RANK-DEPENDENT DP-NOISE BOUNDS FOR FROZEN-A LORA

Anonymous authors
Paper under double-blind review

ABSTRACT

We study differentially private SGD (DP-SGD) for *frozen-A LoRA*, where a random factor $A \in \mathbb{R}^{r\times n}$ is fixed and only $B \in \mathbb{R}^{m\times r}$ is trained. Privacy accounting is dimension independent: the same noise multiplier, Poisson sampling rate, step count, adjacency convention, and accountant yield the same (ε, δ) guarantee. Our main result is instead a utility statement for the restricted B-objective: when $A_{jk} \sim \mathcal{N}(0, 1/n)$ gives $\|A\|_{\mathrm{op}} \le \kappa$, the reparameterization controls both gradient sensitivity and smoothness, and the *DP-noise component* of the stationarity bound scales with mr rather than mn. With inactive clipping, $C_L = \kappa C_F$, and respective positive initial gaps Δ_D, Δ_F, the ratio of the DP-noise terms is bounded by $\kappa^2\sqrt{r\Delta_D/(n\Delta_F)}$. With active clipping, method-specific bias terms remain and we do not claim an unconditional improvement in total error or practical utility. The guarantee controls $\nabla_B L_D$, not unrestricted stationarity in W-space. A Polyak–Łojasiewicz (PL) assumption is used only for excess-risk and optimal-rank results; we prove it for linear LoRA, while the informed-A optimal-rank result is an oracle benchmark. We quantify, without a complete convergence theorem, the product-noise term arising when both factors are trained. A complete 1,800-run controlled synthetic matrix separates restricted error from projection error and demonstrates optimizer-dependent rank orderings; SST-2 experiments remain feasibility checks, not validation of nonlinear PL or the optimal-rank law. A completed 42-run matched-public-start BERT/SST-2 follow-up finds at most 0.153 percentage points mean accuracy difference between frozen and jointly trained factors across six private-stage settings, providing no clear practical advantage for either branch.

1 INTRODUCTION

Fine-tuning pre-trained large language models (LLMs) on sensitive data—medical records, legal documents, personal communications—demands formal privacy guarantees. Differential privacy (DP) (Dwork et al., 2006) provides the gold standard, and DP-SGD (Abadi et al., 2016) is the primary algorithm for private deep learning. However, standard excess risk bounds for DP-SGD (Bassily et al., 2014) scale as $\mathcal{O}(\sqrt{d}/(N\varepsilon))$, where d is the parameter dimension and N is the dataset size. At LLM scale, such dimension-dependent bounds can be uninformative; this does not itself establish practical infeasibility or show that a lower-dimensional parameterization improves total utility.

LoRA (Hu et al., 2022) reduces trainable dimension, and private parameter-efficient fine-tuning has empirical precedent (Yu et al., 2022; Li et al., 2022). Dimension alone does not establish utility: smoothness transfer, clipping bias, and rank-restriction error must also be controlled. FFA-LoRA already fixes one factor in privacy-preserving federated learning (Sun et al., 2024). Malekmohammadi & Farnadi (2024) make formal random-sketching privacy claims; they should not be described as lacking any claimed guarantee. Hu et al. (2026) distinguish vector and matrix Wishart mechanisms, including a negative result without added noise for matrix queries and amplification for a noisy variant. Those mechanisms and release conventions differ from our explicitly noised frozen-A DP-SGD analysis; we do not assume that random projection alone makes a released adapter private.

Contributions and scope. The privacy accountant is unchanged by freezing a LoRA factor. Our analysis instead asks when lower dimension survives the reparameterization in a utility bound. We establish three linked results:

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.

1. Sensitivity and smoothness transfer through the fixed factor A, exposing the effective scale that can preserve or cancel the nominal dimension reduction.
2. A restricted-stationarity bound with an explicit mr DP-noise component and a separate, method-specific bias term when clipping is active.
3. A decomposition of restricted optimization and projection error, supported by a complete controlled matrix and a matched-public-start neural diagnostic.

PL-based excess-risk bounds and a finite-grid oracle rank comparison are conditional extensions. PL is proved only for the linear setting. The jointly trained-factor calculation and PAC-Bayes argument in the appendix do not establish a general utility improvement for nonlinear fine-tuning.

Relation to low-dimensional private optimization. Dimension reduction under DP predates this work. Projected DP-SGD identifies a gradient subspace using public data (Zhou et al., 2021); RGP learns gradient-carrier matrices from historical updates and reconstructs a noisy full-weight update (Yu et al., 2021); private language-model fine-tuning studies several parameter-efficient choices empirically (Yu et al., 2022; Li et al., 2022). Our narrower contribution is a sensitivity and smoothness transfer analysis for a fixed, data-independent LoRA factor under ordinary DP-SGD. Unlike the first two approaches, the optimized and released object remains a frozen-A adapter. We do not claim that low-dimensional private optimization itself is new. Appendix F gives an assumption-by-assumption comparison.

Table 1 summarizes the comparison between full DP fine-tuning and DP-LoRA.

Table 1: Finite-T DP-noise components under inactive clipping and the same accountant-calibrated σ, batch denominator b, and step count T. Here Δ_F, Δ_D are the respective empirical initial gaps. Different stationarity spaces and approximation errors prevent an unconditional utility comparison; $L_B \le \kappa^2 L_W$.

Method	**Trainable params**	**Finite-T DP component**	**Approx. error**
Full DP-SGD	mn	$\mathcal{O}(C_F\sigma\sqrt{\Delta_F L_W mn/T}/b)$	0
DP-LoRA (frozen A)	mr	$\mathcal{O}(C_L\sigma\sqrt{\Delta_D L_B mr/T}/b)$	$\mathcal{E}_{\text{approx}}(r)$

2 PRELIMINARIES

2.1 DIFFERENTIAL PRIVACY AND DP-SGD

Definition 1 (Differential Privacy (Dwork et al., 2006)). *A randomized mechanism $\mathcal{M}: \mathcal{X}^N \to \mathcal{Y}$ satisfies (ε, δ)-differential privacy if for all adjacent datasets D, D' differing in one element and all measurable $S \subseteq \mathcal{Y}$: $\Pr[\mathcal{M}(D) \in S] \le e^{\varepsilon}\Pr[\mathcal{M}(D') \in S] + \delta$.*

Definition 2 (Rényi Differential Privacy (Mironov, 2017)). *A mechanism $\mathcal{M}$ satisfies $(\alpha, \hat{\varepsilon})$-RDP if for all adjacent D, D': $D_\alpha(\mathcal{M}(D)\|\mathcal{M}(D')) \le \hat{\varepsilon}$, where D_α is the Rényi divergence of order $\alpha > 1$.*

DP-SGD (Abadi et al., 2016) privatizes SGD via per-sample gradient clipping and Gaussian noise addition. At each step t: (1) sample a minibatch $\mathcal{S}_t \subseteq [N]$ with Poisson sampling rate $q = b/N$; (2) compute per-sample gradients $g_i = \nabla_\theta \ell(z_i; \theta_t)$; (3) clip: $\tilde{g}_i = g_i \cdot \min(1, C/\|g_i\|_2)$; (4) aggregate with noise: $\bar{g}_t = \frac{1}{b}(\sum_{i\in\mathcal{S}_t}\tilde{g}_i + \xi_t)$ where $\xi_t \sim \mathcal{N}(0, \sigma^2 C^2 I_p)$; (5) update $\theta_{t+1} = \theta_t - \eta\bar{g}_t$.

Lemma 3 (Privacy accounting (Mironov, 2017; Balle et al., 2018)). *DP-SGD with Poisson subsampling rate q, noise multiplier σ, and T iterations satisfies (ε, δ)-DP with*

$$\varepsilon = \inf_{\alpha>1}\left\{T\cdot\rho(\alpha, q, \sigma) + \frac{\log(1/\delta)}{\alpha - 1}\right\}, \tag{1}$$

where ρ denotes a valid per-step RDP bound for the sampled Gaussian mechanism, not a log moment. It must be evaluated using an accountant valid for the stated sampling and adjacency convention. The shorthand q^2/σ^2 scaling requires additional parameter restrictions and is not used here as an unconditional privacy certificate. The accounting is independent of parameter dimension p.

2.2 LOW-RANK ADAPTATION (LORA)

Given a pre-trained weight matrix $W_0 \in \mathbb{R}^{m\times n}$, LoRA (Hu et al., 2022) parameterizes the adapted weight as $W = W_0 + BA$, where $B \in \mathbb{R}^{m\times r}$ and $A \in \mathbb{R}^{r\times n}$ with rank $r \ll \min(m, n)$. The total trainable parameters are $r(m + n)$ if both factors are trained, or mr (resp. rn) if one factor is frozen.

Frozen-A LoRA (FFA-LoRA (Sun et al., 2024)): A is initialized randomly and frozen; only B is trained. This eliminates the bilinear interaction between the two factors and is our primary setting.

Practical implementations often write $W = W_0 + sBA$ with $s = \alpha_{\mathrm{LoRA}}/r$. Our notation absorbs s into A. Accordingly, every occurrence of κ should be read as the effective scale $\kappa_{\mathrm{eff}} = s\,\|A\|_{\mathrm{op}}$; the dimensional advantage is not automatic if this quantity is large.

3 DP-LORA: FRAMEWORK AND SENSITIVITY ANALYSIS

3.1 PROBLEM SETUP

Consider a dataset $D = \{z_1, \ldots, z_N\}$ drawn i.i.d. from distribution $\mathcal{D}$. We fine-tune a single layer[1] with weight $W = W_0 + BA$, where $A \in \mathbb{R}^{r\times n}$ is drawn once, frozen, and independent of D, while $B \in \mathbb{R}^{m\times r}$ is trainable. The empirical and population losses are:

$$L_D(B) = \frac{1}{N}\sum_{i=1}^{N} \ell(z_i; W_0 + BA), \qquad L_{\mathrm{pop}}(B) = \mathbb{E}_{z\sim\mathcal{D}}[\ell(z; W_0 + BA)]. \tag{2}$$

All optimization expectations below are conditional on the fixed dataset D and on the high-probability event $\|A\|_{\mathrm{op}} \le \kappa$ for this single draw of A; they average only over minibatch sampling and DP noise. Expectations over D or A are written explicitly.

We state the following assumptions:

Assumption 1 (Smoothness)**.** *The empirical loss $L_D(W)$ is L_W-smooth; population approximation claims separately assume L_{pop} has the same bound. For the respective objective, $\left\|\nabla^2 L_D(W)\right\|_{\mathrm{op}} \le L_W$ for all W in the domain.*

Poisson variance convention. For independent Bernoulli masks and denominator $b = qN$, the conditional minibatch variance is $q(1-q)b^{-2}\sum_i \|g_i\|_F^2 \le (1-q)G^2/b$ under inactive clipping. Below write $\sigma_{\mathrm{sg}}^2 = (1-q)G^2$ for this conservative bound. It is not the centered variance of one uniformly sampled record; even identical record gradients fluctuate when the Poisson batch size is random.

Assumption 2 (Gradient norm bound)**.** *There exists $G > 0$ such that $\|\nabla_B \ell(z; W_0 + BA)\|_{\mathrm{F}} \le G$ almost surely for all B along the optimization trajectory.*

When Assumption 2 holds with $G \le C$, clipping introduces no bias and $\tilde{g}_i = g_i$. When $G > C$, the clipping bias $\|\mathbb{E}_i[\tilde{g}_i] - \nabla L_D(B)\|_2$ is retained explicitly (see Appendix D).

Update rule. The method is standard DP-SGD in the mr trainable coordinates of B: sample records by independent Poisson masks, clip their B-gradients, add isotropic Gaussian noise to their sum, and divide by the fixed expected batch size b. The frozen A and $B_0 = 0$ define the permitted subspace. Algorithm 1 in the appendix gives the complete update.

3.2 SENSITIVITY ANALYSIS

The key structural result is that the LoRA gradient $\nabla_B \ell = (\nabla_W \ell)A^\top$ lives in $\mathbb{R}^{m\times r}$, a space of dimension $mr \ll mn$.

Lemma 4 (Gradient norm bound)**.** *For any sample z and any B:*

$$\|\nabla_B \ell\|_{\mathrm{F}} \le \|A\|_{\mathrm{op}} \cdot \|\nabla_W \ell\|_{\mathrm{F}}\,. \tag{3}$$

[1]The analysis extends to multiple layers under a joint smoothness assumption; see Appendix H.

Consequently, if $\|A\|_{\mathrm{op}} \le \kappa$ and a clipping bound C suffices for the full gradient $\|\nabla_W \ell\|_{\mathrm{F}} \le C$, then $\|\nabla_B \ell\|_{\mathrm{F}} \le \kappa C$, so the same order of clipping bound applies to the LoRA gradient.

Proof. Right multiplication by $A^\top$ applies the same linear map to every row v of $\nabla_W \ell$. For each row, $\|vA^\top\|_2 \le \|A\|_{\mathrm{op}} \|v\|_2$. Squaring and summing over rows gives $\|(\nabla_W \ell)A^\top\|_{\mathrm{F}} \le \|\nabla_W \ell\|_{\mathrm{F}} \cdot \|A\|_{\mathrm{op}}$. Thus the Frobenius norm measures total row-wise gradient energy, while the operator norm captures the worst amplification by the shared projection. □

Lemma 5 (Operator norm of random A)**.** *Let $A \in \mathbb{R}^{r \times n}$ have i.i.d. entries $\mathcal{N}(0, 1/n)$. Then for any $\delta > 0$, with probability at least $1 - \delta$:*

$$\|A\|_{\mathrm{op}} \le 1 + \sqrt{\frac{r}{n}} + \sqrt{\frac{2\log(2/\delta)}{n}} =: \kappa. \tag{4}$$

In particular, for $r = o(n)$, $\kappa = 1 + o(1)$.

Proof. This follows from the concentration of the operator norm of Gaussian random matrices (Vershynin, 2018). By Gordon's inequality, for A with i.i.d. $\mathcal{N}(0, 1/n)$ entries: $\mathbb{E}[\|A\|_{\mathrm{op}}] \le (\sqrt{r} + \sqrt{n})/\sqrt{n} = 1 + \sqrt{r/n}$. The tail bound follows from the Lipschitz concentration of Gaussian random matrices with constant $1/\sqrt{n}$. □

Lemma 6 (Smoothness of LoRA objective)**.** *Under Assumption 1, conditioned on A with $\|A\|_{\mathrm{op}} \le \kappa$, the loss $L(B) = L_{\mathrm{pop}}(W_0 + BA)$ viewed as a function of $\mathrm{vec}(B) \in \mathbb{R}^{mr}$ is L_B-smooth with:*

$$L_B \le \kappa^2 L_W. \tag{5}$$

Proof. The Hessian of L w.r.t. $\mathrm{vec}(B)$ relates to the Hessian w.r.t. $\mathrm{vec}(W)$ via the chain rule: $\nabla^2_{\mathrm{vec}(B)} L = (A \otimes I_m)\, \nabla^2_{\mathrm{vec}(W)} L\, (A \otimes I_m)^\top$. Therefore $L_B = \left\|\nabla^2_{\mathrm{vec}(B)} L\right\|_{\mathrm{op}} \le \|A \otimes I_m\|^2_{\mathrm{op}} \cdot \left\|\nabla^2_{\mathrm{vec}(W)} L\right\|_{\mathrm{op}} = \|A\|^2_{\mathrm{op}} \cdot L_W \le \kappa^2 L_W$. □

Remark 1. *The initialization $A_{jk} \sim \mathcal{N}(0, 1/n)$ ensures $\kappa \approx 1$ for $r \ll n$, so the bound $L_B \le \kappa^2 L_W$ incurs little worst-case amplification; it does not imply that the two actual smoothness constants are equal. Alternative scalings (e.g., $\mathcal{N}(0, 1/r)$) yield $\kappa \approx \sqrt{n/r}$, canceling the nominal dimensional benefit. With practical LoRA scaling $s = \alpha_{\mathrm{LoRA}}/r$, replace κ throughout by $\kappa_{\mathrm{eff}} = s\,\|A\|_{\mathrm{op}}$.*

4 Main Results

4.1 Privacy guarantee

The privacy guarantee of DP-LoRA follows from standard DP-SGD, since the Gaussian mechanism's privacy depends only on the noise-to-sensitivity ratio, not on the parameter dimension. Stating this explicitly establishes that DP-LoRA matches full fine-tuning's (ε, δ) guarantee when σ, Poisson rate q, step count T, adjacency, and accountant are identical, so any utility improvement comes from the reduced noise magnitude rather than from weaker privacy.

Theorem 7 (Privacy of DP-LoRA)**.** *Algorithm 1 with noise multiplier σ, Poisson subsampling rate $q = b/N$, and T iterations satisfies (ε, δ)-DP when the valid accountant in Lemma 3 returns a budget no greater than ε. Choose σ by that accountant rather than an unrestricted small-q approximation. Its privacy calibration is independent of mr.*

4.2 Convergence with rank-dependent noise

Theorem 8 (Convergence of DP-LoRA)**.** *Under Assumptions 1–2, suppose the fixed draw of $A \in \mathbb{R}^{r \times n}$ satisfies $\|A\|_{\mathrm{op}} \le \kappa$ and $G \le C$ (inactive clipping). Let $\Delta_D = L_D(B_0) - \inf_B L_D(B)$.*

Conditional on D and this event over A, Algorithm 1 with step size $\eta \le 1/L_B$ for T iterations satisfies

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{\mathcal{S},\Xi}\Big[\|\nabla_B L_D(B_t)\|_{\mathrm{F}}^2\Big] \le \frac{2\Delta_D}{\eta T} + \eta L_B\left(\frac{\sigma_{\mathrm{sg}}^2}{b} + \frac{\sigma^2 C^2 \cdot mr}{b^2}\right). \tag{6}$$

With step size $\eta^\star = \min\Big(\frac{1}{L_B}, \sqrt{\frac{2\Delta_D}{TL_B\Psi}}\Big)$ where $\Psi = \frac{\sigma_{\mathrm{sg}}^2}{b} + \frac{\sigma^2 C^2 mr}{b^2}$:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{\mathcal{S},\Xi}\Big[\|\nabla_B L_D(B_t)\|_{\mathrm{F}}^2\Big] \le 2\sqrt{\frac{2\Delta_D \cdot L_B \cdot \Psi}{T}} + \frac{2\Delta_D \cdot L_B}{T}. \tag{7}$$

The exact per-step DP noise contribution is

$$V_{\mathrm{DP}} = \sigma^2 C^2 mr/b^2. \tag{8}$$

In the optimized finite-T upper bound its square-root contribution is

$$2C\sigma\sqrt{2\Delta_D L_B mr/T}/b. \tag{9}$$

Here σ is the value returned by a valid privacy accountant. No unrestricted substitution $\sigma^2 \propto q^2 T/\varepsilon^2$ is made.

Proof outline. With inactive clipping, the sampled gradient is unbiased. Its second moment is the squared population-of-records gradient plus the Poisson-sampling contribution and $\sigma^2 C^2 mr/b^2$. Smoothness and $\eta L_B \le 1$ then give a descent inequality; summing it telescopically yields the stated average-stationarity bound. Optimizing over the admissible step sizes gives the two-regime expression above. The algebra, including the boundary step size, is retained in Appendix N.

Remark 2 (Comparison with full DP fine-tuning)**.** *Let C_F be the full-gradient clipping threshold and set $C_L = \kappa C_F$. The per-step DP-noise energies obey $\mathbb{E}\|\Xi_L\|^2/\mathbb{E}\|\Xi_F\|^2 \le \kappa^2 r/n$. Combining this with $L_B \le \kappa^2 L_W$ gives a ratio at most $\kappa^2\sqrt{r\Delta_D/(n\Delta_F)}$ between the* DP-noise components *of the stationarity bounds when $\Delta_F > 0$. The simpler $\kappa^2\sqrt{r/n}$ factor requires $\Delta_D \le \Delta_F$; otherwise the initial-gap factor must be retained. This statement compares restricted B-stationarity with full W-stationarity and is not an unconditional comparison of practical utility.*

Remark 3 (Stationarity in B-space vs. W-space)**.** *Theorem 8 guarantees an approximate stationary point of $L_D(B) = L_D(W_0 + BA)$. Since $\nabla_B L_D = (\nabla_W L_D)A^\top$ and $r < n$, a small $\|\nabla_B L_D\|_{\mathrm{F}}$ only controls the projection of the full gradient onto the row space of A; it does not imply $\nabla_W L_D \approx 0$. The gap between restricted and unrestricted optima is handled separately through $\mathcal{E}_{\mathrm{approx}}(r, A)$.*

Proposition 9 (Convergence decomposition with active clipping)**.** *With active clipping and $\eta \le 1/L_B$, the finite-T bound becomes:*

$$\frac{1}{T}\sum_{t}\mathbb{E}_{\mathcal{S},\Xi}[\|\nabla_B L_D(B_t)\|_{\mathrm{F}}^2] \le \frac{2\Delta_D}{\eta T} + \eta L_B\left(\frac{(1-q)C^2}{b} + \frac{\sigma^2 C^2 mr}{b^2}\right) + \beta_{\mathrm{clip}}, \tag{10}$$

where $\beta_{\mathrm{clip}} = \sup_t \mathbb{E}\left\|N^{-1}\sum_i \tilde{g}_i - \nabla_B L_D(B_t)\right\|_{\mathrm{F}}^2$ is method specific. Appendix D proves the bound without treating clipped gradients as unbiased. Comparable thresholds do not imply comparable bias or total error.

4.3 Excess risk under the Polyak-Łojasiewicz condition

Theorem 8 establishes stationarity in the non-convex setting. To obtain *excess risk* bounds (function value suboptimality), we introduce a structural condition that bridges gradient norms to function values.

Assumption 3 (Polyak-Łojasiewicz (PL) condition)**.** *There exists $\mu > 0$ such that for all B in the domain: $\|\nabla_B L_D(B)\|_{\mathrm{F}}^2 \ge 2\mu(L_D(B) - L_D(B_D^\star))$, where $B_D^\star = \arg\min_B L_D(B)$.*

The PL condition does not require convexity, but we do not establish it for nonlinear Transformer fine-tuning. We prove it only for linear LoRA in Appendix G (Proposition 15). NTK and local-quadratic arguments may motivate the assumption, but they are not a verification of it in the models studied here. Accordingly, all results below that invoke PL are conditional.

Definition 10 (Approximation error)**.** *Given frozen A, the approximation error at rank r is:*

$$\mathcal{E}_{approx}(r, A) = L_{pop}(B^\star) - L_{pop}(W^\star), \tag{11}$$

where $B^\star = \arg\min_B L_{pop}(W_0 + BA)$ and $W^\star = \arg\min_W L_{pop}(W)$.

Theorem 11 (Excess risk under PL condition)**.** *Under Assumptions 1–3, the output $\hat{B} = B_{\hat{t}}$ of Algorithm 1 (with $\hat{t}$ uniform on $\{0, \dots, T-1\}$) satisfies:*

$$\mathbb{E}_{\mathcal{S},\Xi}[L_D(\hat{B})] - L_D(B_D^\star) \le \frac{1}{2\mu}\left(\frac{2\Delta_D}{\eta T} + \eta L_B \Psi\right), \tag{12}$$

where $\Psi = \sigma_{\text{sg}}^2/b + \sigma^2 C^2 mr/b^2$ as in Theorem 8. Consequently, the total excess population risk decomposes as:

$$\mathbb{E}[L_{\text{pop}}(\hat{B})] - L_{\text{pop}}(W^\star) \le \mathcal{E}_{\text{approx}}(r, A) + \frac{1}{2\mu}\left(\frac{2\Delta_D}{\eta T} + \eta L_B \Psi\right) + \mathcal{E}_{\text{gen}}. \tag{13}$$

The finite-T bound is the conclusion established here. For a fixed privacy-calibrated Ψ, unconstrained step-size minimization gives $\eta_0 = \sqrt{2\Delta_D/(TL_B\Psi)}$; enforce the step-size restriction of Theorem 8 by clipping η_0 to its admissible range. At an admissible interior optimum the optimization term is $\sqrt{2\Delta_D L_B \Psi/T}/\mu$. This argument does not yield an $O(mr/(N^2\varepsilon^2))$ excess-risk rate.

Proof. By the PL condition, $L_D(\hat{B}) - L_D(B_D^\star) \le (2\mu)^{-1}\left\|\nabla_B L_D(\hat{B})\right\|_{\text{F}}^2$. Taking expectations over minibatches and DP noise and applying Theorem 8 gives $\mathbb{E}_{\mathcal{S},\Xi}[L_D(\hat{B}) - L_D(B_D^\star)] \le (2\mu)^{-1}\left(2\Delta_D/(\eta T) + \eta L_B \Psi\right)$. The population risk decomposition follows by adding $\mathcal{E}_{\text{approx}}(r, A) = L_{\text{pop}}(B^\star) - L_{\text{pop}}(W^\star)$ and the generalization gap. The step-size expression follows by differentiating the displayed finite-T bound; it does not square a stationarity guarantee. □

To make the approximation error concrete, we introduce a spectral decay assumption.

Assumption 4 (Spectral decay)**.** *Let $\Delta W^\star = W^\star - W_0$ have singular values $\sigma_1 \ge \sigma_2 \ge \cdots$. Assume $\sigma_j \le c_0 j^{-\alpha}$ for some $\alpha > 1/2$ and $c_0 > 0$.*

Lemma 12 (Approximation error under spectral decay)**.** *Under Assumptions 1 and 4:*

(a) ***Informed*** *A (rows of A span the top-r right singular vectors of $\Delta W^\star$; an oracle baseline, not the algorithm's default):*

$$\mathcal{E}_{\text{approx}}(r, A) \le \frac{L_W}{2}\sum_{j>r}\sigma_j^2 \le \frac{L_W c_0^2}{2(2\alpha-1)} \cdot r^{1-2\alpha}. \tag{14}$$

(b) ***Random*** *A (entries i.i.d. $\mathcal{N}(0, 1/n)$). Under spectral decay:*

$$\mathbb{E}_A[\mathcal{E}_{\text{approx}}(r, A)] \le \frac{L_W}{2} \cdot \frac{n-r}{n} \sum_{j=1}^{\min(m,n)} \sigma_j^2 = \frac{L_W}{2} \cdot \frac{n-r}{n} \cdot \|\Delta W^\star\|_{\text{F}}^2. \tag{15}$$

When $\Delta W^\star$ has effective rank $r_{\text{eff}} \ll n$ (i.e., $\|\Delta W^\star\|_{\text{F}}^2 \approx \sum_{j=1}^{r_{\text{eff}}} \sigma_j^2$), this simplifies to $\mathcal{O}(L_W \|\Delta W^\star\|_{\text{F}}^2)$ for $r \ll n$, which is approximately constant in this regime. This approximation term alone does not determine an optimal rank: all rank-dependent constants in the finite-T bound must be retained, as stated in Remark 4.

Proof. For any fixed A, the best B minimizing $\|BA - \Delta W^\star\|_{\text{F}}^2$ is $B^\star = \Delta W^\star A^\top (AA^\top)^{-1}$, giving residual $\Delta W^\star(I_n - \Pi_A)$ where $\Pi_A = A^\top(AA^\top)^{-1}A$ is the projection onto the row space of A. By L_W-smoothness: $\mathcal{E}_{\text{approx}}(r, A) \le \frac{L_W}{2}\|\Delta W^\star(I_n - \Pi_A)\|_{\text{F}}^2$.

Part (a): When A's rows span the top-r right singular vectors, $I_n - \Pi_A$ projects onto the remaining $n - r$ directions. Thus $\|\Delta W^\star(I_n - \Pi_A)\|_{\text{F}}^2 = \sum_{j>r}\sigma_j^2$. The integral bound follows from $\sum_{j>r} j^{-2\alpha} \le \int_r^\infty x^{-2\alpha}dx = r^{1-2\alpha}/(2\alpha-1)$.

Part (b): For random A, the rows of A span a uniformly random r-dimensional subspace of $\mathbb{R}^n$ (after normalization). For any fixed unit vector $v \in \mathbb{R}^n$, $\mathbb{E}[\|(I_n - \Pi_A)v\|_2^2] = (n-r)/n$. Linearity of expectation over the columns of $\Delta W^\star$ yields the result. □

Corollary 13 (Finite-grid oracle rank selection)**.** *For a finite admissible rank set $\mathcal{R}$, let $U(r, \eta)$ denote the approximation term from Lemma 12(a) plus the finite-T optimization term in Theorem 11. Use the actual rank-dependent L_B, μ, Δ_D, Ψ and any generalization term being claimed. Then an upper-bound selection rule is*

$$r_{\mathrm{grid}} \in \arg\min_{r\in\mathcal{R}} \min_{\eta \text{ admissible}} U(r, \eta). \tag{16}$$

This is a finite-grid oracle comparison. Its constants depend on the task and rank, so it does not give a data-free rank prescription.

Proof. For every admissible pair the finite-T inequality holds. Evaluating those bounds and selecting their minimum gives the stated upper-bound rule. □

Remark 4 (Random-A rank selection)**.** *The approximation upper bound decreases linearly with r, but the corrected optimized finite-T noise term need not increase linearly. Its smoothness, PL constant, initial gap and clipping threshold may also change with rank. No universal monotonicity or $r = 1$ threshold follows. Use the complete rank-dependent oracle bound in Corollary 13.*

Interpretation. Evaluating a rank bound requires its approximation term, smoothness, PL constant, initial gap and clipping threshold. The informed-A construction is an oracle benchmark, and the stationarity result supplies no matching minimax lower bound for this loss class.

Generalization scope. The supplementary Gibbs-predictor bound (Appendix M) separates approximation, optimization, and generalization terms; it is not an unconditional utility guarantee.

5 Controlled mechanism evidence

The complete 1,800-run matrix in Appendix L separates restricted optimization error from projection error. Their sum is an identity for the training objective, not a decomposition of held-out test risk. Figure 1 shows the fixed reference slice with all five seeds and a zero-update baseline. Under the tested SGD setting, a lower rank can outperform full adaptation while both remain worse than no update; under the tested AdamW setting, the rank ordering changes. The optimizers follow separately declared update rules, so this is not a comparison of their optimally tuned performance.

The matrix therefore tests the mechanism's boundary as well as its noise component. It does not validate a nonlinear PL condition or remove clipping and approximation error. The neural matched-public-start experiments use AdamW and have active clipping, whereas the inactive-clipping convergence result concerns ordinary SGD. Valid privacy accounting does not make these optimization assumptions interchangeable.

Restricted error and projection error. For the regularized empirical training objective, write f_A^* for the restricted optimum and f_{full}^* for the full-space optimum. Then

$$f(BA_T) - f_{\mathrm{full}}^* = \underbrace{f(BA_T) - f_A^*}_{\text{restricted optimization error}} + \underbrace{f_A^* - f_{\mathrm{full}}^*}_{\text{projection error}}.$$

The largest numerical discrepancy in this identity is 1.78×10^{-15}. The strongly convex reference solver has certificate $\|\nabla f_A\|^2/(2\mu_A) \le 5.44 \times 10^{-15}$ over all runs. Optima for identical feasible sets vary by at most 5.33×10^{-15} across clipping, privacy, optimizer and scale settings. This is a training-objective decomposition; it is not an additive decomposition of held-out risk. All 30,000 recorded trace points in the 600 unclipped runs have exactly zero clipping fraction and clipping bias. Analytic Gaussian gradient-energy values recomputed from the registered configurations match $32r\sigma^2C^2/64^2$; this is a formula check, not an independent empirical noise-moment estimate. The trace checks validate the specified no-clipping regime.

Why retain the zero-update reference? A comparison against full adaptation alone can make a restricted method appear useful even when both training procedures increase the objective relative to their common starting point. The zero-update reference detects that possibility without spending a private optimization step. Its inclusion does not resolve model-selection privacy: choosing an

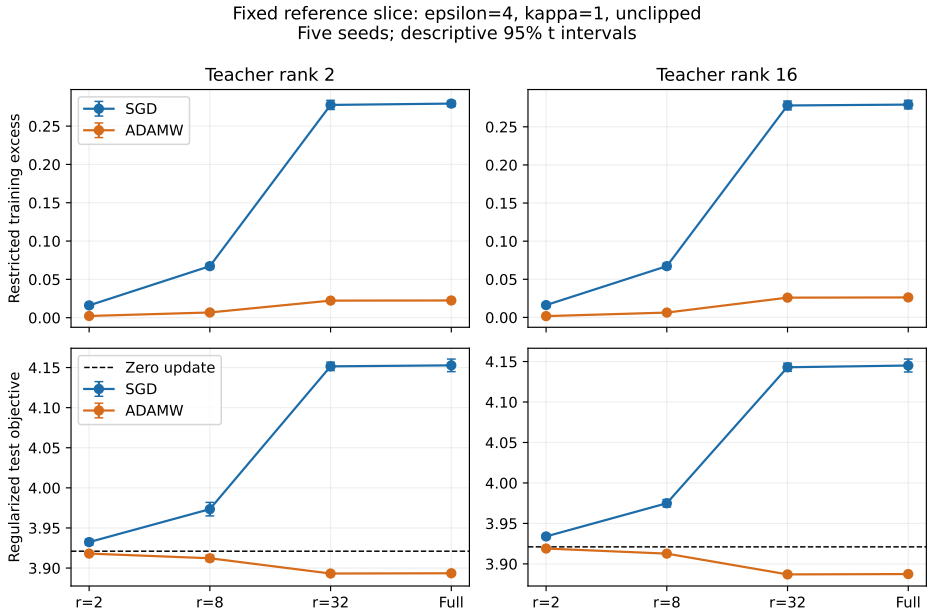

Figure 1: Fixed reference slice of the complete matrix. Intervals are descriptive 95% t_4 intervals over five seeds and are not multiplicity-adjusted. Each seed changes the synthetic data, random basis and training randomness; paired comparisons retain the same seed. These intervals are not confirmatory tests.

Table 2: Matched-public-start BERT/SST-2. All arms start at 88.53% (772/872). Seed columns give joint-minus-frozen accuracy in percentage points for seeds 51604/5/6. The final column is a retrospective pointwise 95% interval from 10,000 paired question-and-seed bootstrap draws; three seeds provide limited uncertainty information.

Rank	ε	Seed 4	Seed 5	Seed 6	Mean	Interval (pp)
2	1	+0.000	+0.000	+0.000	+0.000	$[+0.000, +0.000]$
2	4	+0.115	+0.000	+0.344	+0.153	$[+0.000, +0.459]$
8	1	+0.000	+0.229	+0.229	+0.153	$[+0.000, +0.459]$
8	4	-0.115	-0.115	+0.229	+0.000	$[-0.229, +0.306]$
32	1	+0.115	+0.344	-0.573	-0.038	$[-0.650, +0.459]$
32	4	+0.115	+0.229	-0.229	+0.038	$[-0.382, +0.459]$

adapter or rank after consulting private outcomes can itself require accounting. In this controlled study, it instead provides a fixed interpretive anchor for the prespecified matrix. The relevant comparisons are therefore restricted versus full adaptation and each method versus its own shared starting point, alongside the decomposition into optimization and projection error. A smaller DP-noise term addresses only one of these quantities.

5.1 Matched-public-start neural diagnostic

The 42 BERT/SST-2 runs use one common public warm start, disjoint private-stage data, matched minibatches and three paired seeds for each rank/privacy setting. Every arm initially classifies 772 of 872 validation sentences correctly (88.53%). Across the six private settings, mean gains from this starting point range from -0.115 to $+0.535$ percentage points. Thus the final accuracies mostly reflect the shared public stage, and the small branch contrasts do not establish useful gains from private adaptation.

Table 2 retains every seed contrast. Its intervals resample the same validation questions and paired training seeds across arms; they are retrospective and pointwise. At rank 2 and $\varepsilon = 1$, both branches

classify every retained question identically for each seed, producing a degenerate bootstrap contrast. That property of the saved predictions is not an equivalence guarantee for new runs or data. The artifact also reports each arm's change from the shared initial predictor.

All neural runs use AdamW, and 9.00–20.10% of sampled examples have gradients above the clipping threshold. These runs test branch behavior with active clipping; they do not instantiate the ordinary-SGD, inactive-clipping theorem. Appendix A.2 gives the data contract, accounting and full accuracy table. The single-seed zero-noise controls remain descriptive.

6 LIMITATIONS

The main convergence guarantee is specific to frozen-A LoRA and controls stationarity only in the permitted B-subspace. Standard LoRA usually trains both factors; Appendix I identifies its direct and product-noise terms but does not prove a complete convergence theorem. The PL-based excess-risk and optimal-rank results are conditional: PL is proved here only for a linear model, and the informed-A construction is an oracle benchmark. With active clipping, method-specific bias terms prevent the DP-noise comparison from becoming an unconditional comparison of total error or practical utility. The single-layer formulas extend to multiple layers only under a joint smoothness assumption.

The experiments are small diagnostics rather than end-to-end private LLM fine-tuning. In the submitted SST-2 setup, the privacy guarantee excludes the non-private training of W_0 on the same subset. The BERT follow-up instead uses disjoint public warm-start and private data, but its public stage supplies much of the accuracy and its three ranks do not reveal an interior optimum. Rank or model selection on private data, repeated private runs, and the release of multiple selected adapters require additional privacy accounting; releasing one adapter from one accounted run with publicly fixed hyperparameters is post-processing; private selection is not automatically covered.

6.1 A REGIME WITH INACTIVE DATA-GRADIENT CLIPPING

For an identifiable convex diagnostic, let $x \in \mathbb{R}^{32}$ satisfy $\|x\|_2 \le 1$, let $y \in \{0,1\}^{32}$, and use the record-level loss $\ell(W; x, y) = 32^{-1/2} \sum_j \mathrm{BCE}(\sigma((Wx)_j), y_j)$. The complete (x, y) is one privacy record. Then $\nabla_W \ell = (\sigma(Wx) - y)x^\top/\sqrt{32}$ has Frobenius norm at most one. For $W = W_0 + BA$ and $AA^\top = \kappa^2 I$, the data gradient with respect to B has norm at most κ. Thus clipping at $C = \kappa$ is inactive in exact arithmetic. Adding $\lambda\|W_0 + BA\|_F^2/2$ gives strong convexity $\mu_B = \lambda\kappa^2$ and smoothness at most $\kappa^2(\lambda + 1/(4\sqrt{32}))$. The deterministic regularizer gradient is $\lambda(W_0 + BA)A^\top$, not λB. It can be added after the private data gradient is clipped, sampled, and noised; its public parameters must not be selected using unaccounted private information. Only the data gradient is bounded; the regularized objective gradient need not be.

With conditional clipping bias b_t and centered gradient noise ζ_t, SGD obeys

$$\mathbb{E}_t\|B_{t+1} - B^*\|_F^2 = \|B_t - B^* - \eta(\nabla f(B_t) + b_t)\|_F^2 + \eta^2\mathbb{E}_t\|\zeta_t\|_F^2.$$

For expected batch size b, the Gaussian contribution alone is $32r\,\sigma^2C^2/b^2$; sampling variance and clipping bias remain separate. The diagnostic in Appendix L tests these mechanisms. These SGD identities do not establish AdamW/AdamBC guarantees or a PL condition for BERT.

7 CONCLUSION

Frozen-A LoRA and full fine-tuning have the same dimension-independent DP accounting when the sampling scheme, rate, step count, adjacency, and noise multiplier are fixed. The utility distinction comes from the restricted optimization problem: after controlling the projection's sensitivity and smoothness, the DP-noise component depends on mr rather than mn. This explains when reduced trainable dimension survives the reparameterization, while the clipping, approximation, and unrestricted-stationarity qualifications delimit what the comparison does not show. The complete controlled matrix in Appendix L demonstrates why restricted optimization error must be reported alongside projection error and a zero-update baseline: lower rank can reduce the former without yielding useful adaptation, and rank ordering changes across the tested optimizers. Extending the analysis to jointly trained LoRA and establishing useful nonlinear conditions remain open.

Reproducibility Statement

The main text and appendix specify the fixed-subspace assumptions, clipping-bias terms, objective, and proof steps. Retained scripts and result records support the rank and mechanism diagnostics; the supplementary experiment details describe the historical embedding-MLP evaluation. That evaluation is not a full Transformer LoRA replication. Reproduction must preserve the stated clipping mechanism, parameterization, and privacy assumptions; varying rank alone does not isolate all of these factors.

AI Use Statement

OpenAI Codex was used to aid and polish manuscript writing and to assist with literature retrieval and discovery, including finding related work. The authors are responsible for the final manuscript and citations.

References

Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 308–318, 2016.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pp. 464–473, 2014.

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Yaxi Hu, Johanna Düngler, Bernhard Schölkopf, and Amartya Sanyal. LoRA and privacy: When random projections help (and when they don't). *arXiv preprint arXiv:2601.21719*, 2026.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. In *International Conference on Learning Representations (ICLR)*, 2022.

Saber Malekmohammadi and Golnoosh Farnadi. LoRA provides differential privacy by design via random sketching. *arXiv preprint arXiv:2409.17538*, 2024.

David A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the 12th Annual Conference on Computational Learning Theory (COLT)*, pp. 164–170, 1999.

Ilya Mironov. Rényi differential privacy. In *IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275, 2017.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving LoRA in privacy-preserving federated learning. In *International Conference on Learning Representations (ICLR)*, 2024.

Roman Vershynin. *High-Dimensional Probability*. Cambridge University Press, 2018.

Da Yu, Huishuai Zhang, Wei Chen, Jian Yin, and Tie-Yan Liu. Large scale private learning via low-rank reparametrization. In *International Conference on Machine Learning (ICML)*, pp. 12208–12218, 2021.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A. Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Yingxue Zhou, Zhiwei Steven Wu, and Arindam Banerjee. Bypassing the ambient dimension: Private SGD with gradient subspace identification. In *International Conference on Learning Representations (ICLR)*, 2021.

.1 ALGORITHM

Algorithm 1 is ordinary DP-SGD applied to the mr trainable coordinates of B. The random factor A is sampled once before training and is never updated. This fixed linear parameterization is the structural restriction used throughout the main theorems.

Algorithm 1 DP-LoRA (Frozen A)

Require: Dataset $D = \{z_1, \dots, z_N\}$, pre-trained $W_0 \in \mathbb{R}^{m\times n}$, rank r, noise multiplier σ, clipping bound C, learning rate η, iterations T, batch size b.

Initialize $A \in \mathbb{R}^{r\times n}$ with i.i.d. entries $A_{jk} \sim \mathcal{N}(0, 1/n)$; freeze A.
Initialize $B_0 = \mathbf{0} \in \mathbb{R}^{m\times r}$.
for $t = 0, 1, \dots, T-1$ **do**
  Sample batch $\mathcal{S}_t \subseteq [N]$ via Poisson sampling with rate $q = b/N$.
  for each $i \in \mathcal{S}_t$ **do**
    Compute $g_i = \nabla_B \ell(z_i; W_0 + B_t A) = (\nabla_W \ell(z_i; W_0 + B_t A)) \cdot A^\top \in \mathbb{R}^{m\times r}$.
    Clip: $\tilde{g}_i = g_i \cdot \min\left(1, \frac{C}{\|g_i\|_{\mathrm{F}}}\right)$.
  end for
  $\bar{g}_t = \frac{1}{b}\left(\sum_{i\in\mathcal{S}_t} \tilde{g}_i + \Xi_t\right)$, where $\Xi_t \sim \mathcal{N}(0, \sigma^2 C^2 I_{mr})$ (reshapeable to $\mathbb{R}^{m\times r}$).
  $B_{t+1} = B_t - \eta \bar{g}_t$.
end for
return B_T (or $B_{\hat{t}}$ with $\hat{t}$ uniformly random from $\{0, \dots, T-1\}$).

A SST-2 EXPERIMENT DETAILS

Table 3: SST-2 validation accuracy (%) for the submitted MLP-over-embeddings experiment. Pre-trained baseline (no DP): 81.0%. Results average two seeds; the clustered values and trained output head limit conclusions about rank dependence.

ε	Full (p=98k)	r=2 (514)	r=8 (1.3k)	r=32 (4.5k)	r=64 (8.5k)	r=128 (17k)
1	82.0±0.0	81.8±0.3	82.0±0.0	81.5±0.7	81.9±0.1	81.5±0.7
4	81.2±1.4	82.0±0.0	82.3±0.4	81.0±0.0	81.9±0.4	81.5±1.1
8	81.4±0.6	82.0±0.0	81.9±0.4	81.6±0.6	81.9±0.3	81.7±0.4
20	81.6±0.4	82.0±0.0	81.8±0.6	81.4±0.1	81.5±0.1	82.0±0.0

We fine-tune a 2-layer MLP (hidden dim 128) on SST-2 using GPT-2 embeddings (768-dim). The model is first pre-trained non-privately to 81% accuracy, then DP-LoRA fine-tuned with $N = 5000$ training samples, $\varepsilon \in \{1, 4, 8, 20\}$, $\delta = 10^{-5}$, and 15 epochs. For the LoRA variant, $A \in \mathbb{R}^{r\times 768}$ is initialized with $\mathcal{N}(0, 1/768)$ entries and frozen; B and a rank-independent output head are trained with DP-SGD. Because W_0 is pre-trained non-privately on the same subset, the reported ε applies only to the subsequent DP stage conditional on W_0. The ceiling effect, two seeds, and additional output head preclude a meaningful test of the rank law.

A.1 EARLIER UNMATCHED-WARM-START BERT CHECK

We additionally ran BERT-base on raw SST-2 text, adapting query and value projections in all 12 Transformer layers. A public stage used 10,000 records, followed by DP-SGD on a disjoint private set of 50,000 records. Both variants used the same split, optimizer, hyperparameters, three seeds, and $(\varepsilon, \delta) = (4, 10^{-5})$. However, each LoRA variant underwent its own public warm-start, so the weights at the start of the private stage were not matched.

Table 4: BERT-base SST-2 validation accuracy (mean $\pm$ standard deviation over three seeds). Public warm-starts differ across variants; this is a feasibility comparison, not a matched-initialization comparison or a test of PL or the oracle rank law.

Rank	Frozen A	Jointly trained A/B
2	87.92 ± 0.90	90.75 ± 0.57
8	89.41 ± 0.75	90.83 ± 0.28
32	90.79 ± 0.19	76.99 ± 18.44

At rank 32, the jointly trained runs obtained 89.33, 90.71, and a near-chance result. The failed seed was already near chance after the non-private public warm-start, before DP-SGD. Thus the observed rank-32 failure cannot be attributed specifically to DP noise or private-stage factor training. The separately trained public stages confound that interpretation, even though the full pipeline was less stable for this configuration. The study is too small to establish a general stability difference, and none of these runs proves that the frozen-A theorem extends to jointly trained LoRA.

A.2 Matched-public-start protocol

The preceding feasibility comparison does not isolate private-stage factor training. The matched follow-up uses one shared BERT/classifier state trained on 2,000 public benchmark sentences for two epochs before constructing any LoRA arm. Normalized sentence hashes separate this public subset, a 10,000-example private-stage simulation subset, an independent 32-example engineering subset, and the 872-example validation split. Validation texts and conflicting training labels are excluded before selecting the fixed training subsets. All ranks, branches, and seeds share the same public state; dropout is disabled for this controlled comparison.

The frozen matrix contains ranks 2/8/32, frozen-A and joint-A, B branches, $\varepsilon \in \{1, 4\}$ at $\delta = 10^{-5}$, and three paired seeds (36 runs). Six additional clipped, zero-noise controls use one seed and are descriptive. Every private-stage update samples each record independently with probability $q = 0.0064$, clips its joint gradient over all trainable LoRA and classifier parameters to norm one, adds independent Gaussian noise to the summed gradient, and divides by the fixed expected batch size 64. Empty batches still perform a noise update and advance the accountant. We specify 469 updates with AdamW, learning rate 5×10^{-4} and zero weight decay. The RDP calibration uses add/remove-one-record adjacency; it is not a claim for replacement adjacency at the same noise level. Public benchmark PRNG seeds and diagnostics support research reproduction, not a production release guarantee for actual private records.

All 42 formal runs completed, with 469 updates each; the saved evaluation and trace audit passes. AdamW is used as a utility diagnostic, not as a direct empirical test of the SGD convergence theorem. Three seeds and three ranks will not by themselves establish an interior optimum, the PL assumption, or a general advantage over FFA-LoRA.

A.3 Completed matched-public-start results

Table 5: SST-2 validation accuracy (percent), averaged over three paired seeds. Difference is joint minus frozen, in percentage points.

Rank	ε	Frozen A	Joint	Difference
2	1	88.42	88.42	0.00
2	4	88.61	88.76	0.15
8	1	88.72	88.88	0.15
8	4	89.07	89.07	0.00
32	1	89.03	88.99	-0.04
32	4	88.95	88.99	0.04

The absolute mean branch difference is at most 0.153 percentage points across these six settings. These small, mixed differences do not establish a practical advantage for either branch, nor an interior

optimal rank. The single-seed zero-noise controls give joint-minus-frozen differences of $+1.606$, $+0.688$, and -0.459 percentage points for ranks 2, 8, and 32; they have no seed-level uncertainty estimate and are not private releases.

Paired branches share the same public initialization and all 469 sampled minibatches. Reported RDP epsilons are 0.9999682 and 3.9999465, below their targets. Results use one public benchmark, one public warm start, and AdamW; they cannot validate the SGD theorem's assumptions or establish a new privacy–utility frontier relative to FFA-LoRA.

B Proof of Theorem 8 (Full Details)

We provide the complete proof with explicit constants.

Proof. **Step 1: Descent lemma.** Condition on the dataset D and the high-probability event for the fixed draw of A. By L_B-smoothness of $L_D(B)$:

$$L_D(B_{t+1}) \le L_D(B_t) + \langle \nabla_B L_D(B_t), B_{t+1} - B_t \rangle_{\mathrm{F}} + \frac{L_B}{2} \|B_{t+1} - B_t\|_{\mathrm{F}}^2 . \tag{17}$$

With the update $B_{t+1} = B_t - \eta \bar{g}_t$:

$$L_D(B_{t+1}) \le L_D(B_t) - \eta \langle \nabla_B L_D(B_t), \bar{g}_t \rangle_{\mathrm{F}} + \frac{\eta^2 L_B}{2} \|\bar{g}_t\|_{\mathrm{F}}^2 . \tag{18}$$

Step 2: Decompose the noisy gradient. The noisy gradient is $\bar{g}_t = \bar{g}_t^{\text{clean}} + \xi_t/b$ where $\bar{g}_t^{\text{clean}} = \frac{1}{b}\sum_{i \in \mathcal{S}_t} \tilde{g}_i$ and $\xi_t \sim \mathcal{N}(0, \sigma^2 C^2 I_{mr})$.

Under Assumption 2 with $G \le C$ (no clipping bias): $\mathbb{E}_{\mathcal{S}}[\bar{g}_t^{\text{clean}}] = \nabla_B L_D(B_t)$, and $\mathbb{E}_{\mathcal{S}}[\|\bar{g}_t^{\text{clean}} - \nabla_B L_D(B_t)\|_{\mathrm{F}}^2] \le \sigma_{\text{sg}}^2/b$.

Step 3: Compute expectations.

$$\begin{aligned} \mathbb{E}_{\mathcal{S},\Xi}[\langle \nabla_B L_D(B_t), \bar{g}_t \rangle] &= \langle \nabla_B L_D(B_t), \mathbb{E}_{\mathcal{S}}[\bar{g}_t^{\text{clean}}] \rangle + \langle \nabla_B L_D(B_t), \mathbb{E}_{\Xi}[\xi_t/b] \rangle \\ &= \|\nabla_B L_D(B_t)\|_{\mathrm{F}}^2 . \end{aligned} \tag{19}$$

$$\begin{aligned} \mathbb{E}[\|\bar{g}_t\|_{\mathrm{F}}^2] &= \mathbb{E}[\|\bar{g}_t^{\text{clean}}\|_{\mathrm{F}}^2] + \mathbb{E}[\|\xi_t/b\|_{\mathrm{F}}^2] + 2\mathbb{E}[\langle \bar{g}_t^{\text{clean}}, \xi_t/b \rangle] \\ &= \mathbb{E}[\|\bar{g}_t^{\text{clean}}\|_{\mathrm{F}}^2] + \frac{\sigma^2 C^2 mr}{b^2} + 0 \\ &\le \|\nabla_B L_D(B_t)\|_{\mathrm{F}}^2 + \frac{\sigma_{\text{sg}}^2}{b} + \frac{\sigma^2 C^2 mr}{b^2}. \end{aligned} \tag{20}$$

Here we used: (i) ξ_t is independent of $\bar{g}_t^{\text{clean}}$; (ii) $\mathbb{E}[\|\xi_t\|_{\mathrm{F}}^2] = \sigma^2 C^2 \cdot mr$; and (iii) $\mathbb{E}_{\mathcal{S}}[\|\bar{g}_t^{\text{clean}}\|_{\mathrm{F}}^2] \le \|\nabla_B L_D(B_t)\|_{\mathrm{F}}^2 + \sigma_{\text{sg}}^2/b$.

Step 4: Combine.

$$\begin{aligned} \mathbb{E}_{\mathcal{S},\Xi}[L_D(B_{t+1})] \le \mathbb{E}_{\mathcal{S},\Xi}[L_D(B_t)] &- \eta \left(1 - \frac{\eta L_B}{2}\right) \mathbb{E}_{\mathcal{S},\Xi}[\|\nabla_B L_D(B_t)\|_{\mathrm{F}}^2] \\ &+ \frac{\eta^2 L_B}{2} \left(\frac{\sigma_{\text{sg}}^2}{b} + \frac{\sigma^2 C^2 mr}{b^2}\right). \end{aligned} \tag{21}$$

For $\eta \le 1/L_B$: $1 - \eta L_B/2 \ge 1/2$. Therefore:

$$\mathbb{E}_{\mathcal{S},\Xi}[L_D(B_{t+1})] \le \mathbb{E}_{\mathcal{S},\Xi}[L_D(B_t)] - \frac{\eta}{2}\mathbb{E}_{\mathcal{S},\Xi}[\|\nabla_B L_D(B_t)\|_{\mathrm{F}}^2] + \frac{\eta^2 L_B}{2}\Psi, \tag{22}$$

where $\Psi = \sigma_{\text{sg}}^2/b + \sigma^2 C^2 mr/b^2$.

Step 5: Telescope. Summing from $t = 0$ to $T - 1$:

$$\mathbb{E}_{\mathcal{S},\Xi}[L_D(B_T)] \leq L_D(B_0) - \frac{\eta}{2}\sum_{t=0}^{T-1}\mathbb{E}_{\mathcal{S},\Xi}[\|\nabla_B L_D(B_t)\|_{\mathrm{F}}^2] + \frac{T\eta^2 L_B}{2}\Psi. \tag{23}$$

Rearranging:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{\mathcal{S},\Xi}[\|\nabla_B L_D(B_t)\|_{\mathrm{F}}^2] \leq \frac{2\Delta_D}{\eta T} + \eta L_B \Psi. \tag{24}$$

Step 6: Optimize step size. The unconstrained optimizer is $\eta_0 = \sqrt{2\Delta_D/(TL_B\Psi)}$. After capping at $1/L_B$ and applying the two-case argument in the main proof:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{\mathcal{S},\Xi}[\|\nabla_B L_D(B_t)\|_{\mathrm{F}}^2] \leq 2\sqrt{\frac{2\Delta_D L_B \Psi}{T}} + \frac{2\Delta_D L_B}{T}. \tag{25}$$

Step 7: Retain the accounted noise. Use σ from Lemma 3; the DP component is $2C\sigma\sqrt{2\Delta_D L_B mr/T}/b$. The previously substituted small-q rate did not have established validity conditions, so no asymptotic rate is deduced from that substitution. □

C Scope of Minimax Comparisons

A restricted-coordinate dimension alone is insufficient to certify that a convex population lower bound matches the empirical stationarity bound here. A matching minimax claim would require a lower bound for the same objective, parameter class, privacy model, and stationarity criterion.

D Clipping Bias Analysis

Proof of Proposition 9. Conditional on the current iterate, let $g = \nabla L_D(B_t)$, $u = N^{-1}\sum_i \tilde{g}_i$, and $e = u - g$. The Poisson-plus-noise conditional variance is at most $V = (1-q)C^2/b + \sigma^2 C^2 mr/b^2$. Smoothness and $\eta L_B \leq 1$ give

$$\begin{aligned}
\mathbb{E}_t L_D(B_{t+1}) &\leq L_D(B_t) - \eta\langle g, u\rangle + \tfrac{\eta^2 L_B}{2}(\|u\|_F^2 + V) \\
&= L_D(B_t) - \tfrac{\eta}{2}\|g\|_F^2 + \tfrac{\eta}{2}\|e\|_F^2 - \tfrac{\eta}{2}(1-\eta L_B)\|u\|_F^2 + \tfrac{\eta^2 L_B}{2}V \\
&\leq L_D(B_t) - \tfrac{\eta}{2}\|g\|_F^2 + \tfrac{\eta}{2}\|e\|_F^2 + \tfrac{\eta^2 L_B}{2}V.
\end{aligned}$$

Taking total expectations, telescoping, and dividing by $\eta T/2$ proves the stated squared-bias bound. This also handles cancellation among unclipped record gradients, which need not be preserved by clipping.

When Assumption 2 does not hold with $G \leq C$, clipping introduces bias. We bound this bias for LoRA gradients.

Lemma 14 (Clipping bias bound)**.** *Let $g = \nabla_B \ell(z; W_0 + BA)$ and $\tilde{g} = g \cdot \min(1, C/\|g\|_{\mathrm{F}})$. Then:*

$$\|\mathbb{E}[\tilde{g}] - \nabla L(B)\|_{\mathrm{F}} \leq \mathbb{E}\left[\|g\|_{\mathrm{F}} \cdot \mathbf{1}_{\|g\|_{\mathrm{F}} > C}\right]. \tag{26}$$

Moreover, under Lemma 4, $\|g\|_{\mathrm{F}} \leq \kappa\|\nabla_W \ell\|_{\mathrm{F}}$, so setting $C = \kappa C_W$ (where C_W would be the clipping bound for full gradients) ensures only an upper bound on the LoRA clipping event, not equal clipping bias.

Proof. $\mathbb{E}[\tilde{g}] - \mathbb{E}[g] = \mathbb{E}[g(\min(1, C/\|g\|_{\mathrm{F}}) - 1)] = -\mathbb{E}[g \cdot (1 - C/\|g\|_{\mathrm{F}})_+]$.

Taking norms: $\|\mathbb{E}[\tilde{g}] - \mathbb{E}[g]\|_{\mathrm{F}} \leq \mathbb{E}[\|g\|_{\mathrm{F}} \cdot (1 - C/\|g\|_{\mathrm{F}})_+] = \mathbb{E}[(\|g\|_{\mathrm{F}} - C)_+] \leq \mathbb{E}[\|g\|_{\mathrm{F}} \cdot \mathbf{1}_{\|g\|_{\mathrm{F}} > C}]$. □

E Optimal Rank for Random A

The earlier linear-noise rank threshold relied on the withdrawn squared PL rate. It is not retained. The random-subspace approximation calculation remains valid under its stated smoothness and independence hypotheses; it must be combined with the full finite-T optimization bound.

F Assumption-by-Assumption Comparison with Related Methods

Table 6: Methods differ in how a low-dimensional structure is obtained and in what is optimized or released; their guarantees are therefore not interchangeable.

Method	**Subspace source**	**Optimized/released object**	**Mechanism and scope**
Standard DP-SGD	None	Full parameter update	Per-example clipping and Gaussian noise; general baseline.
Projected DP-SGD (Zhou et al., 2021)	Gradient subspace estimated from public data	Projected update	Formal DP optimization with a public-data assumption.
RGP (Yu et al., 2021)	Gradient-carrier matrices estimated from past updates	Reconstructed noisy full-weight update	Low-rank reparameterization designed for large-scale private learning.
Private PEFT (Yu et al., 2022; Li et al., 2022)	Chosen parameter-efficient architecture	Adapter or selected parameters	Primarily empirical privacy–utility evidence across fine-tuning choices.
This work	Fixed random A, independent of private data	Frozen-A adapter B	Ordinary DP-SGD; formal sensitivity and smoothness transfer only for frozen A.

G Polyak-Łojasiewicz Condition for Linear LoRA

Proposition 15 (PL holds for linear LoRA)**.** *Consider linear regression with LoRA: $L(B) = \frac{1}{N}\left\|Y - X(W_0 + BA)^\top\right\|_{\mathrm{F}}^2$ where $X \in \mathbb{R}^{N\times n}$ is the data matrix with empirical covariance $\Sigma_x = X^\top X/N$. If $\Sigma_x \succ 0$ and $A \in \mathbb{R}^{r\times n}$ has full row rank, then $L(B)$ satisfies the PL condition with constant:*

$$\mu = 2\sigma_{\min}^2(A) \cdot \lambda_{\min}(\Sigma_x). \tag{27}$$

For Gaussian A and $r \le n$, full row rank holds almost surely, hence $\mu > 0$. This does not give a dimension-uniform lower bound near $r = n$: a positive finite-dimensional smallest singular value can still be arbitrarily small.

Proof. The loss is quadratic in B: $L(B) = \frac{1}{N}\left\|(Y - XW_0^\top) - XA^\top B^\top\right\|_{\mathrm{F}}^2$. Let $Z = XA^\top \in \mathbb{R}^{N\times r}$. Then $\nabla^2_{\mathrm{vec}(B)} L = \frac{2}{N}(Z^\top Z \otimes I_m) = 2(A\Sigma_x A^\top \otimes I_m)$. Since $\Sigma_x \succ 0$ and A has full row rank, $A\Sigma_x A^\top \succ 0$ with smallest eigenvalue at least $\sigma^2_{\min}(A)\lambda_{\min}(\Sigma_x)$. The Hessian is PD, so L is strongly convex (hence PL) with $\mu = 2\sigma^2_{\min}(A)\lambda_{\min}(\Sigma_x)$. □

H Multi-layer Extension

For a model with ℓ LoRA-adapted layers, each with weight $W^{(l)} = W_0^{(l)} + B^{(l)}A^{(l)}$ where $B^{(l)} \in \mathbb{R}^{m_l\times r}$ and $A^{(l)} \in \mathbb{R}^{r\times n_l}$:

- Total trainable parameters: $p = \sum_{l=1}^{\ell} m_l \cdot r = r\sum_l m_l$.
- DP noise is $\xi \sim \mathcal{N}(0, \sigma^2 C^2 I_p)$, with $\mathbb{E}[\|\xi\|_2^2] = \sigma^2 C^2 p$.
- Theorem 8 applies with mr replaced by $p = r\sum_l m_l$.

- For full fine-tuning: $d = \sum_l m_l n_l$. If all layers have the same dimensions $m \times n$, the ratio is $p/d = r\ell m/(\ell mn) = r/n$.

The multi-layer calculation requires a global smoothness assumption on the loss with respect to all frozen-A LoRA parameters jointly. Under that assumption, the stationarity proof carries through with mr replaced by p; this does not establish PL for a multi-layer nonlinear model.

I Both-Trained LoRA: Bilinear Noise Analysis

When both A and B are trained, the parameter vector is $\theta = (\mathrm{vec}(B), \mathrm{vec}(A)) \in \mathbb{R}^{r(m+n)}$. The DP noise is $\xi = (\xi_B, \xi_A) \sim \mathcal{N}(0, \sigma^2 C^2 I_{r(m+n)})$, contributing $\mathbb{E}[\|\xi\|_2^2] = \sigma^2 C^2 r(m+n)$ to the variance.

The additional complication arises in the smoothness analysis. The update to W is: $\Delta W = (B + \eta(\bar{g}_B + \xi_B/b))(A + \eta(\bar{g}_A + \xi_A/b)) - BA$.

Expanding: $\Delta W = \eta(B \cdot \bar{g}_A + \bar{g}_B \cdot A) + \eta(B \cdot \xi_A/b + \xi_B \cdot A/b) + \eta^2(\bar{g}_B + \xi_B/b)(\bar{g}_A + \xi_A/b)$.

The bilinear noise term $\eta^2 \xi_B \xi_A / b^2$ has the following expected squared norm. Here $\xi_B \in \mathbb{R}^{m \times r}$ and $\xi_A \in \mathbb{R}^{r \times n}$ with i.i.d. $\mathcal{N}(0, \sigma^2 C^2)$ entries, and ξ_B, ξ_A are independent:

$$\begin{aligned}
\mathbb{E}[\|\xi_B \xi_A\|_{\mathrm{F}}^2] &= \mathbb{E}[\mathrm{tr}(\xi_A^\top \xi_B^\top \xi_B \xi_A)] = \mathrm{tr}(\mathbb{E}[\xi_A^\top \mathbb{E}[\xi_B^\top \xi_B] \xi_A]) \\
&= \mathrm{tr}(\mathbb{E}[\xi_A^\top \cdot m\sigma^2 C^2 I_r \cdot \xi_A]) = m\sigma^2 C^2 \cdot \mathbb{E}[\|\xi_A\|_{\mathrm{F}}^2] \\
&= m\sigma^2 C^2 \cdot rn \cdot \sigma^2 C^2 = mnr \cdot \sigma^4 C^4. \qquad (28)
\end{aligned}$$

Thus $\mathbb{E}[\|\xi_B \xi_A / b^2\|_{\mathrm{F}}^2] = mnr \cdot \sigma^4 C^4 / b^4$, and the contribution to a descent calculation from the η^2-order bilinear noise is:

$$\frac{\eta^4 L_{AB}}{2} \cdot \frac{mnr \cdot \sigma^4 C^4}{b^4}. \qquad (29)$$

The direct first-order noise dimension is smaller than the full mn dimension only when $r(m+n) < mn$. Relative to full-space first-order noise, the product-noise diagnostic scales as $\eta^2 r \sigma^2 C^2 / b^2$ and need not be negligible. A complete convergence result would additionally require uniform control of $\|A_t\|_{\mathrm{op}}$ and $\|B_t\|_{\mathrm{op}}$ and joint smoothness along the trajectory. This appendix supplies only the noise calculation, not such a theorem.

J Strongly Convex Case

The earlier strongly-convex and convex population-rate shortcuts are not retained. Global strong convexity and uniformly bounded gradients on an unbounded Gaussian trajectory are incompatible assumptions. A constrained projected algorithm needs its own analysis; stationarity of an unprojected algorithm alone does not imply the stated final-iterate convex risk rate.

K Numerical Diagnostics and Feasibility Checks

We use synthetic problems to check the algebraic scaling and a small SST-2 study to check that the frozen-A procedure can be run in practice. These experiments do not establish the PL assumption or the oracle optimal-rank law. Full details and code are in the supplementary material.

Setup. We consider a quadratic loss $L(B) = \frac{1}{2}\|BA - \Delta W^\star\|_{\mathrm{F}}^2$ with $m = n = 32$, rank-r frozen $A \sim \mathcal{N}(0, I/n)$, and $\Delta W^\star$ having singular values $\sigma_j = j^{-1.5}$ (spectral decay with $\alpha = 1.5$). We run DP-SGD with $N = 500$ samples, $T = 300$ iterations, batch size 32, and clipping bound $C = 3$.

Results (Figure 2). In panel (a), the best tested ranks in the actual DP-SGD runs are 1, 1, and 32 for $\varepsilon = 1, 2, 4$, respectively. These runs use random A, whereas Corollary 13 assumes informed A, so they are not a direct test of that corollary. Panel (b) verifies the noise construction: its empirical energy ratio agrees with r/n. Panel (c) is only an algebraic consistency check because its grid search minimizes the same surrogate used to derive $r^\star$.

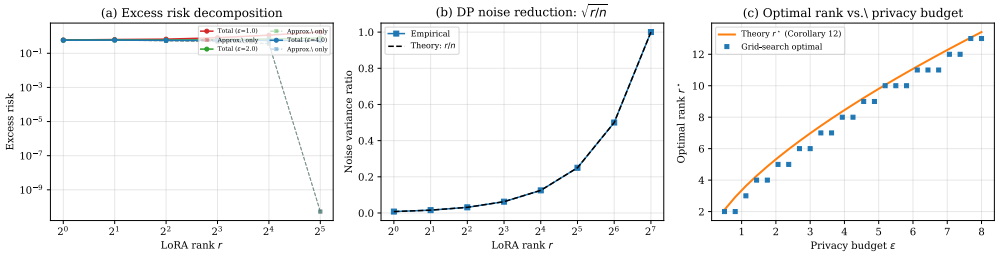

Figure 2: Synthetic diagnostics. **(a)** Excess risk from actual random-A DP-SGD runs across ranks and privacy budgets. **(b)** Direct check of the constructed Gaussian-noise energy ratio. **(c)** Formula check: both the closed form and grid search optimize the same analytic surrogate, so their agreement is not independent validation.

SST-2 checks. The submitted experiment fine-tunes a two-layer MLP over fixed GPT-2 embeddings. Its accuracies cluster near 81–82%, leaving a ceiling effect that makes rank differences difficult to interpret. Moreover, W_0 was trained non-privately on the same subset: the reported ε covers only the subsequent DP stage conditional on W_0, not that pre-training. Appendix A records this privacy boundary and a BERT-base comparison using disjoint public warm-start and private partitions. The latter is a multi-layer Transformer feasibility check, not evidence for PL, an interior optimum, or the rank formula.

L Complete controlled mechanism matrix

We completed the preregistered 1,200 DP and 600 zero-noise runs: two teacher ranks, ten parameterizations (nine frozen-A combinations and a full-space control), three clipping ratios, two optimizers, and five independent seeds. The DP branches target $\varepsilon \in \{1, 4\}$ at $\delta = 10^{-5}$ with Poisson sampling rate $64/2048$ for 1,000 updates. Every run retains all 50 trace points. A separate development seed selected one AdamW learning rate, 10^{-3}, by averaging the certified restricted excess over the fixed development grid. SGD uses the single conservative B-space step size specified by the largest $\kappa = 2$ smoothness bound. Formal outcomes do not select hyperparameters.

Lower restricted error does not establish better total utility. The following fixed reference slice uses $\kappa = 1$, clipping ratio one and $\varepsilon = 4$; entries average all five seeds. The full matrix and 108 paired rank comparisons are retained, rather than selecting the best formal cell.

Teacher	Optimizer	Rank	Projection gap	Restricted excess	Test objective
2	adamw	2	0.04794	0.00222	3.91806
2	sgd	2	0.04794	0.01603	3.93238
2	adamw	8	0.03700	0.00680	3.91224
2	sgd	8	0.03700	0.06723	3.97346
2	adamw	32	0.00000	0.02227	3.89318
2	sgd	32	0.00000	0.27762	4.15151
2	adamw	Full	0.00000	0.02246	3.89353
2	sgd	Full	0.00000	0.27936	4.15274
16	adamw	2	0.06016	0.00170	3.91897
16	sgd	2	0.06016	0.01615	3.93391
16	adamw	8	0.04842	0.00633	3.91267
16	sgd	8	0.04842	0.06723	3.97493
16	adamw	32	-0.00000	0.02593	3.88700
16	sgd	32	-0.00000	0.27800	4.14294
16	adamw	Full	0.00000	0.02618	3.88744
16	sgd	Full	0.00000	0.27913	4.14507

At zero update, both the training and held-out regularized objectives equal $\sqrt{32}\log 2 = 3.92103257$ exactly. In this slice, all SGD variants are worse than zero update, even though lower ranks improve on the noisy full-space SGD control. AdamW exhibits a different ordering: the full-rank held-out

objectives are lower than those at rank two. The optimizers have different step-size rules, so this is not a claim of optimizer superiority after matched hyperparameter optimization. It does show why a smaller restricted error or noise-energy term alone does not imply useful overall adaptation.

Interpretation of interventions. We archive all DP-minus-zero-noise comparisons at matched settings and all clipped-minus-unclipped comparisons in the zero-Gaussian-noise branch. Endpoint differences include changes in optimization and clipping trajectories; they are not separable additive Gaussian and clipping contributions to loss. Likewise, varying κ at fixed B-space step size changes the effective W-space update. It is not evidence for a coordinate-invariant advantage. The complete experiment supports an inspectable mechanism analysis, not a universal rank recommendation, a BERT PL condition, or an AdamBC theorem. Repeated executions at a fixed configuration and seed contribute only one scientific replicate; retention does not depend on the observed outcome.

M Generalization Bound

Theorem 16 (PAC-Bayes generalization bound for DP-LoRA)**.** *Let $P = \mathcal{N}(0, \lambda^2 I_{mr})$ be a data-independent prior over $\mathrm{vec}(B) \in \mathbb{R}^{mr}$. Let $Q = \mathcal{N}(\mathrm{vec}(B_T), \gamma^2 I_{mr})$ be a posterior centered at the DP-LoRA output. Assume $\ell(z; W_0 + BA) \in [0,1]$ for all z, B, and that ℓ is L_ℓ-Lipschitz in $\mathrm{vec}(B)$. Then with probability at least $1-\delta$ over the training data:*

$$L_{\mathrm{pop}}(B_T) \le L_D(B_T) + 2L_\ell \gamma \sqrt{mr} + \sqrt{\frac{\mathrm{KL}(Q\|P) + \log(2\sqrt{N}/\delta)}{2N}}, \tag{30}$$

where the term $2L_\ell\gamma\sqrt{mr}$ accounts for both empirical and population gaps between the Gibbs predictor $\mathbb{E}_{B\sim Q}[\ell(\cdot; B)]$ and the deterministic predictor $\ell(\cdot; B_T)$, and

$$\mathrm{KL}(Q\|P) = \frac{1}{2}\left(\frac{mr\cdot\gamma^2}{\lambda^2} + \frac{\|B_T\|_{\mathrm{F}}^2}{\lambda^2} - mr - mr\log\frac{\gamma^2}{\lambda^2}\right). \tag{31}$$

Proof sketch. Apply the PAC-Bayes–McAllester bound (McAllester, 1999) to the Gibbs predictor, then transfer to B_T via Lipschitz continuity: $|L_{\mathrm{pop}}(B_T) - \mathbb{E}_{B\sim Q}[L_{\mathrm{pop}}(B)]| \le L_\ell\gamma\sqrt{mr}$. The empirical Gibbs-to-mean transfer contributes a second $L_\ell\gamma\sqrt{mr}$. Prior hyperparameters and W_0, A must be independent of the fine-tuning sample. No unconditional r/n generalization improvement follows because the norms and Lipschitz constants can differ. □

Numerical diagnostics. Synthetic and SST-2 feasibility checks, including their privacy boundaries, are reported in Appendix K.

N Main descent calculation

Proof. By L_B-smoothness of L_D:

$$\begin{aligned} L_D(B_{t+1}) &\le L_D(B_t) + \langle \nabla L_D(B_t), B_{t+1} - B_t\rangle_{\mathrm{F}} + \frac{L_B}{2}\|B_{t+1} - B_t\|_{\mathrm{F}}^2 \\ &= L_D(B_t) - \eta\langle \nabla L_D(B_t), \bar{g}_t\rangle_{\mathrm{F}} + \frac{\eta^2 L_B}{2}\|\bar{g}_t\|_{\mathrm{F}}^2 . \end{aligned} \tag{32}$$

Taking the conditional expectation over Ξ_t and $\mathcal{S}_t$, with D and A fixed, and using $G \le C$:

$$\mathbb{E}_{\mathcal{S},\Xi}[\langle \nabla L_D(B_t), \bar{g}_t\rangle_{\mathrm{F}}] = \|\nabla L_D(B_t)\|_{\mathrm{F}}^2 . \tag{33}$$

For the second moment of the noisy gradient:

$$\begin{aligned} \mathbb{E}[\|\bar{g}_t\|_{\mathrm{F}}^2] &= \mathbb{E}\left[\left\|\frac{1}{b}\sum_{i\in\mathcal{S}_t}\tilde{g}_i + \frac{\Xi_t}{b}\right\|_{\mathrm{F}}^2\right] \\ &= \|\nabla L_D(B_t)\|_{\mathrm{F}}^2 + \frac{q(1-q)}{b^2}\sum_i \|g_i\|_{\mathrm{F}}^2 + \frac{\sigma^2 C^2\cdot mr}{b^2} \\ &\le \|\nabla L_D(B_t)\|_{\mathrm{F}}^2 + \frac{\sigma_{\mathrm{sg}}^2}{b} + \frac{\sigma^2 C^2\cdot mr}{b^2}. \end{aligned} \tag{34}$$

The DP noise term $\sigma^2 C^2 mr/b^2$ arises because $\Xi_t \sim \mathcal{N}(0, \sigma^2 C^2 I_{mr})$ has $\mathbb{E}[\|\Xi_t\|_{\mathrm{F}}^2] = \sigma^2 C^2 \cdot mr$. Substituting (34) into (32) and rearranging:

$$\mathbb{E}_{\mathcal{S},\Xi}[L_D(B_{t+1})] \leq \mathbb{E}_{\mathcal{S},\Xi}[L_D(B_t)] - \eta\left(1 - \frac{\eta L_B}{2}\right)\mathbb{E}_{\mathcal{S},\Xi}[\|\nabla L_D(B_t)\|_{\mathrm{F}}^2] + \frac{\eta^2 L_B}{2}\left(\frac{\sigma_{\mathrm{sg}}^2}{b} + \frac{\sigma^2 C^2 mr}{b^2}\right). \tag{35}$$

For $\eta \leq 1/L_B$, we have $1 - \eta L_B/2 \geq 1/2$. Telescoping over $t = 0, \ldots, T-1$:

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}_{\mathcal{S},\Xi}[\|\nabla L_D(B_t)\|_{\mathrm{F}}^2] \leq \frac{2\Delta_D}{\eta T} + \eta L_B \Psi. \tag{36}$$

For completeness, set $a = 2\Delta_D L_B/T$ and $\eta_0 = \sqrt{2\Delta_D/(TL_B\Psi)}$. If $\eta_0 \leq 1/L_B$, direct substitution gives $2\sqrt{a\Psi}$. Otherwise $\Psi < a$ and the boundary choice $\eta^\star = 1/L_B$ gives $a + \Psi \leq a + \sqrt{a\Psi} \leq a + 2\sqrt{a\Psi}$. This proves (7) in both cases. The DP component (9) follows by inserting its exact variance into the finite-T bound and using $L_B \leq \kappa^2 L_W$. □