ON THE CONVERGENCE OF LORA UNDER GENERALIZED SMOOTHNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-rank adaptation (LoRA) has become a popularly used parameter-efficient approach for fine-tuning large pretrained models via replacing the full weight update with a factorized update $\Delta W = BA$. Despite its practical utility, existing convergence theory for LoRA remains limited: some prior analyses handle this factorized optimization problem by imposing boundedness assumptions on the LoRA factors, an additional assumption not guaranteed by LoRA updates. More recent results remove this assumption but still rely on L-smoothness with respect to the weight variable W. In practice, empirical studies of neural network training show that the Hessian norm can grow with the gradient norm, violating the L-smoothness assumption on the bounded Hessian norm. Therefore, we analyze LoRA-GD and LoRA-SGD under generalized smoothness without assuming the bounded factor norms. For LoRA-GD, we prove an $O(\epsilon^{-4})$ upper bound and the first $\Omega(\epsilon^{-4})$ lower bound, even when the stepsize is chosen adaptively from the optimization history. For LoRA-SGD with unbiased stochastic gradients of bounded variance, we obtain an $O(\epsilon^{-8})$ high-probability guarantee. We also prove an $\Omega(\epsilon^{-8})$ lower bound at fixed confidence for adaptive stepsize policies $\eta_t \in [0, 1]$ that may depend on the observed optimization history and the current stochastic gradient. Experiments support our theoretical analysis: the Hessian norm grows with the gradient norm, the factor norms grow during optimization, and this growth amplifies stochastic factor-gradient noise.

1 INTRODUCTION

Low-rank adaptation (LoRA) has become a widely used parameter-efficient approach for improving the large pretrained models, reducing the number of trainable parameters by representing the weight update as a low-rank product $\Delta W = BA$ (Hu et al., 2022). However, this representation introduces difficulty to the convergence analysis. Let $F(X) := \mathcal{L}(W_{\text{base}} + X)$ denote the objective with respect to the weight variable X, and let $J(B, A) := F(BA)$ denote the corresponding factorized objective. Since the map $(B, A) \mapsto BA$ is bilinear, the gradient with respect to each factor depends on the other factor. As a result, the local smoothness of J depends on the factor norms, thus J is generally not globally smooth in (B, A) even when F is L-smooth in X.

Some previous convergence analyses skip this difficulty by imposing explicit boundedness assumptions on the LoRA factors (Jiang et al., 2024), an additional assumption not guaranteed by LoRA updates. Recent works began to remove this assumption: Mu & Klabjan (2025) established non-asymptotic convergence of LoRA gradient descent without bounded factors, and Wang et al. (2026) further obtained an $O(\epsilon^{-4})$ iteration complexity. These results, however, still assume that F is globally L-smooth on the weight variable X, which imposes a uniform bound on the operator norm of the Hessian $\|\nabla^2 F(X)\|_{\text{op}}$. Empirical studies of neural-network training instead show that the Hessian norm tends to increase with the gradient norm (Zhang et al., 2020). Motivated by this observation, generalized smoothness relaxes this uniform bound by allowing the Hessian norm to grow with the gradient norm (Li et al., 2023), providing a more realistic alternative to global L-smoothness. For example, $\|\nabla^2 F(X)\|_{\text{op}} \le L_0 + L_1\|\nabla F(X)\|_F$. Therefore, we ask: *how do LoRA-GD and LoRA-SGD converge under generalized smoothness without assuming the bounded factor norms?*

The main challenge in conducting LoRA analysis under generalized smoothness is that our analysis cannot rely on a uniform priori bound on either the Hessian norm of F or the factor norms. Under

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.

generalized smoothness, the Hessian norm of F can grow with $\|\nabla F(X)\|_F$. As discussed above, the local smoothness of the factorized objective also depends on the factor norms. Therefore, without assuming bounded factors, the analysis must control the gradient norm $\|\nabla F(X)\|_F$ and the factor norms jointly. The stochastic setting introduces an additional challenge. Even when the stochastic weight-gradient noise has uniformly bounded variance, the noise in two LoRA factor gradients is multiplied by A and B, respectively. Thus, as the factor norms grow, the factor-gradient noise grows correspondingly. These two issues, *gradient-factor coupling* and *factor-amplified stochastic noise*, are the main difficulties in analyzing LoRA-GD and LoRA-SGD, respectively.

In this paper, we first investigate the convergence complexity of LoRA-GD. Under the sub-quadratic generalized smoothness, we show that LoRA-GD reaches ϵ-stationarity of J in $O(\epsilon^{-4})$ iterations without assuming bounded factor norms. This retains the $O(\epsilon^{-4})$ rate in (Wang et al., 2026) while extending the analysis from global L-smoothness to generalized smoothness. More importantly, we prove the first lower bound for LoRA-GD, $\Omega(\epsilon^{-4})$ (matching upper bound), on a rank-one LoRA instance with a 1-smooth, lower-bounded weight-space objective F. The lower bound holds even when the stepsize is chosen adaptively from the optimization history. Together, these results establish a tight $\Theta(\epsilon^{-4})$ worst-case complexity for this class of LoRA-GD methods.

We further analyze LoRA-SGD. Under finite-variance stochastic gradients and sub-quadratic generalized smoothness, we prove that LoRA-SGD achieves ϵ-stationarity of J in $O(\epsilon^{-8})$ iterations with high probability, without assuming bounded factor norms. The same result also yields an $O(\epsilon^{-10})$ complexity for the expected best-iterate squared gradient. We then prove an $\Omega(\epsilon^{-8})$ constant-confidence lower bound on a rank-one LoRA instance whose weight-space objective F is globally 1-smooth and lower bounded. The lower bound allows the stepsize $\eta_t \in [0, 1]$ to depend on the full observed trajectory and the current stochastic gradient, and thus covers bounded global gradient clipping and other bounded scalar normalizations.

Our main contributions can be summarized as follows:

- **Tight convergence complexity of LoRA-GD.** Under sub-quadratic generalized smoothness, we show that LoRA-GD reaches ϵ-stationarity of J in $O(\epsilon^{-4})$ iterations without assuming bounded factor norms. We further prove the first $\Omega(\epsilon^{-4})$ lower bound when the stepsize is chosen adaptively from the optimization history, establishing a tight $\Theta(\epsilon^{-4})$ worst-case complexity for this class.
- **Convergence complexity of LoRA-SGD.** Under finite-variance stochastic gradients and sub-quadratic generalized smoothness, we derive an $O(\epsilon^{-8})$ high-probability upper bound without assuming bounded factor norms, together with an $\Omega(\epsilon^{-8})$ lower bound for constant-confidence convergence over adaptive stepsizes satisfying $\eta_t \in [0, 1]$. The high-probability upper bound further implies an $O(\epsilon^{-10})$ expected best-iterate complexity.
- **Empirical validation.** We empirically examine three phenomena central to our analysis: the relationship between the Hessian and gradient norms motivating generalized smoothness, the growth of the LoRA factor norms during optimization, and the amplification of stochastic factor-gradient noise as the factors grow.

2 Related Work

LoRA convergence theory. For full-gradient LoRA-GD, Mu & Klabjan (2025) gave a non-asymptotic guarantee without bounded factors, and Wang et al. (2026) improved the complexity to $O(\epsilon^{-4})$ under global L-smoothness. The latter also studied stochastic LoRA: LoRA-NSGDM combines joint normalization with momentum and achieves $O(\epsilon^{-8})$ stochastic oracle complexity under unbiased finite-variance gradients, while LoRA-STORM improves this to $O(\epsilon^{-6})$ under additional mean-square smoothness. In contrast, under generalized smoothness and without bounded factors, we obtain an $O(\epsilon^{-4})$ upper bound for simultaneous LoRA-GD and an $O(\epsilon^{-8})$ high-probability upper bound for simultaneous LoRA-SGD, together with $\Omega(\epsilon^{-4})$ and $\Omega(\epsilon^{-8})$ lower bounds for adaptive stepsizes. Other work studied randomized LoRA updates (Malinovsky et al., 2024; Sokolov et al., 2025) or the LoRA loss landscape (Kim et al., 2025). Appendix H provides a detailed comparison of assumptions, update rules, stationarity guarantees, and lower-bound scopes.

Optimization under generalized smoothness. Zhang et al. (2020) studied clipped and normalized gradient descent under (L_0, L_1)-smoothness, where local smoothness can grow linearly with the

gradient norm. Li et al. (2023) generalized this relation to broader functions of the gradient norm and establish convergence guarantees for gradient descent, stochastic gradient descent, and Nesterov's accelerated gradient method. These results assume that the smoothness condition is stated for the same parameters that the algorithm updates. In LoRA, however, smoothness is assumed for the weight variable, while gradient descent updates its two low-rank factors. Because the smoothness of this factorized objective also depends on the factor norms, the existing generalized-smoothness guarantees do not directly apply to LoRA-GD and LoRA-SGD.

Lower bounds for nonconvex stationarity. Classical lower bounds study algorithms that optimize a smooth nonconvex objective directly in its own parameter space. For globally smooth objectives, deterministic first-order methods require $\Omega(\epsilon^{-2})$ queries in the worst case (Carmon et al., 2020), while under unbiased stochastic gradients with bounded variance, stochastic first-order methods require $\Omega(\epsilon^{-4})$ queries (Arjevani et al., 2023). LoRA has a different structure: even when the weight-space objective F is globally smooth, the factorized objective $J(B, A) = F(BA)$ optimized by LoRA is generally not globally smooth in (B, A). We show that this LoRA factorization leads to stronger barriers for the LoRA methods considered here: $\Omega(\epsilon^{-4})$ for deterministic LoRA-GD and $\Omega(\epsilon^{-8})$ for constant-confidence LoRA-SGD.

3 Problem Setup

We use $\langle U, V\rangle := \mathrm{tr}(U^\top V)$ for the Frobenius inner product, $\|\cdot\|_F$ for the Frobenius norm, and $\|\mathcal{A}\|_{\mathrm{op}} := \sup_{\|U\|_F=1} \|\mathcal{A}[U]\|_F$ for the operator norm of a linear map $\mathcal{A}$.

LoRA objective and stationarity. Let $W_{\mathrm{base}} \in \mathbb{R}^{m\times n}$ be a frozen weight matrix and $\mathcal{L} : \mathbb{R}^{m\times n} \to \mathbb{R}$ a differentiable loss. For a trainable weight update X, define $F(X) := \mathcal{L}(W_{\mathrm{base}} + X)$. LoRA parameterizes $X = BA$[1] with $B \in \mathbb{R}^{m\times r}$ and $A \in \mathbb{R}^{r\times n}$, and optimizes $J(B, A) := F(BA)$. By the chain rule,

$$\nabla_B J(B, A) = \nabla F(BA)A^\top, \qquad \nabla_A J(B, A) = B^\top \nabla F(BA). \tag{1}$$

We measure convergence with respect to the LoRA factors.

Definition 3.1 (ϵ-stationarity)**.** A pair (B, A) is ϵ-stationary if

$$\|\nabla J(B, A)\|_F^2 := \|\nabla_B J(B, A)\|_F^2 + \|\nabla_A J(B, A)\|_F^2 \le \epsilon^2.$$

For a sequence of factor iterates $\{(B_t, A_t)\}_{t\ge 0}$, we measure convergence by the best iterate among the first T iterations:

$$\mathcal{G}_T := \min_{0\le t<T} \|\nabla J(B_t, A_t)\|_F.$$

Thus, an ϵ-stationary iterate is reached within T iterations whenever $\mathcal{G}_T \le \epsilon$.

For the same sequence, define

$$X_t := B_t A_t, \qquad J_t := F(X_t), \qquad G_t := \|\nabla F(X_t)\|_F, \qquad R_t := \|A_t\|_F^2 + \|B_t\|_F^2.$$

Assumption 3.2 (Lower boundedness)**.** $F_\star := \inf_X F(X) > -\infty$. Let $\Delta := J_0 - F_\star$.

Generalized smoothness with respect to weight matrix. All smoothness assumptions are imposed on F as a function of the weight update X, rather than on the factorized objective J.

We use the following C^2 version of generalized ℓ-smoothness (Li et al., 2023).

Definition 3.3 (Generalized ℓ-smoothness)**.** Let $\ell : [0, \infty) \to (0, \infty)$ be continuous and nondecreasing. We call $F \in C^2$ generalized ℓ-smooth if

$$\|\nabla^2 F(X)\|_{\mathrm{op}} \le \ell(\|\nabla F(X)\|_F) \qquad \text{for all } X.$$

We use the generalized ℓ-smoothness framework of Li et al. (2023). The choices $\ell(u) \equiv L$ and $\ell(u) = L_0 + L_1 u$ recover L-smoothness and (L_0, L_1)-smoothness (Zhang et al., 2020), respectively.

[1] We omit the fixed LoRA scaling factor for notational simplicity; it can be absorbed by rescaling the factors and the stepsize.

Assumption 3.4 (Sub-quadratic generalized smoothness)**.** F is generalized ℓ-smooth with $\ell(u) = o(u^2)$ as $u \to \infty$.

Why generalized smoothness? Classical convergence analyses often assume global L-smoothness, which requires a single constant L such that $\|\nabla^2 F(X)\|_{\mathrm{op}} \le L$ for every X. This requires a uniform bound on the Hessian norm with respect to the weight matrix X. However, the empirical studies of deep neural networks show a different behavior: the local smoothness can vary substantially along the training trajectory and tend to increase with the gradient norm (Zhang et al., 2020; Crawshaw et al., 2022). These observations motivated the (L_0, L_1)-smoothness condition

$$\|\nabla^2 F(X)\|_{\mathrm{op}} \le L_0 + L_1 \|\nabla F(X)\|_F,$$

which allows the Hessian norm to grow with the gradient norm (Zhang et al., 2020). More generally, Li et al. (2023) replaced the affine bound by

$$\|\nabla^2 F(X)\|_{\mathrm{op}} \le \ell(\|\nabla F(X)\|_F),$$

where ℓ is nondecreasing and satisfies $\ell(u) = o(u^2)$. This condition contains global L-smoothness as the special case $\ell(u) \equiv L$ and (L_0, L_1)-smoothness as the special case $\ell(u) = L_0 + L_1 u$.

The assumption $\ell(u) = o(u^2)$ still provides enough control for convergence. Together with lower boundedness, generalized smoothness gives the following gradient bound in terms of the objective gap (Li et al., 2023):

$$\|\nabla F(X)\|_F^2 \le 2\ell(2\|\nabla F(X)\|_F)\big(F(X) - F_\star\big).$$

Hence, if $F(X) - F_\star \le \Delta$ and $G = \|\nabla F(X)\|_F$, then $G^2 \le 2\ell(2G)\Delta$. Since $\ell(2G) = o(G^2)$, this inequality cannot hold for arbitrarily large G.

Stochastic first-order oracle. For the stochastic setting, given X and a random sample ξ, the oracle returns $g(X;\xi) \in \mathbb{R}^{m\times n}$. Let $\mathcal{F}_t$ denote the history before drawing ξ_t.

Assumption 3.5 (Unbiased finite-variance oracle)**.** For some $\sigma \ge 0$,

$$\mathbb{E}[g(X_t;\xi_t) \mid \mathcal{F}_t] = \nabla F(X_t), \qquad \mathbb{E}\big[\|g(X_t;\xi_t) - \nabla F(X_t)\|_F^2 \,\big|\, \mathcal{F}_t\big] \le \sigma^2.$$

4 LoRA-GD under Generalized Smoothness

We first consider LoRA gradient descent (LoRA-GD), where the two factors are updated simultaneously:

$$B_{t+1} = B_t - \eta_t \nabla_B J(B_t, A_t), \quad A_{t+1} = A_t - \eta_t \nabla_A J(B_t, A_t), \tag{2}$$

where $\eta_t \ge 0$ is the stepsize.

4.1 Upper Bound

Under generalized smoothness, the local Hessian norm grows with the weight-gradient norm $G_t := \|\nabla F(X_t)\|_F$, i.e., $\|\nabla^2 F(X_t)\|_{\mathrm{op}} \le \ell(G_t)$, while the change in $X_t = B_t A_t$ grows with the factor scale R_t. This gradient-factor coupling makes descent depend on both G_t and R_t, so both quantities must be controlled. We show that the simple schedule $\eta_t = \gamma/\sqrt{t+1}$ achieves this without assuming bounded factors.

Theorem 4.1 (Convergence under generalized smoothness)**.** *Suppose Assumptions 3.2 and 3.4 hold. Let*

$$\overline{G} := \sup\{u \ge 0 : u^2 \le 2\ell(2u)\Delta\}, \qquad \overline{L} := \ell(2\overline{G}), \qquad K := \max\{1/32,\, \overline{G},\, 2\overline{L}(R_0 + 2\Delta)\}.$$

For any $\gamma > 0$ satisfying $\gamma + 2\gamma^2 \le 1/(16K)$, LoRA-GD with $\eta_t = \gamma/\sqrt{t+1}$ satisfies, for every $T \ge 1$,

$$\min_{0\le t<T} \|\nabla J(B_t, A_t)\|_F^2 \le \frac{\Delta}{\gamma(\sqrt{T+1} - 1)}. \tag{3}$$

Consequently, $O(\epsilon^{-4})$ iterations suffice to reach an ϵ-stationary iterate.

Remark 4.2 (Specialization to (L_0, L_1)-smoothness). The sub-quadratic condition guarantees that $\overline{G} < \infty$. For $\ell(u) = L_0 + L_1 u$,

$$\overline{G} = 2L_1\Delta + \sqrt{4L_1^2\Delta^2 + 2L_0\Delta}, \qquad \overline{L} = L_0 + 2L_1\overline{G}.$$

In particular, when $L_1 = 0$, this reduces to $\overline{G} = \sqrt{2L_0\Delta}$ and $\overline{L} = L_0$.

Proof sketch. The main difficulty is the gradient-factor coupling: larger factor norms amplify the change in $X_t = B_tA_t$, while generalized smoothness makes the local Hessian norm depend on the weight-gradient norm. Assume inductively that all previous steps satisfy the descent inequality. Then $J_t \le J_0$, so $F(X_t) - F_\star \le \Delta$. This keeps the weight-gradient norm G_t uniformly bounded, and generalized smoothness then bounds the local Hessian norm. The total objective decrease also limits how fast the factor norms can grow, yielding $R_t = O(\sqrt{t})$. Since $\eta_t = \Theta(t^{-1/2})$, its decay matches the $O(\sqrt{t})$ growth of R_t, so the next update remains small enough for the local descent inequality to hold. This closes the induction. Summing the descent inequalities over $t < T$ gives the $O(T^{-1/4})$ stationarity rate. Appendix B.3 gives the full proof.

When the number of iterations T is fixed in advance, the same analysis also applies to a constant stepsize.

Corollary 4.3 (Constant stepsize for a fixed number of iterations)**.** *Under the assumptions of Theorem 4.1, let the number of iterations $T \ge 1$ be fixed in advance. For any $\gamma > 0$ satisfying $\gamma + \gamma^2 \le 1/(16K)$, LoRA-GD with $\eta_t = \gamma/\sqrt{T}$ for $t = 0, \dots, T-1$ satisfies*

$$\min_{0 \le t < T} \|\nabla J(B_t, A_t)\|_F^2 \le \frac{2\Delta}{\gamma\sqrt{T}}.$$

Hence this schedule also reaches an ϵ-stationary iterate in $O(\epsilon^{-4})$ iterations.

4.2 Lower Bound

We next investigate whether the adaptive stepsizes can improve the $O(\epsilon^{-4})$ iteration rate for reaching ϵ-stationarity. At iteration t, let

$$\mathcal{O}_t := \big(B_t, A_t, X_t, F(X_t), \nabla F(X_t), \nabla_B J(B_t, A_t), \nabla_A J(B_t, A_t)\big),$$
$$\mathcal{H}_t := \big(\mathcal{O}_0, \eta_0, \dots, \mathcal{O}_{t-1}, \eta_{t-1}, \mathcal{O}_t\big)$$

denote the current observation and the full optimization history, with $\mathcal{H}_0 = (\mathcal{O}_0)$. A deterministic causal stepsize rule chooses $\eta_t = \Phi_t(\epsilon, \mathcal{H}_t) \in [0, \infty)$ using no evaluations of F or ∇F outside the generated iterates. The same stepsize is used for both factors. This class includes predetermined and history-dependent stepsize schedules. We show that even with history-dependent stepsizes, there exist smooth LoRA instances that require $\Omega(\epsilon^{-4})$ iterations to reach ϵ-stationarity.

Theorem 4.4 (Lower bound for adaptive stepsizes)**.** *Consider rank-one LoRA initialized at $B_0 = 0$ and $A_0 = u^\top$, where $u \in \mathbb{R}^n$ is any unit vector. For every causal stepsize rule above and every $0 < \epsilon \le 1/2$, there exists a lower-bounded objective $F \in C^2(\mathbb{R}^{m\times n})$ such that*

$$F(0) - F_\star = \frac{25}{2}, \qquad \|\nabla^2 F(X)\|_{\mathrm{op}} \le 1 \quad \textit{for all } X,$$

and

$$\|\nabla J(B_t, A_t)\|_F = 2\epsilon \qquad \textit{for all } 0 \le t < \left\lfloor \frac{1}{16\epsilon^4} \right\rfloor. \tag{4}$$

Thus reaching ϵ-stationarity requires at least $\big\lfloor 1/(16\epsilon^4) \big\rfloor = \Omega(\epsilon^{-4})$ iterations.

Importantly, the hard instance is globally 1-smooth, a special case of the generalized smoothness class. Thus, the $\Omega(\epsilon^{-4})$ lower bound is not caused by the weaker generalized-smoothness assumption, but by the bilinear LoRA factorization.

Proof sketch. The construction exploits the gradient-factor coupling along a rank-one trajectory $B_t = b_t v$ and $A_t = a_t u^\top$, with $R_t = a_t^2 + b_t^2$. At each iteration, we set the weight-gradient magnitude to $G_t = 2\epsilon/\sqrt{R_t}$, so the chain rule gives $\|\nabla J(B_t, A_t)\|_F = G_t\sqrt{R_t} = 2\epsilon$ for every $0 \le t < \lfloor 1/(16\epsilon^4) \rfloor$. As the weight-gradient magnitude decreases, the growing factor scale prevents the factor-gradient norm from decreasing, keeping it fixed at 2ϵ. Appendix C constructs a single lower-bounded C^2, globally 1-smooth objective whose values and weight gradients at the visited points $X_t = B_tA_t$ reproduce this trajectory, while $F(0) - F_\star = 25/2$ remains independent of ϵ.

5 LoRA-SGD under Generalized Smoothness

We now consider stochastic optimization under the same generalized smoothness assumption. At iteration t, we draw ξ_t and apply the stochastic weight gradient $g(X_t;\xi_t)$ through the LoRA chain rule:

$$B_{t+1} = B_t - \eta_t g(X_t;\xi_t)A_t^\top, \quad A_{t+1} = A_t - \eta_t B_t^\top g(X_t;\xi_t). \tag{5}$$

The same stochastic weight gradient is therefore used to update both factors. For the upper-bound analysis, the stepsize is fixed before the stochastic samples are drawn.

We first characterize the stochastic error in factor space and then establish high-probability, expected best-iterate, and lower-bound guarantees for LoRA-SGD.

5.1 Factor-Amplified Stochastic Noise

Let $E_t := g(X_t;\xi_t) - \nabla F(X_t)$. By the LoRA chain rule, the stochastic terms in the B- and A-gradients are $E_tA_t^\top$ and $B_t^\top E_t$, respectively. We denote them jointly by $\zeta_t := (E_tA_t^\top, B_t^\top E_t)$. By Assumption 3.5,

$$\mathbb{E}\big[\|\zeta_t\|_F^2 \mid \mathcal{F}_t\big] \le R_t\,\mathbb{E}\big[\|E_t\|_F^2 \mid \mathcal{F}_t\big] \le \sigma^2 R_t. \tag{6}$$

Thus, although $\mathbb{E}[\|E_t\|_F^2 \mid \mathcal{F}_t] \le \sigma^2$ uniformly, the corresponding conditional second moment for the factor gradient noise is bounded by $\sigma^2 R_t$ and can therefore grow with R_t.

For nonconvex SGD with unbiased stochastic gradients of bounded variance, Li et al. (2023) established the classical $O(\epsilon^{-4})$ gradient complexity with high probability under sub-quadratic generalized smoothness. In LoRA-SGD, by contrast, the conditional second moment of the stochastic factor gradient noise is bounded by $\sigma^2 R_t$, which depends on the current factor norm. Our analysis therefore need control the growth of R_t together with the stochastic terms over the optimization trajectory.

5.2 Stochastic Upper Bound

Eq. (6) shows that $\mathbb{E}[\|\zeta_t\|_F^2 \mid \mathcal{F}_t] \le \sigma^2 R_t$. Thus, as the factors grow, the stepsize must be small enough to control both the stochastic terms and the growth of the factor norm. The following theorem shows that a constant stepsize chosen for a fixed horizon T is sufficient.

Theorem 5.1 (High-probability convergence of LoRA-SGD)**.** *Suppose Assumptions 3.2, 3.4, and 3.5 hold. Fix $q \in (0,1)$ and $T \ge 1$. There exist positive constants c, C, depending only on Δ, R_0, σ, and ℓ, such that, with the constant stepsize*

$$\eta_t \equiv \eta = c\min\left\{\sqrt{\frac{q}{T}},\, q^{1/4}T^{-3/4}\right\},$$

we have, with probability at least $1-q$,

$$\frac{1}{T}\sum_{t=0}^{T-1}\|\nabla J(B_t,A_t)\|_F^2 \le C\left((qT)^{-1/4} + (qT)^{-1/2}\right). \tag{7}$$

Consequently, for any $q \in (0,1)$, $O(q^{-1}\epsilon^{-8})$ iterations suffice to achieve an ϵ-stationary iterate with probability at least $1-q$.

Corollary 5.2 (Expected best-iterate convergence)**.** *Under the assumptions of Theorem 5.1, let $T \ge 2$, choose $q = T^{-1/5}$, and use the corresponding constant stepsize $\eta = cT^{-4/5}$. Then*

$$\mathbb{E}\left[\min_{0\le t<T}\|\nabla J(B_t,A_t)\|_F^2\right] \le C\Big(T^{-1/5} + T^{-2/5}\Big) + T^{-1/5}\|\nabla J(B_0,A_0)\|_F^2,$$

where C is the same constant as in Theorem 5.1. In particular, $\mathbb{E}\big[\min_{0\le t<T}\|\nabla J(B_t,A_t)\|_F^2\big] = O(T^{-1/5})$. Consequently, $O(\epsilon^{-10})$ iterations suffice to ensure $\mathbb{E}\big[\min_{0\le t<T}\|\nabla J(B_t,A_t)\|_F^2\big] \le \epsilon^2$.

Notably, Theorem 5.1 does not assume an a priori bound on the factor norms. Their growth is controlled along the trajectory as part of the analysis. Corollary 5.2 follows directly from this high-probability guarantee by decomposing the expectation over the high-probability event and its complement. The proof is given in Appendix D.

Infinite expected gradient norms. Corollary 5.2 bounds the expected minimum squared gradient norm along the trajectory. This does not ensure that the expected gradient norm is finite at every iterate. Appendix E gives a scalar rank-one LoRA example with a globally smooth objective F and finite-variance stochastic gradients for which $\mathbb{E}\|\nabla J(B_2, A_2)\|_F = \infty$. This does not contradict the corollary: the minimum squared gradient norm is always bounded by $\|\nabla J(B_0, A_0)\|_F^2$.

5.3 Stochastic Lower Bound

We further check whether adapting the stepsize to the observed stochastic trajectory can improve the $O(\epsilon^{-8})$ fixed-confidence rate of Theorem 5.1. Let $g_t := g(X_t; \xi_t)$, and let

$$\mathcal{H}_t^{\mathrm{sgd}} := \big(B_0, A_0, g_0, \eta_0, \ldots, B_{t-1}, A_{t-1}, g_{t-1}, \eta_{t-1}, B_t, A_t\big),$$

with $\mathcal{H}_0^{\mathrm{sgd}} = (B_0, A_0)$. After observing g_t, the algorithm may choose

$$\eta_t = \Phi_t\Big(\epsilon, \mathcal{H}_t^{\mathrm{sgd}}, g_t, U_t\Big), \qquad 0 \le \eta_t \le 1,$$

where U_t is an independent random seed used by the stepsize policy. Thus, the stepsize may depend on the entire observed trajectory and on the current stochastic gradient.

Theorem 5.3 (Stochastic lower bound for adaptive stepsizes)**.** *There exist universal constants $c > 0$ and $\epsilon_0 > 0$ such that, for every $0 < \epsilon \le \epsilon_0$, there exists a rank-one LoRA problem with $r = n = 1$ and $m = O(\epsilon^{-4})$ whose weight-space objective $F \in C^2$ satisfies*

$$\|\nabla^2 F(X)\|_{\mathrm{op}} \le 1, \qquad F(0) - F_\star = 8.$$

The stochastic weight-gradient oracle satisfies

$$\mathbb{E}[g(X_t; \xi_t) \mid \mathcal{F}_t] = \nabla F(X_t), \qquad \mathbb{E}\big[\|g(X_t; \xi_t) - \nabla F(X_t)\|_F^2 \mid \mathcal{F}_t\big] \le 1.$$

Starting from $B_0 = 0 \in \mathbb{R}^{m\times 1}$ and $A_0 = 1$, for every stepsize policy defined above,

$$\mathbb{P}(\|\nabla J(B_t, A_t)\|_F > \epsilon \text{ for all } 0 \le t \le T) \ge \frac{3}{4} \qquad \text{for every } T \le c\epsilon^{-8}. \tag{8}$$

Consequently, any such policy requires $\Omega(\epsilon^{-8})$ stochastic gradient evaluations to produce an ϵ-stationary iterate with constant probability.

Importantly, the adaptive class in Theorem 5.3 allows the adaptive stepsize to depend on the current stochastic gradient. It therefore includes bounded global gradient clipping and other bounded normalization methods. Hence, such a technique alone cannot improve the $\Omega(\epsilon^{-8})$ worst-case rate.

6 Experiments

We empirically examine three phenomena that arise in our analysis: the relation between the Hessian and gradient norms motivating generalized smoothness, the growth of the LoRA factors during optimization, and the amplification of stochastic factor-gradient noise as the factors grow. Additional experimental details are provided in Appendix G.

6.1 Empirical motivation for generalized smoothness

Our generalized-smoothness condition allows the Hessian norm to grow with the weight-gradient norm. We examine this relation during LoRA fine-tuning of DistilBERT on IMDb and ViT-B/16 on CIFAR-100 by measuring $\|\nabla_X F\|_F$ and $\|\nabla_X^2 F\|_{\mathrm{op}}$ along the training trajectories. We also compare a constant fit L with the gradient-dependent fit $L_0 + L_1\|\nabla_X F\|_F$. For each trajectory shown in Fig. 1, both fits are estimated using the other two seeds. Fitting details are given in Appendix G.1.

As shown in Fig. 1, the Hessian norm varies substantially during training and tends to be larger when the gradient norm is larger in both settings. Across all three choices of evaluation seed, the gradient-dependent fit has a smaller normalized mean fitted value than the constant fit. Averaged across these evaluations, it is 34% smaller for DistilBERT/IMDb and 21% smaller for ViT-B/16/CIFAR-100. These results support using a gradient-dependent curvature model to describe the observed LoRA training trajectories. The complete results and analysis are reported in Appendix G.1.

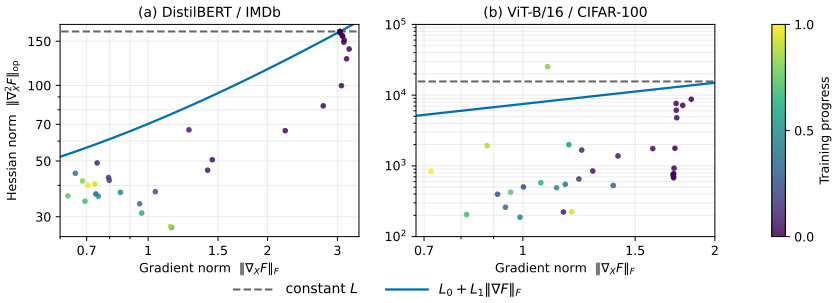

Figure 1: **Hessian norm versus gradient norm during LoRA fine-tuning.** We plot the Hessian norm $\|\nabla_X^2 F\|_{\mathrm{op}}$ against the weight-gradient norm $\|\nabla_X F\|_F$ along optimization trajectories of (a) DistilBERT on IMDb and (b) ViT-B/16 on CIFAR-100. The dashed horizontal line shows the constant fit L, and the solid line shows the gradient-dependent fit $L_0 + L_1\|\nabla_X F\|_F$; both are estimated using the other two seeds. Color indicates training progress, normalized from 0 (start) to 1 (end).

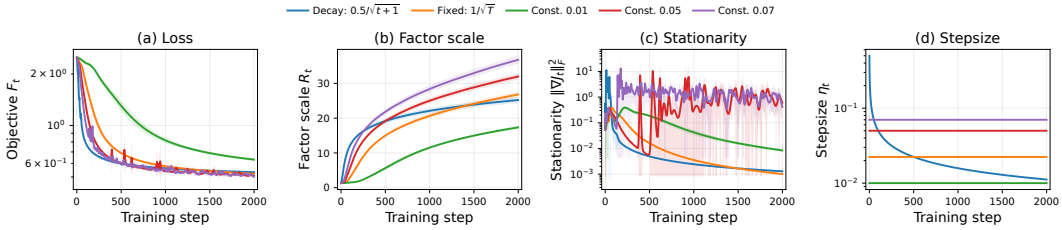

Figure 2: **LoRA-GD dynamics under different stepsizes.** Rank-4 LoRA-GD for CIFAR-10 logistic regression using frozen ImageNet-pretrained ResNet-18 features. We compare the decaying stepsize $\eta_t = 0.5/\sqrt{t+1}$, the fixed-horizon stepsize $\eta_t = 1/\sqrt{T}$ with $T = 2000$, and constant stepsizes $\eta \in \{0.01, 0.05, 0.07\}$. The panels show (a) the objective $F_t := F(X_t)$, (b) the factor scale R_t, (c) the squared factor-gradient norm $\|\nabla J(B_t, A_t)\|_F^2$, and (d) the stepsize η_t. Curves show the mean over five paired runs; shaded regions indicate $\pm$ one standard deviation.

6.2 Factor dynamics of LoRA-GD

We next study the factor growth that appears in our convergence analysis. We use rank-4 LoRA-GD for CIFAR-10 logistic regression on frozen ImageNet-pretrained ResNet-18 features, which allows us to compute the full gradient exactly. Figure 2 compares the decaying and fixed-horizon stepsizes from our analysis with several constant stepsizes. While all choices decrease the loss, their factor dynamics differ substantially. In particular, the factor scale $R_t = \|A_t\|_F^2 + \|B_t\|_F^2$ grows during training, with larger stepsizes leading to substantially larger factor norms. The stationarity measure $\|\nabla J(B_t, A_t)\|_F^2$ also differs sharply: the decaying and fixed-horizon stepsizes continue to decrease it, whereas the larger constant stepsizes remain much higher and oscillate throughout training. Thus, similar loss values can correspond to very different values of R_t and $\|\nabla J(B_t, A_t)\|_F^2$, motivating the explicit control of factor growth in our analysis.

6.3 Factor growth amplifies stochastic gradient noise

We next study how growing LoRA factors amplify stochastic-gradient noise. On CIFAR-10 logistic regression with frozen ImageNet-pretrained ResNet-18 features, we freeze checkpoints along a LoRA-SGD trajectory. Let $\mathbf{G}_t := \nabla F(X_t)$ denote the full weight gradient, $\widehat{\mathbf{G}}_t^{(k)}$ the gradient from the k-th sampled minibatch, and $\mathbf{E}_t^{(k)} := \widehat{\mathbf{G}}_t^{(k)} - \mathbf{G}_t$ the corresponding empirical weight-gradient error. We estimate the empirical weight-gradient and factor-gradient noise as

$$\widehat{\sigma}_{W,t}^2 = \frac{1}{M}\sum_{k=1}^{M} \|\mathbf{E}_t^{(k)}\|_F^2, \qquad \widehat{\sigma}_{F,t}^2 = \frac{1}{M}\sum_{k=1}^{M} \left(\|\mathbf{E}_t^{(k)} A_t^\top\|_F^2 + \|B_t^\top \mathbf{E}_t^{(k)}\|_F^2\right).$$

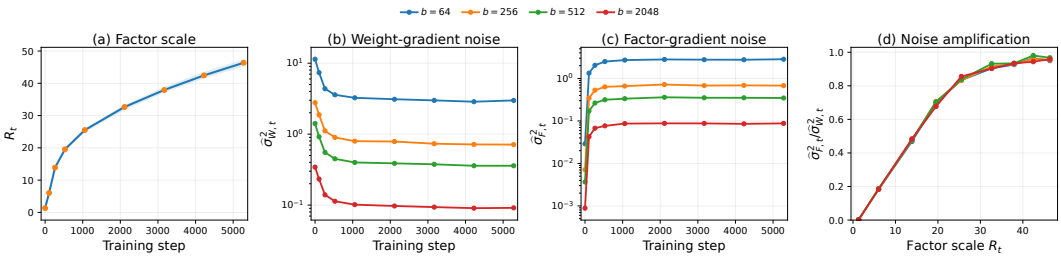

Figure 3: **Gradient-noise amplification with growing LoRA factors.** On CIFAR-10 logistic regression with frozen ImageNet-pretrained ResNet-18 features with a fixed training minibatch size of 512, we estimate gradient noise at checkpoints along LoRA-SGD using diagnostic minibatch sizes $b \in \{64, 256, 512, 2048\}$. We report (a) factor scale R_t, (b) empirical weight-gradient noise $\widehat{\sigma}^2_{W,t}$, (c) empirical factor-gradient noise $\widehat{\sigma}^2_{F,t}$, and (d) their ratio $\widehat{\sigma}^2_{F,t}/\widehat{\sigma}^2_{W,t}$ versus R_t.

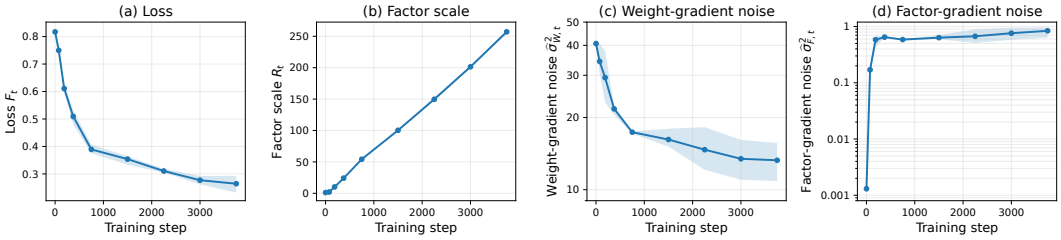

Figure 4: **Factor growth and gradient-noise dynamics during ViT fine-tuning.** We fine-tune ViT-B/16 on CIFAR-100 with LoRA-SGD and report (a) loss F_t, (b) factor scale R_t, (c) empirical weight-gradient noise $\widehat{\sigma}^2_{W,t}$, and (d) empirical factor-gradient noise $\widehat{\sigma}^2_{F,t}$. Shaded regions indicate $\pm$ one standard deviation.

We use $M = 200$ sampled minibatches for $b \in \{64, 256, 512, 2048\}$. As shown in Fig. 3, the weight-gradient noise decreases during training and is smaller for larger batches, whereas the factor-gradient noise increases as R_t grows. Panel (d) shows this relation more directly: $\widehat{\sigma}^2_{F,t}/\widehat{\sigma}^2_{W,t}$ increases with R_t and follows a similar trend across all four batch sizes. This behavior is consistent with Eq. (6), where the factor-gradient noise is controlled by the weight-gradient noise multiplied by the factor scale R_t.

6.4 Validation on ViT fine-tuning

We finally test whether factor growth and stochastic-noise amplification also appear when LoRA is used to fine-tune a pretrained vision transformer. We fine-tune ViT-B/16 on CIFAR-100 with LoRA-SGD and track the loss, factor scale, weight-gradient noise, and factor-gradient noise. As shown in Fig. 4, the factor scale R_t grows substantially during training, consistent with the factor growth observed in Fig. 2. At the same time, the weight-gradient noise decreases, whereas the factor-gradient noise increases by several orders of magnitude. This shows the same noise behavior as Fig. 3: as the LoRA factors grow, smaller weight-gradient noise can still correspond to larger factor-gradient noise. Thus, both factor growth and the increase in factor-gradient noise also appear during ViT fine-tuning.

7 Conclusion

We studied the convergence of LoRA under sub-quadratic generalized smoothness without assuming bounded factor norms. For LoRA-GD, we established a tight $\Theta(\epsilon^{-4})$ complexity, even when the stepsize is chosen adaptively from the optimization history. For LoRA-SGD with finite-variance stochastic gradients, we derived a high-probability $O(\epsilon^{-8})$ upper bound and a fixed-confidence $\Omega(\epsilon^{-8})$ lower bound. Experiments support our theoretical analysis: the Hessian norm grows with the gradient norm, the factor norms grow during optimization, and this growth amplifies stochastic factor-gradient noise.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, identifying potentially relevant literature, and code generation. All citations included in the paper were independently located and verified by the authors using DBLP rather than taken directly from AI-generated outputs. AI-generated code was reviewed, tested, and validated by the authors before being used in the experiments. The authors take full responsibility for the manuscript, citations, code, and reported results.

REPRODUCIBILITY STATEMENT

Complete proofs of all theoretical results are provided in the appendix. Experimental settings, data splits, model and optimization configurations, hyperparameter selection procedures, and additional results are documented in Appendix G.

REFERENCES

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake E. Woodworth. Lower bounds for non-convex stochastic optimization. *Math. Program.*, 199(1): 165–214, 2023. doi: 10.1007/S10107-022-01822-7. URL https://doi.org/10.1007/s10107-022-01822-7.

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Math. Program.*, 184(1):71–120, 2020. doi: 10.1007/S10107-019-01406-Y. URL https://doi.org/10.1007/s10107-019-01406-y.

Michael Crawshaw, Mingrui Liu, Francesco Orabona, Wei Zhang, and Zhenxun Zhuang. Robustness to unbounded smoothness of generalized signsgd. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/40924475a9bf768bdac3725e67745283-Abstract-Conference.html.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

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 *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Zhanhong Jiang, Nastaran Saadati, Aditya Balu, Minh Pham, Joshua Russell Waite, Nasla Saleem, Chinmay Hegde, and Soumik Sarkar. A unified convergence theory for large language model efficient fine-tuning. In *OPT 2024: Optimization for Machine Learning*, 2024. URL https://openreview.net/forum?id=f0lq26eITJ.

Junsu Kim, Jaeyeon Kim, and Ernest K. Ryu. LoRA training provably converges to a low-rank global minimum or it fails loudly (But it probably won't fail). In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30224–30247. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/kim25n.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and non-convex optimization under generalized smoothness. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/7e8bb8d17bb1cb24dfe972a2f8ff2500-Abstract-Conference.html.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1015/.

Grigory Malinovsky, Umberto Michieli, Hasan Abed Al Kader Hammoud, Taha Ceritli, Hayder Elesedy, Mete Ozay, and Peter Richtárik. Randomized asymmetric chain of lora: The first meaningful theoretical framework for low-rank adaptation. *CoRR*, abs/2410.08305, 2024. doi: 10.48550/ARXIV.2410.08305. URL https://doi.org/10.48550/arXiv.2410.08305.

Siqiao Mu and Diego Klabjan. On the convergence rate of lora gradient descent. *CoRR*, abs/2512.18248, 2025. doi: 10.48550/ARXIV.2512.18248. URL https://doi.org/10.48550/arXiv.2512.18248.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL http://arxiv.org/abs/1910.01108.

Igor Sokolov, Abdurakhmon Sadiev, Yury Demidovich, Fawaz S. Al-Qahtani, and Peter Richtárik. Bernoulli-lora: A theoretical framework for randomized low-rank adaptation. *CoRR*, abs/2508.03820, 2025. doi: 10.48550/ARXIV.2508.03820. URL https://doi.org/10.48550/arXiv.2508.03820.

Ru Wang, Chengchang Liu, and John C. S. Lui. On the convergence of stochastic low-rank adaptation. *CoRR*, abs/2607.21975, 2026. doi: 10.48550/ARXIV.2607.21975. URL https://doi.org/10.48550/arXiv.2607.21975.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=BJgnXpVYwS.

A TECHNICAL TOOLS UNDER GENERALIZED SMOOTHNESS

For the factorized objective, we write $\nabla_B J_t := \nabla_B J(B_t, A_t)$ and $\nabla_A J_t := \nabla_A J(B_t, A_t)$, with $\|\nabla J_t\|_F^2 := \|\nabla_B J_t\|_F^2 + \|\nabla_A J_t\|_F^2$ as in Definition 3.1.

A.1 LOCAL DESCENT UNDER GENERALIZED SMOOTHNESS

Lemma A.1 (Local descent under generalized smoothness)**.** *Suppose Assumption 3.4 holds. Fix X and let $G \geq \|\nabla F(X)\|_F$. If S satisfies*

$$\|S\|_F \leq \frac{G}{\ell(2G)}, \tag{9}$$

then, for every $\tau \in [0,1]$,

$$\|\nabla F(X + \tau S)\|_F \leq 2G. \tag{10}$$

Moreover,

$$F(X+S) \leq F(X) + \langle \nabla F(X), S \rangle + \frac{\ell(2G)}{2}\|S\|_F^2. \tag{11}$$

Proof. If $G = 0$, then (9) forces $S = 0$, and both claims follow. Assume $G > 0$ and define $g(\tau) := \|\nabla F(X + \tau S)\|_F$. Suppose, for contradiction, that $g(\widehat{\tau}) > 2G$ for some $\widehat{\tau} \in (0,1]$. Since $g(0) \leq G < 2G$, continuity yields a first $\tau_\star \in (0, \widehat{\tau})$ such that $g(\tau_\star) = 2G$. For every $s \in [0, \tau_\star]$, we have $g(s) \leq 2G$. The fundamental theorem of calculus, applied to the matrix-valued gradient along the segment, gives

$$\nabla F(X + \tau_\star S) = \nabla F(X) + \int_0^{\tau_\star} \nabla^2 F(X + sS)[S]\, ds.$$

Taking Frobenius norms and using $\|\nabla^2 F(X+sS)[S]\|_F \leq \|\nabla^2 F(X+sS)\|_{\mathrm{op}}\|S\|_F$, followed by generalized smoothness and monotonicity of ℓ, yields

$$\begin{aligned} g(\tau_\star) &\leq G + \int_0^{\tau_\star} \|\nabla^2 F(X+sS)\|_{\mathrm{op}}\, \|S\|_F\, ds \\ &\leq G + \tau_\star \ell(2G)\|S\|_F \leq G + \tau_\star G < 2G, \end{aligned}$$

where the last inequality uses $\tau_\star < \widehat{\tau} \leq 1$. This contradicts $g(\tau_\star) = 2G$, proving (10).

Combining (10) with Assumption 3.4 gives

$$\|\nabla^2 F(X + \tau S)\|_{\mathrm{op}} \leq \ell(2G), \qquad \tau \in [0,1].$$

For each $\tau \in [0,1]$, Cauchy–Schwarz and the operator-norm bound imply

$$\big\langle S, \nabla^2 F(X+\tau S)[S] \big\rangle \leq \|S\|_F \|\nabla^2 F(X+\tau S)[S]\|_F \leq \ell(2G)\|S\|_F^2.$$

Taylor's theorem with integral remainder and $\int_0^1 (1-\tau)\, d\tau = 1/2$ therefore give

$$\begin{aligned} F(X+S) &= F(X) + \langle \nabla F(X), S \rangle \\ &\quad + \int_0^1 (1-\tau)\, \big\langle S, \nabla^2 F(X+\tau S)[S] \big\rangle\, d\tau \\ &\leq F(X) + \langle \nabla F(X), S \rangle + \frac{\ell(2G)}{2}\|S\|_F^2, \end{aligned}$$

which is (11). □

A.2 GRADIENT AND HESSIAN BOUNDS FROM THE OBJECTIVE GAP

The next lemma bounds $\|\nabla F(X)\|_F$ in terms of the objective gap $F(X) - F_\star$.

Lemma A.2 (Gradient bound from the objective gap)**.** *Under Assumptions 3.2 and 3.4, every X satisfies*

$$\|\nabla F(X)\|_F^2 \leq 2\ell(2\|\nabla F(X)\|_F)\, \big(F(X) - F_\star\big). \tag{12}$$

Proof. Set $G := \|\nabla F(X)\|_F$. The case $G = 0$ is immediate. For $G > 0$, let $L := \ell(2G) > 0$ and $S := -L^{-1}\nabla F(X)$. Then $\|S\|_F = G/L$, so (9) holds with equality. For this choice of S,

$$\langle \nabla F(X), S\rangle = -\frac{G^2}{L}, \qquad \|S\|_F^2 = \frac{G^2}{L^2}.$$

Substituting both identities into Lemma A.1 gives

$$F(X+S) \le F(X) - \frac{G^2}{L} + \frac{L}{2}\frac{G^2}{L^2} = F(X) - \frac{G^2}{2L}.$$

Since $F_\star \le F(X+S)$, we obtain

$$F(X) - F_\star \ge F(X) - F(X+S) \ge \frac{G^2}{2L}.$$

Therefore,

$$G^2 \le 2L\big(F(X) - F_\star\big) = 2\ell(2G)\big(F(X) - F_\star\big),$$

which is (12). □

For $U \ge 0$, define the U-sublevel set

$$\mathcal{S}_U := \{X : F(X) - F_\star \le U\}.$$

We first bound the gradient uniformly over $\mathcal{S}_U$. Define

$$\overline{G}(U) := \sup\left\{u \ge 0 : u^2 \le 2\ell(2u)U\right\}. \tag{13}$$

Proposition A.3 (Gradient bound on a sublevel set)**.** *Under Assumptions 3.2 and 3.4, $\overline{G}(U)$ is finite for every $U \ge 0$:*

$$\overline{G}(U) < \infty. \tag{14}$$

Moreover, for every X,

$$F(X) - F_\star \le U \quad \Longrightarrow \quad \|\nabla F(X)\|_F \le \overline{G}(U). \tag{15}$$

Proof. The set in (13) contains 0. If $U = 0$, the defining inequality reduces to $u^2 \le 0$, so $\overline{G}(U) = 0$. Now suppose $U > 0$. Since $\ell(u) = o(u^2)$ by Assumption 3.4,

$$\frac{2U\ell(2u)}{u^2} = 8U\frac{\ell(2u)}{(2u)^2} \longrightarrow 0 \qquad (u \to \infty).$$

Thus there exists $u_0 < \infty$ such that $2U\ell(2u) < u^2$ whenever $u \ge u_0$. Therefore the set in (13) is bounded, which proves (14).

Now let X satisfy $F(X) - F_\star \le U$ and set $G := \|\nabla F(X)\|_F$. Lemma A.2 gives

$$G^2 \le 2\ell(2G)\big(F(X) - F_\star\big) \le 2\ell(2G)U.$$

Thus G belongs to the set in (13), so $G \le \overline{G}(U)$. □

Define

$$\overline{L}(U) := \ell(2\overline{G}(U)). \tag{16}$$

For every $X \in \mathcal{S}_U$, Proposition A.3 gives $\|\nabla F(X)\|_F \le \overline{G}(U)$. Assumption 3.4 and monotonicity of ℓ therefore imply

$$\|\nabla^2 F(X)\|_{\mathrm{op}} \le \ell(\|\nabla F(X)\|_F) \le \ell(2\|\nabla F(X)\|_F) \le \overline{L}(U). \tag{17}$$

Thus both the gradient norm and the Hessian operator norm are uniformly bounded on $\mathcal{S}_U$.

Specialization to (L_0, L_1)-smoothness. Under the condition

$$\|\nabla^2 F(X)\|_{\mathrm{op}} \le L_0 + L_1\|\nabla F(X)\|_F \qquad \text{for all } X, \tag{18}$$

where $L_0 > 0$ and $L_1 \ge 0$, Assumption 3.4 holds with $\ell(u) = L_0 + L_1 u$. In this case, (13) becomes

$$u^2 - 4L_1 U u - 2L_0 U \le 0.$$

Completing the square gives

$$(u - 2L_1 U)^2 \le 4L_1^2 U^2 + 2L_0 U.$$

Thus every admissible u is at most $2L_1U + \sqrt{4L_1^2U^2 + 2L_0U}$, and this value itself satisfies the inequality. Hence

$$\overline{G}(U) = 2L_1 U + \sqrt{4L_1^2 U^2 + 2L_0 U}, \tag{19}$$

so $\overline{L}(U) = L_0 + 2L_1\overline{G}(U)$. If $L_1 = 0$, (19) reduces to $\overline{G}(U) = \sqrt{2L_0U}$, the standard gradient bound for an L_0-smooth lower-bounded objective.

Remark A.4 (Role of sub-quadratic growth). The condition $\ell(u) = o(u^2)$ is used only to guarantee $\overline{G}(U) < \infty$. Once this finiteness is established, (15) follows from Lemma A.2, while (17) follows from monotonicity of ℓ.

For the initial gap $\Delta = J_0 - F_\star$, we write $\overline{G} := \overline{G}(\Delta)$ and $\overline{L} := \overline{L}(\Delta)$. Then Proposition A.3 and (17) show that every iterate with $J_t \le J_0$ satisfies

$$G_t \le \overline{G}, \qquad \ell(2G_t) \le \overline{L}. \tag{20}$$

B Proof of the LoRA-GD Upper Bound

Suppose Assumptions 3.2 and 3.4 hold. Consider the iterates in (2), with the same nonnegative stepsize for both factors. Define

$$c := \frac{1}{16}, \qquad L_t := \ell(2G_t). \tag{21}$$

We establish descent under the stepsize conditions

$$\eta_t G_t \le c, \qquad \eta_t L_t R_t \le c. \tag{22}$$

B.1 LoRA identities and one-step descent

Factor gradients and the weight update. The chain rule gives

$$\nabla_B J_t = \nabla F(X_t) A_t^\top, \qquad \nabla_A J_t = B_t^\top \nabla F(X_t). \tag{23}$$

In particular, the product inequality $\|UV\|_F \le \|U\|_F\|V\|_F$ gives

$$\|\nabla_B J_t\|_F \le G_t\|A_t\|_F, \qquad \|\nabla_A J_t\|_F \le G_t\|B_t\|_F.$$

Squaring and adding these inequalities yields

$$\begin{aligned} \|\nabla J_t\|_F^2 &= \|\nabla_B J_t\|_F^2 + \|\nabla_A J_t\|_F^2 \\ &\le G_t^2\big(\|A_t\|_F^2 + \|B_t\|_F^2\big) = G_t^2 R_t, \\ \|\nabla J_t\|_F &\le G_t\sqrt{R_t}. \end{aligned} \tag{24}$$

Write the factor increments as

$$\Delta B_t := -\eta_t \nabla_B J_t, \qquad \Delta A_t := -\eta_t \nabla_A J_t.$$

Expanding $X_{t+1} = (B_t + \Delta B_t)(A_t + \Delta A_t)$ gives

$$X_{t+1} - X_t = B_t \Delta A_t + \Delta B_t A_t + \Delta B_t \Delta A_t. \tag{25}$$

Using (25), the bound $\|UV\|_F \le \|U\|_F\|V\|_F$, and the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}\|X_{t+1} - X_t\|_F &\le \|B_t\|_F\|\Delta A_t\|_F + \|A_t\|_F\|\Delta B_t\|_F + \|\Delta B_t\|_F\|\Delta A_t\|_F \\ &\le \sqrt{R_t}\sqrt{\|\Delta A_t\|_F^2 + \|\Delta B_t\|_F^2} + \frac{1}{2}\left(\|\Delta A_t\|_F^2 + \|\Delta B_t\|_F^2\right) \\ &= \eta_t\sqrt{R_t}\|\nabla J_t\|_F + \frac{\eta_t^2}{2}\|\nabla J_t\|_F^2.\end{aligned}$$

Substituting $\|\nabla J_t\|_F \le G_t\sqrt{R_t}$ from (24) yields

$$\begin{aligned}\|X_{t+1} - X_t\|_F &\le \eta_t\sqrt{R_t}\|\nabla J_t\|_F\left(1 + \frac{\eta_t G_t}{2}\right) \\ &\le \eta_t G_t R_t\left(1 + \frac{\eta_t G_t}{2}\right).\end{aligned} \tag{26}$$

The following lemma gives the descent estimate used in Section 4.1.

Lemma B.1 (One-step descent)**.** *If* (22) *holds at iteration* t*, then*

$$J_{t+1} \le J_t - \frac{\eta_t}{2}\|\nabla J_t\|_F^2. \tag{27}$$

Proof. If $G_t = 0$, then (24) gives $\nabla J_t = 0$. Both factors remain unchanged, so (27) holds with equality. Suppose $G_t > 0$. By (26) and (22),

$$L_t\|X_{t+1} - X_t\|_F \le \eta_t L_t R_t G_t\left(1 + \frac{\eta_t G_t}{2}\right) \le cG_t\left(1 + \frac{c}{2}\right) < G_t.$$

Consequently,

$$\|X_{t+1} - X_t\|_F \le \frac{G_t}{\ell(2G_t)}, \tag{28}$$

so Lemma A.1 applies to the step $X_{t+1} - X_t$ and gives

$$J_{t+1} \le J_t + \langle \nabla F(X_t), X_{t+1} - X_t\rangle + \frac{L_t}{2}\|X_{t+1} - X_t\|_F^2. \tag{29}$$

For the first two terms in (25), we use the Frobenius inner-product identities

$$\langle H, BU\rangle = \langle B^\top H, U\rangle, \qquad \langle H, VA\rangle = \langle HA^\top, V\rangle.$$

Applying these identities with $H = \nabla F(X_t)$, and then substituting the factor increments, gives

$$\begin{aligned}\langle \nabla F(X_t), B_t\Delta A_t + \Delta B_t A_t\rangle &= \langle \nabla_A J_t, \Delta A_t\rangle + \langle \nabla_B J_t, \Delta B_t\rangle \\ &= -\eta_t\|\nabla J_t\|_F^2.\end{aligned}$$

For the remaining term, $\Delta B_t\Delta A_t$, we have

$$\begin{aligned}|\langle \nabla F(X_t), \Delta B_t\Delta A_t\rangle| &\le G_t\|\Delta B_t\|_F\|\Delta A_t\|_F \\ &\le \frac{\eta_t^2 G_t}{2}\|\nabla J_t\|_F^2.\end{aligned}$$

Therefore

$$\langle \nabla F(X_t), X_{t+1} - X_t\rangle \le -\eta_t\left(1 - \frac{\eta_t G_t}{2}\right)\|\nabla J_t\|_F^2.$$

The quadratic remainder is bounded by the first inequality in (26):

$$\frac{L_t}{2}\|X_{t+1} - X_t\|_F^2 \le \frac{\eta_t^2 L_t R_t}{2}\left(1 + \frac{\eta_t G_t}{2}\right)^2\|\nabla J_t\|_F^2.$$

Substituting these bounds into (29) yields

$$J_{t+1} \le J_t - \eta_t\left[1 - \frac{\eta_t G_t}{2} - \frac{\eta_t L_t R_t}{2}\left(1 + \frac{\eta_t G_t}{2}\right)^2\right]\|\nabla J_t\|_F^2. \tag{30}$$

Under (22), the coefficient in brackets is at least

$$1 - \frac{c}{2} - \frac{c}{2}\left(1 + \frac{c}{2}\right)^2 \ge 1 - \frac{3c}{2} = \frac{29}{32} > \frac{1}{2},$$

where we used $c = 1/16$ and $(1 + c/2)^2 \le 2$. Multiplying by the nonnegative quantity $\eta_t\|\nabla J_t\|_F^2$ proves (27). □

B.2 CUMULATIVE DESCENT AND FACTOR-GROWTH CONTROL

Let $\Delta := J_0 - F_\star$, and let $\overline{G}$ and $\overline{L}$ be the corresponding bounds in (13) and (16). For $t \geq 0$, define

$$S_t := \sum_{s=0}^{t-1} \eta_s, \qquad S_0 := 0. \tag{31}$$

Lemma B.2 (Cumulative descent and factor bounds)**.** *Fix an integer $t \geq 0$ and suppose* (22) *holds for all $0 \leq s < t$. Then $J_t \leq J_0$ and*

$$\sum_{s=0}^{t-1} \eta_s \|\nabla J_s\|_F^2 \leq 2(J_0 - J_t) \leq 2\Delta. \tag{32}$$

In particular,

$$G_t \leq \overline{G}, \qquad L_t \leq \overline{L}. \tag{33}$$

Moreover,

$$\sqrt{R_t} \leq \sqrt{R_0} + \sqrt{2\Delta S_t}, \tag{34}$$

and therefore

$$\begin{aligned} R_t &\leq 2R_0 + 4\Delta S_t \\ &\leq 2(R_0 + 2\Delta)(1 + S_t). \end{aligned} \tag{35}$$

If $S_t > 0$, then

$$\min_{0 \leq s < t} \|\nabla J_s\|_F^2 \leq \frac{2\Delta}{S_t}. \tag{36}$$

Proof. Summing the descent inequality in Lemma B.1 from $s = 0$ to $t - 1$ gives

$$J_t \leq J_0 - \frac{1}{2} \sum_{s=0}^{t-1} \eta_s \|\nabla J_s\|_F^2.$$

Since every summand is nonnegative, this first gives $J_t \leq J_0$. Rearranging, and then using $J_t \geq F_\star$, gives

$$\sum_{s<t} \eta_s \|\nabla J_s\|_F^2 \leq 2(J_0 - J_t) \leq 2(J_0 - F_\star) = 2\Delta,$$

which is (32). Since $J_t \leq J_0$, (20) gives (33).

The triangle inequality for the simultaneous factor updates gives

$$\begin{aligned} \sqrt{R_{s+1}} &= \sqrt{\|B_s - \eta_s \nabla_B J_s\|_F^2 + \|A_s - \eta_s \nabla_A J_s\|_F^2} \\ &\leq \sqrt{R_s} + \eta_s \|\nabla J_s\|_F. \end{aligned}$$

Summing over $s < t$ telescopes the factor-norm increments. For the remaining sum, apply Cauchy–Schwarz to $\sqrt{\eta_s}$ and $\sqrt{\eta_s}\|\nabla J_s\|_F$; this remains valid when some stepsizes are zero. Together with (32), this gives

$$\begin{aligned} \sqrt{R_t} &\leq \sqrt{R_0} + \sum_{s=0}^{t-1} \eta_s \|\nabla J_s\|_F \\ &\leq \sqrt{R_0} + \sqrt{S_t} \sqrt{\sum_{s=0}^{t-1} \eta_s \|\nabla J_s\|_F^2} \\ &\leq \sqrt{R_0} + \sqrt{2\Delta S_t}, \end{aligned}$$

This proves (34). Squaring and using $(a + b)^2 \leq 2a^2 + 2b^2$ gives

$$R_t \leq 2R_0 + 4\Delta S_t \leq (2R_0 + 4\Delta)(1 + S_t) = 2(R_0 + 2\Delta)(1 + S_t).$$

The second inequality adds only the nonnegative terms $4\Delta + 2R_0 S_t$, proving (35).

If $S_t > 0$, nonnegativity of the stepsizes gives

$$S_t \min_{0 \leq s < t} \|\nabla J_s\|_F^2 \leq \sum_{s=0}^{t-1} \eta_s \|\nabla J_s\|_F^2 \leq 2\Delta.$$

Dividing by S_t proves (36). □

A sufficient stability condition. For a predetermined sequence, each η_t is fixed before optimization. We verify (22) at iteration t by induction: descent at iterations $s < t$ bounds G_t and R_t at the current iterate.

With $c = 1/16$ as in (21), define

$$K := \max\left\{1/32, \overline{G}, 2\overline{L}(R_0 + 2\Delta)\right\}. \tag{37}$$

This agrees with the definition in Theorem 4.1 and satisfies $K \geq 1/32 > 0$.

Lemma B.3 (Sufficient condition for predetermined stepsizes)**.** *Let $\{\eta_t\}_{t\geq 0}$ be a predetermined nonnegative sequence. If*

$$\eta_t(1 + S_t) \leq \frac{c}{K} \qquad \textit{for every } t \geq 0, \tag{38}$$

then (22) *holds at every iteration. Consequently, for every $t \geq 0$,*

$$J_{t+1} \leq J_t - \frac{\eta_t}{2}\|\nabla J_t\|_F^2, \tag{39}$$

and

$$G_t \leq \overline{G}. \tag{40}$$

The factor bound (34) *holds for every $t \geq 0$. For every integer $T \geq 1$ with $S_T > 0$, the stationarity bound* (36) *also holds.*

Proof. Fix $t \geq 0$ and suppose (22) holds for all $s < t$. For $t = 0$, the induction hypothesis is vacuous. Applying Lemma B.2 to the updates with $s < t$ gives

$$\sum_{s=0}^{t-1} \eta_s \|\nabla J_s\|_F^2 \leq 2\Delta, \tag{41}$$

as well as

$$G_t \leq \overline{G}, \qquad L_t \leq \overline{L},$$

and

$$R_t \leq 2(R_0 + 2\Delta)(1 + S_t).$$

Since $1 + S_t \geq 1$, condition (38) and the definition of K yield

$$\eta_t G_t \leq \overline{G}\,\eta_t(1 + S_t) \leq \frac{c\overline{G}}{K} \leq c,$$

$$\eta_t L_t R_t \leq 2\overline{L}(R_0 + 2\Delta)\eta_t(1 + S_t) \leq \frac{2c\overline{L}(R_0 + 2\Delta)}{K} \leq c.$$

Thus (22) holds at iteration t, completing the induction for all $t \geq 0$.

Lemma B.1 gives (39). Applying Lemma B.2 for each t yields (40) and (34), as well as (36) when $S_T > 0$. □

B.3 Proof of the inverse-square-root stepsize theorem

Proof of Theorem 4.1. For $\eta_t = \gamma/\sqrt{t+1}$ and $t \geq 1$, the function $x \mapsto x^{-1/2}$ is decreasing. Thus $k^{-1/2} \leq \int_{k-1}^{k} x^{-1/2}\,dx$ for each $k \geq 1$, where the integral at zero is finite. Summing gives

$$S_t = \gamma \sum_{k=1}^{t} k^{-1/2} \leq \gamma \int_0^t x^{-1/2}\,dx = 2\gamma\sqrt{t},$$

while $S_0 = 0$. Therefore, for every $t \geq 0$,

$$\eta_t(1 + S_t) \leq \frac{\gamma}{\sqrt{t+1}}(1 + 2\gamma\sqrt{t}) \leq \gamma + 2\gamma^2 \leq \frac{c}{K}.$$

Lemma B.3 applies. In particular, $G_t \leq \overline{G}$ and

$$R_t \leq 2R_0 + 4\Delta S_t \leq 2R_0 + 8\Delta\gamma\sqrt{t}.$$

For the lower bound, monotonicity gives $k^{-1/2} \geq \int_k^{k+1} x^{-1/2}\,dx$. Hence

$$S_T \geq \gamma \int_1^{T+1} x^{-1/2}\,dx = 2\gamma(\sqrt{T+1} - 1).$$

Combining this inequality with (36) yields

$$\min_{0\leq t<T} \|\nabla J_t\|_F^2 \leq \frac{\Delta}{\gamma(\sqrt{T+1} - 1)},$$

This is (3). For $T \geq 1$, rationalizing the denominator gives

$$\sqrt{T+1} - 1 = \frac{T}{\sqrt{T+1}+1} \geq \frac{\sqrt{T}}{\sqrt{2}+1}.$$

Taking square roots therefore yields

$$\min_{0\leq t<T} \|\nabla J_t\|_F \leq \sqrt{\frac{(\sqrt{2}+1)\Delta}{\gamma}}\, T^{-1/4} = O(T^{-1/4}).$$

More directly, the right-hand side of (3) is at most ϵ^2 whenever $\sqrt{T+1} - 1 \geq \Delta/(\gamma\epsilon^2)$. Solving for T, it suffices to choose an integer

$$T \geq \max\left\{1, \left(1 + \frac{\Delta}{\gamma\epsilon^2}\right)^2 - 1\right\}$$

to obtain an ϵ-stationary iterate. With the problem parameters and γ fixed, this requires $O(\epsilon^{-4})$ iterations. □

B.4 Fixed-horizon constant stepsize

Proof of Corollary 4.3. Fix an integer $T \geq 1$ and take $\eta_t = \gamma/\sqrt{T}$ for $0 \leq t < T$. Then $S_t = t\gamma/\sqrt{T}$ for $0 \leq t \leq T$, and

$$\eta_t(1 + S_t) = \frac{\gamma}{\sqrt{T}}\left(1 + \frac{t\gamma}{\sqrt{T}}\right) \leq \gamma + \gamma^2 \leq \frac{c}{K}, \qquad 0 \leq t < T.$$

Extend the sequence by setting $\eta_t = 0$ for $t \geq T$. Lemma B.3 then applies, and $S_T = \gamma\sqrt{T}$ gives

$$\min_{0\leq t<T} \|\nabla J_t\|_F^2 \leq \frac{2\Delta}{\gamma\sqrt{T}}. \tag{42}$$

Taking square roots gives $\min_{t<T} \|\nabla J_t\|_F \leq \sqrt{2\Delta/\gamma}\, T^{-1/4}$. The squared-gradient bound is at most ϵ^2 whenever $\sqrt{T} \geq 2\Delta/(\gamma\epsilon^2)$. Squaring this nonnegative inequality shows that an integer

$$T \geq \max\left\{1, \frac{4\Delta^2}{\gamma^2\epsilon^4}\right\}$$

suffices to obtain an ϵ-stationary iterate. □

C Proof of the LoRA-GD Lower Bound

We prove Theorem 4.4 by constructing a C^2 objective for any fixed initialization scale $\rho > 0$. The theorem follows by setting $\rho = 1$.

C.1 Stepsize policies and lower-bound statement

Consider rank-one simultaneous LoRA gradient descent with $B_t \in \mathbb{R}^{m\times 1}$ and $A_t \in \mathbb{R}^{1\times n}$:

$$B_{t+1} = B_t - \eta_t \nabla_B J(B_t, A_t), \qquad A_{t+1} = A_t - \eta_t \nabla_A J(B_t, A_t), \tag{43}$$

where $J(B, A) = F(BA)$.

Set $X_t := B_t A_t$. Before choosing η_t, the policy observes

$$\mathcal{O}_t := \big(B_t,\, A_t,\, X_t,\, F(X_t),\, \nabla F(X_t),\\ \nabla_B J(B_t, A_t),\, \nabla_A J(B_t, A_t)\big)$$

and retains all previous observations and stepsizes. The available history is

$$\mathcal{H}_t := \big(\mathcal{O}_0, \eta_0,\, \mathcal{O}_1, \eta_1,\, \ldots,\, \mathcal{O}_{t-1}, \eta_{t-1},\, \mathcal{O}_t\big),$$

with $\mathcal{H}_0 := (\mathcal{O}_0)$.

A deterministic causal stepsize policy is a sequence of deterministic maps $\{\Phi_t\}_{t\geq 0}$ satisfying

$$\eta_t = \Phi_t(\epsilon, \mathcal{H}_t) \in [0, \infty), \tag{44}$$

with a finite output for every finite history. The policy accesses F only through the observations in $\mathcal{H}_t$ and cannot evaluate F or ∇F at other points when choosing η_t. Both updates use the same nonnegative scalar in (43). This class includes predetermined schedules and deterministic rules $\eta_t = \alpha_t(G_t, R_t)$ with finite, nonnegative outputs.

Proposition C.1 (C^2 lower bound for causal stepsizes)**.** *Fix unit vectors $u \in \mathbb{R}^n$, $v \in \mathbb{R}^m$ and $\rho > 0$, and initialize $B_0 = 0$, $A_0 = \rho u^\top$. Let*

$$\kappa_{\mathrm{area}} = 6 \cdot 2^{2/3}, \qquad C_\rho = \kappa_{\mathrm{area}}(1 + \rho^3)^{1/3} + \frac{1}{2}.$$

For every deterministic causal stepsize policy and every $0 < \epsilon \leq \frac{1}{2}\min\{1, \rho^3\}$, there exists $F \in C^2(\mathbb{R}^{m\times n})$ satisfying $\|\nabla^2 F(X)\|_{\mathrm{op}} \leq 1$ for all X, $F_\star = 0$, and $F(0) = C_\rho$, such that

$$\|\nabla J(B_t, A_t)\|_F = 2\epsilon, \qquad 0 \leq t < N := \left\lfloor \frac{1}{16\epsilon^4} \right\rfloor. \tag{45}$$

Thus at least $N = \Omega(\epsilon^{-4})$ iterations are required to reach ϵ-stationarity. The objective may depend on the policy and on ϵ.

We first construct a trajectory consistent with the stepsize policy, then interpolate its weight gradients while controlling their integral. Extending this interpolant gives a global objective that reproduces every observation in the history.

C.2 Construction of the hard trajectory

Fix a policy and an admissible ϵ as in Proposition C.1. Set $\delta = 2\epsilon$, so $0 < \delta \leq \min\{1, \rho^3\}$ and $N = \lfloor \delta^{-4} \rfloor \geq 1$. Lemma C.2, proved in Appendix C.3, provides the interpolant on each interval.

Start with $a_0 = \rho$, $b_0 = 0$, and define

$$D_t = a_t^2 + b_t^2, \quad x_t = a_t b_t, \quad g_t = \frac{\delta}{\sqrt{D_t}}, \quad B_t = b_t v, \quad A_t = a_t u^\top.$$

We construct h one interval at a time, maintaining $h \geq 0$, $h(x_t) = g_t$, and $h'(x_t) = 0$. Start with $h(0) = g_0$ and $h'(0) = 0$. When the policy chooses η_t, for $0 \leq t < N$, h has already been fixed on $[0, x_t]$. We report the value

$$\lambda_t := C_\rho - \int_0^{x_t} h(s)\, ds,$$

the weight gradient $-g_t v u^\top$, and the factor gradients $-g_t a_t v$ and $-g_t b_t u^\top$. These responses determine $\mathcal{O}_t$ and hence $\mathcal{H}_t$ before the policy selects η_t. We will construct an objective that produces all these responses. Once the policy chooses η_t, set $q_t = \eta_t g_t$ and define

$$a_{t+1} = a_t + q_t b_t, \qquad b_{t+1} = b_t + q_t a_t. \tag{46}$$

Expanding the product and the sum of squares gives

$$x_{t+1} = (a_t + q_t b_t)(b_t + q_t a_t) = (1 + q_t^2) x_t + q_t D_t,$$
$$D_{t+1} = (a_t + q_t b_t)^2 + (b_t + q_t a_t)^2 = (1 + q_t^2) D_t + 4 q_t x_t.$$

Thus

$$x_{t+1} - x_t = q_t D_t + q_t^2 x_t, \qquad D_{t+1} = (1+q_t^2)D_t + 4q_t x_t. \tag{47}$$

Since $q_t \geq 0$, the recursion preserves $a_t \geq \rho > 0$ and $b_t \geq 0$. Both increments in (47) are therefore nonnegative. In particular, $D_t \geq \rho^2$, and $D_t - 2x_t = (a_t - b_t)^2 \geq 0$. Also,

$$g_t = \frac{\delta}{\sqrt{D_t}} \leq \frac{\rho^3}{\sqrt{D_t}} \leq \frac{D_t^{3/2}}{\sqrt{D_t}} = D_t.$$

If $q_t = 0$, the factors and all observations are unchanged and no interval is added. If $q_t > 0$, the bound $2x_t \leq D_t$ yields

$$D_t \leq D_{t+1} = (1+q_t^2)D_t + 4q_t x_t \leq (1+q_t)^2 D_t.$$

Taking reciprocal square roots and multiplying by $\delta > 0$ gives $g_t/(1+q_t) \leq g_{t+1} \leq g_t$. Subtracting from g_t gives

$$0 \leq g_t - g_{t+1} \leq \frac{g_t q_t}{1+q_t} < g_t q_t \leq D_t q_t \leq x_{t+1} - x_t. \tag{48}$$

For $q_t > 0$, this strict inequality allows us to apply Lemma C.2 on $[x_t, x_{t+1}]$ with

$$(a, b, d, e) = (g_t, g_{t+1}, x_{t+1} - x_t, 1/(4N)).$$

The interpolant depends only on these four quantities. Its zero endpoint derivatives match those of the preceding interpolant. We leave h unchanged on $[0, x_t]$. The integral over the new interval determines λ_{t+1} before the next stepsize is chosen. A zero step leaves the observations unchanged but is still recorded in the history.

Each finite stepsize gives finite $a_{t+1}, b_{t+1}, x_{t+1}$, while $D_{t+1} \geq \rho^2 > 0$ ensures that g_{t+1} is finite and positive. Each new interpolant is bounded on a finite interval and therefore has a finite integral. The recursion defines all observations through iteration N and all stepsizes through iteration $N-1$, without altering previous responses or evaluating Φ_N.

C.3 Smooth interpolation and integral bounds

To obtain $F \in C^2$, we construct a C^1 interpolant of the scalar derivative. The endpoint derivatives vanish, allowing adjacent pieces to join while preserving all previously reported values and gradients.

Lemma C.2 (C^1 interpolation with an integral bound)**.** *Let $a \geq b \geq 0$, $d > a - b$, and $e > 0$. There is a nonnegative, piecewise quadratic $H \in C^1([0,d])$ such that*

$$H(0) = a, \quad H(d) = b, \quad H'(0) = H'(d) = 0, \quad 0 \leq H \leq a, \quad \|H'\|_\infty \leq 1,$$

and, for $p(s) = \max\{a - s, b - (d-s), 0\}$,

$$\int_0^d H(s)\,ds \leq \int_0^d p(s)\,ds + e \leq \min\{ad, (a^2+b^2)/2\} + e. \tag{49}$$

The interpolant depends only on (a, b, d, e).

Proof. Set

$$\theta = \min\left\{\frac{1}{4}, \frac{d-a+b}{2d}, \frac{e}{d^2}\right\}, \qquad \kappa = 1 - \theta,$$

and define

$$p_\kappa(s) = \max\{a - \kappa s, b - \kappa(d-s), 0\}.$$

Since $d > a - b$ and $e > 0$, we have $0 < \theta \leq 1/4$. Moreover,

$$\kappa d \geq \frac{d+a-b}{2} > a - b.$$

Thus $p_\kappa(0) = a$, $p_\kappa(d) = b$, and $0 \leq p_\kappa \leq a$. On every affine piece, its slope has magnitude at most κ.

To make the derivative vanish at the endpoints of each affine piece, define

$$S_\theta(u) = \frac{1}{1-\theta}\int_0^u \min\left\{\frac{v}{\theta}, 1, \frac{1-v}{\theta}\right\} dv, \qquad 0 \le u \le 1.$$

The integrand equals v/θ on $[0,\theta]$, 1 on $[\theta, 1-\theta]$, and $(1-v)/\theta$ on $[1-\theta, 1]$. It is continuous, and its integral is

$$\frac{\theta}{2} + (1-2\theta) + \frac{\theta}{2} = 1-\theta.$$

Consequently, S_θ is C^1 and piecewise quadratic. Its normalization and the bounds on its integrand give

$$S_\theta(0) = 0, \qquad S_\theta(1) = 1, \qquad S'_\theta(0) = S'_\theta(1) = 0,$$
$$0 \le S_\theta \le 1, \qquad 0 \le S'_\theta \le \kappa^{-1}, \quad \int_0^1 S_\theta(u)\,du = \frac{1}{2}.$$

The integrand is symmetric about $1/2$, so $S_\theta(1-u) = 1 - S_\theta(u)$. Integrating this equality and changing variables on its left-hand side gives

$$\int_0^1 S_\theta(u)\,du = 1 - \int_0^1 S_\theta(u)\,du.$$

Moving the second integral to the left yields $2\int_0^1 S_\theta(u)\,du = 1$, proving the last identity.

On each maximal affine interval $[r, s]$ of p_κ, set

$$H(x) = p_\kappa(r) + [p_\kappa(s) - p_\kappa(r)] S_\theta\left(\frac{x-r}{s-r}\right).$$

This function agrees with p_κ at the endpoints, stays between the endpoint values, and has zero derivative at both endpoints. For $r < x < s$,

$$|H'(x)| = \left|\frac{p_\kappa(s) - p_\kappa(r)}{s-r}\right| S'_\theta\left(\frac{x-r}{s-r}\right) \le \kappa\,\kappa^{-1} = 1.$$

The finitely many pieces therefore join to form a nonnegative, piecewise quadratic C^1 function with the required values and derivatives.

The identity $\int_0^1 S_\theta = 1/2$ also gives

$$\int_r^s H(x)\,dx = \frac{s-r}{2}\big(p_\kappa(r) + p_\kappa(s)\big) = \int_r^s p_\kappa(x)\,dx.$$

For $0 \le x \le d$, the two affine terms change by

$$(a - \kappa x) - (a - x) = \theta x, \qquad [b - \kappa(d-x)] - [b - (d-x)] = \theta(d-x).$$

Each change lies in $[0, \theta d]$, and the third term remains zero. Taking the maximum therefore gives $0 \le p_\kappa(x) - p(x) \le \theta d$. Summing the integral identities over all affine pieces, and using $\theta d^2 \le e$, yields

$$\int_0^d H(x)\,dx = \int_0^d p_\kappa(x)\,dx \le \int_0^d p(x)\,dx + \theta d^2 \le \int_0^d p(x)\,dx + e.$$

Finally, both affine terms defining p are at most a, so $\int_0^d p(x)\,dx \le ad$. For the other bound, let $(r)_+ := \max\{r, 0\}$ and use $p(x) \le (a-x)_+ + (b-(d-x))_+$. Changing variables in the second integral gives

$$\begin{aligned}\int_0^d p(x)\,dx &\le \int_0^d (a-x)_+\,dx + \int_0^d (b-x)_+\,dx \\ &\le \int_0^a (a-x)\,dx + \int_0^b (b-x)\,dx = \frac{a^2+b^2}{2}.\end{aligned}$$

Combining the two bounds proves (49). □

Uniform integral bound. Let $d_t = x_{t+1} - x_t$, and let $I_t^{(0)}$ be the integral of the piecewise-linear reference function $\max\{g_t - s, g_{t+1} - (d_t - s), 0\}$ over $[0, d_t]$. Set $I_t^{(0)} = 0$ when $d_t = 0$. There are at most N nonempty intervals, each adding at most $1/(4N)$ to the reference integral. On a nonempty interval, Lemma C.2 bounds its reference integral by

$$I_t^{(0)} \le \min\left\{ g_t d_t, \frac{g_t^2 + g_{t+1}^2}{2} \right\} \le \min\{g_t d_t, g_t^2\},$$

where the last inequality uses $g_{t+1} \le g_t$. Summing the interpolation errors gives at most $N/(4N) = 1/4$, hence

$$A_{\text{path}} := \int_0^{x_N} h(s)\, ds \le \sum_{t=0}^{N-1} I_t^{(0)} + \frac{1}{4}, \qquad I_t^{(0)} \le \min\{g_t d_t, g_t^2\}. \tag{50}$$

To bound the sum without an upper bound on d_t, set

$$w_t = \rho^2 + 2x_t, \qquad y_t = w_t^{3/2}.$$

Adding $D_t \ge \rho^2$ and $D_t \ge 2x_t$ gives $2D_t \ge \rho^2 + 2x_t = w_t$. Therefore $g_t \le \sqrt{2}\,\delta/\sqrt{w_t}$ and $g_t^2 \le 2\delta^2/w_t$. Also,

$$y_{t+1} - y_t = \int_{x_t}^{x_{t+1}} 3\sqrt{\rho^2 + 2x}\, dx \ge 3\sqrt{w_t}\, d_t.$$

Using $w_t = y_t^{2/3}$, we obtain

$$g_t d_t \le \frac{\sqrt{2}\,\delta d_t}{\sqrt{w_t}} \le \frac{\sqrt{2}}{3}\delta \frac{y_{t+1} - y_t}{y_t^{2/3}},$$

$$g_t^2 \le \frac{2\delta^2}{y_t^{2/3}}.$$

Define the truncated increments and a second nondecreasing sequence by

$$m_t = \min\{y_{t+1} - y_t, \delta\}, \qquad z_0 = \rho^3, \qquad z_{t+1} = z_t + m_t.$$

Since $\sqrt{2}/3 \le 2$, the two bounds on $I_t^{(0)}$ imply

$$I_t^{(0)} \le \min\left\{ \frac{2\delta(y_{t+1} - y_t)}{y_t^{2/3}}, \frac{2\delta^2}{y_t^{2/3}} \right\} = \frac{2\delta m_t}{y_t^{2/3}}.$$

We have $z_0 = y_0$ and $m_t \le y_{t+1} - y_t$, so $z_t \le y_t$. Moreover, $0 \le m_t \le \delta \le \rho^3 \le z_t$, hence $z_{t+1} \le 2z_t$. For $m_t > 0$, the mean value theorem gives

$$z_{t+1}^{1/3} - z_t^{1/3} \ge \frac{m_t}{3z_{t+1}^{2/3}} \ge \frac{m_t}{3 \cdot 2^{2/3} z_t^{2/3}}.$$

The same inequality holds with both sides zero when $m_t = 0$. Consequently,

$$I_t^{(0)} \le \frac{2\delta m_t}{y_t^{2/3}} \le \frac{2\delta m_t}{z_t^{2/3}} \le \kappa_{\text{area}}\delta\big(z_{t+1}^{1/3} - z_t^{1/3}\big).$$

Summing the last inequality telescopes the cube-root increments:

$$\sum_{t=0}^{N-1} I_t^{(0)} \le \kappa_{\text{area}}\delta \sum_{t=0}^{N-1} (z_{t+1}^{1/3} - z_t^{1/3}) = \kappa_{\text{area}}\delta(z_N^{1/3} - \rho).$$

Since each $m_t \le \delta$ and $N \le \delta^{-4}$, $z_N \le \rho^3 + N\delta \le \rho^3 + \delta^{-3}$. Thus

$$\delta z_N^{1/3} \le (1 + \delta^3\rho^3)^{1/3} \le (1 + \rho^3)^{1/3},$$

where the last inequality uses $\delta \le 1$. Combining these bounds with (50) and the definition $C_\rho = \kappa_{\text{area}}(1 + \rho^3)^{1/3} + 1/2$ gives

$$\begin{aligned} A_{\text{path}} &\le \kappa_{\text{area}}\delta\big(z_N^{1/3} - \rho\big) + \frac{1}{4} \\ &\le \kappa_{\text{area}}\big((1 + \delta^3\rho^3)^{1/3} - \delta\rho\big) + \frac{1}{4} \le C_\rho - \frac{1}{4} - \kappa_{\text{area}}\delta\rho. \end{aligned} \tag{51}$$

Repeated visits contribute $d_t = I_t^{(0)} = m_t = 0$ and hence do not affect the sums. The term $\kappa_{\text{area}}\delta\rho$ ensures that the remaining integral is sufficient for the extension constructed next.

C.4 GLOBAL OBJECTIVE AND TRAJECTORY CONSISTENCY

Construction of the global objective. For $g > 0$, define

$$Q_g(s) = \begin{cases} g - s^2/(2g), & 0 \le s \le g, \\ (2g - s)^2/(2g), & g \le s \le 2g, \\ 0, & s \ge 2g. \end{cases}$$

The values agree at $s = g$ and $s = 2g$, and

$$Q_g'(s) = \begin{cases} -s/g, & 0 \le s \le g, \\ (s - 2g)/g, & g \le s \le 2g, \\ 0, & s \ge 2g. \end{cases}$$

Thus Q_g is C^1, with $Q_g'(g) = -1$ and $Q_g'(0) = Q_g'(2g) = 0$. It is nonnegative, satisfies $Q_g(0) = g$ and $\|Q_g'\|_\infty \le 1$, and has integral

$$\begin{aligned} \int_0^\infty Q_g(s)\, ds &= \int_0^g \left(g - \frac{s^2}{2g} \right) ds + \int_g^{2g} \frac{(2g - s)^2}{2g}\, ds \\ &= \left(g^2 - \frac{g^2}{6} \right) + \frac{g^2}{6} = g^2. \end{aligned}$$

Set $g = g_N$ and $A_{\mathrm{rem}} = C_\rho - A_{\mathrm{path}} = \lambda_N$. Since $D_N \ge \rho^2$ and $\delta \le \rho^3$,

$$g^2 = \frac{\delta^2}{D_N} \le \frac{\delta^2}{\rho^2} = \delta\rho\, \frac{\delta}{\rho^3} \le \delta\rho.$$

On the other hand, (51) gives

$$A_{\mathrm{rem}} = C_\rho - A_{\mathrm{path}} \ge \frac{1}{4} + \kappa_{\mathrm{area}}\delta\rho \ge g^2,$$

because $\kappa_{\mathrm{area}} > 1$. Hence the remaining integral is large enough to include the piece Q_g, whose integral is g^2. The excess is supplied by a constant segment of height $g > 0$ and finite length $\tau = (A_{\mathrm{rem}} - g^2)/g \ge 0$. Leave h unchanged on $[0, x_N]$ and extend it by

$$h(x) = \begin{cases} Q_{g_0}(-x), & x \le 0, \\ g, & x_N \le x \le x_N + \tau, \\ Q_g(x - x_N - \tau), & x \ge x_N + \tau. \end{cases}$$

At each junction with the existing interpolant or the constant segment, the values agree and both one-sided derivatives vanish. If $x_N = 0$, all stepsizes $\eta_0, \dots, \eta_{N-1}$ are zero by (47). Hence $g_N = g_0$, and the left and right extensions agree at the only visited point. The resulting h is continuously differentiable, vanishes outside $[-2g_0, x_N + \tau + 2g]$, and satisfies

$$h \ge 0, \qquad \|h'\|_\infty \le 1, \qquad h(x_t) = g_t, \qquad h'(x_t) = 0 \quad (0 \le t \le N).$$

The integral over the right extension is $g\tau + g^2 = A_{\mathrm{rem}}$, so

$$\int_0^\infty h(s)\, ds = A_{\mathrm{path}} + A_{\mathrm{rem}} = C_\rho.$$

Define

$$\varphi(x) = C_\rho - \int_0^x h(s)\, ds, \qquad F(W) = \varphi(v^\top W u). \tag{52}$$

Because h is continuously differentiable, $\varphi \in C^2$ and $\varphi'' = -h'$. The integral identity gives

$$\varphi(x) = \int_x^\infty h(s)\, ds \ge 0 \qquad \text{for every } x \in \mathbb{R}.$$

For $x \ge x_N + \tau + 2g$, we have $\varphi(x) = 0$ and $F(xvu^\top) = \varphi(x) = 0$. Thus $F_\star = 0$ and $F(0) = C_\rho$.

By the chain rule,

$$\nabla F(W) = -h(v^\top W u) v u^\top, \qquad \nabla^2 F(W)[Z] = -h'(v^\top W u)(v^\top Z u) v u^\top.$$

The Hessian is continuous because h' is continuous. The unit-norm assumptions imply $\|vu^\top\|_F = \|v\|_F \|u\|_F = 1$. Moreover,

$$|v^\top Z u| = |\langle vu^\top, Z\rangle| \le \|vu^\top\|_F \|Z\|_F = \|Z\|_F.$$

Thus

$$\|\nabla^2 F(W)[Z]\|_F \le |h'(v^\top W u)|\, |v^\top Z u|\, \|vu^\top\|_F \le \|Z\|_F.$$

Hence $F \in C^2$ and $\|\nabla^2 F(W)\|_{\mathrm{op}} \le 1$ for every W. In particular, F has globally 1-Lipschitz gradient and satisfies Assumption 3.4 with $\ell \equiv 1$.

Consistency with the constructed trajectory. Apply the same policy, with the same accuracy ϵ and initialization, to the fixed objective F. Denote its iterates by $\widetilde{B}_t, \widetilde{A}_t$. We show by induction that they equal $b_t v, a_t u^\top$ and that the full histories agree through every iteration $t \le N$.

At $t = 0$, the initial factors agree. Suppose the factors at iteration t and all preceding observations and stepsizes agree. Then $\widetilde{X}_t = \widetilde{B}_t \widetilde{A}_t = x_t v u^\top$. The preserved integral on $[0, x_t]$ and the equality $h(x_t) = g_t$ give

$$F(\widetilde{X}_t) = \lambda_t, \qquad \nabla F(\widetilde{X}_t) = -g_t v u^\top,$$
$$\nabla_B J(\widetilde{B}_t, \widetilde{A}_t) = -g_t a_t v, \quad \nabla_A J(\widetilde{B}_t, \widetilde{A}_t) = -g_t b_t u^\top.$$

Thus the current observations agree, and together with the induction hypothesis they imply $\widetilde{\mathcal{H}}_t = \mathcal{H}_t$. For $t < N$, determinism yields

$$\widetilde{\eta}_t = \Phi_t(\epsilon, \widetilde{\mathcal{H}}_t) = \Phi_t(\epsilon, \mathcal{H}_t) = \eta_t.$$

Substituting this stepsize and the factor gradients into (43) gives

$$\widetilde{B}_{t+1} = (b_t + \eta_t g_t a_t) v = b_{t+1} v,$$
$$\widetilde{A}_{t+1} = (a_t + \eta_t g_t b_t) u^\top = a_{t+1} u^\top,$$

by (46), completing the induction. At $t = N$, only the observations need to agree; no stepsize at iteration N is required. Equality of the histories suffices, so the policy need not be continuous.

C.5 Proof of the main lower bound

Proof of Proposition C.1. The construction above gives a lower-bounded C^2 objective with $F_\star = 0$, $F(0) = C_\rho$, and Hessian operator norm at most one. By trajectory consistency, for every $0 \le t < N$,

$$\|\nabla J(B_t, A_t)\|_F^2 = \| - g_t a_t v\|_F^2 + \| - g_t b_t u^\top\|_F^2 = g_t^2 D_t = \delta^2.$$

Thus the factor-gradient norm equals $\delta = 2\epsilon$ at every such iterate. Define

$$T_\epsilon := \inf\{t \ge 0 : \|\nabla J(B_t, A_t)\|_F \le \epsilon\}, \qquad \inf \varnothing := +\infty,$$

so that $T_\epsilon \ge N = \lfloor 1/(16\epsilon^4) \rfloor$. Because $\epsilon \le 1/2$, the quantity inside the floor is at least one, so $N \ge 1/(32\epsilon^4)$. The objective depends only on the fixed policy and target accuracy, whereas C_ρ depends only on the initialization scale. □

Proof of Theorem 4.4. Set $\rho = 1$ in Proposition C.1. The initialization becomes $B_0 = 0$, $A_0 = u^\top$, the accuracy range is $0 < \epsilon \le 1/2$, and

$$C_1 = 6 \cdot 2^{2/3}(1 + 1^3)^{1/3} + \frac{1}{2} = \frac{25}{2}.$$

The C^2 objective constructed in Section C.4 satisfies

$$F_\star = 0, \qquad F(0) = \frac{25}{2}, \qquad \|\nabla^2 F(X)\|_{\mathrm{op}} \le 1 \quad \text{for every } X.$$

By (45), its iterates satisfy $\|\nabla J(B_t, A_t)\|_F = 2\epsilon$ for every $0 \le t < N$. Thus $T_\epsilon \ge N = \Omega(\epsilon^{-4})$, including $T_\epsilon = +\infty$ when no iterate reaches ϵ-stationarity. This proves the theorem. □

D Proof of the LoRA-SGD Upper Bound

Suppose Assumptions 3.2, 3.4, and 3.5 hold. Consider the LoRA-SGD iterates in (5).

D.1 Stochastic LoRA identities and noise amplification

Write

$$V_t := (B_t, A_t), \qquad p_t := \nabla J(V_t), \qquad E_t := g(X_t; \xi_t) - \nabla F(X_t),$$

and let $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_t]$. Then

$$\mathbb{E}_t[E_t] = 0, \qquad \mathbb{E}_t\|E_t\|_F^2 \le \sigma^2. \tag{53}$$

We use a constant stepsize chosen before any oracle samples are drawn. Define the stochastic factor-gradient error

$$\zeta_t = (E_t A_t^\top, B_t^\top E_t).$$

The factors A_t, B_t are $\mathcal{F}_t$-measurable, so

$$\mathbb{E}_t[E_t A_t^\top] = (\mathbb{E}_t E_t) A_t^\top = 0, \qquad \mathbb{E}_t[B_t^\top E_t] = B_t^\top (\mathbb{E}_t E_t) = 0.$$

Also, the product inequality gives

$$\|\zeta_t\|_F^2 = \|E_t A_t^\top\|_F^2 + \|B_t^\top E_t\|_F^2 \le (\|A_t\|_F^2 + \|B_t\|_F^2)\|E_t\|_F^2 = R_t\|E_t\|_F^2.$$

Thus $V_{t+1} = V_t - \eta(p_t + \zeta_t)$ and

$$\mathbb{E}_t\zeta_t = 0, \qquad \|\zeta_t\|_F^2 \le R_t\|E_t\|_F^2, \qquad \mathbb{E}_t\|\zeta_t\|_F^2 \le \sigma^2 R_t. \tag{54}$$

The chain rule also gives $\|p_t\|_F \le \|\nabla F(X_t)\|_F\sqrt{R_t}$. These bounds control the stochastic terms up to a stopping time defined by thresholds on the objective gap and factor norm.

D.2 High-probability trajectory control

Define the following constants, independent of q, T, ϵ:

$$\begin{gathered}\overline{G}(U) = \sup\{u \ge 0 : u^2 \le 2\ell(2u)U\}, \quad G = 1 + \overline{G}(\Delta + 1), \quad L = \ell(2G), \\ D = 4(R_0 + \Delta + 2), \quad s = \max\{1, \sigma\}, \quad W = G + 2s, \quad K_0 = G + 2LD, \\ a = \min\left\{1, \frac{1}{2W}, \frac{G}{4DLW}, \frac{1}{4K_0}, \frac{1}{16DGs}, \frac{1}{8\sqrt{DK_0}s}, \frac{1}{\sqrt{8D}s}\right\}.\end{gathered} \tag{55}$$

By Proposition A.3, $\overline{G}(U) < \infty$, and the weight-gradient norm is at most G whenever $J - F_\star \le \Delta + 1$.

Lemma D.1 (High-probability trajectory bounds)**.** *Under the stated assumptions, let $q \in (0, 1)$ and $T \ge 1$ be an integer. Set*

$$\eta = \frac{a}{2}\min\left\{\sqrt{\frac{q}{T}}, q^{1/4}T^{-3/4}\right\}. \tag{56}$$

With probability at least $1 - q$, simultaneously for all $0 \le k \le T$,

$$J_k - F_\star + \frac{\eta}{2}\sum_{t<k}\|\nabla J_t\|_F^2 \le \Delta + \frac{1}{2}, \tag{57}$$

$$\sqrt{R_k} \le \sqrt{R_0} + 1 + \sqrt{(2\Delta + 1)\eta k}. \tag{58}$$

Proof. Stepsize bounds. Write $S = \eta T$ and $z = \sqrt{qT}$. The stepsize definition gives $S = (a/2)\min\{z, \sqrt{z}\}$. If $z \le 1$, then $S = az/2$ and $S^2 = a^2z^2/4 \le a^2z/4$. If $z \ge 1$, then $S = a\sqrt{z}/2 \le az/2$ and $S^2 = a^2z/4$. In both cases, $a \le 1$ gives

$$S(1 + S) \le \left(\frac{a}{2} + \frac{a^2}{4}\right)z \le a\sqrt{qT}. \tag{59}$$

Set

$$\mathcal{R} = D(1 + S), \quad K = G + L\mathcal{R}, \quad M = 2s\sqrt{T/q}, \quad K_c = G + LD \le K_0.$$

Since $S \geq 0$ and $\sqrt{T/q} \geq 1$,

$$K = G + LD(1+S) \leq K_c(1+S), \qquad G + M \leq (G+2s)\sqrt{T/q} = W\sqrt{T/q}.$$

Using $\eta \leq (a/2)\sqrt{q/T}$ and (59), we obtain

$$\begin{aligned}\eta(G+M) &\leq aW/2 \leq 1/4,\\ \eta K &\leq K_c \frac{S(1+S)}{T} \leq aK_c\sqrt{q/T} \leq 1/4,\\ \eta\mathcal{R}(G+M) &\leq DW\frac{S(1+S)}{T}\sqrt{T/q} \leq DaW.\end{aligned}$$

Here the first two upper bounds follow from $a \leq 1/(2W)$ and $a \leq 1/(4K_0)$, respectively. In particular, $1 + \eta(G+M)/2 \leq 9/8 \leq 5/4$. The bound $a \leq G/(4DLW)$ therefore gives

$$\eta\mathcal{R}(G+M)\left(1 + \frac{\eta(G+M)}{2}\right) \leq \frac{5}{4}DaW \leq \frac{5G}{16L} < \frac{G}{L}. \tag{60}$$

One-step descent. Suppose $J_t - F_\star \leq \Delta + 1$, $R_t \leq \mathcal{R}$, and $\|E_t\|_F \leq M$. Then $\|\nabla F(X_t)\|_F \leq G$ and $\|g(X_t;\xi_t)\|_F \leq G + M$. Write $H_t = p_t + \zeta_t$ for the sampled factor gradient. The chain rule gives

$$\|H_t\|_F^2 = \|g(X_t;\xi_t)A_t^\top\|_F^2 + \|B_t^\top g(X_t;\xi_t)\|_F^2 \leq R_t\|g(X_t;\xi_t)\|_F^2.$$

Expanding the product of the two updated factors as in (25) yields

$$\begin{aligned}\|\Delta X_t\|_F &\leq \eta\sqrt{R_t}\|H_t\|_F + \frac{\eta^2}{2}\|H_t\|_F^2\\ &\leq \eta\sqrt{R_t}\|H_t\|_F\left(1 + \frac{\eta}{2}\|g(X_t;\xi_t)\|_F\right)\\ &\leq \eta\mathcal{R}(G+M)\left(1 + \frac{\eta(G+M)}{2}\right) < G/L.\end{aligned}$$

Thus Lemma A.1 applies with gradient threshold G and $L = \ell(2G)$.

The terms linear in the factor increments contribute $-\eta\langle p_t, H_t\rangle$ to the weight-space inner product. The bilinear term satisfies

$$|\langle \nabla F(X_t), \Delta B_t \Delta A_t\rangle| \leq G\|\Delta B_t\|_F\|\Delta A_t\|_F \leq \frac{\eta^2 G}{2}\|H_t\|_F^2.$$

The Taylor remainder is bounded by

$$\frac{L}{2}\|\Delta X_t\|_F^2 \leq \frac{\eta^2 LR_t}{2}\left(\frac{5}{4}\right)^2\|H_t\|_F^2 = \frac{25\eta^2 LR_t}{32}\|H_t\|_F^2.$$

Since $G/2 + (25/32)LR_t \leq G + L\mathcal{R} = K$, these bounds give

$$\begin{aligned}J_{t+1} &\leq J_t - \eta\|p_t\|_F^2 - \eta\langle p_t, \zeta_t\rangle + \eta^2 K\|p_t + \zeta_t\|_F^2\\ &\leq J_t - (\eta - 2\eta^2 K)\|p_t\|_F^2 - \eta\langle p_t, \zeta_t\rangle + 2\eta^2 K\|\zeta_t\|_F^2.\end{aligned}$$

The second inequality uses $\|p_t+\zeta_t\|_F^2 \leq 2\|p_t\|_F^2 + 2\|\zeta_t\|_F^2$. Finally, $\eta K \leq 1/4$ implies $\eta - 2\eta^2 K \geq \eta/2$, so

$$J_{t+1} \leq J_t - \frac{\eta}{2}\|p_t\|_F^2 - \eta\langle p_t, \zeta_t\rangle + 2\eta^2 K\|\zeta_t\|_F^2. \tag{61}$$

Stopped stochastic sums. Define

$$\tau = \min\big(\{t \in \{0,\dots,T\} : J_t - F_\star > \Delta + 1 \text{ or } R_t > \mathcal{R}\} \cup \{T+1\}\big), \qquad I_t = \mathbf{1}_{\{t<\tau\}}.$$

The initial iterate satisfies both thresholds, so $\tau \geq 1$. The indicator I_t is $\mathcal{F}_t$-measurable, before E_t is sampled. Therefore $I_t\zeta_t$ and $I_t\langle p_t, \zeta_t\rangle$ have zero conditional means. By (54),

$$\begin{aligned}\mathbb{E}_t\|I_t\zeta_t\|_F^2 &= I_t\mathbb{E}_t\|\zeta_t\|_F^2 \leq I_t\sigma^2 R_t \leq \sigma^2\mathcal{R},\\ \mathbb{E}_t|I_t\langle p_t, \zeta_t\rangle|^2 &\leq I_t\|p_t\|_F^2\mathbb{E}_t\|\zeta_t\|_F^2 \leq I_t\sigma^2 G^2 R_t^2 \leq \sigma^2 G^2\mathcal{R}^2.\end{aligned}$$

These deterministic upper bounds make the stopped increments square integrable, without requiring moments of the unstopped factors. Set

$$\mathsf{M}_k = \eta \sum_{t<k} I_t \langle p_t, \zeta_t \rangle, \quad \mathsf{Z}_k = \eta \sum_{t<k} I_t \zeta_t, \quad \mathsf{Q} = 2\eta^2 K \sum_{t<T} I_t \|\zeta_t\|_F^2.$$

Martingale increments at distinct times are orthogonal in expectation. Indeed, conditioning on the later history makes the later increment's conditional mean zero. Expanding the squares of the sums therefore gives

$$\mathbb{E}|\mathsf{M}_T|^2 \le \eta^2 T \sigma^2 G^2 \mathcal{R}^2 = D^2 G^2 \sigma^2 \frac{S^2(1+S)^2}{T} \le D^2 G^2 \sigma^2 a^2 q,$$
$$\mathbb{E}\|\mathsf{Z}_T\|_F^2 \le \eta^2 T \sigma^2 \mathcal{R} = D\sigma^2 \frac{S^2(1+S)}{T} \le D\sigma^2 a^2 q,$$
$$\mathbb{E}\mathsf{Q} \le 2\eta^2 K T \sigma^2 \mathcal{R} \le 2DK_c\sigma^2 \frac{S^2(1+S)^2}{T} \le 2DK_c\sigma^2 a^2 q.$$

The last inequalities use (59); for the second line, we also use $1 + S \le (1+S)^2$.

Probability estimates. The processes $|\mathsf{M}_k|^2$ and $\|\mathsf{Z}_k\|_F^2$ are nonnegative submartingales. Their maximal inequality, followed by the preceding moment bounds, gives

$$\mathbb{P}\left(\max_{k\le T} |\mathsf{M}_k| > \frac{1}{4}\right) \le 16\mathbb{E}|\mathsf{M}_T|^2 \le 16D^2G^2\sigma^2a^2q \le q/16,$$
$$\mathbb{P}\left(\max_{k\le T} \|\mathsf{Z}_k\|_F > 1\right) \le \mathbb{E}\|\mathsf{Z}_T\|_F^2 \le D\sigma^2a^2q \le q/8.$$

Here we used $a \le 1/(16DGs)$, $a \le 1/(\sqrt{8D}s)$, and $\sigma \le s$. Markov's inequality for the nonnegative random variable Q gives

$$\mathbb{P}(\mathsf{Q} > 1/4) \le 4\mathbb{E}\mathsf{Q} \le 8DK_c\sigma^2a^2q \le q/8,$$

by $a \le 1/(8\sqrt{DK_0}s)$ and $K_c \le K_0$. Finally, predictability of I_t and conditional Markov's inequality give

$$\mathbb{P}(\exists t < T : t < \tau,\ \|E_t\|_F > M)$$
$$\le \sum_{t<T} \mathbb{E}[I_t\, \mathbb{P}(\|E_t\|_F > M \mid \mathcal{F}_t)] \le \sum_{t<T} \frac{\sigma^2 \mathbb{E} I_t}{M^2} \le \frac{T\sigma^2}{M^2} \le q/4.$$

Let $\mathcal{E}$ be the event on which all four thresholds hold. A union bound gives $\mathbb{P}(\mathcal{E}^c) \le q/16 + q/8 + q/8 + q/4 = 9q/16 \le q$. The martingale estimates were obtained for the stopped increments without conditioning on the current noise being small; the noise bound is used only when applying the pathwise descent inequality on $\mathcal{E}$.

Excluding a stopping event. Fix a sample path in $\mathcal{E}$ and an integer $k \le \min\{\tau, T\}$. Every update with $t < k$ starts before τ, so both state thresholds hold and $\|E_t\|_F \le M$. Summing (61) gives

$$J_k - F_\star + \frac{\eta}{2}\sum_{t<k} \|p_t\|_F^2 \le \Delta - \mathsf{M}_k + 2\eta^2 K \sum_{t<k} I_t \|\zeta_t\|_F^2$$
$$\le \Delta + \frac{1}{4} + \mathsf{Q} \le \Delta + \frac{1}{2}.$$

This proves (57) up to $\min\{\tau, T\}$, including $k = \tau$ if $\tau \le T$: the update at $t = \tau - 1$ still starts before the stopping time. Since $J_k - F_\star \ge 0$, we also have $\eta \sum_{t<k} \|p_t\|_F^2 \le 2\Delta + 1$.

For these same indices, $I_t = 1$ when $t < k$, so the exact update identity is $V_k = V_0 - \eta \sum_{t<k} p_t - \mathsf{Z}_k$. The triangle inequality and weighted Cauchy–Schwarz give

$$\sqrt{R_k} = \|V_k\|_F \le \sqrt{R_0} + \|\mathsf{Z}_k\|_F + \eta \sum_{t<k} \|p_t\|_F$$
$$\le \sqrt{R_0} + 1 + \sqrt{\eta k}\sqrt{\eta \sum_{t<k} \|p_t\|_F^2}$$
$$\le \sqrt{R_0} + 1 + \sqrt{(2\Delta+1)\eta k},$$

which is (58). Applying $(u+v)^2 \le 2u^2 + 2v^2$ twice yields

$$R_k \le 2(\sqrt{R_0}+1)^2 + 2(2\Delta+1)\eta k \le 4R_0 + 4 + (4\Delta+2)\eta k.$$

Since $k \le T$ and $D = 4(R_0 + \Delta + 2)$, the gap between the factor threshold and this upper bound is at least

$$\begin{aligned}\mathcal{R} - [4R_0 + 4 + (4\Delta+2)\eta k] &\ge D(1+S) - [4R_0 + 4 + (4\Delta+2)S] \\ &= 4\Delta + 4 + (4R_0+6)S > 0.\end{aligned}$$

Thus $R_k < \mathcal{R}$, while the objective bound gives $J_k - F_\star \le \Delta + 1/2 < \Delta + 1$. If $\tau \le T$, these two strict inequalities at $k = \tau$ contradict the definition of τ. Therefore $\tau = T + 1$ on $\mathcal{E}$, and both conclusions hold simultaneously for all $0 \le k \le T$. □

D.3 Proof of the main SGD upper bound

Proof of Theorem 5.1. Take $c = a/2$ in (56) and let $U_0 = 2\Delta + 1$. On the event in Lemma D.1, (57) and the nonnegative objective gap imply

$$\eta \sum_{t=0}^{T-1} \|\nabla J_t\|_F^2 \le U_0.$$

Since $\eta T = c \min\{(qT)^{1/2}, (qT)^{1/4}\}$, it follows that

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla J_t\|_F^2 &\le \frac{U_0}{c} \max\{(qT)^{-1/2}, (qT)^{-1/4}\} \\ &\le \frac{U_0}{c}\big((qT)^{-1/4} + (qT)^{-1/2}\big).\end{aligned}$$

This is (7) with $C = U_0/c$. Both constants depend only on Δ, R_0, σ, and ℓ. For $0 < \epsilon \le 1$, an integer horizon satisfying

$$T \ge \frac{16 U_0^4}{a^4 q \epsilon^8} \tag{62}$$

gives

$$qT \ge \left(\frac{2U_0}{a\epsilon^2}\right)^4 \ge 1,$$

since $U_0 \ge 1$, $a \le 1$, and $\epsilon \le 1$. For $qT \ge 1$, the smaller term defining ηT is $(qT)^{1/4}$, so

$$\eta T = \frac{a}{2}(qT)^{1/4} \ge \frac{U_0}{\epsilon^2}.$$

Consequently, on the same event,

$$\min_{t<T} \|\nabla J_t\|_F^2 \le \frac{1}{T}\sum_{t<T} \|\nabla J_t\|_F^2 \le \frac{U_0}{\eta T} \le \epsilon^2.$$

Thus at least one iterate is ϵ-stationary with probability at least $1 - q$. This proves the $O(q^{-1}\epsilon^{-8})$ iteration bound. □

Proof of Corollary 5.2. Let $\mathcal{E}$ denote the event in Theorem 5.1. On $\mathcal{E}$,

$$\min_{0\le t<T} \|\nabla J(B_t, A_t)\|_F^2 \le \frac{1}{T}\sum_{t=0}^{T-1} \|\nabla J(B_t, A_t)\|_F^2 \le C\left((qT)^{-1/4} + (qT)^{-1/2}\right).$$

Since the minimum includes $t = 0$, for every realization,

$$\min_{0\le t<T} \|\nabla J(B_t, A_t)\|_F^2 \le \|\nabla J(B_0, A_0)\|_F^2.$$

Therefore,

$$\begin{aligned}\mathbb{E}\left[\min_{0\le t<T}\|\nabla J(B_t,A_t)\|_F^2\right] &= \mathbb{E}\left[\min_{0\le t<T}\|\nabla J(B_t,A_t)\|_F^2\mathbf{1}_{\mathcal{E}}\right] \\ &\quad + \mathbb{E}\left[\min_{0\le t<T}\|\nabla J(B_t,A_t)\|_F^2\mathbf{1}_{\mathcal{E}^c}\right] \\ &\le C\left((qT)^{-1/4}+(qT)^{-1/2}\right)+q\|\nabla J(B_0,A_0)\|_F^2,\end{aligned}$$

where we used $\mathbb{P}(\mathcal{E}^c)\le q$.

Taking $q=T^{-1/5}$ gives

$$(qT)^{-1/4}=T^{-1/5},\qquad (qT)^{-1/2}=T^{-2/5},$$

and hence

$$\mathbb{E}\left[\min_{0\le t<T}\|\nabla J(B_t,A_t)\|_F^2\right]\le C\left(T^{-1/5}+T^{-2/5}\right)+T^{-1/5}\|\nabla J(B_0,A_0)\|_F^2.$$

For this choice of q, the stepsize in Theorem 5.1 becomes

$$\eta = c\min\left\{\sqrt{\frac{q}{T}}, q^{1/4}T^{-3/4}\right\} = c\min\{T^{-3/5},T^{-4/5}\} = cT^{-4/5}.$$

Thus

$$\mathbb{E}\left[\min_{0\le t<T}\|\nabla J(B_t,A_t)\|_F^2\right]=O(T^{-1/5}),$$

and $T=O(\epsilon^{-10})$ suffices to make the expected best-iterate squared gradient at most ϵ^2. □

E Infinite Expected Gradient Norms

Corollary 5.2 bounds the expected minimum squared gradient norm along the trajectory. This does not ensure that the expected gradient norm is finite at every iterate. The following example shows that, even for a globally 1-smooth objective F, unbiased stochastic gradients of variance one can produce an infinite expected factor-gradient norm after two LoRA-SGD updates.

Proposition E.1 (Infinite expected gradient norms under finite-variance noise)**.** *Consider scalar rank-one LoRA with*

$$F(x)=\frac{1}{2}(x-1)^2,\qquad J(b,a)=F(ab),\qquad b_0=0,\quad a_0=1.$$

Let the stochastic weight gradient be

$$\widehat{h}_t = a_tb_t-1+\xi_t,$$

where $\{\xi_t\}$ are independent with symmetric density

$$p_\xi(z)=\frac{3}{2(1+|z|)^4}.$$

Then $F\in C^\infty$, $F_\star=0$, $F''\equiv 1$, and

$$\mathbb{E}\xi_t=0,\qquad \mathbb{E}\xi_t^2=1,\qquad \mathbb{E}|\xi_t|^3=\mathbb{E}|\xi_t|^4=+\infty.$$

Suppose $0<\eta_0,\eta_1<\infty$ almost surely, with each stepsize chosen before the current noise sample is drawn. In particular, η_1 may depend on the entire history through time 1. Then

$$\mathbb{E}[J(b_2,a_2)]=+\infty,\qquad \mathbb{E}\|\nabla J(b_2,a_2)\|_F=+\infty,\qquad \mathbb{E}\|\nabla J(b_2,a_2)\|_F^2=+\infty.$$

Proof. First, the density is normalized because

$$\int_{\mathbb{R}}p_\xi(z)\,dz = 3\int_0^\infty (1+z)^{-4}\,dz = 1.$$

Symmetry and the finite first absolute moment give $\mathbb{E}\xi_t = 0$. For the second moment, the substitution $u = (1+z)^{-1}$ gives

$$\mathbb{E}\xi_t^2 = 3\int_0^\infty \frac{z^2}{(1+z)^4}\,dz = 3\int_0^1 (1-u)^2\,du = 1.$$

For $z \geq 1$, the moment integrand satisfies $z^m/(1+z)^4 \geq z^{m-4}/16$. Hence the absolute moments of orders $m = 3, 4$, and 6 are infinite. The stated properties of F follow from $F'(x) = x - 1$ and $F''(x) = 1$.

At $t = 0$, the stochastic weight gradient is $-1 + \xi_0$. The simultaneous updates and $b_0 = 0, a_0 = 1$ give

$$a_1 = a_0 - \eta_0 b_0(-1+\xi_0) = 1, \qquad b_1 = b_0 - \eta_0 a_0(-1+\xi_0) = \eta_0(1-\xi_0).$$

Since ξ_0 has a continuous density and $0 < \eta_0 < \infty$, b_1 is finite and nonzero almost surely. Condition on $\mathcal{F}_1$, including the information used to choose η_1. The fresh noise ξ_1 retains its stated distribution. For almost every such history, write

$$a = a_1 = 1, \quad b = b_1 \neq 0, \quad \eta = \eta_1 > 0, \quad d = ab - 1, \quad h = d + \xi_1.$$

All these quantities except h and ξ_1 are now fixed and finite. The simultaneous update is

$$b_+ = b - \eta a h, \qquad a_+ = a - \eta b h.$$

Multiplying these expressions gives

$$a_+ b_+ = ab - \eta(a^2+b^2)h + \eta^2 ab\,h^2. \tag{63}$$

Set $c = \eta^2 ab \neq 0$. Substituting $h = d + \xi_1$ and subtracting one yields

$$a_+ b_+ - 1 = c\xi_1^2 + [2cd - \eta(a^2+b^2)]\xi_1 + [d - \eta(a^2+b^2)d + cd^2].$$

The last two terms have degree at most one in ξ_1, whereas $c \neq 0$. Thus there is a finite history-dependent threshold H_1 such that

$$|a_+ b_+ - 1| \geq \frac{|c|}{2}|\xi_1|^2 \qquad \text{whenever } |\xi_1| \geq H_1.$$

For example, the threshold can be chosen so that the sum of the absolute values of the linear and constant terms is at most $|c|\,|\xi_1|^2/2$. It follows that

$$J(b_+, a_+) = \frac{1}{2}(a_+ b_+ - 1)^2 \geq \frac{c^2}{8}|\xi_1|^4 \qquad (|\xi_1| \geq H_1).$$

Removing a bounded interval does not make an infinite fourth moment finite. Since $c^2/8 > 0$ for almost every conditioned history,

$$\mathbb{E}[J(b_2, a_2) \mid \mathcal{F}_1] = +\infty \qquad \text{almost surely.}$$

For the factor norm, write the updated pair as

$$\begin{pmatrix} b_+ \\ a_+ \end{pmatrix} = \begin{pmatrix} b - \eta a d \\ a - \eta b d \end{pmatrix} - \eta\xi_1 \begin{pmatrix} a \\ b \end{pmatrix}.$$

The reverse triangle inequality gives

$$\sqrt{a_+^2 + b_+^2} \geq \eta\sqrt{a^2+b^2}\,|\xi_1| - \sqrt{(b-\eta a d)^2 + (a - \eta b d)^2}.$$

The subtracted term is fixed and finite. Hence there is a finite $H_2 \geq H_1$ such that

$$\sqrt{a_+^2 + b_+^2} \geq \frac{\eta}{2}\sqrt{a^2+b^2}\,|\xi_1| \qquad (|\xi_1| \geq H_2).$$

Using the exact factor-gradient formula, we obtain

$$\begin{aligned}\|\nabla J(b_+, a_+)\|_F &= |a_+ b_+ - 1|\sqrt{a_+^2 + b_+^2} \\ &\geq \frac{|c|\eta\sqrt{a^2+b^2}}{4}|\xi_1|^3 \qquad (|\xi_1| \geq H_2).\end{aligned}$$

The positive coefficient and the infinite third moment imply an infinite conditional expectation for the norm. Squaring this inequality and using the infinite sixth moment gives the same conclusion for its square. Finally, all three quantities are nonnegative, so the tower property for extended nonnegative expectations applies even when their expectations are infinite. Taking expectation over $\mathcal{F}_1$ proves all three claims. □

The minimum in Corollary 5.2 includes the initial iterate. For the initialization above and every $T \geq 1$,

$$0 \leq \min_{0 \leq t < T} \|\nabla J(b_t, a_t)\|_F^2 \leq \|\nabla J(b_0, a_0)\|_F^2 = 1.$$

Thus, the expected minimum squared gradient norm remains finite even though the expected gradient norm at $t = 2$ is infinite. When the stepsize is chosen as in Corollary 5.2, the expected minimum squared gradient norm is $O(T^{-1/5})$.

F Proof of the LoRA-SGD Lower Bound

F.1 Stepsize policies and lower-bound statement

We prove Theorem 5.3 for the simultaneous updates in (5). After observing the current stochastic gradient g_t, the stepsize may be chosen as

$$\eta_t = \Phi_t(\epsilon, \mathcal{H}_t^{\mathrm{sgd}}, g_t, U_t), \qquad 0 \leq \eta_t \leq 1,$$

where $\mathcal{H}_t^{\mathrm{sgd}}$ is the full observed history defined in Section 5.3, and U_t is independent internal randomness. Both factor updates use the same scalar stepsize and oracle sample.

The construction combines a sequential coordinate chain with a scalar component that controls factor growth. The chain limits the activation of new coordinates. Together, the two components keep the factor-gradient norm above ϵ until the chain is complete. For each sufficiently small ϵ, the same objective and oracle are fixed before optimization and apply to every policy in this class. The rank, initialization, initial gap, smoothness bound, and variance bound are fixed, while the dimension is $O(\epsilon^{-4})$.

F.2 Sequential coordinate-chain construction

Define

$$\Psi(u) = \begin{cases} 0, & u \leq 0, \\ 10u^3 - 15u^4 + 6u^5, & 0 < u < 1, \\ 1, & u \geq 1, \end{cases} \qquad \Phi(u) = 2 + \arctan u, \qquad P = 2 + \pi/2. \tag{64}$$

For an integer $K \geq 1$, put $u_0 = 1$ and

$$Z_K(u) = -\sum_{i=1}^{K} \Psi(u_{i-1})\Phi(u_i), \qquad u \in \mathbb{R}^K. \tag{65}$$

We establish the chain properties needed below for the functions in (64).

Lemma F.1 (Chain properties)**.** *The function Z_K belongs to C^2 and has the following properties, uniformly in K:*

$$-KP \leq Z_K(u) \leq 0, \quad \inf Z_K = -KP, \quad Z_K(0) = -2; \tag{66}$$

$$-9 \leq \partial_i Z_K(u) \leq 0, \quad \|\nabla^2 Z_K(u)\|_{\mathrm{op}} \leq 32. \tag{67}$$

If $u_{k+1} = \cdots = u_K = 0$, all gradient coordinates with $i > k+1$ are zero. If $u \geq 0$ coordinatewise and $u_K = 0$, then some coordinate satisfies $|\partial_i Z_K(u)| \geq 1/2$.

Proof. The values and the first two derivatives of the polynomial in (64) match the constant pieces at 0 and 1. Thus $\Psi \in C^2$, while $\Phi \in C^\infty$, so $Z_K \in C^2$. On $(0, 1)$,

$$\Psi'(u) = 30u^2(1-u)^2, \qquad \Psi''(u) = 60u(1-u)(1-2u).$$

Since $u(1-u) \leq 1/4$, the first derivative is at most $15/8$. The extrema of the second derivative occur at $u = (3 \pm \sqrt{3})/6$, where its magnitude is $10/\sqrt{3}$. Also,

$$\Phi'(u) = \frac{1}{1+u^2}, \qquad \Phi''(u) = \frac{-2u}{(1+u^2)^2}.$$

The inequality $2|u| \le 1 + u^2$ implies $|\Phi''(u)| \le 1$. Consequently,

$$0 \le \Psi \le 1, \quad 0 \le \Psi' \le 15/8 < 2, \quad |\Psi''| \le 10/\sqrt{3} < 6,$$
$$0 < \Phi < P < 4, \quad 0 < \Phi' \le 1, \quad |\Phi''| \le 1.$$

In addition, $\Phi'(u) \ge 1/2$ for $0 \le u \le 1$.

Every summand of Z_K lies in $[-P, 0]$, which gives $-KP \le Z_K \le 0$. If all coordinates equal R and $R \to \infty$, then every activation equals one and every $\Phi(R)$ tends to P. Hence $\inf Z_K = -KP$. At the origin only the first summand is nonzero, so $Z_K(0) = -\Psi(1)\Phi(0) = -2$.

For each i, only summands i and $i+1$ can depend on u_i. Differentiating those summands gives

$$\partial_i Z_K(u) = -\Psi(u_{i-1})\Phi'(u_i) - \mathbf{1}_{\{i<K\}}\Psi'(u_i)\Phi(u_{i+1}). \tag{68}$$

Both terms are nonpositive, and their combined magnitude is at most $1 + 2 \cdot 4 = 9$. Differentiating once more gives

$$\partial^2_{ii} Z_K(u) = -\Psi(u_{i-1})\Phi''(u_i) - \mathbf{1}_{\{i<K\}}\Psi''(u_i)\Phi(u_{i+1}),$$
$$\partial^2_{i,i+1} Z_K(u) = -\Psi'(u_i)\Phi'(u_{i+1}) \qquad (i < K).$$

All other entries outside the three central diagonals are zero. Thus each diagonal entry has magnitude at most $1 + 6 \cdot 4 = 25$, and each off-diagonal entry has magnitude at most 2. Every absolute row sum is at most $25 + 2 + 2 = 29$. To pass from row sums to the operator norm, let $Hv = \lambda v$ be an eigenpair of the symmetric Hessian, and choose i with $|v_i| = \max_j |v_j| > 0$. Then

$$|\lambda|\,|v_i| \le \sum_j |H_{ij}|\,|v_j| \le \left(\sum_j |H_{ij}|\right)|v_i| \le 29|v_i|.$$

Dividing by $|v_i|$ bounds every eigenvalue magnitude by 29. For a symmetric matrix, the operator norm is the largest eigenvalue magnitude, so $\|\nabla^2 Z_K(u)\|_{\mathrm{op}} \le 29 < 32$.

If $u_{k+1} = \cdots = u_K = 0$ and $i > k+1$, then $u_{i-1} = u_i = 0$. Both terms in (68) vanish because $\Psi(0) = \Psi'(0) = 0$, proving the support property. Finally, suppose $u \ge 0$ and $u_K = 0$, and take the first i such that $u_i < 1$. Such an index exists, and $u_{i-1} \ge 1$ when $i > 1$; for $i = 1$ we have $u_0 = 1$ by definition. Thus $\Psi(u_{i-1}) = 1$ and $\Phi'(u_i) \ge 1/2$. The other term in (68) is nonpositive, so $\partial_i Z_K(u) \le -1/2$. □

Choose $\nu > 0$ so that the following definitions satisfy $K \ge 1$ and $\varpi < 1$:

$$K = \left\lfloor \frac{1}{128\nu^2} \right\rfloor, \qquad \alpha = 32\nu^2, \qquad \beta = 32\nu, \qquad \varpi = 81\nu^2, \qquad C_\nu(y) = \alpha\big[Z_K(y/\beta) + KP\big]. \tag{69}$$

The scaling factors satisfy $\alpha/\beta = \nu$ and $\alpha/\beta^2 = 1/32$, so the chain rule gives

$$\partial_i C_\nu(y) = \nu\,\partial_i Z_K(y/\beta), \qquad \nabla^2 C_\nu(y) = \frac{1}{32}\nabla^2 Z_K(y/\beta).$$

Moreover, $K \le 1/(128\nu^2)$ implies

$$C_\nu(0) = 32\nu^2(KP - 2) \le 32\nu^2 KP \le P/4 < 1.$$

The value and derivative bounds in Lemma F.1 therefore yield

$$\inf C_\nu = 0, \quad C_\nu \ge 0, \quad C_\nu(0) \le 1, \quad \|\nabla^2 C_\nu\|_{\mathrm{op}} \le 1, \quad -9\nu \le \partial_i C_\nu \le 0. \tag{70}$$

If $y \ge 0$ and $y_K = 0$, some coordinate of $\nabla C_\nu(y)$ has magnitude at least $\nu/2$.

Stochastic oracle. For any query y, let $k(y) = \max(\{i : y_i \ne 0\} \cup \{0\})$. When $k < K$, set $m = k + 1$ and draw $Z \sim \mathrm{Bernoulli}(\varpi)$. Return the exact chain gradient except for its mth coordinate:

$$g_i^C(y; Z) = \begin{cases} (Z/\varpi)\partial_m C_\nu(y), & i = m, \\ \partial_i C_\nu(y), & i \ne m. \end{cases} \tag{71}$$

If $k = K$, return the exact gradient. For the additional coordinates introduced below, the oracle also returns the exact gradient. This Borel-measurable probability kernel is determined by the query point and fixed before optimization begins. It need not be the gradient of a smooth sample loss.

Lemma F.2 (Variance and coordinate activation)**.** *The oracle is unbiased and its conditional variance is at most one. For the rank-one specialization $A_t = a_t \in \mathbb{R}$ of* (5)*, initialize all chain coordinates of B_0 at zero. Suppose $a_t > 0$ throughout the horizon and let k_t be the last nonzero chain-coordinate index. For any choice of finite scalar stepsizes,*

$$k_t \le \sum_{s<t} Z_s. \tag{72}$$

In particular,

$$\mathbb{P}(k_t < K \text{ for all } 0 \le t \le T) \ge 1 - \frac{\varpi T}{K}. \tag{73}$$

The constructions below satisfy $a_t > 0$, so the factor and the product have the same nonzero chain coordinates.

Proof. Fix the current query point and first suppose $k < K$. Since $\mathbb{E}Z = \varpi$, the masked coordinate satisfies

$$\mathbb{E}[g_m^C \mid y] = \frac{\mathbb{E}Z}{\varpi}\partial_m C_\nu(y) = \partial_m C_\nu(y).$$

Every other coordinate is exact, so the oracle is unbiased. Only the masked coordinate contributes to the error, and

$$\begin{aligned}\mathbb{E}\|g^C - \nabla C_\nu(y)\|_F^2 &= \mathbb{E}(Z/\varpi - 1)^2 \,|\partial_m C_\nu(y)|^2 \\ &= \left[\varpi(\varpi^{-1} - 1)^2 + (1 - \varpi)\right] |\partial_m C_\nu(y)|^2 \\ &= (\varpi^{-1} - 1)|\partial_m C_\nu(y)|^2 \le \frac{81\nu^2}{\varpi} = 1.\end{aligned}$$

These are pointwise bounds, so they also hold conditionally on the history at a random query. When $k = K$, the oracle is exact and has zero error variance.

Since $a_t > 0$, the nonzero chain-coordinate indices of the product $a_t B_t$ agree with those of B_t. At a state with last nonzero index k_t, the support property of the chain makes all sampled gradient coordinates beyond $k_t + 1$ zero. If $Z_t = 0$, coordinate $k_t + 1$ also has zero sampled gradient. The coordinatewise update $B_{t+1,i} = B_{t,i} - \eta_t a_t g_{t,i}^C$ then keeps every coordinate $i > k_t$ zero. If $Z_t = 1$, only coordinate $k_t + 1$ can be newly activated. A simultaneous change in the scalar factor a_t cannot make a zero coordinate of B_{t+1} nonzero. Therefore, for each realization,

$$k_{t+1} \le k_t + Z_t, \qquad k_t \le k_0 + \sum_{s<t} Z_s = \sum_{s<t} Z_s.$$

The argument does not restrict how η_t depends on Z_t. When $k_t = K$, continue to draw independent Bernoulli variables without using them in the oracle, so the same recursion remains valid.

Set $N_T = \sum_{s<T} Z_s$. If $N_T < K$, then for every $t \le T$, $k_t \le \sum_{s<t} Z_s \le N_T < K$. Hence

$$\mathbb{P}(k_t < K \text{ for all } 0 \le t \le T) \ge \mathbb{P}(N_T < K) \ge 1 - \frac{\mathbb{E}N_T}{K} = 1 - \frac{\varpi T}{K},$$

where the second inequality is Markov's inequality and $\mathbb{E}N_T = \varpi T$ because the Bernoulli draws are fresh. This proves (72) and (73). □

F.3 Factor amplification

Let $0 < \epsilon \le 1/64$, put $\delta = 2\epsilon$, and use

$$\nu = 8\delta^2 = 32\epsilon^2 \tag{74}$$

in (69). Then $K \ge 128$ and $0 < \varpi < 1$. In addition to the chain coordinates, introduce a coordinate x with $L_\delta = (\delta^{-2} - 1)/2$ and $\chi(v) = 1 - 3v^2 + 2v^3$ on $[0, 1]$. Define

$$h_\delta(x) = \begin{cases} \delta(2 - e^x), & x \le 0, \\ \delta(1 + 2x)^{-1/2}, & 0 \le x \le L_\delta, \\ \delta(1 + 2x)^{-1/2}\chi(x - L_\delta), & L_\delta < x < L_\delta + 1, \\ 0, & x \ge L_\delta + 1, \end{cases} \qquad f_\delta(x) = \int_x^\infty h_\delta(z)\, dz. \tag{75}$$

The scalar component bounds the factor-gradient norm from below before the factor grows sufficiently large. Afterward, the amplified chain gradient provides the required lower bound.

Lemma F.3 (Smoothness and the initial objective gap)**.** *The function f_δ is nonnegative, belongs to C^2, satisfies $|f_\delta''| \le 1$, and has infimum zero and $f_\delta(0) \le 1$. For $x \ge 0$, $0 \le h_\delta(x) \le \delta$. Let*

$$D = 8 - f_\delta(0) - C_\nu(0), \qquad P_D(z) = \frac{D}{2}\left(1 + \cos\frac{z}{\sqrt{D}}\right). \tag{76}$$

Then $6 \le D \le 8$, and the objective on $\mathbb{R}^{K+2}$

$$F(x, y, z) = f_\delta(x) + C_\nu(y) + P_D(z) \tag{77}$$

has $F \in C^2$, $\|\nabla^2 F\|_{\mathrm{op}} \le 1$, $\inf F = 0$, and $F(0) = 8$.

Proof. At zero, the left and right formulas for h_δ both have value δ and derivative $-\delta$. On $[0, L_\delta]$, put $r(x) = \delta(1 + 2x)^{-1/2}$. Then

$$r'(x) = -\delta(1 + 2x)^{-3/2}, \quad r(L_\delta) = \delta^2, \quad r'(L_\delta) = -\delta^4.$$

On the cutoff interval, $h_\delta(x) = r(x)\chi(x - L_\delta)$. The identities $\chi(0) = 1$, $\chi(1) = 0$, and $\chi'(0) = \chi'(1) = 0$ show that both values and derivatives match at L_δ and $L_\delta + 1$. Since $0 \le \chi \le 1$ and $|\chi'| \le 3/2$,

$$|h_\delta'(x)| \le |r'(x)|\,|\chi(x - L_\delta)| + r(x)|\chi'(x - L_\delta)| \le \delta^4 + \frac{3}{2}\delta^2 < 1$$

on the cutoff interval. On $x \le 0$, the derivative has magnitude $\delta e^x \le \delta$; on $[0, L_\delta]$ it has magnitude at most δ. Thus $h_\delta \in C^1$ and $|h_\delta'| \le 1$.

The function h_δ is nonnegative, bounded on each finite interval, and zero to the right of $L_\delta + 1$. Therefore the integral defining $f_\delta(x)$ is finite for every real x, and $f_\delta' = -h_\delta$, $f_\delta'' = -h_\delta'$. It follows that $f_\delta \in C^2$, $f_\delta \ge 0$, and $|f_\delta''| \le 1$. The function equals zero for $x \ge L_\delta + 1$, so its infimum is zero. Also,

$$\begin{aligned} f_\delta(0) &= \int_0^{L_\delta} r(x)\,dx + \int_{L_\delta}^{L_\delta+1} r(x)\chi(x - L_\delta)\,dx \\ &\le \delta(\sqrt{1 + 2L_\delta} - 1) + \delta^2 = 1 - \delta + \delta^2 \le 1. \end{aligned}$$

The cutoff integral is at most δ^2 because its length is one, $r(x) \le \delta^2$ there, and $\chi \le 1$. The formulas also give $0 \le h_\delta(x) \le \delta$ for $x \ge 0$.

Since $0 \le f_\delta(0), C_\nu(0) \le 1$, the definition of D gives $6 \le D \le 8$. Differentiation of the auxiliary term yields

$$P_D'(z) = -\frac{\sqrt{D}}{2}\sin(z/\sqrt{D}), \qquad P_D''(z) = -\frac{1}{2}\cos(z/\sqrt{D}).$$

Thus $P_D \ge 0$, $P_D(0) = D$, $P_D'(0) = 0$, $\inf P_D = P_D(\pi\sqrt{D}) = 0$, and $|P_D''| \le 1/2$. The full Hessian is block diagonal:

$$\nabla^2 F(x, y, z) = \mathrm{diag}\big(f_\delta''(x), \nabla^2 C_\nu(y), P_D''(z)\big).$$

Its operator norm is the maximum of the three block norms, hence at most one. Each summand is C^2 and nonnegative. To verify $\inf F = 0$, fix $x = L_\delta + 1$ and $z = \pi\sqrt{D}$, and let all coordinates of y/β tend to $+\infty$. The first and third terms are then zero, while the chain term tends to zero by Lemma F.1. Finally, $F(0) = f_\delta(0) + C_\nu(0) + D = 8$. □

Use the masked oracle (71), returning the exact gradients for the scalar and auxiliary components. Set $r = n = 1$ and initialize $B_0 = 0 \in \mathbb{R}^{(K+2)\times 1}$ and $A_0 = 1$. Write $B = (b, c, z_B)$ and $A = a$, so $x = ab$, $y = ac$, and $z = az_B$; in particular, $b_0 = 0$, $c_0 = 0$, and $z_{B,0} = 0$. If $z_{B,t} = 0$, then the product coordinate is $az_{B,t} = 0$ and its sampled weight gradient is $P_D'(0) = 0$. Hence $z_{B,t+1} = z_{B,t} - \eta_t a_t P_D'(0) = 0$, and this coordinate also contributes $z_{B,t}P_D'(0) = 0$ to the update of a_t. Induction from $z_{B,0} = 0$ therefore keeps the auxiliary coordinate inactive. All sampled chain-gradient components are nonpositive. Set $v_t = -g_t^C \ge 0$. The factor updates become

$$b^+ = b + \eta a h_\delta(ab), \qquad c^+ = c + \eta a v, \qquad a^+ = a + \eta\big(b h_\delta(ab) + c^\top v\big). \tag{78}$$

Thus $a_t \ge 1$, $b_t, c_t \ge 0$ for every finite nonnegative stepsize, including a stepsize selected after observing the current sample.

Lemma F.4 (Factor amplification and nonstationarity)**.** *Suppose the updates before time t satisfy $\eta_s h_\delta(x_s) \le 1$. Then $a_t \ge b_t$. If also $k_t < K$, the true factor gradient satisfies $\|\nabla J(V_t)\|_F > \epsilon$.*

Proof. Subtracting the first update in (78) from the third gives

$$a^+ - b^+ = (1 - \eta h_\delta(ab))(a - b) + \eta c^\top v \ge 0 \tag{79}$$

whenever $a \ge b$ and $\eta h_\delta(ab) \le 1$. The inequality uses $\eta \ge 0$, $c, v \ge 0$, and the nonnegative coefficient $1 - \eta h_\delta(ab)$. Since $a_0 - b_0 = 1$, induction over the updates before time t gives $a_t \ge b_t$.

First suppose $0 \le x_t \le L_\delta$. The true factor derivatives in the scalar coordinates are

$$\partial_b J = -a_t h_\delta(x_t), \qquad \partial_a J = -b_t h_\delta(x_t) + c_t^\top \nabla C_\nu(y_t).$$

Because $c_t \ge 0$ and $\nabla C_\nu(y_t) \le 0$ coordinatewise, $\partial_a J \le -b_t h_\delta(x_t) \le 0$. There is therefore no cancellation between these terms, and $|\partial_a J| \ge b_t h_\delta(x_t)$. Together with $|\partial_b J| = a_t h_\delta(x_t)$, this gives

$$\|\nabla J(V_t)\|_F^2 \ge \delta^2 \frac{a_t^2 + b_t^2}{1 + 2x_t} \ge \frac{\delta^2}{2} > \epsilon^2. \tag{80}$$

For the middle inequality, $a_t \ge 1$ gives $a_t^2 + b_t^2 \ge 1$, while $(a_t - b_t)^2 \ge 0$ gives $a_t^2 + b_t^2 \ge 2x_t$. Averaging these two lower bounds yields $a_t^2 + b_t^2 \ge (1 + 2x_t)/2$. The last inequality uses $\delta = 2\epsilon$.

Now suppose $x_t \ge L_\delta$. Since $a_t \ge b_t \ge 0$, $a_t^2 \ge a_t b_t = x_t$. Also $\delta^2 \le 1/2$, so

$$L_\delta = \frac{1 - \delta^2}{2\delta^2} \ge \frac{1}{4\delta^2}, \qquad a_t \ge \sqrt{x_t} \ge \sqrt{L_\delta} \ge \frac{1}{2\delta}.$$

The condition $k_t < K$ means the last chain coordinate is zero, and $y_t = a_t c_t \ge 0$. The scaled chain property therefore gives $\|\nabla C_\nu(y_t)\|_\infty \ge \nu/2$. Since the derivative with respect to the chain factor is $\nabla_c J = a_t \nabla C_\nu(y_t)$, we have

$$\|\nabla J(V_t)\|_F \ge a_t \|\nabla C_\nu(y_t)\|_\infty \ge \frac{\nu}{4\delta} = 2\delta > \epsilon. \tag{81}$$

Here the first inequality uses that the full gradient norm dominates any coordinate magnitude, and the equality uses $\nu = 8\delta^2$. The two cases cover every nonnegative x_t. □

F.4 Proof of the main SGD lower bound

We prove the lower bound under the larger stepsize cap $1/(2\epsilon)$, which includes the policy class in the main theorem.

Theorem F.5 (Lower bound for bounded adaptive stepsizes)**.** *For each $0 < \epsilon \le 1/64$, the construction above has dimension $K + 2 = O(\epsilon^{-4})$, initial gap $F(0) - \inf F = 8$, and initial squared factor norm $R_0 = 1$. The objective has Hessian operator norm at most one, and the oracle has conditional variance at most one. Starting from the rank-one initialization $B_0 = 0$, $A_0 = 1$, consider any policy satisfying*

$$0 \le \eta_t \le \frac{1}{2\epsilon}, \tag{82}$$

with arbitrary dependence on the iteration index, optimization history, current sample, and independent randomness. For every integer $T \ge 0$, the iterates satisfy

$$\mathbb{P}\big(\|\nabla J(V_t)\|_F > \epsilon \text{ for all } 0 \le t \le T\big) \ge 1 - \frac{\varpi T}{K}. \tag{83}$$

In particular, this probability is at least $3/4$ when $T \le K/(4\varpi)$. Since $K/(4\varpi) = \Omega(\epsilon^{-8})$, this proves the lower bound, including for $0 \le \eta_t \le 1$.

Proof. The nonnegative recursions give $x_t \ge 0$ for every realization. Since $h_\delta(x_t) \le \delta$ and $\eta_t \le 1/(2\epsilon) = 1/\delta$,

$$0 \le \eta_t h_\delta(x_t) \le \eta_t \delta \le 1.$$

Thus the order invariant holds at every iteration. On the event $k_t < K$ for all $0 \le t \le T$, Lemma F.4 applies at each such t and gives $\|\nabla J(V_t)\|_F > \epsilon$. This is the event inclusion

$$\{k_t < K \text{ for all } t \le T\} \subseteq \{\|\nabla J(V_t)\|_F > \epsilon \text{ for all } t \le T\}.$$

Taking probabilities and applying (73) proves (83). In particular, $T \le K/(4\varpi)$ implies $1 - \varpi T/K \ge 3/4$. The factor inequalities hold for every realization of the oracle samples; only the Bernoulli count requires a probability estimate. Thus the policy may change its stepsize or internal state after a zero response without affecting the argument.

For the iteration lower bound, $1/(128\nu^2) \ge 2$ and $\lfloor u \rfloor \ge u/2$ for $u \ge 2$ imply $K \ge 1/(256\nu^2)$. Substituting $\varpi = 81\nu^2$ gives

$$\frac{K}{4\varpi} \ge \frac{1/(256\nu^2)}{4 \cdot 81\nu^2} = \frac{1}{82944\nu^4} = \frac{1}{82944 \cdot 32^4 \epsilon^8}. \tag{84}$$

The last equality uses $\nu = 32\epsilon^2$. Also $K \le 1/(128\nu^2) = O(\epsilon^{-4})$, so the dimension has the stated order. The smoothness, gap, and oracle bounds follow from Lemmas F.3 and F.2, and $R_0 = \|B_0\|_F^2 + |A_0|^2 = 1$. On the event in (83), every generated iterate through T is nonstationary, so no output chosen from those iterates can be ϵ-stationary. □

Proof of Theorem 5.3. Take $\epsilon_0 = 1/64$ and use Theorem F.5. For every $0 < \epsilon \le \epsilon_0$, the class $0 \le \eta_t \le 1$ is contained in the larger class $0 \le \eta_t \le 1/(2\epsilon)$. The construction has rank one, dimension $K + 2 = O(\epsilon^{-4})$, initial gap 8, global smoothness at most one, and conditional oracle variance at most one. Moreover, (84) allows the universal choice $c = (82944 \cdot 32^4)^{-1}$. Then $T \le c\epsilon^{-8}$ implies $T \le K/(4\varpi)$, and hence

$$\mathbb{P}(\|\nabla J(V_t)\|_F > \epsilon \text{ for all } 0 \le t \le T) \ge \frac{3}{4}.$$

The theorem allows the stepsize policy to use the full first-order history, the current oracle response, and independent internal randomness. This proves Theorem 5.3. □

G Experimental Details

This section provides the experimental settings and measurement procedures for Figs. 1–4.

Shared training settings. Only the LoRA factors are optimized in all LoRA experiments; all other model parameters are frozen. For the CIFAR-10 logistic-regression experiments on frozen ResNet-18 features, the base classifier W_0 and its bias are also frozen. We use gradient descent or SGD without momentum, weight decay, or gradient clipping. Tuned hyperparameters are selected before the reported multi-seed evaluations and then kept fixed across all seeds. Real-model experiments are run on NVIDIA H200 GPUs with 141 GB of HBM3e memory.

G.1 Hessian–gradient measurements

Models and data. For Fig. 1, we initialize DistilBERT (Sanh et al., 2019) from `distilbert/distilbert-base-uncased-finetuned-sst-2-english` and fine-tune it on IMDb (Maas et al., 2011), and initialize ViT-B/16 (Dosovitskiy et al., 2021) from `google/vit-base-patch16-224-in21k` and fine-tune it on CIFAR-100 (Krizhevsky, 2009). Each setting uses 20,000 training examples and 2,000 validation examples. DistilBERT inputs are truncated to 256 tokens, while ViT inputs are resized to 224×224 without data augmentation. The pretrained DistilBERT classification head remains frozen. For ViT, we first train only the classification head for one epoch using Adam with learning rate 10^{-3} and batch size 32, and then freeze it before LoRA training.

We apply rank-8 LoRA to the query and value projections in every transformer block. We initialize $B_0 = 0$ and rescale each adapted A_0 to have Frobenius norm one. DistilBERT has 12 adapted matrices and ViT has 24. Only the LoRA factors are updated, using plain SGD with batch size 32. The learning rates are 0.1 for DistilBERT and 0.3 for ViT. Each trajectory contains 1,875 optimizer updates, and we use three seeds per setting. The real-model runs use float32; TF32 is disabled for both training and Hessian diagnostics.

Gradient and Hessian measurements. We evaluate the gradient and Hessian at 33 pre-specified checkpoints:

$$\{0, 1, 2, 3, 5, 6, 9, 12, 16, 22, 30, 40, 55, 74, 101, 137, 187, 281, 374, 468, 562, 656, 750, 843, 937, 1054, 1171, 1288, 1406, 1523, 1640, 1757, 1874\}.$$

For each model–dataset setting, we fix one diagnostic batch of 128 examples and reuse it for all three seeds at every checkpoint. Both $\|\nabla_X F\|_F$ and $\|\nabla_X^2 F\|_{\mathrm{op}}$ are measured on this batch, so changes across checkpoints reflect the optimization trajectory rather than changes in the sampled examples. The diagnostic loss is averaged over these 128 examples. To take derivatives with respect to the LoRA weight update X, rather than the factors (A, B), we replace BA in each adapted matrix by an independent dense variable X during the diagnostic forward pass. We compute $\|\nabla_X F\|_F$ directly and estimate $\|\nabla_X^2 F\|_{\mathrm{op}}$ jointly over all adapted matrices using Hessian–vector products and 15-step Lanczos with full re-orthogonalization. We use the largest absolute Ritz value as the operator-norm estimate, thereby approximating $\max_i |\lambda_i|$ for the symmetric Hessian rather than only its largest positive eigenvalue. We do not form the dense Hessian explicitly. All 99 checkpoints in each setting are retained in the analysis.

Displayed trajectories. For each setting, Fig. 1 shows the seed whose within-run Spearman correlation between the gradient and Hessian norms is the median of the three values reported in Table 3. This selects seed 0 for DistilBERT and seed 1 for ViT. The choice is made solely by this criterion and is independent of the fitting results below.

Constant and gradient-dependent fits. At diagnostic checkpoint i, let

$$g_i := \|\nabla_X F(X_i)\|_F, \qquad \widehat{h}_i := \|\widehat{\nabla_X^2 F(X_i)}\|_{\mathrm{op}},$$

where $\widehat{h}_i$ denotes the Lanczos estimate described above. We compare a constant fit with a gradient-dependent fit of the form $\widehat{L}_{0,s} + \widehat{L}_{1,s} g_i$, with nonnegative fitted coefficients. For each evaluation seed s, both fits use the 66 checkpoints from the other two seeds, denoted by $\mathcal{I}_{-s}$, and are evaluated on the 33 checkpoints of seed s, denoted by $\mathcal{I}_s$. The evaluation trajectory is not used to choose any fitting parameter. Both fits target 95% empirical coverage on the fitting trajectories, meaning that at least 95% of the measured Hessian norms are no larger than the fitted values. This percentage is a fitting criterion, not a probabilistic confidence level.

For values $z_1, \ldots, z_n$ sorted as $z_{(1)} \le \cdots \le z_{(n)}$, we use the conservative empirical quantile

$$Q_{0.95}(z_1, \ldots, z_n) := z_{(1+\lceil 0.95(n-1) \rceil)}.$$

The constant fit is

$$\widehat{L}_s = Q_{0.95}\Big(\{\widehat{h}_i : i \in \mathcal{I}_{-s}\}\Big).$$

For a candidate slope $b \ge 0$, define

$$\widehat{L}_{0,s}(b) = \max\Big\{0,\ Q_{0.95}\Big(\{\widehat{h}_i - b g_i : i \in \mathcal{I}_{-s}\}\Big)\Big\}.$$

The candidate slopes are 0 and 40 logarithmically spaced values between $10^{-6} b_{\max,s}$ and $10 b_{\max,s}$, where

$$b_{\max,s} = \frac{\max_{i \in \mathcal{I}_{-s}} \widehat{h}_i}{\min_{i \in \mathcal{I}_{-s}} g_i}.$$

All measured g_i are positive. Among the candidate slopes attaining at least 95% coverage on the fitting points, we choose $\widehat{L}_{1,s}$ to minimize

$$\frac{1}{66} \sum_{i \in \mathcal{I}_{-s}} \Big(\widehat{L}_{0,s}(b) + b g_i\Big),$$

and set $\widehat{L}_{0,s} := \widehat{L}_{0,s}(\widehat{L}_{1,s})$. Ties are resolved by choosing the smaller slope and then the smaller intercept. The slope grid is determined solely from the two fitting seeds, so the evaluation seed does not affect either the candidate slopes or the fitted parameters. Table 1 reports the resulting parameters.

Table 1: **Parameters of the constant and gradient-dependent fits.** Seed s is used only for evaluation; both fits use the other two seeds. The displayed main-text trajectories use DistilBERT seed 0 and ViT seed 1.

Setting	Seed s	$\widehat{L}_s$	$\widehat{L}_{0,s}$	$\widehat{L}_{1,s}$
DistilBERT / IMDb	0	164.02	24.49	45.79
	1	164.02	24.49	45.79
	2	164.00	34.70	42.45
ViT-B/16 / CIFAR-100	0	9407.55	1885.00	3964.64
	1	15617.51	2029.84	7465.97
	2	16549.86	2820.04	7465.97

Evaluation measures. For evaluation seed s, define

$$U_{\mathrm{const},i} := \widehat{L}_s, \qquad U_{\mathrm{grad},i} := \widehat{L}_{0,s} + \widehat{L}_{1,s} g_i.$$

We report the fraction of evaluation points satisfying each fit, its mean fitted value, and the mean amount by which the measurements exceed it. To make the latter two quantities comparable in scale, we normalize both by

$$h_{\mathrm{ref},s} := \mathrm{median}_{i \in \mathcal{I}_{-s}} \widehat{h}_i,$$

which is computed only from the fitting seeds. For $m \in \{\mathrm{const}, \mathrm{grad}\}$, define

$$\mathrm{Cov}_{m,s} = \frac{1}{33} \sum_{i \in \mathcal{I}_s} \mathbf{1}\{\widehat{h}_i \le U_{m,i}\},$$

$$\mathrm{Size}_{m,s} = \frac{1}{33 h_{\mathrm{ref},s}} \sum_{i \in \mathcal{I}_s} U_{m,i},$$

$$\mathrm{Exc}_{m,s} = \frac{1}{33 h_{\mathrm{ref},s}} \sum_{i \in \mathcal{I}_s} (\widehat{h}_i - U_{m,i})_+, \qquad (z)_+ := \max\{z, 0\}.$$

Here $\mathrm{Cov}_{m,s}$ is empirical coverage, $\mathrm{Size}_{m,s}$ is the normalized mean fitted value, and $\mathrm{Exc}_{m,s}$ is the normalized mean exceedance. The same $h_{\mathrm{ref},s}$ is used for both fits in each evaluation. Smaller $\mathrm{Size}_{m,s}$ indicates a lower fitted value, but is interpreted together with coverage and exceedance rather than as an improvement on its own. Each checkpoint receives equal weight; the measurements are not uniformly spaced in training time.

Table 2: **Evaluation on the seed excluded from fitting.** Coverage is the percentage of evaluation checkpoints satisfying the fit, Size is the normalized mean fitted value, and Exceedance is the normalized mean amount above the fit. Mean rows average the three evaluations equally.

		Constant fit			Gradient-dependent fit		
Setting	Seed	Cov. ↑	Size ↓	Exc. ↓	Cov. ↑	Size ↓	Exc. ↓
DistilBERT / IMDb	0	100.0	3.458	0.0000	97.0	2.189	0.0001
	1	97.0	3.618	0.0003	97.0	2.447	0.0004
	2	90.9	3.444	0.0012	93.9	2.305	0.0022
	Mean	96.0	3.507	0.0005	96.0	2.314	0.0009
ViT-B/16 / CIFAR-100	0	90.9	9.279	1.1649	90.9	7.042	1.1966
	1	97.0	13.507	0.2517	97.0	10.414	0.3939
	2	100.0	20.512	0.0000	100.0	16.654	0.0000
	Mean	96.0	14.433	0.4722	96.0	11.370	0.5301

Results across seeds. Both fits attain the same mean evaluation coverage of 96.0% in each setting (Table 2). Define

$$\overline{\mathrm{Size}}_m := \frac{1}{3} \sum_{s=0}^{2} \mathrm{Size}_{m,s}.$$

The ratios $\overline{\text{Size}}_{\text{grad}}/\overline{\text{Size}}_{\text{const}}$ are 0.660 for DistilBERT/IMDb and 0.788 for ViT-B/16/CIFAR-100. Thus, the normalized mean fitted value is approximately 34% and 21% smaller, respectively. These percentages use the ratio of the mean normalized values, not the mean of the three per-seed ratios.

For DistilBERT, the fitted parameters vary little across seeds and the normalized mean exceedance is below 0.003 in every evaluation. The ViT fits vary more: $\widehat{L}_{1,s}$ ranges from approximately 3965 to 7466, and the normalized mean exceedance rises from 0.472 for the constant fit to 0.530 for the gradient-dependent fit after averaging over seeds. The smaller average fitted value on ViT therefore comes with larger exceedance, even though average coverage is unchanged. For the two trajectories displayed in Fig. 1, the gradient-dependent mean fitted values are 0.633 and 0.771 times the corresponding constant values, respectively.

The affine form $L_0 + L_1 u$ is sub-quadratic for finite L_0, L_1, since $(L_0 + L_1 u)/u^2 \to 0$ as $u \to \infty$. The comparison evaluates this form on measured trajectories.

Correlation across seeds. Figure 5 shows the scatter plots for all seeds, and Table 3 reports their Spearman correlations. These correlations describe the monotone association between the two measured norms and complement the constant/gradient-dependent comparison.

Table 3: **Across-seed Hessian–gradient correlations.** Here ρ_s denotes the within-run Spearman correlation between $\|\nabla_X F\|_F$ and $\|\nabla_X^2 F\|_{\text{op}}$ for seed s.

Setting	ρ_0	ρ_1	ρ_2
DistilBERT / IMDb	0.726	0.776	0.715
ViT-B/16 / CIFAR-100	0.533	0.606	0.728

G.2 CIFAR-10 FIXED-FEATURE LoRA-GD

Fixed-feature objective. For Fig. 2, we train a 10-class multinomial logistic-regression classifier on fixed CIFAR-10 (Krizhevsky, 2009) features extracted once with an ImageNet (Deng et al., 2009)-pretrained ResNet-18 (He et al., 2016) (checkpoint `microsoft/resnet-18`). We split the CIFAR-10 training set into 45,000 training examples and 5,000 validation examples. Images are deterministically resized to 224×224 and normalized with the ImageNet mean and standard deviation; no data augmentation is used. The resulting feature vectors are 512-dimensional and are kept fixed throughout optimization. All computations in this fixed-feature experiment use float64.

LoRA-GD and stepsizes. We parameterize the classifier weight as $W = W_0 + BA$ with rank 4. The base classifier is a PyTorch linear layer with 512 input features and 10 outputs. Its weight W_0 and bias are initialized independently from $\mathcal{U}(-1/\sqrt{512},\, 1/\sqrt{512})$ and then frozen. We initialize the entries of A_0 from the same uniform distribution and set $B_0 = 0$. Because the classifier is linear on the cached features, every LoRA-GD update uses the exact full gradient over all 45,000 training examples. Each run lasts $T = 2000$ updates. We compare

$$\eta_t = \frac{0.5}{\sqrt{t+1}}, \qquad \eta_t = \frac{1}{\sqrt{T}} \quad (T = 2000), \qquad \eta \in \{0.01, 0.05, 0.07\}.$$

For $T = 2000$, the fixed-horizon schedule uses the constant stepsize $1/\sqrt{2000} \approx 0.0224$. Within each seed, all five stepsize choices use the same classifier and LoRA initialization. Figure 2 reports the pointwise mean over five paired runs, with shaded regions showing one sample standard deviation. For these schedules, the theory provides the initial stepsize scale, and we tune the multiplicative coefficient only in a neighborhood of this value using separate tuning runs. This procedure selects $\eta_t = 1/\sqrt{T}$ and $\eta_t = 0.5/\sqrt{t+1}$.

G.3 STOCHASTIC-GRADIENT NOISE MEASUREMENTS

LoRA-SGD training. For Fig. 3, we train the same rank-4 LoRA logistic-regression classifier on the frozen CIFAR-10 ResNet-18 features used in Fig. 2, but replace exact full-gradient updates with minibatch SGD. The training batch size is 512, and each run lasts 60 epochs, or 5,280 updates.

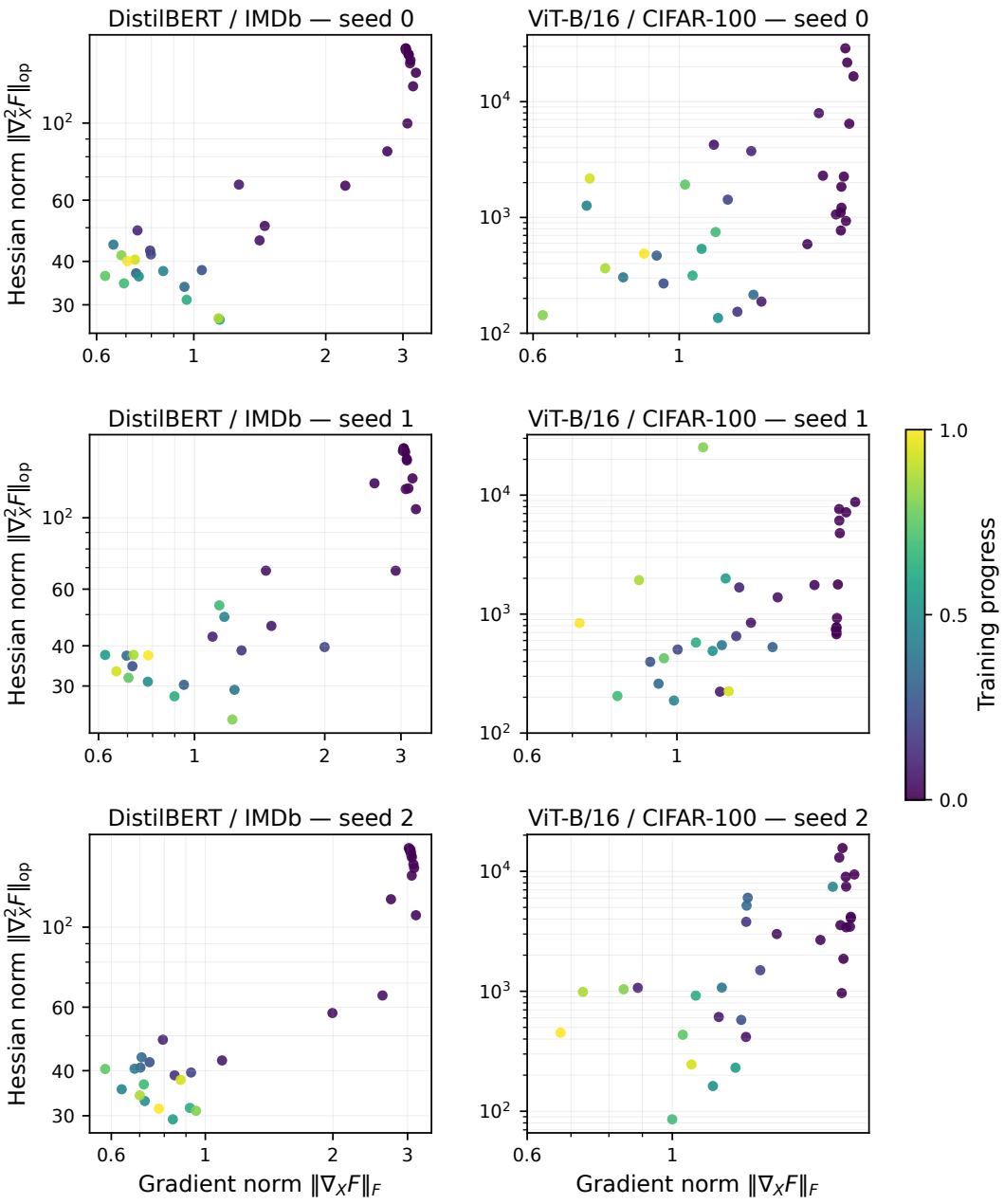

Figure 5: **Gradient and Hessian norms across all seeds.** Each row corresponds to one seed for DistilBERT on IMDb (left) and ViT-B/16 on CIFAR-100 (right). We plot the estimated Hessian norm $\|\nabla_X^2 F\|_{\text{op}}$ against the weight-gradient norm $\|\nabla_X F\|_F$. Color indicates training progress, normalized from 0 (start) to 1 (end).

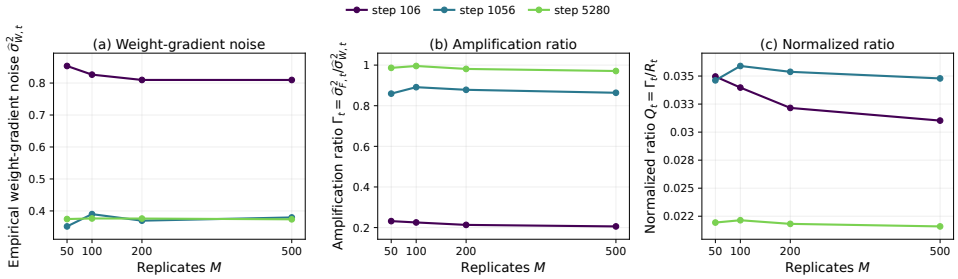

Figure 6: **Noise measurements versus the number of sampled minibatches.** For seed 0 and batch size $b = 512$, we vary M at optimizer steps 106, 1056, and 5280. We report the empirical weight-gradient noise, the amplification ratio $\Gamma_t = \widehat{\sigma}^2_{F,t}/\widehat{\sigma}^2_{W,t}$, and $Q_t = \Gamma_t/R_t$.

To choose the learning rate, we run $\eta \in \{0.03, 0.05, 0.07\}$ for 500 updates on three tuning seeds. None of these runs diverges, and $\eta = 0.05$ gives the lowest mean validation loss. We therefore use $\eta = 0.05$ for the five seeds shown in Fig. 3.

Noise measurement. We measure the gradient noise at optimizer steps

$$\{0, 106, 264, 528, 1056, 2112, 3168, 4224, 5280\},$$

which correspond to training progress $\{0, 0.02, 0.05, 0.10, 0.20, 0.40, 0.60, 0.80, 1.00\}$. At each step, we keep A_t and B_t fixed and draw $M = 200$ minibatches, with replacement, from the 45,000 training examples for each $b \in \{64, 256, 512, 2048\}$. Let $\mathbf{G}_t := \nabla F(X_t)$ be the exact gradient on all 45,000 training examples and $\widehat{\mathbf{G}}_t^{(k)}$ the gradient on the k-th sampled minibatch. With $\mathbf{E}_t^{(k)} := \widehat{\mathbf{G}}_t^{(k)} - \mathbf{G}_t$, we compute

$$\widehat{\sigma}^2_{W,t} = \frac{1}{M}\sum_{k=1}^{M} \|\mathbf{E}_t^{(k)}\|_F^2, \qquad \widehat{\sigma}^2_{F,t} = \frac{1}{M}\sum_{k=1}^{M} \left(\|\mathbf{E}_t^{(k)} A_t^\top\|_F^2 + \|B_t^\top \mathbf{E}_t^{(k)}\|_F^2\right).$$

The first quantity measures the minibatch error in the weight gradient. The second measures the error induced in the two LoRA-factor gradients by the same minibatch error.

Across-seed averaging and M sensitivity. For this sensitivity analysis, we write

$$\Gamma_t := \frac{\widehat{\sigma}^2_{F,t}}{\widehat{\sigma}^2_{W,t}}, \qquad Q_t := \frac{\Gamma_t}{R_t}.$$

In Fig. 3, panel (a) shows the mean R_t over five seeds at each checkpoint. Panels (b)–(d) show the mean over the same five seeds separately for each batch size b. To choose M, we repeat the $b = 512$ calculation on seed 0 with $M \in \{50, 100, 200, 500\}$ at steps 106, 1056, and 5280. Across these three checkpoints, the largest relative difference between the $M = 200$ and $M = 500$ values among the three quantities in Fig. 6 is 3.7%. We therefore use $M = 200$ in Fig. 3.

G.4 ViT fine-tuning

Model and LoRA setup. For Fig. 4, we fine-tune `google/vit-base-patch16-224-in21k` on CIFAR-100. Before LoRA training, we train only the classification head for one epoch using Adam with learning rate 10^{-3} and batch size 32, and then freeze the head. We apply rank-32 LoRA to the query, key, value, and output projections in all 12 transformer blocks, giving 48 adapted matrices. We set $B_0 = 0$ and initialize the entries of A_0 independently as $A_{0,ij} \sim \mathcal{N}(0, (10^{-3})^2)$. Since each adapted input dimension is 768,

$$\mathbb{E}[R_0] = 48 \times 32 \times 768 \times 10^{-6} \approx 1.1796,$$

which matches the initial factor scale in Fig. 4.

Training and reporting. We use SGD with learning rate 0.3 and batch size 16 for three epochs (3,750 updates), with three seeds. We select 256 examples once for each run and reuse the same examples at every measurement checkpoint for the reported gradient-noise calculations. Figure 4 shows the pointwise mean over the three runs; shaded regions show one sample standard deviation.

Gradient-noise measurement. Computing the exact full-dataset gradient is prohibitively expensive for this model. At each frozen checkpoint, we therefore evaluate $M = 200$ minibatch gradients with batch size $b = 16$ and use their empirical mean

$$\overline{\mathbf{G}}_t = \frac{1}{M}\sum_{k=1}^{M} \widehat{\mathbf{G}}_t^{(k)}$$

as the reference gradient. We then compute

$$\widehat{\sigma}_{W,t}^2 = \frac{1}{M}\sum_{k=1}^{M} \|\widehat{\mathbf{G}}_t^{(k)} - \overline{\mathbf{G}}_t\|_F^2,$$

$$\widehat{\sigma}_{F,t}^2 = \frac{1}{M}\sum_{k=1}^{M}\Big(\|(\widehat{\mathbf{G}}_t^{(k)} - \overline{\mathbf{G}}_t)A_t^\top\|_F^2 + \|B_t^\top(\widehat{\mathbf{G}}_t^{(k)} - \overline{\mathbf{G}}_t)\|_F^2\Big).$$

All M minibatch gradients are evaluated with the LoRA factors fixed at the same checkpoint. Thus, unlike Fig. 3, which uses the exact gradient over all 45,000 training examples as its reference, Fig. 4 uses the empirical mean of the sampled minibatch gradients.

H Comparison with Existing LoRA Convergence Theory

We compare our results with those of Mu & Klabjan (2025) and Wang et al. (2026). All results below assume that F is lower bounded and do not impose an a priori bound on the LoRA factor norms. Smoothness and stochastic-oracle assumptions concern F, whereas stationarity is measured by the factor gradient $\nabla J(B, A)$. Complexities count full or stochastic gradient evaluations of F, with problem-dependent constants suppressed.

Deterministic LoRA-GD. Table 4 compares guarantees for simultaneous gradient updates of both factors. The target is $\min_{0\le t<T} \|\nabla J(B_t, A_t)\|_F \le \epsilon$.

Table 4: Deterministic LoRA-GD: assumptions, stepsizes used for the upper bounds, and full-gradient complexities. A dash indicates that the cited work does not establish a matching lower bound.

Reference	Smoothness of F	Stepsize for the upper bound	Upper bound	Matching lower bound
Mu & Klabjan (2025)	Global L-smoothness	Adaptive; depends on the current factor and weight-gradient norms	$\exp\{O(\epsilon^{-2})\}$	—
Wang et al. (2026)	Global L-smoothness	Same adaptive rule as Mu & Klabjan (2025)	$O(\epsilon^{-4})$	—
This work	Sub-quadratic generalized smoothness	Predetermined: $\eta_t = \gamma/\sqrt{t+1}$	$O(\epsilon^{-4})$	$\Omega(\epsilon^{-4})$

Our upper bound retains the $O(\epsilon^{-4})$ rate of Wang et al. (2026) under sub-quadratic generalized smoothness and uses a predetermined stepsize schedule. In addition, Theorem 4.4 proves a matching lower bound for deterministic stepsize rules that choose a stepsize from the observed optimization history. This lower bound holds on a rank-one instance with a globally 1-smooth objective F; it therefore does not rely on the weaker smoothness assumption used in our upper bound.

Stochastic LoRA. All stochastic results in Table 5 assume an unbiased weight-gradient oracle with bounded variance. LoRA-NSGDM averages the stochastic factor gradients using momentum and normalizes the joint direction. LoRA-STORM uses a normalized recursive gradient estimate, with the same fresh sample evaluated at two consecutive iterates; it additionally requires mean-square smoothness of the weight-gradient oracle (Wang et al., 2026). Our upper bound instead analyzes simultaneous LoRA-SGD without momentum or normalization, using a constant stepsize chosen for the horizon T and failure probability q.

Table 5: Stochastic LoRA: assumptions, stationarity guarantees, and stochastic-gradient complexities. Here $\nabla J_t = \nabla J(B_t, A_t)$, and I is sampled uniformly from $\{0, \dots, T-1\}$. The lower bound has the policy and confidence restrictions specified below. A dash indicates that no matching lower bound is established in the corresponding result.

Method	Smoothness assumptions	Stationarity and upper complexity	Matching lower bound
LoRA-NSGDM (Wang et al., 2026)	Global L-smoothness of F	$\mathbb{E}\|\nabla J_I\|_F \le \epsilon$; $O(\epsilon^{-8})$ evaluations	—
LoRA-STORM (Wang et al., 2026)	Mean-square smoothness of the oracle, which implies global L-smoothness of F	$\mathbb{E}\|\nabla J_I\|_F \le \epsilon$; $O(\epsilon^{-6})$ evaluations	—
LoRA-SGD (Theorem 5.1)	Sub-quadratic generalized smoothness of F	With probability at least $1-q$, $\frac{1}{T}\sum_{t<T}\|\nabla J_t\|_F^2 \le \epsilon^2$; $O(q^{-1}\epsilon^{-8})$ evaluations	$\Omega(\epsilon^{-8})$
LoRA-SGD (Corollary 5.2)	Sub-quadratic generalized smoothness of F	$\mathbb{E}[\min_{t<T}\|\nabla J_t\|_F^2] \le \epsilon^2$; $O(\epsilon^{-10})$ evaluations	—

The stochastic upper bounds differ in both the algorithm and the guarantee. Wang et al. (2026) control the expected gradient norm at a uniformly sampled iterate. Theorem 5.1 controls the average squared gradient norm with high probability under generalized smoothness, and Corollary 5.2 controls the expected minimum squared gradient norm using $q = T^{-1/5}$ and the corresponding stepsize. This is not a guarantee for a uniformly sampled output, and the $O(\epsilon^{-10})$ bound is not claimed to be tight.

Scope of the stochastic lower bound. Theorem 5.3 permits the shared scalar $\eta_t \in [0, 1]$ to depend on the observed history, the current stochastic gradient, and independent internal randomness. For a fixed failure probability such as $q = 1/4$, its $\Omega(\epsilon^{-8})$ lower bound matches the ϵ-dependence of Theorem 5.1. This class includes global clipping and shared-scalar normalization of the current stochastic factor gradient whenever the effective scalar remains in $[0, 1]$. It does not cover the momentum or recursive gradient estimates used by LoRA-NSGDM and LoRA-STORM.

Infinite expected gradient norms. The negative result for plain LoRA-SGD in Wang et al. (2026) concerns gradient norms at fixed iterates: with positive stepsizes chosen before the current sample, finite-variance noise can yield an infinite expected gradient norm. Appendix E gives a related example. Infinite expectations do not rule out high-probability control. They are also compatible with the expected best-iterate guarantee, since

$$\min_{0 \le t < T} \|\nabla J(B_t, A_t)\|_F^2 \le \|\nabla J(B_0, A_0)\|_F^2$$

for every trajectory. Thus, an infinite expected gradient norm at a fixed iterate does not imply an infinite expectation for the minimum squared gradient norm along the trajectory.