2| \le t_A t_B + t_A\mathbb{E}[W] + t_B\mathbb{E}[\sigma_{\text{cond}}^2] \le \tau.$$

Consequently,

$$\mathbb{P}(|\Delta_{12}| \ge \frac{\delta}{4}) \le \mathbb{P}(|\|\Sigma_{12}\|_2^2 - \mathbb{E}\|\Sigma_{12}\|_2^2| \ge \tau) \le \mathbb{P}(\mathcal{E}).$$

Bounding $\mathbb{P}(|W - \mathbb{E}[W]| \ge t_B)$. We evaluate the concentration of the standard χ_{d-1}^2 variable W. We apply the Hanson-Wright inequality (Vershynin, 2018, Theorem 6.2.2),[17] there exists $C_B > 0$ such that

$$\mathbb{P}(|W - \mathbb{E}[W]| \le t_B) \le 2\exp\left(-C_B \min\left\{\frac{t_B^2}{(d-1)}, t_B\right\}\right). \tag{43}$$

Substituting $t_B = \min\{\frac{\tau}{3\mathbb{E}[\sigma_{\text{cond}}^2]}, \sqrt{\frac{\tau}{3}}\}$ yields two possible regimes for the tail. Because $\mathbb{E}[\sigma_{\text{cond}}^2]\mathbb{E}[W] = \mathbb{E}\|\Sigma_{12}\|_2^2 \le U_{12}^2$ and $\mathbb{E}[W] = d-1$, we have $\frac{\tau}{3\mathbb{E}[\sigma_{\text{cond}}^2]} \ge \frac{\delta U_{12}/8}{3U_{12}^2/(d-1)} = \frac{\delta(d-1)}{24U_{12}}$. Consequently, the exponent in (43) is lower-bounded by

$$E_B := C_B \min\left\{\frac{\delta^2(d-1)}{576U_{12}^2}, \frac{\delta U_{12}}{24(d-1)}, \frac{\delta(d-1)}{24U_{12}}, \sqrt{\frac{\delta U_{12}}{24}}\right\}.$$

Bounding $\mathbb{P}(|\sigma_{cond}^2 - \mathbb{E}[\sigma_{cond}^2]| \ge t_A)$. We expand σ_{cond}^2 as the sum of a centered quadratic form and a linear Gaussian term. Let $v_1 := \frac{(1-\alpha)\sigma_{\text{DP}}^2}{n-2f}$ and $v_2 := \frac{\alpha(1-\alpha)\sigma_{\text{DP}}^2}{n-2f}$. We define $V := [X_1; z_1] \in \mathbb{R}^{n-2f+1}$ and $M_A := \text{diag}(v_1^2 P_{n-2f}, v_2^2)$. The non-centrality stems from $(z_1 - \beta\sqrt{n-2f})^2 = z_1^2 - 2\beta\sqrt{n-2f}z_1 + \beta^2(n-2f)$. Thus,

$$\sigma_{\text{cond}}^2 - \mathbb{E}[\sigma_{\text{cond}}^2] = \underbrace{V^\top M_A V - \mathbb{E}[V^\top M_A V]}_{:=Q} \underbrace{-2c_2^2\beta\sqrt{n-2f}z_1}_{:=L}.$$

To bound $\mathbb{P}(|Q+L| \ge t_A)$, we apply the union bound $\mathbb{P}(|Q| \ge t_A/2) + \mathbb{P}(|L| \ge t_A/2)$. For the quadratic form Q, we apply the Hanson-Wright inequality, with $\|M_A\|_{\text{F}}^2 = v_1^4(n-2f-1) + v_2^4$ and $\|M_A\|_{\text{sp}} = v_1^2$. Substituting $t_A = \min\{\frac{\tau}{3(d-1)}, \sqrt{\frac{\tau}{3}}\} = \min\{\frac{\delta U_{12}}{24(d-1)}, \sqrt{\frac{\delta U_{12}}{24}}\}$, the exponent for $\mathbb{P}(|Q| \ge t_A/2)$ is

$$E_Q := C_A \min\left\{\frac{\delta^2 U_{12}^2}{2304\|M_A\|_F^2(d-1)^2}, \frac{\delta U_{12}}{96\|M_A\|_F^2}, \frac{\delta U_{12}}{48v_1^2(d-1)}, \frac{\sqrt{\delta U_{12}}}{2\sqrt{24}v_1^2}\right\}.$$

For $L \sim \mathcal{N}(0, 4v_2^4\beta^2(n-2f))$, we apply the Gaussian tail bound $\exp(-\frac{(t_A/2)^2}{2\sigma_L^2})$,

$$E_L := \min\left\{\frac{\delta^2 U_{12}^2}{18432v_2^4\beta^2(n-2f)(d-1)^2}, \frac{\delta U_{12}}{768v_2^4\beta^2(n-2f)}\right\}.$$

[17] With the theorem notation, we have $A = I_{d-1}$, hence $\|I_{d-1}\|_{\text{F}}^2 = d-1$ and $\|I_{d-1}\|_{\text{sp}} = 1$. Moreover the sub-Gaussian norm K is a constant for standard normals, hence we absorb it in the constants.

Wrapping-up. Applying the union bound, we have

$$\begin{aligned}\mathbb{P}\left(|\Delta_{12}| \geq \frac{\delta}{4}\right) &\leq \mathbb{P}(|W - \mathbb{E}[W]| \geq t_B) + \mathbb{P}(|Q| \geq t_A/2) + \mathbb{P}(|L| \geq t_A/2) \\ &\leq 2e^{-E_B} + 2e^{-E_Q} + 2e^{-E_L} \\ &\leq 6\exp\left(-\min\{E_B, E_Q, E_L\}\right).\end{aligned}$$

Wrapping the high probability bounds. Crucially, conditioned on the intersection of the success events ($E_{2,t} \cap E_{3,t}$), we have

$$\lambda_{\max}(\Sigma_{S_t}) \leq \lambda_{\max}(M_{\text{exp}}) + \delta \leq (\Lambda - 2\delta) + \delta = \Lambda - \delta < \lambda_{\max}(\Sigma_{\mathcal{H}}).$$

This proves that the attack systematically lures SMEA with high probability, masking their attack beneath the honest spectral norm. Finally, we search for the bottleneck exponent

$$E_{\text{SMEA}} := \min\{E_2, E_{11}, E_{22}, E_{12}\},$$

where

$$\begin{aligned}E_2 &:= c_2 \frac{\delta^2(n-f)}{\sigma_{\text{DP}}^2 \Lambda}, \\ E_{11} &:= \tilde{c}_4 \min\left\{\frac{\delta^2(n-f)^2}{64\sigma_{\text{DP}}^4(n-2f)}, \frac{\delta(n-f)}{8\sigma_{\text{DP}}^2}, \frac{\delta^2(n-f)^3}{128f(n-f-1)\beta^2\sigma_{\text{DP}}^4}\right\}, \\ E_{22} &:= \tilde{c}_5 \frac{\delta^2(n-f)}{\sigma_{\text{DP}}^2 \Lambda}, \\ E_{12} &:= \min\{E_B, E_Q, E_L\},\end{aligned}$$

Recall that

$$\mathbb{P}(E_{2,t}^{\mathsf{c}}) \leq e^{-E_2},$$

and

$$E_{3,t}^{\mathsf{c}} \subset \left\{|\Delta_{11}| \geq \frac{\delta}{4}\right\} \cup \left\{|\Delta_{22}| \geq \frac{\delta}{4}\right\} \cup \left\{|\Delta_{12}| \geq \frac{\delta}{4}\right\}.$$

Applying the union bound yields

$$\mathbb{P}(E_{3,t}^{\mathsf{c}}) \leq 4e^{-E_{11}} + 2e^{-E_{22}} + 6e^{-E_{12}}.$$

Consequently, we finally have

$$\mathbb{P}(E_{2,t}^{\mathsf{c}} \cap E_{2,t}^{\mathsf{c}}) \geq 1 - \mathbb{P}(E_{2,t}^{\mathsf{c}}) - \mathbb{P}(E_{3,t}^{\mathsf{c}}) \geq 1 - 13e^{-E_{\text{SMEA}}}.$$

The scaling of E_{SMEA}. Observe that, substituting $\delta = c_\delta \Lambda$, all exponents are proportional to

$$R := \frac{\Lambda(n-f)}{\sigma_{\text{DP}}^2},$$

where, using (30) and Assumption C.3 gives

$$R = c_\Lambda^2 \left(\sqrt{d} + \sqrt{n-f-1}\right)^2 \geq c_\Lambda^2 \max\{(1+\sqrt{\mu})^2(n-f-1), (1+\sqrt{\max\{\nu^{-1}-1, 0\}})^2 d\}. \tag{44}$$

Indeed, $E_2 = c_2 c_\delta^2 R$, $E_{22} = \tilde{c}_5 c_\delta^2 R$. Moreover, using (28) the definition of β,

$$E_{11} = \tilde{c}_4 \min\left\{\frac{c_\delta^2}{64}\frac{1}{(1-\alpha)}\frac{R}{n-f}, \frac{c_\delta}{8}, \frac{c_\delta^2}{128c^2}(1-\alpha)\frac{n-f}{n-f-1}\right\} R.$$

Note that $1-\alpha > \frac{1}{2}$, $\frac{1}{1-\alpha} \geq \frac{1}{1-\alpha_{\min}}$, $\frac{R}{n-f} = \frac{\Lambda}{\sigma_{\text{DP}}^2} \geq c_\Lambda^2 \frac{d}{n-f} \geq c_\Lambda^2 \mu$, hence

$$E_{11} \geq \frac{\tilde{c}_4 c_\delta}{8} \min\left\{\frac{c_\delta c_\Lambda^2 \mu}{8(1-\alpha_{\min})}, 1, \frac{c_\delta}{32c^2}\right\} R.$$

For $E_{12} = \min\{E_B, E_Q, E_L\}$, to recover $E_{12} \in \Omega(R)$, we actually require to revisit how we fix t_A and t_B as we treat them symmetrically but the yield different concentration. Indeed, fixing $\frac{t_B^2}{d-1} = \frac{t_A^2}{\|M_A\|_{\text{F}}}$ and $t_A t_B = \frac{\tau}{3}$ yields $E_{12} \in \Omega(R)$.

Finally, as explained in the beginning of the paragraph, under E_1 and the trajectory produced by S' (i.e., with $g_t^{(1)\prime} = \frac{1}{m}\sum_{j=1}^{m} z^{(1,j)\prime} + \xi_t^{(1)} = -\frac{C}{m}e_1 + \xi_t^{(1)}$), we do an analogous analysis as above, in the opposite direction (i.e., $-e_1$). That is, due to Algorithm 3, the Byzantine participants successfully steer the optimization trajectory in the negative direction. Otherwise, in the complementary event E_1^{c}, the influence of Byzantine participants is considered negligible, and any potential contribution to instability from this case is disregarded.

Consequence for the lower bound. We proceed analogously to the corresponding paragraph in the proof of the case $d = 1$, Theorem 3.1. Indeed,

$$\mathbb{P}(E) \geq \mathbb{P}(E_1)\left(1 - 13e^{-E_{\text{SMEA}}}\right)^{T-n_{\text{test}}} \in \Omega(\mathbb{P}(E_1)) \text{ if on only if } T \in \mathcal{O}\left(e^{E_{\text{SMEA}}}\right), \tag{45}$$

which is guaranteed by Assumption C.3. Moreover, because the dataset shift $-\frac{C}{m}e_1$ and the Byzantine vectors $\pm\beta\sigma_{\text{DP}}e_1$ are aligned along e_1, the expected parameter divergence can be evaluated by its marginal drift along e_1, yielding

$$\|\mathbb{E}\left[\theta_T - \theta'_T \mid E\right]\|_2 = \sup_{u\in\mathbb{S}^{d-1}} \mathbb{E}\left[(\theta_T - \theta'_T)^\intercal u \mid E\right] \geq |\mathbb{E}\left[(\theta_T - \theta'_T)^\intercal e_1 \mid E\right]|.$$

We then bound $|\mathbb{E}\left[|(\theta_T - \theta'_T)^\intercal e_1|E\right]|$ analogously to the case $d = 1$, Theorem 3.1.

Finally, the "**Gluing lower bounds**" argument follows exactly from the one in Theorem 3.1. □

D Experimental evaluation details of section 4

We report table of the expected train and test accuracies under varying noise multipliers (σ) and Byzantine participants (f) shown in the main text. They are reported as mean ± standard error.

Table 1: Figure 3 (a) expected train and test accuracies under varying noise multipliers (σ) and Byzantine participants (f). Reported as mean ± standard error.

Noise (σ)	$f = 1$		$f = 2$		$f = 3$	
	Train Acc.	Test Acc.	Train Acc.	Test Acc.	Train Acc.	Test Acc.
32	$10.41_{\pm 0.18}$	$10.44_{\pm 0.17}$	$9.74_{\pm 0.17}$	$9.72_{\pm 0.17}$	$9.60_{\pm 0.22}$	$9.67_{\pm 0.20}$
16	$12.30_{\pm 0.23}$	$12.24_{\pm 0.21}$	$9.60_{\pm 0.21}$	$9.61_{\pm 0.19}$	$9.27_{\pm 0.17}$	$9.34_{\pm 0.18}$
8	$16.22_{\pm 0.22}$	$16.04_{\pm 0.21}$	$9.43_{\pm 0.27}$	$9.47_{\pm 0.26}$	$8.88_{\pm 0.21}$	$8.93_{\pm 0.20}$
4	$22.37_{\pm 0.19}$	$21.93_{\pm 0.19}$	$8.10_{\pm 0.22}$	$8.05_{\pm 0.22}$	$6.56_{\pm 0.16}$	$6.68_{\pm 0.17}$
2	$31.83_{\pm 0.22}$	$31.02_{\pm 0.20}$	$6.67_{\pm 0.19}$	$6.73_{\pm 0.18}$	$2.69_{\pm 0.11}$	$2.85_{\pm 0.11}$
1	$44.43_{\pm 0.19}$	$43.00_{\pm 0.17}$	$5.01_{\pm 0.17}$	$5.12_{\pm 0.18}$	$2.14_{\pm 0.08}$	$2.24_{\pm 0.09}$
0.5	$54.13_{\pm 0.10}$	$52.13_{\pm 0.10}$	$4.26_{\pm 0.25}$	$4.34_{\pm 0.24}$	$1.44_{\pm 0.09}$	$1.52_{\pm 0.09}$
0.25	$59.23_{\pm 0.09}$	$56.86_{\pm 0.07}$	$4.35_{\pm 0.28}$	$4.41_{\pm 0.28}$	$0.79_{\pm 0.04}$	$0.90_{\pm 0.05}$
0.125	$61.22_{\pm 0.06}$	$58.74_{\pm 0.06}$	$11.30_{\pm 0.57}$	$11.12_{\pm 0.55}$	$0.47_{\pm 0.03}$	$0.53_{\pm 0.02}$

Table 2: Figure 3 (b) expected train and test accuracy under varying noise multipliers and heterogeneity $\alpha = \beta$ Li et al. (2020). Reported as mean $\pm$ standard srror.

Noise (σ)	$\alpha = 1$		$\alpha = 5$		$\alpha = 10$	
	Train Acc.	Test Acc.	Train Acc.	Test Acc.	Train Acc.	Test Acc.
32	$12.31_{\pm 0.46}$	$12.34_{\pm 0.46}$	$12.70_{\pm 0.67}$	$12.69_{\pm 0.67}$	$17.97_{\pm 1.19}$	$17.94_{\pm 1.19}$
16	$19.49_{\pm 0.61}$	$19.52_{\pm 0.62}$	$19.26_{\pm 0.70}$	$19.21_{\pm 0.70}$	$41.41_{\pm 1.18}$	$41.38_{\pm 1.17}$
8	$38.66_{\pm 0.71}$	$38.61_{\pm 0.71}$	$33.04_{\pm 0.70}$	$32.95_{\pm 0.70}$	$73.72_{\pm 1.02}$	$73.62_{\pm 1.01}$
4	$62.93_{\pm 0.52}$	$62.87_{\pm 0.53}$	$49.58_{\pm 0.65}$	$49.51_{\pm 0.65}$	$84.97_{\pm 0.44}$	$84.87_{\pm 0.43}$
2	$83.49_{\pm 0.37}$	$83.43_{\pm 0.37}$	$66.02_{\pm 0.65}$	$65.92_{\pm 0.66}$	$89.12_{\pm 0.18}$	$88.99_{\pm 0.17}$
1	$90.16_{\pm 0.08}$	$90.07_{\pm 0.07}$	$74.34_{\pm 0.44}$	$74.19_{\pm 0.43}$	$90.83_{\pm 0.03}$	$90.69_{\pm 0.03}$
0.5	$90.99_{\pm 0.04}$	$90.83_{\pm 0.04}$	$76.22_{\pm 0.28}$	$75.97_{\pm 0.28}$	$90.83_{\pm 0.01}$	$90.70_{\pm 0.01}$
0.25	$91.40_{\pm 0.05}$	$91.11_{\pm 0.04}$	$75.49_{\pm 0.17}$	$75.26_{\pm 0.17}$	$90.73_{\pm 0.00}$	$90.61_{\pm 0.00}$
0.125	$91.61_{\pm 0.03}$	$91.25_{\pm 0.03}$	$73.84_{\pm 0.23}$	$73.66_{\pm 0.22}$	$90.71_{\pm 0.00}$	$90.59_{\pm 0.00}$

Table 3: Figure 3 (c) expected train and test accuracies (%) under varying noise multipliers (σ) and dimensionality (d). These values, reported as mean $\pm$ standard error, are from Figure 3, right panel.

Noise (σ)	$d = 12$		$d = 25$		$d = 50$	
	Train Acc.	Test Acc.	Train Acc.	Test Acc.	Train Acc.	Test Acc.
32	$10.77_{\pm 0.21}$	$10.73_{\pm 0.21}$	$12.41_{\pm 0.32}$	$12.32_{\pm 0.32}$	$10.41_{\pm 0.18}$	$10.44_{\pm 0.17}$
16	$12.98_{\pm 0.22}$	$12.95_{\pm 0.22}$	$18.43_{\pm 0.39}$	$18.32_{\pm 0.40}$	$12.30_{\pm 0.23}$	$12.24_{\pm 0.21}$
8	$16.44_{\pm 0.29}$	$16.27_{\pm 0.27}$	$29.03_{\pm 0.44}$	$28.67_{\pm 0.45}$	$16.22_{\pm 0.22}$	$16.04_{\pm 0.21}$
4	$23.51_{\pm 0.32}$	$23.28_{\pm 0.31}$	$43.21_{\pm 0.44}$	$42.52_{\pm 0.45}$	$22.37_{\pm 0.19}$	$21.93_{\pm 0.19}$
2	$34.67_{\pm 0.33}$	$34.22_{\pm 0.32}$	$59.17_{\pm 0.27}$	$58.13_{\pm 0.28}$	$31.83_{\pm 0.22}$	$31.02_{\pm 0.20}$
1	$48.85_{\pm 0.23}$	$48.16_{\pm 0.23}$	$71.27_{\pm 0.10}$	$70.01_{\pm 0.11}$	$44.43_{\pm 0.19}$	$43.00_{\pm 0.17}$
0.5	$57.68_{\pm 0.14}$	$56.93_{\pm 0.14}$	$75.20_{\pm 0.06}$	$73.90_{\pm 0.06}$	$54.13_{\pm 0.10}$	$52.13_{\pm 0.10}$
0.25	$60.80_{\pm 0.10}$	$60.11_{\pm 0.09}$	$76.14_{\pm 0.04}$	$74.82_{\pm 0.03}$	$59.23_{\pm 0.09}$	$56.86_{\pm 0.07}$
0.125	$62.25_{\pm 0.07}$	$61.51_{\pm 0.07}$	$76.60_{\pm 0.03}$	$75.25_{\pm 0.03}$	$61.22_{\pm 0.06}$	$58.74_{\pm 0.06}$

Table 4: Figure 4 expected train and test accuracy under varying noise multipliers. Reported as mean $\pm$ standard error across data splits.

Noise (σ)	Train Acc. (%)	Test Acc. (%)
32	$10.15_{\pm 0.10}$	$10.12_{\pm 0.13}$
16	$10.37_{\pm 0.20}$	$10.30_{\pm 0.19}$
8	$14.52_{\pm 0.30}$	$14.10_{\pm 0.22}$
4	$31.15_{\pm 0.56}$	$29.95_{\pm 0.38}$
2	$54.88_{\pm 0.51}$	$52.65_{\pm 0.38}$
1	$72.59_{\pm 0.32}$	$70.00_{\pm 0.14}$
0.5	$79.88_{\pm 0.38}$	$77.12_{\pm 0.16}$
0.25	$82.41_{\pm 0.28}$	$79.64_{\pm 0.12}$
0.125	$82.79_{\pm 0.43}$	$80.06_{\pm 0.19}$

Table 5: Figure 5 expected train and test accuracy under varying noise multipliers. Reported as mean $\pm$ standard error across data splits.

Noise (σ)	Train Acc. (%)	Test Acc. (%)
32	$10.10_{\pm 0.13}$	$10.08_{\pm 0.10}$
16	$9.90_{\pm 0.03}$	$10.02_{\pm 0.07}$
8	$10.10_{\pm 0.15}$	$10.13_{\pm 0.15}$
4	$11.12_{\pm 0.23}$	$10.66_{\pm 0.22}$
2	$18.87_{\pm 0.50}$	$17.83_{\pm 0.46}$
1	$30.10_{\pm 0.85}$	$26.74_{\pm 0.71}$
0.5	$32.21_{\pm 1.53}$	$27.31_{\pm 0.71}$
0.25	$36.41_{\pm 1.75}$	$30.67_{\pm 1.08}$
0.125	$34.45_{\pm 0.87}$	$29.74_{\pm 0.75}$