A LOCAL THEORY OF GRADIENT-BASED RANK-ONE RECOVERY IN NONLINEAR LOW-RANK MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sequential rank-one methods build a low-rank update one component at a time. In the linear case, the residual, which is the part of the target update not yet recovered, directly reveals the next component. Nonlinearity hides that residual, but we can still compute the loss gradient from training data. Can this gradient reveal a useful direction? Motivated by sequential rank-one LoRA, we study this question through a direction-locked recovery procedure in a single-layer nonlinear teacher–student model. This provides a nonlinear counterpart to classical Hotelling deflation for memory-bounded fine-tuning and tasks with unknown target rank. At each stage k, the procedure extracts a rank-one direction $\hat{u}_k\hat{v}_k^\top$ from the gradient of the squared-error adaptation loss $\mathcal{L}_N$, fits a scalar amplitude s_k, and freezes the update $U_k = s_k\hat{u}_k\hat{v}_k^\top$ before proceeding. For a Gaussian Leaky-ReLU teacher–student model with rank-$r^\star$ target and slope $\alpha \in [0, 1)$, we prove that the rescaled negative population gradient equals the current residual plus a nonlinearity-induced bias controlled by a sector angle. The empirical gradient on a fresh batch concentrates around this surrogate at rate $\gamma_N = \widetilde{O}(\sqrt{d/N})$. Under sector control and gap-relative bounds on nonlinear bias, sampling error, and accumulated drift, this yields local per-stage recovery. After r stages, the conditional multi-stage bound controls the final parameter error by the oracle rank-r tail plus accumulated local errors. Sharpness examples show the worst-case need for spectral separation in top-dyad recovery and covariance control in the frozen-feature bridge, while a complementary gap-free bound gives sufficient conditions for one-stage parameter-residual decrease. Our formal contribution is a local residual-surrogate theorem for this direction-locked procedure. Empirically, we test the surrogate mechanism, observe residual progress outside the top-dyad sufficient regime, and illustrate the tradeoff between rank and active optimizer-state memory on frozen image features.

1 INTRODUCTION

Sequential rank-one learning is the family of methods that build a low-rank object one rank-one piece at a time, each piece chosen to account for what the earlier pieces left behind. It is appealing for three reasons that recur across applications. First, a sequential procedure produces an intermediate-rank frontier along the same training path, without retraining separate models from scratch; this gives a one-run rank sweep, although we do not provide a practical stopping rule for unknown target rank. Second, only one rank-one piece is being optimized at any moment, so the active optimizer-state memory for the current stage stays bounded by a single rank-one factor pair regardless of the final rank. This is an active-state advantage only: total memory, gradient computation, top-dyad extraction, and storage of the frozen accumulated rank-one updates from earlier stages are not reduced. Third, the analysis tradition for such procedures is mature (Liao et al., 2024; 2025; Alizadeh Vandchali et al., 2025): errors made at one stage feed into the residual the next stage operates on, which makes them a natural fit for deflation-style perturbation analysis going back to Hotelling (1933).

Low-Rank Adaptation (LoRA, Hu et al., 2022) reparameterizes a fine-tuned weight as $\mathbf{W}_0 + BA^\top$ with a frozen base $\mathbf{W}_0 \in \mathbb{R}^{d\times m}$ and trainable factors $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{m\times r}$, so the adapter $\Delta = BA^\top$ has rank at most r. Standard LoRA and its variants (Dettmers et al., 2023; Zhang et al., 2023; Valipour et al., 2023; Ding et al., 2023; Benedek & Wolf, 2024; Liu et al., 2024; Meng et al., 2024; Wang et al., 2024; Zhao et al., 2024; He et al., 2025; Zhang et al., 2025) train the rank-r

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.

factorization jointly. We study a single-layer model in which the same recipe is run sequentially: one rank-one piece at a time. Beyond the generic motivations above, the LoRA-specific consideration that distinguishes the sequential variant is that the joint factorization $\Delta = BA^\top$ has a rotational ambiguity: any orthogonal $Q \in \mathbb{R}^{r\times r}$ takes (B, A) to (BQ, AQ) without changing Δ, whereas fixing each rank-one direction at the time it is added selects a specific representative of the factorization.

Problem statement. We work with a single weight matrix passed through Leaky ReLU. The teacher is $\widetilde{y} = \sigma_\alpha(\mathbf{x}^\top(\mathbf{W}_0 + \Delta))$ on Gaussian inputs $\mathbf{x} \sim \mathcal{N}(0, I_d)$, where $\sigma_\alpha(z) = \max\{z, \alpha z\}$ is Leaky ReLU with slope $\alpha \in [0, 1)$, the base weight $\mathbf{W}_0$ is frozen, and the target adapter $\Delta \in \mathbb{R}^{d\times m}$ has rank $r^\star$. The *adaptation loss* is the empirical squared error between teacher and student outputs,

$$\mathcal{L}_N(\mathbf{W}) = \frac{1}{2N}\sum_{i=1}^{N} \left\|\sigma_\alpha(\mathbf{x}_i^\top \mathbf{W}) - \sigma_\alpha(\mathbf{x}_i^\top(\mathbf{W}_0 + \Delta))\right\|_2^2 .$$

A sequential rank-one algorithm builds the learned adapter $\widehat{\Delta}^{(r)}$ as a sum $\widehat{\Delta}^{(r)} = \sum_{k=1}^{r} U_k$ of r rank-one updates $U_k = s_k \hat{u}_k \hat{v}_k^\top$, one per stage. At stage k it performs three operations. (i) It looks at the gradient of $\mathcal{L}_N$ at the current iterate $\mathbf{W}_0 + \widehat{\Delta}^{(k-1)}$; we call the matrix $\Delta - \widehat{\Delta}^{(k-1)}$ the *residual*, since it is what the adapter still has to recover. The analysis additionally tracks an *oracle residual* $\Delta_k^\star$ along an idealized path that subtracts the exact best rank-one approximation at each stage; the actual and oracle residuals coincide when the algorithm's updates match the oracle. (ii) It picks a unit-norm pair $(\hat{u}_k, \hat{v}_k)$ from this loss gradient: concretely, the top left and right singular vectors of a rescaled gradient, with $\hat{u}_k \in \mathbb{R}^d$ left and $\hat{v}_k \in \mathbb{R}^m$ right. (iii) It fits the per-stage scale s_k, a single scalar, to set the magnitude of the new rank-one update $U_k = s_k \hat{u}_k \hat{v}_k^\top$. The update is then frozen and the algorithm proceeds to stage $k{+}1$ on the new residual. After r stages the learned adapter has the same $BA^\top$ structure as joint LoRA but is built one rank-one piece at a time. We analyze the *direction-locked* variant of this procedure: once the rank-one direction $(\hat{u}_k, \hat{v}_k)$ is read off the gradient, only the scalar amplitude s_k is optimized; the direction itself is not refit. Free two-factor optimization, which can rotate away from the gradient dyad, is not covered by the per-stage theorem; Corollary A.1 provides only a deterministic diagnostic decomposition with a measured rotation term, and the paper does not prove that this rotation term is small.

What changes with a nonlinearity. In the isotropic linear population model (replace σ_α by the identity), the negative loss gradient equals the parameter residual, so its leading singular vectors identify the next rank-one residual direction. Alizadeh Vandchali et al. (2025) give matching error-propagation bounds for sequential rank-one regression in this linear setting. With nonlinearity, this exact identity is lost, although we can still compute the loss gradient from training examples. We therefore use a rescaled gradient as a surrogate for the unknown residual and extract its leading rank-one direction. Our analysis asks when this direction approximates the next oracle component by separating three effects. First, the rescaled negative *population* gradient $-\nabla\mathcal{L}_{\mathrm{pop}}/\mu_\alpha$ (with $\mu_\alpha = (1+\alpha^2)/2$) equals the residual plus an activation-induced bias whose size is controlled by a sector angle between teacher and student columns. Second, the empirical gradient on a fresh batch differs from this population surrogate by sampling error. Reading a rank-one direction off the empirical gradient therefore reads the residual only up to that bias and the sampling error. Third, the residual update is still algebraically additive, $R_{k+1}^W = R_k^W - U_k$, but errors in earlier updates create mismatch between the actual residual and the oracle deflation path. Stage errors therefore propagate through both the residual recursion and the nonlinearity-and-sampling map from residual to surrogate, requiring us to revisit the linear-case analysis of Alizadeh Vandchali et al. (2025).

The sequential rank-one idea is classical. Hotelling deflation (Hotelling, 1933) is the original procedure: extract the top eigenpair, subtract its outer product from the matrix, and recurse. Mackey (2008) extended deflation to sparse PCA, where the inexactness of each stage forced a careful look at how errors compound. Subsequent work has sharpened that picture: Liao et al. (2024) analyze error propagation for inexact Hotelling deflation in PCA, and Alizadeh Vandchali et al. (2025) carry the same analysis through to sequential rank-one regression in the linear setting. A parallel line studies multi-component versions: Gemp et al. (2021; 2022) cast PCA as a multi-player game, and Liao et al. (2025) give the first provable parallel deflation algorithm for distributed PCA. Within parameter-efficient fine-tuning, ReLoRA (Lialin et al., 2023) restarts low-rank updates and merges them, which is sequential rank-r rather than sequential rank-one but shares the philosophy. Our work picks up exactly where Alizadeh Vandchali et al. (2025) stop: their analysis is for linear regression; we ask the same questions when a Leaky-ReLU sits between the input and the loss.

Contributions.

- We develop a residual-surrogate framework for gradient-based rank-one recovery, separating a model-independent perturbation principle (Theorem 4.1) from the nonlinear bias (Lemma 4.3) and fresh-batch sampling error (Lemma 4.4) of a Gaussian Leaky-ReLU model.
- We derive local per-stage recovery (Proposition 4.5) and conditional multi-stage guarantees (Proposition 4.6, Theorem A.4) for direction-locked updates, together with sharpness examples (Proposition 4.2) for spectral separation and frozen-feature covariance control. The all-stage certificate uses oracle quantities and is a post-hoc analytical tool, not an observable stopping rule.
- We establish complementary gap-free residual-progress guarantees (Proposition 4.8) and, under a dimension-dependent near-linearity condition, population contraction from arbitrary initialization (Corollary C.4) using isotropic second moments and exact spectral amplitudes.

2 Related Work

Greedy low-rank and atomic methods. The closest mathematical neighbors are methods that select rank-one directions from a residual or from first-order information: deflation (Hotelling, 1933; Mackey, 2008), matching pursuit (Mallat & Zhang, 1993), Frank-Wolfe and atomic decomposition (Jaggi & Sulovský, 2010; Jaggi, 2013; Shalev-Shwartz et al., 2011; Lee & Bresler, 2009), and low-rank optimization over convex surrogates (Journée et al., 2010), with related analyses of LoRA expressivity and matrix-factorization dynamics (Zeng & Lee, 2024; Li et al., 2021). These works typically study a linear inverse problem, a convex matrix objective, or a directly observed residual. Our setting is different: the residual is hidden, and inferring it through a nonlinear teacher-student gradient brings in sector angle, sampling noise, activation bias, and spectral gaps as additional ingredients in the rank-one recovery guarantee.

Low-rank fine-tuning and gradient/spectral subspace selection. A related primitive appears in low-rank fine-tuning: choose, grow, prune, initialize, or project low-rank subspaces using gradient or spectral structure. LoRA (Hu et al., 2022) and related PEFT methods (Houlsby et al., 2019; Lester et al., 2021; Li & Liang, 2021; Ben Zaken et al., 2022; Dettmers et al., 2023) adapt a frozen weight through low-rank, prompt, prefix, quantization, or bias updates; rank-adaptive variants (Valipour et al., 2023; Zhang et al., 2023; Ding et al., 2023; Benedek & Wolf, 2024; Liu et al., 2024; Lialin et al., 2023) grow, prune, or freeze low-rank capacity online; and gradient- or spectrum-driven initializations (Meng et al., 2024; Wang et al., 2024; Zhao et al., 2024; He et al., 2025) use first-order or spectral structure to pick or project the low-rank subspace. These works use gradients and spectra as algorithmic objects in deep models. We study a narrower teacher-student setting that isolates the local question of when a gradient actually reveals the next rank-one residual direction.

3 Setup and Algorithm

Let $\sigma_\alpha(z) = \max\{z, \alpha z\}$ with $\alpha \in [0, 1)$, applied coordinatewise. Inputs satisfy $\mathbf{x} \sim \mathcal{N}(0, I_d)$. The teacher is

$$\widetilde{y} = \sigma_\alpha(\mathbf{x}^\top \widetilde{\mathbf{W}}^\star), \qquad \widetilde{\mathbf{W}}^\star = \mathbf{W}_0 + \Delta, \qquad \Delta \in \mathbb{R}^{d\times m}, \quad \operatorname{rank}(\Delta) = r^\star.$$

Given N samples $(\mathbf{x}_i, \widetilde{y}_i)$, define

$$\mathcal{L}_N(\mathbf{W}) = \frac{1}{2N} \sum_{i=1}^{N} \left\| \sigma_\alpha(\mathbf{x}_i^\top \mathbf{W}) - \sigma_\alpha(\mathbf{x}_i^\top \widetilde{\mathbf{W}}^\star) \right\|_2^2.$$

The population loss is denoted $\mathcal{L}_{\text{pop}}(\mathbf{W}) = \mathbb{E}[\mathcal{L}_N(\mathbf{W})]$. At the kink of σ_α, empirical gradients use an arbitrary fixed subgradient value in $[\alpha, 1]$; all concentration statements below use the same convention. Away from the kink this is the ordinary gradient.

The sequential learner writes $\widehat{\Delta}^{(r)} = \sum_{k=1}^{r} U_k$, where each $U_k = b_k a_k^\top$ has rank at most one. When the rank-one direction is locked to the gradient's top dyad, we write $U_k = s_k \hat{u}_k \hat{v}_k^\top$ with unit-norm $\hat{u}_k \in \mathbb{R}^d$, $\hat{v}_k \in \mathbb{R}^m$ and a single scalar amplitude s_k; this is the form analyzed by Algorithm 1 and Proposition 4.5. The general $b_k a_k^\top$ form is used in Corollary A.1, which gives only a deterministic

Algorithm 1 Fresh-batch direction-locked sequential rank-one recovery

Require: base weight $\mathbf{W}_0$, target rank r, independent stage batches $\mathcal{B}_1, \dots, \mathcal{B}_r$, scalar-solver steps $t_1, \dots, t_r$

$\mathbf{W}^{\langle 1 \rangle} \leftarrow \mathbf{W}_0$
for $k = 1, \dots, r$ **do**
  $G_k \leftarrow \nabla \mathcal{L}_{\mathcal{B}_k}(\mathbf{W}^{\langle k \rangle})$ and $A_k \leftarrow -G_k/\mu_\alpha$, where $\mu_\alpha = (1+\alpha^2)/2$
  $(\hat{u}_k, \hat{v}_k) \leftarrow$ top left/right singular vectors of A_k
  Initialize $s_k \leftarrow \sigma_1(A_k)$
  Refine s_k for t_k scalar steps on $\mathcal{L}_{\mathcal{B}_k}(\mathbf{W}^{\langle k \rangle} + s_k \hat{u}_k \hat{v}_k^\top)$ ($t_k = 0$ keeps the SVD-derived s_k)
  $U_k \leftarrow s_k \hat{u}_k \hat{v}_k^\top$ and $\mathbf{W}^{\langle k+1 \rangle} \leftarrow \mathbf{W}^{\langle k \rangle} + U_k$
end for
return $\widehat{\Delta}^{(r)} = \sum_{k=1}^{r} U_k$

diagnostic decomposition with a measured rotation term for free-factor solvers that may rotate away from the gradient dyad. Let

$$\mathbf{W}^{\langle k \rangle} = \mathbf{W}_0 + \sum_{j<k} U_j, \qquad R_k^W = \Delta - \sum_{j<k} U_j.$$

The oracle residual is $\Delta_1^\star = \Delta$ and $\Delta_{k+1}^\star = \Delta_k^\star - U_k^\star$, where $U_k^\star$ is the best rank-one approximation of $\Delta_k^\star$. Write

$$\mathcal{T}_k^\star = \sigma_1(\Delta_k^\star) - \sigma_2(\Delta_k^\star).$$

Singular values beyond rank are zero. If $\sigma_1(\Delta_k^\star) = \sigma_2(\Delta_k^\star)$ (including $\Delta_k^\star = 0$), then $\mathcal{T}_k^\star = 0$ and the top-dyad perturbation guarantees do not apply at that stage.

The scalar budgets t_k are optimization hyperparameters, not part of the theory. Appendix B gives scalar-solver details for the ablations. If $r > r^\star$ (over-specified rank), the algorithm continues for all r stages, but the perturbation results in Section 4 apply only at stages $k \le r^\star$ where $\mathcal{T}_k^\star > 0$.

4 Theory

The theory separates the perturbation principle from the model-specific gradient calculation. A spectral gap separates the next oracle component from competing directions. When the surrogate error is small relative to this gap, Theorem 4.1 bounds the resulting error in the recovered direction. In the Gaussian Leaky-ReLU model, Lemmas 4.3 and 4.4 bound sector-controlled nonlinear bias and fresh-batch sampling error. Applying the perturbation principle to these bounds and accumulated drift yields the one-stage bound (Proposition 4.5) under its gap-relative condition. A smooth matched model supplies another surrogate through oracle Hessian preconditioning. Combining the one-stage bound with additive propagation (Proposition 4.6) yields the conservative conditional multi-stage certificate (Theorem A.4). Separately, Proposition 4.8 gives sufficient conditions for decreasing the actual parameter residual without requiring recovery of a specific oracle dyad or a spectral gap.

Throughout, C, C', C_0, C_1, C_2 denote positive universal constants independent of d, m, r, α, the teacher spectrum, and δ. Constants are fixed within each statement and retained when that statement is invoked. The generic constant C may change from line to line in proofs.

4.1 Rank-one recovery from a residual surrogate

Theorem 4.1 (Surrogate-to-dyad principle)**.** *Let A_k be any matrix and assume $\mathcal{T}_k^\star > 0$. Let $u_{1,k}$ and $v_{1,k}$ be a corresponding pair of leading left and right unit singular vectors of $\Delta_k^\star$. If*

$$\|A_k - \Delta_k^\star\|_{\mathrm{F}} \le \varepsilon_k \le \frac{1}{4}\mathcal{T}_k^\star,$$

then the top singular dyad $\hat{u}_k \hat{v}_k^\top$ of A_k obeys

$$\left\|\hat{u}_k \hat{v}_k^\top - u_{1,k} v_{1,k}^\top\right\|_{\mathrm{F}} \le C \frac{\varepsilon_k}{\mathcal{T}_k^\star}.$$

If additionally $U_k = s_k \hat{u}_k \hat{v}_k^\top$ *with* $|s_k - \sigma_1(A_k)| \le \varepsilon_{k,\mathrm{amp}}$*, then*

$$\|U_k - U_k^\star\|_{\mathrm{F}} \le \varepsilon_{k,\mathrm{amp}} + C' \frac{\sigma_1(\Delta_k^\star)}{\mathcal{T}_k^\star} \varepsilon_k.$$

This theorem is the abstract perturbation step behind the paper. The remaining results show when a gradient-based surrogate A_k is close to $\Delta_k^\star$. The following proposition records the matching sharpness: the gap dependence in Theorem 4.1 cannot be removed in worst case, and a companion construction shows the analogous covariance condition for the frozen-feature bridge of Appendix D.

Proposition 4.2 (Sharpness of gap and covariance requirements)**.** *The gap and covariance terms in the surrogate and bridge bounds cannot be removed in worst case.*

1. *Let* $\Delta = \sigma_1 u_1 v_1^\top + \sigma_2 u_2 v_2^\top$*, where* u_1, u_2 *and* v_1, v_2 *are orthonormal,* $\sigma_1 > \sigma_2 \ge 0$*, and* $g = \sigma_1 - \sigma_2$*. For any* $0 < \varepsilon < (\sigma_1 + \sigma_2)/2$*, there exists* E *with* $\|E\|_2 = g/2 + \varepsilon$ *and* $\|E\|_{\mathrm{F}} = \sqrt{2}(g/2 + \varepsilon)$ *such that* $\Delta + E$ *has top singular dyad* $u_2 v_2^\top$*, not* $u_1 v_1^\top$*.*
2. *More generally, in the deterministic perturbation model* $A = \Delta + E$*,* $\|E\|_2 \le \varepsilon$*, no estimator of the top dyad from* A *can improve the gap scaling uniformly over rank-two residuals with gap* g*: for a universal* $c > 0$*,*

$$\inf_{\Phi} \sup_{\Delta, E} \left\| \Phi(\Delta + E) - u_1 v_1^\top \right\|_{\mathrm{F}} \ge c \min\{1, \varepsilon/g\},$$

   where the supremum is over rank-two Δ *with* $\sigma_1(\Delta) - \sigma_2(\Delta) = g$ *and* E *with* $\|E\|_2 \le \varepsilon$*.*
3. *Consider the frozen-feature squared-loss bridge of Appendix D, in which a linear head is fit on frozen features with covariance* $\Sigma = \mathbb{E}[zz^\top]$ *and the population negative gradient at residual* R *is* ΣR*. Take* $R = e_1 e_1^\top + \rho e_2 e_2^\top$ *with* $0 < \rho < 1$ *and* $\Sigma = \mathrm{diag}(1, \lambda)$ *with* $\lambda > 1/\rho$*. Then the residual top dyad is* $e_1 e_1^\top$*, but the population negative gradient* ΣR *has top dyad* $e_2 e_2^\top$*. Moreover* $\|\Sigma - I\|_2 = \lambda - 1$ *can be arbitrarily close to* $(1-\rho)/\rho$*, which is of the same order as the residual gap when* ρ *is close to one.*

Thus top-dyad recovery cannot be certified from perturbation size without a gap-scale condition, and the frozen-feature bridge genuinely requires covariance control.

4.2 Gaussian Leaky-ReLU scalar-gradient surrogate

Lemma 4.3 (Population gradient decomposition)**.** *For the Leaky-ReLU teacher-student loss, suppose every nonzero teacher column has a nonzero learner column:* $w_j \neq 0$ *whenever* $\widetilde{w}_j^\star \neq 0$*. Then*

$$\nabla \mathcal{L}_{\mathrm{pop}}(\mathbf{W}) = \mu_\alpha (\mathbf{W} - \widetilde{\mathbf{W}}^\star) + \Psi_\alpha(\mathbf{W}), \qquad \mu_\alpha = \frac{1 + \alpha^2}{2}.$$

If, in addition, every nonzero teacher column satisfies $\angle(w_j, \widetilde{w}_j^\star) \le \theta_0 < \pi/2$*, then*

$$\|\Psi_\alpha(\mathbf{W})\|_{\mathrm{F}} \le (1-\alpha)^2 C_0 \theta_0 \left\| \mathbf{W} - \widetilde{\mathbf{W}}^\star \right\|_{\mathrm{F}}, \qquad C_0 = \frac{1}{2\pi} \sqrt{1 + \frac{4}{\pi^2}}.$$

Lemma 4.4 (Gradient concentration on a fixed or fresh-batch path)**.** *Fix a path* $\{\mathbf{W}_t\}_{t=0}^T \subset \mathbb{R}^{d \times m}$ *independent of the sample used to estimate the gradients. In particular, the statement applies conditionally at a stage when a fresh batch is drawn after the previous stages have fixed the current iterate. In this lemma,* $\nabla \mathcal{L}_N$ *denotes the selected empirical gradient obtained by the fixed kink convention above, and* $\nabla \mathcal{L}_{\mathrm{pop}}$ *denotes the expectation of the same selected gradient; at differentiability points the selected and ordinary population gradients coincide. For some universal constants* C_1, C_2*, if*

$$N \ge C_1 d \log\left(\frac{10 dm(T+1)}{\delta} \right),$$

then with probability at least $1 - \delta$*,*

$$\max_{0 \le t \le T} \|\nabla \mathcal{L}_N(\mathbf{W}_t) - \nabla \mathcal{L}_{\mathrm{pop}}(\mathbf{W}_t)\|_{\mathrm{F}} \le \gamma_N \max_{0 \le t \le T} \left\| \mathbf{W}_t - \widetilde{\mathbf{W}}^\star \right\|_{\mathrm{F}},$$

where

$$\gamma_N = C_2 \sqrt{\frac{d \log\left(\frac{10 dm(T+1)}{\delta} \right)}{N}}.$$

The residual-relative bound uses noiseless teacher outputs: each sample gradient vanishes at $\mathbf{W} = \widetilde{\mathbf{W}}^\star$. With label noise, finite-batch gradient fluctuations generally remain at zero parameter residual, requiring an additional noise-dependent term.

Proposition 4.5 (One-stage perturbation bound). *Assume $\mathcal{T}_k^\star > 0$, assume the sector condition of Lemma 4.3 holds at $\mathbf{W}^{\langle k \rangle}$, and assume the batch used for the stage-k gradient is independent of $\mathbf{W}^{\langle k \rangle}$ as in Lemma 4.4. Work on the concentration event of Lemma 4.4 and set*

$$\eta_\alpha(\theta_0, \gamma_N) = \frac{(1-\alpha)^2 C_0 \theta_0 + \gamma_N}{\mu_\alpha}.$$

Let $A_k = -\nabla \mathcal{L}_{\mathcal{B}_k}(\mathbf{W}^{\langle k \rangle})/\mu_\alpha$. If

$$\eta_\alpha \left\|R_k^W\right\|_{\mathrm{F}} + \left\|R_k^W - \Delta_k^\star\right\|_{\mathrm{F}} \le \frac{\mathcal{T}_k^\star}{4},$$

then the top singular dyad $\hat{u}_k \hat{v}_k^\top$ of A_k obeys

$$\left\|\hat{u}_k \hat{v}_k^\top - u_{1,k} v_{1,k}^\top\right\|_{\mathrm{F}} \le C \frac{\eta_\alpha \left\|R_k^W\right\|_{\mathrm{F}} + \left\|R_k^W - \Delta_k^\star\right\|_{\mathrm{F}}}{\mathcal{T}_k^\star},$$

where $u_{1,k}, v_{1,k}$ are the leading singular vectors of $\Delta_k^\star$. Moreover, if $U_k = s_k \hat{u}_k \hat{v}_k^\top$ and $\varepsilon_{k,\mathrm{amp}} = |s_k - \sigma_1(A_k)|$, then

$$\|U_k - U_k^\star\|_{\mathrm{F}} \le \varepsilon_{k,\mathrm{amp}} + C' \frac{\eta_\alpha \left\|R_k^W\right\|_{\mathrm{F}} + \left\|R_k^W - \Delta_k^\star\right\|_{\mathrm{F}}}{\mathcal{T}_k^\star} \sigma_1(\Delta_k^\star).$$

Locality and the sector condition. Proposition 4.5 is a *local* stage certificate: Algorithm 1 does not enforce the sector condition, and the proposition only certifies a stage when $\mathbf{W}^{\langle k \rangle}$ stays in a small-sector neighborhood of the teacher and the surrogate is accurate relative to $\mathcal{T}_k^\star$.

Reading the sufficient condition. The sufficient condition gives two errors a shared budget of one quarter of the oracle gap, $\mathcal{T}_k^\star/4$. First, $\eta_\alpha \left\|R_k^W\right\|_{\mathrm{F}}$ bounds the gradient surrogate's error relative to the current residual. Small sector angles limit teacher–student activation-pattern disagreement. The nonlinear-bias bound shrinks as $\alpha \to 1$, while larger N reduces sampling error. Second, $\left\|R_k^W - \Delta_k^\star\right\|_{\mathrm{F}}$ measures mismatch with the ideal deflation path from earlier updates and is zero at stage 1. Larger batches at the current stage do not remove nonlinear bias or earlier mismatch. The amplitude mismatch $\varepsilon_{k,\mathrm{amp}}$ measures deviation from the spectral amplitude, not scalar optimization error. Refinement can lower the empirical loss while increasing this additive term in the update bound. The quarter-gap condition is sufficient, not necessary, for the stated per-stage dyad error bound. Its failure does not preclude small final parameter error.

Stage-wise sector check. The sector assumption in Proposition 4.5 is automatic if $m_{\min} = \min_{j:\widetilde{w}_j^\star \neq 0} \left\|\widetilde{w}_j^\star\right\|_2 > 0$ and $\left\|R_k^W\right\|_{\mathrm{F}} = \left\|\widetilde{\mathbf{W}}^\star - \mathbf{W}^{\langle k \rangle}\right\|_{\mathrm{F}} \le m_{\min} \sin\theta_0$; Appendix C.1 gives the one-line proof and the corresponding all-stage envelope check.

Proposition 4.5 analyzes the direction-locked update $U_k = s_k \hat{u}_k \hat{v}_k^\top$ produced by Algorithm 1. A stage optimizer that freely updates both factors b_k, a_k can rotate away from the gradient dyad and is not covered; Corollary A.1 (App. A) records a deterministic diagnostic decomposition with a measured rotation term and does not prove that this rotation term is small. The same surrogate-to-dyad principle also applies to smooth matched activations through oracle Hessian preconditioning (Proposition A.2, App. A), which scopes the mechanism beyond Leaky-ReLU.

Proofs of Theorem 4.1 and Proposition A.2 are in Appendix A.

4.3 Multi-stage propagation and the all-stage certificate

Proposition 4.6 (Additive end-to-end propagation). *For any sequence of rank-one increments $U_1, \ldots, U_r$,*

$$\left\|\Delta - \widehat{\Delta}^{(r)}\right\|_{\mathrm{F}} \le \left(\sum_{j>r} \sigma_j(\Delta)^2\right)^{1/2} + \sum_{k=1}^{r} \|U_k - U_k^\star\|_{\mathrm{F}}.$$

Consequently, if Proposition 4.5 applies at every stage, the final parameter error is bounded by the oracle rank-r tail plus the sum of the stage-wise perturbation bounds.

In the displayed bound, the oracle tail is the smallest Frobenius error achievable by a rank-at-most-r approximation to Δ. It is nonincreasing in r and vanishes for $r \geq r^\star$. Along a fixed trajectory, the accumulated update-error sum is nondecreasing. Neither term dominates in general.

Proposition 4.7 (Prediction loss from parameter error)**.** *For any* $\mathbf{W}$ *and any* $\alpha \in [0, 1]$,

$$\mathcal{L}_{\mathrm{pop}}(\mathbf{W}) \leq \frac{1}{2} \left\| \mathbf{W} - \widetilde{\mathbf{W}}^\star \right\|_{\mathrm{F}}^2 .$$

Consequently, the parameter-error bound in Proposition 4.6 implies the corresponding population-prediction bound obtained by squaring its right hand side and multiplying by $1/2$.

Conservative multi-stage certification. Theorem A.4 bounds parameter error by the oracle rank-r tail plus its recursively defined envelope d_{r+1}. Write $\rho_k^\star = \sigma_1(\Delta_k^\star)/\mathcal{T}_k^\star$, $\kappa_k^\star = \|\Delta_k^\star\|_{\mathrm{F}}/\mathcal{T}_k^\star$, and $e_k = d_k/\mathcal{T}_k^\star$, with η_k the stage-wise value of η_α in Proposition 4.5. The certificate requires the concentration events, sector control, and $\eta_k(\kappa_k^\star + e_k) + e_k \leq 1/4$ at every stage. Thus gap-normalized surrogate error and accumulated mismatch share a $1/4$ budget. Although Proposition 4.6 sums stage errors additively, the recursive envelope amplifies past mismatch by $1 + C'\rho_k^\star(1 + \eta_k)$ per stage and can grow geometrically. Fresh-batch sample and sector certification conditions are in Corollaries A.5–A.6; they are not algorithmically enforced and do not give generic convergence. Corollary A.3 separately covers ε-optimal population scalar refitting on fixed directions.

The following elementary bound concerns actual-residual progress, not top-dyad recovery.

Proposition 4.8 (Gap-free residual progress)**.** *Let* $A = R + E \in \mathbb{R}^{d\times m}$*, let* $D = uv^\top$ *be a leading singular dyad of* A *with unit singular vectors, and set* $\sigma = \sigma_1(A)$*. For any* $s \geq 0$*,*

$$\|R\|_{\mathrm{F}}^2 - \|R - sD\|_{\mathrm{F}}^2 \geq s\big(2\sigma - s - 2\|E\|_2\big).$$

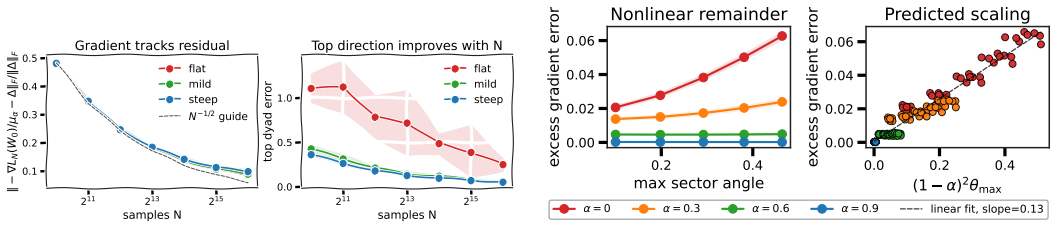
Thus $\|R - sD\|_{\mathrm{F}} < \|R\|_{\mathrm{F}}$ *whenever* $0 < s < 2(\sigma - \|E\|_2)$.

Apply this with $R = R_k^W$, $A = A_k$, and $s = s_k \geq 0$ (automatic for $t_k = 0$, not guaranteed after scalar refinement). The more conservative condition $0 < s < 2(\sigma - \|E\|_{\mathrm{F}})$ also suffices. This certifies residual decrease, not monotonic prediction loss; neither certificate is an observable stopping rule. The proof and near-best bound are in Appendix A.5. Under a dimension-dependent near-linearity condition, Corollary C.4 additionally gives population residual contraction from arbitrary initialization under isotropy, using exact spectral amplitudes and no sector condition.

5 Experiments

The synthetic experiments follow the proof: we test the stage-1 rescaled-gradient surrogate (Lemmas 4.3–4.4), check the direct surrogate condition of Theorem 4.1 at each stage of Algorithm 1 on a rank-8 teacher, and isolate the residual spectral gap to test the top-dyad error scaling predicted by Theorem 4.1. Appendix B adds calibration, sample-scaling, rank-mismatch, activation-slope, scalar-solver, and robustness diagnostics. Appendix D probes the residual-surrogate mechanism on non-Gaussian frozen features. Whitening lowers the measured perturbation ratio and top-dyad error, consistent with the covariance-control motivation of the linear squared-loss bridge (Proposition D.1), although the plotted $1/4$ threshold is not reached. In DeBERTa-v3-base, gradient-selected frozen directions outperform random directions in all ten paired seeds at each tested training-set size on SST-2 and MNLI, with matched adapter capacity and optimization-step budgets but not total compute (Appendix E). These experiments do not verify the main theorem.

Stage-1 residual surrogate. At $\mathbf{W}_0$ we compare the rescaled empirical gradient to the true residual as we sweep the fresh batch size N (Figure 1(a)). Both the relative gradient–residual error and the top-dyad error decrease with N at the $\sqrt{d/N}$ rate of Lemma 4.4, consistent with the stage-1 conclusion of Theorem 4.1. In Figure 1(b) we probe the proportional dependence in the remainder bound of Lemma 4.3. After we subtract the α=1 sampling baseline, the excess gradient–residual error scales linearly with $(1-\alpha)^2\theta_{\max}$. We use an excess-error proxy and therefore test the predicted scaling, not the literal constant C_0.



(a) Gradient as residual.

(b) Nonlinear remainder.

Figure 1: Mechanism checks at $\mathbf{W}_0$ for steep/mild/flat teacher spectra (large/moderate/small singular-value gaps). **(a) Gradient as residual.** Left: relative gradient–residual error $\|-\nabla\mathcal{L}_N(\mathbf{W}_0)/\mu_\alpha - \Delta\|_{\mathrm{F}}/\|\Delta\|_{\mathrm{F}}$ decreases with the fresh batch size N; the dashed curve is an $N^{-1/2}$ rate guide for the sampling term in Lemma 4.4, not a finite-sample upper bound. Right: top singular dyad error at the same iterate decreases with N, consistent with the stage-1 conclusion of Theorem 4.1 composed with Proposition 4.5. **(b) Nonlinear remainder.** Left: after subtracting the $\alpha{=}1$ sampling baseline, the excess relative gradient–residual error grows with the maximum sector angle for fixed α, with the gap closing as $\alpha \to 1$. Right: pooled scatter against $(1-\alpha)^2\theta_{\max}$ across $(\alpha, \|\Delta\|_{\mathrm{F}})$ settings, with linear fit (slope 0.13). The figure tests the proportional dependence in the remainder bound of Lemma 4.3, $\|\Psi_\alpha\|_{\mathrm{F}} \propto (1-\alpha)^2\theta_0\|\mathbf{W} - \widetilde{\mathbf{W}}^\star\|_{\mathrm{F}}$, using an excess-error proxy; it is consistent with the predicted scaling but does not certify the literal constant C_0.

Residual progress beyond the oracle test. We reanalyze the 24 runs underlying Figure 2 (three spectra $\times$ eight random seeds), each with eight rank-one stages, giving 192 stage updates in total. None of these stages passes the plotted $1/4$ oracle perturbation test, yet all satisfy $0 < s_k < 2(\sigma_1(A_k) - \left\|A_k - R_k^W\right\|_{\mathrm{F}})$, the conservative sufficient condition of Proposition 4.8, including after scalar refinement. All tested stages decrease the actual parameter residual. After eight stages, mean final relative parameter errors are 0.0351, 0.0647, and 0.0650 for steep, mild, and flat spectra, respectively (Table 2, fresh batches). Thus failure of the oracle test can coexist with certified residual progress and small final parameter error. Appendix B.3 gives the calculation; it uses the known teacher, not an observable stopping rule or a guarantee for other runs. For comparison, the separated rank-3 calibration in Figure 3 passes the plotted oracle test at every stage (App. B).

Gap is the dominant lever. Theorem 4.1 predicts that the top-dyad error scales as the gap-normalized perturbation ratio, and Proposition 4.2 shows that the gap requirement cannot be removed in worst case. Figure 4 (App. B) tests this prediction with a controlled sweep that varies only the residual spectral gap and finds the top-dyad error grows roughly linearly with the perturbation ratio.

End-to-end sample dependence closes the chain. The preceding experiments probe individual links in the theoretical argument. Appendix B.4 reports the fresh stage-batch size N sweep on the rank-8 run (Figure 6) and compares fresh and reused batches. Larger N tightens the maximum-over-stages perturbation ratio (right), as Lemma 4.4 predicts, and improves both final parameter recovery (left) and held-out prediction loss (middle). We compare the observed log-log slope to the $N^{-1/2}$ guide of Lemma 4.4. Agreement is qualitative because final error also includes nonlinear bias, gap effects, scalar slack, and accumulated past-stage mismatch. Read together, these experiments move from the gradient-as-residual identity (Lemma 4.3, Lemma 4.4) through stage-wise dyad recovery (Proposition 4.5) to additive end-to-end propagation (Proposition 4.6, Proposition 4.7).

6 Discussion

The main message is theoretical: in nonlinear low-rank models, a gradient reveals the next rank-one residual direction when the chosen gradient surrogate is accurate relative to the residual spectral gap. Sector control keeps the gradient identity tight, fresh batches keep concentration unconditional, and gap separation keeps the SVD perturbation argument quantitative. Proposition 4.2 shows that the gap requirement and the frozen-feature covariance requirement cannot be removed in worst case.

Implication for the LoRA initialization step. The formal contribution analyzes a local question shared with gradient- and spectrum-based LoRA initialization schemes (PiSSA, LoRA-GA, LoRA-

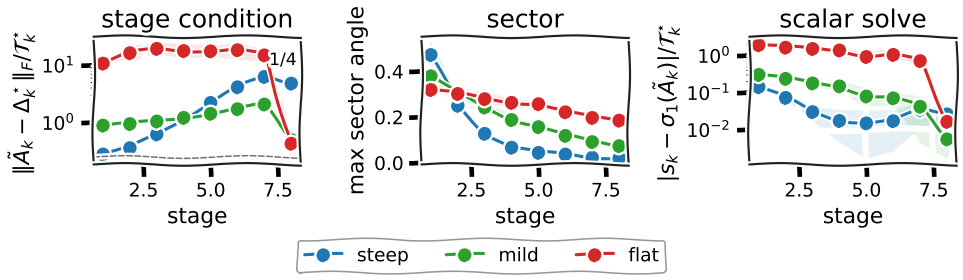


Figure 2: Per-stage diagnostics for the analyzed direction-locked algorithm on a harder rank-8 run, steep/mild/flat teacher spectra (large suite). Left: empirical perturbation ratio $\|A_k - \Delta_k^\star\|_{\mathrm{F}}/\mathcal{T}_k^\star$ at each stage; the dashed line marks the $1/4$ sufficient threshold of Theorem 4.1. Middle: maximum sector angle along teacher columns, the quantity that controls the remainder bound in Lemma 4.3. Right: scalar-amplitude mismatch $|s_k - \sigma_1(A_k)|/\mathcal{T}_k^\star$, the gap-normalized version of the additive slack $\varepsilon_{k,\mathrm{amp}}$ in Proposition 4.5. Rank-8 is a diagnostic stress test: all tested stages have perturbation ratio above $1/4$ (minimum 0.272). Steep is closest to the threshold at stage 1, whereas mild and flat are closest at stage 8. These stages cannot be certified by Proposition 4.5, whose hypotheses imply this direct test; failure does not preclude progress. The direct test certifies the per-stage dyad bound for the rank-3 calibration with large N in Figure 3.

One, GaLore): when does the top dyad of the current loss gradient approximate the next residual direction? Proposition 4.5 gives a sufficient local condition for that selection step in our single-layer Gaussian Leaky-ReLU model. The scalar-solver ablations are consistent with the local theory: with fixed directions, line search, scalar Adam, and scalar backfitting behave nearly identically, while the residual spectral gap dominates stagewise error. They do not extend the certificate to post-initialization free two-factor optimization. The noiseless linear-head bridge provides a special case with no requirement for a nearby initialization. With population-whitened features and squared loss, rank-one population-gradient updates with exact spectral amplitudes recover a best rank-r approximation without Gaussianity (Proposition D.1). The frozen-feature experiments remain illustrative and do not establish superiority over simultaneous rank-r training.

7 Limitations and Future Work

This paper is not a convergence theorem for LoRA in deep networks; the formal contribution is a local residual-surrogate theorem for the gradient-dyad selection step in a direction-locked rank-one procedure on a single-layer Gaussian Leaky-ReLU teacher–student model. The finite-sample top-dyad recovery theorems are local, gap-dependent, and model-specific. They assume Gaussian inputs, a single Leaky-ReLU layer, a fresh independent batch per stage, sector control around the teacher, and a separated oracle spectrum at every stage; Lemma C.1 converts an envelope on $\left\|R_k^W\right\|_{\mathrm{F}}$ into the sector condition along the path, but Algorithm 1 does not enforce it. Gaussianity underlies the sector-controlled calculation in Lemma 4.3 and the concentration argument in Lemma 4.4. The separate near-linear population result in Corollary C.4 replaces Gaussianity by isotropic second moments, but does not extend the finite-sample top-dyad recovery guarantees. The per-stage theorem covers the direction-locked update only. Corollary A.1 provides a deterministic diagnostic decomposition for free two-factor optimization with a measured rotation term that we do not prove to be small; free-factor optimization is post-initialization and outside the theorem. Extending the analysis to free two-factor LoRA requires bounding this rotation. The all-stage certificates (Theorem A.4, Corollary A.5, Corollary A.6) use oracle quantities $\mathcal{T}_k^\star, \kappa_k^\star, \rho_k^\star, m_{\min}$ and are post-hoc analytical tools, not observable stopping rules; the envelope is conservative and does not yield a generic convergence theorem from sample size alone. The procedure produces a one-run rank frontier, but we do not provide a practical stopping rule for unknown target rank, and the memory benefit concerns active optimizer-state memory only: total memory, gradient computation, top-dyad extraction, and storage of frozen updates are not reduced. The smooth extension uses an oracle Hessian preconditioner. Reused-batch adaptive concentration, multilayer feature coupling, cross-stage representation drift, and deep-network LoRA convergence remain outside scope; the frozen-feature image experiments are bridge evidence for the residual-surrogate mechanism, not a deep fine-tuning benchmark.

REPRODUCIBILITY STATEMENT

The model, its assumptions, and the direction-locked algorithm are specified in Section 3. Proofs of the main theoretical results are provided in Appendix A, with auxiliary results and the frozen-feature bridge developed in Appendices C and D. Appendix B.1 gives the synthetic suites, sample budgets, seed counts, and optimization settings. Appendices D.5 and E give the frozen-feature and transformer protocols, including data preparation, training budgets, evaluation metrics, and variability across seeds. The frozen-feature and transformer experiments use publicly available benchmark datasets and pretrained checkpoints.

AI USE STATEMENT

We used large language models to aid or polish writing, for finding related work, and for interpreting experimental results. AI assistance also supported refining and checking our mathematical arguments. We reviewed all AI-assisted material, verified the mathematical arguments and proofs, checked the reported numerical results against the recorded experimental outputs, and manually checked the cited references against their original sources. We take responsibility for the final manuscript, including its text, mathematical claims, experimental results, citations, and associated artifacts.

REFERENCES

Mahtab Alizadeh Vandchali, Fangshuo Liao, and Anastasios Kyrillidis. One rank at a time: Cascading error dynamics in sequential learning. *arXiv preprint arXiv:2505.22602*, 2025.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 1–9. Association for Computational Linguistics, 2022.

Nadav Benedek and Lior Wolf. PRILoRA: Pruned and rank-increasing low-rank adaptation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 252–263. Association for Computational Linguistics, 2024.

Youngmin Cho and Lawrence K. Saul. Kernel methods for deep learning. In *Advances in Neural Information Processing Systems*, volume 22, 2009.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Ning Ding, Xingtai Lv, Qiaosen Wang, Yulin Chen, Bowen Zhou, Zhiyuan Liu, and Maosong Sun. Sparse low-rank adaptation of pre-trained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4133–4145. Association for Computational Linguistics, 2023.

Ian Gemp, Brian McWilliams, Claire Vernade, and Thore Graepel. EigenGame: PCA as a Nash equilibrium. In *International Conference on Learning Representations*, 2021.

Ian Gemp, Brian McWilliams, Claire Vernade, and Thore Graepel. EigenGame unloaded: When playing games is better than optimizing. In *International Conference on Learning Representations*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Yutong He, Pengrui Li, Yipeng Hu, Chuyan Chen, and Kun Yuan. Subspace optimization for large language models with convergence guarantees. In *International Conference on Machine Learning*, 2025.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1314–1324, 2019.

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

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 427–435. PMLR, 2013.

Martin Jaggi and Marek Sulovský. A simple algorithm for nuclear norm regularized problems. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 471–478, 2010.

Michel Journée, Francis Bach, P.-A. Absil, and Rodolphe Sepulchre. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, 20(5):2327–2351, 2010.

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

Kiryung Lee and Yoram Bresler. ADMiRA: Atomic decomposition for minimum rank approximation. *arXiv preprint arXiv:0905.0044*, 2009.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3045–3059. Association for Computational Linguistics, 2021.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 4582–4597. Association for Computational Linguistics, 2021.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2021.

Vladislav Lialin, Namrata Shivagunde, Sherin Muckatira, and Anna Rumshisky. ReLoRA: High-rank training through low-rank updates. *arXiv preprint arXiv:2307.05695*, 2023.

Fangshuo Liao, Junhyung Lyle Kim, Cruz Barnum, and Anastasios Kyrillidis. On the error-propagation of inexact Hotelling's deflation for principal component analysis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 29720–29747, 2024.

Fangshuo Liao, Wenyi Su, and Anastasios Kyrillidis. Provable model-parallel distributed principal component analysis with parallel deflation. In *Conference on Parsimony and Learning*, 2025.

Zeyu Liu, Souvik Kundu, Anni Li, Junrui Wan, Lianghao Jiang, and Peter A. Beerel. AFLoRA: Adaptive freezing of low rank adaptation in parameter efficient fine-tuning of large models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 161–167. Association for Computational Linguistics, 2024.

Lester Mackey. Deflation methods for sparse PCA. In *Advances in Neural Information Processing Systems*, volume 21, 2008.

Stéphane G. Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. PiSSA: Principal singular values and singular vectors adaptation of large language models. In *Advances in Neural Information Processing Systems*, 2024.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, pp. 722–729, 2008.

Shai Shalev-Shwartz, Alon Gonen, and Ohad Shamir. Large-scale convex minimization with a low-rank constraint. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 329–336, 2011.

Mojtaba Valipour, Mehdi Rezagholizadeh, Ivan Kobyzev, and Ali Ghodsi. DyLoRA: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3274–3287. Association for Computational Linguistics, 2023.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Shaowen Wang, Linxi Yu, and Jian Li. LoRA-GA: Low-rank adaptation with gradient approximation. In *Advances in Neural Information Processing Systems*, 2024.

Yuchen Zeng and Kangwook Lee. The expressive power of low-rank adaptation. In *International Conference on Learning Representations*, 2024.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, 2023.

Yuanhe Zhang, Fanghui Liu, and Yudong Chen. LoRA-One: One-step full gradient could suffice for fine-tuning large language models, provably and efficiently. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61121–61143. PMLR, 2024.

A Proofs of Main Theory Results

A.1 Proof of Theorem 4.1

Proof. Weyl's inequality gives

$$|\sigma_i(A_k) - \sigma_i(\Delta_k^\star)| \le \|A_k - \Delta_k^\star\|_2 \le \varepsilon_k.$$

Since $\varepsilon_k \le \mathcal{T}_k^\star/4$, the leading singular subspace remains separated. Wedin's sin-theta theorem applied to the left and right singular subspaces yields

$$\min_{\xi\in\{\pm1\}} \left(\|\hat{u}_k - \xi u_{1,k}\|^2 + \|\hat{v}_k - \xi v_{1,k}\|^2\right)^{1/2} \le C\varepsilon_k/\mathcal{T}_k^\star.$$

Therefore

$$\left\|\hat{u}_k\hat{v}_k^\top - u_{1,k}v_{1,k}^\top\right\|_{\mathrm{F}} \le \|\hat{u}_k - \xi u_{1,k}\| + \|\hat{v}_k - \xi v_{1,k}\| \le C\varepsilon_k/\mathcal{T}_k^\star.$$

For the scaled update,

$$\begin{aligned}\left\|s_k\hat{u}_k\hat{v}_k^\top - \sigma_1(\Delta_k^\star)u_{1,k}v_{1,k}^\top\right\|_{\mathrm{F}} \le{}& |s_k - \sigma_1(\Delta_k^\star)| \\ &+ \sigma_1(\Delta_k^\star)\left\|\hat{u}_k\hat{v}_k^\top - u_{1,k}v_{1,k}^\top\right\|_{\mathrm{F}}.\end{aligned}$$

Another application of Weyl gives $|\sigma_1(A_k) - \sigma_1(\Delta_k^\star)| \le \varepsilon_k$, so $|s_k - \sigma_1(\Delta_k^\star)| \le \varepsilon_{k,\mathrm{amp}} + \varepsilon_k$. Since $\mathcal{T}_k^\star \le \sigma_1(\Delta_k^\star)$, the additive ε_k is absorbed into $C'\sigma_1(\Delta_k^\star)\varepsilon_k/\mathcal{T}_k^\star$. □

A.2 Proof of Proposition 4.2

For the perturbation claim, set $t = g/2 + \varepsilon$ and

$$E = t(u_2 v_2^\top - u_1 v_1^\top).$$

The two rank-one terms have orthogonal left and right singular directions, so $\|E\|_2 = t$ and $\|E\|_{\mathrm{F}} = \sqrt{2}t$. Moreover

$$\Delta + E = (\sigma_1 - t)u_1 v_1^\top + (\sigma_2 + t)u_2 v_2^\top.$$

Since $t = g/2 + \varepsilon$,

$$\sigma_2 + t = \frac{\sigma_1 + \sigma_2}{2} + \varepsilon > \frac{\sigma_1 + \sigma_2}{2} - \varepsilon = \sigma_1 - t.$$

The upper bound on ε ensures $\sigma_1 - t > 0$, so the leading singular dyad of $\Delta + E$ is $u_2 v_2^\top$. At $\varepsilon = 0$ the two leading singular values tie, so the top dyad is not uniquely determined.

For the minimax perturbation claim, it suffices to work in two dimensions with symmetric positive semidefinite matrices. Let $g > 0$ be fixed and write $q_\theta = (\cos\theta, \sin\theta)$ and $q_\theta^\perp = (-\sin\theta, \cos\theta)$. For any $\theta \in (0, \pi/4]$, define

$$\Delta_\theta = (1+g) q_\theta q_\theta^\top + q_\theta^\perp q_\theta^{\perp\top}, \qquad \Delta_{-\theta} = (1+g) q_{-\theta} q_{-\theta}^\top + q_{-\theta}^\perp q_{-\theta}^{\perp\top}.$$

Both matrices have singular-value gap g, and their top dyads are $D_\theta = q_\theta q_\theta^\top$ and $D_{-\theta} = q_{-\theta} q_{-\theta}^\top$. Since

$$\Delta_\theta - \Delta_{-\theta} = g(D_\theta - D_{-\theta}), \qquad \|D_\theta - D_{-\theta}\|_2 = \sin(2\theta),$$

the common observation $A = (\Delta_\theta + \Delta_{-\theta})/2$ is consistent with both teachers using perturbations of operator norm $g\sin(2\theta)/2$. Choose θ so that $\sin(2\theta) \le \min\{1, 2\varepsilon/g\}$ and this norm is at most ε. For any estimator $\Phi(A)$, the triangle inequality gives

$$\max_{\xi \in \{\theta, -\theta\}} \|\Phi(A) - D_\xi\|_{\mathrm{F}} \ge \frac{1}{2}\|D_\theta - D_{-\theta}\|_{\mathrm{F}} = \frac{\sqrt{2}}{2}\sin(2\theta).$$

Choosing θ with $\sin(2\theta)$ a fixed constant multiple of $\min\{1, \varepsilon/g\}$ proves the displayed lower bound.

For the covariance claim,

$$\Sigma R = e_1 e_1^\top + \lambda\rho e_2 e_2^\top.$$

The residual R has leading singular value 1 with dyad $e_1 e_1^\top$, because $\rho < 1$. If $\lambda > 1/\rho$, then $\lambda\rho > 1$, so the population negative gradient ΣR has leading dyad $e_2 e_2^\top$. Finally, $\|\Sigma - I\|_2 = \lambda - 1$, and taking $\lambda \downarrow 1/\rho$ makes this quantity approach $1/\rho - 1 = (1-\rho)/\rho$. If $\rho = 1 - g$, this equals $g/(1-g)$, which is of the same order as the residual gap g for small g.

A.3 Proof of Lemma 4.3

The loss separates over output coordinates, so it is enough to prove the claim for one column and then sum in Frobenius norm. Fix a teacher column $\mathbf{w}^\star$ and define

$$\ell_\alpha(\mathbf{w}) = \frac{1}{2}\mathbb{E}\big[\big(\sigma_\alpha(\mathbf{x}^\top\mathbf{w}) - \sigma_\alpha(\mathbf{x}^\top\mathbf{w}^\star)\big)^2\big], \qquad h = \mathbf{w} - \mathbf{w}^\star.$$

If $\mathbf{w}^\star = 0$, then

$$\ell_\alpha(\mathbf{w}) = \frac{1}{2}\mathbb{E}[\sigma_\alpha(\mathbf{x}^\top\mathbf{w})^2] = \frac{\mu_\alpha}{2}\|\mathbf{w}\|^2,$$

so $\nabla\ell_\alpha(\mathbf{w}) = \mu_\alpha\mathbf{w} = \mu_\alpha(\mathbf{w} - \mathbf{w}^\star)$ and the remainder is zero. We may therefore assume $\mathbf{w}^\star \neq 0$. The sector condition is not needed for the identity; the nonzero-column hypothesis gives $\mathbf{w} \neq 0$. The sector condition is used only for the final remainder bound. Values of σ'_α at zero are irrelevant for the population identities below; for empirical gradients one may choose any value in $[\alpha, 1]$ at the kink.

Write $\sigma_\alpha(z) = \alpha z + (1-\alpha)\sigma_0(z)$ with $\sigma_0(z) = \max\{z, 0\}$. Expanding the square gives

$$\ell_\alpha(\mathbf{w}) = \frac{\alpha^2}{2}\mathbb{E}[(\mathbf{x}^\top h)^2] + \alpha(1-\alpha)\mathbb{E}[(\mathbf{x}^\top h)(\sigma_0(\mathbf{x}^\top\mathbf{w}) - \sigma_0(\mathbf{x}^\top\mathbf{w}^\star))] + (1-\alpha)^2\ell_0(\mathbf{w}),$$

where

$$\ell_0(\mathbf{w}) = \frac{1}{2}\mathbb{E}[(\sigma_0(\mathbf{x}^\top\mathbf{w}) - \sigma_0(\mathbf{x}^\top\mathbf{w}^\star))^2]$$

is the corresponding ReLU loss. The first term has gradient $\alpha^2 h$. For the mixed term, use the Gaussian half-space identity

$$\mathbb{E}[(\mathbf{x}^\top a)\sigma_0(\mathbf{x}^\top b)] = \frac{1}{2}a^\top b,$$

which follows by conditioning on $\mathbf{x}^\top b$ and using symmetry. Therefore

$$\mathbb{E}[(\mathbf{x}^\top h)(\sigma_0(\mathbf{x}^\top \mathbf{w}) - \sigma_0(\mathbf{x}^\top \mathbf{w}^\star))] = \frac{1}{2}h^\top(\mathbf{w} - \mathbf{w}^\star) = \frac{1}{2}\|h\|^2,$$

and the mixed term contributes $\alpha(1-\alpha)h$ to the gradient.

It remains to record the ReLU gradient exactly. Let $a = \|\mathbf{w}\|$, $b = \|\mathbf{w}^\star\|$, $\theta = \angle(\mathbf{w}, \mathbf{w}^\star)$, $\rho = \cos\theta$, $s = \sin\theta$, $u = \mathbf{w}/a$, and write

$$\mathbf{w}^\star/b = \rho u + sq, \qquad q \perp u, \quad \|q\| = 1.$$

The ReLU kernel for Gaussian inputs is

$$K(\mathbf{w}, \mathbf{w}^\star) = \mathbb{E}[\sigma_0(\mathbf{x}^\top \mathbf{w})\sigma_0(\mathbf{x}^\top \mathbf{w}^\star)] = \frac{ab}{2\pi}\{\sin\theta + (\pi - \theta)\cos\theta\} \quad \text{(Cho \& Saul, 2009)}.$$

Since $\mathbb{E}[\sigma_0(\mathbf{x}^\top \mathbf{w})^2] = a^2/2$, we have $\ell_0(\mathbf{w}) = a^2/4 + \text{const} - K(\mathbf{w}, \mathbf{w}^\star)$. Set $f(\theta) = \sin\theta + (\pi - \theta)\cos\theta$. Then $f'(\theta) = -(\pi - \theta)\sin\theta$ and $\nabla_{\mathbf{w}}\theta = -q/a$, so

$$\nabla_{\mathbf{w}} K(\mathbf{w}, \mathbf{w}^\star) = \frac{b}{2\pi}\{f(\theta)u + (\pi - \theta)sq\}.$$

Thus

$$\nabla \ell_0(\mathbf{w}) = \frac{1}{2}\mathbf{w} - \frac{b}{2\pi}\{(\sin\theta + (\pi - \theta)\rho)u + (\pi - \theta)sq\}.$$

Subtracting $\frac{1}{2}h = \frac{1}{2}(a - b\rho)u - \frac{1}{2}bsq$ gives

$$\nabla \ell_0(\mathbf{w}) = \frac{1}{2}h + \Psi_0(\mathbf{w}), \qquad \Psi_0(\mathbf{w}) = \frac{b}{2\pi}\left[(\rho\theta - \sin\theta)u + \theta sq\right].$$

Combining the three contributions,

$$\nabla \ell_\alpha(\mathbf{w}) = \left(\alpha^2 + \alpha(1-\alpha) + \frac{(1-\alpha)^2}{2}\right)h + (1-\alpha)^2\Psi_0(\mathbf{w}) = \mu_\alpha h + (1-\alpha)^2\Psi_0(\mathbf{w}),$$

where $\mu_\alpha = (1 + \alpha^2)/2$.

We now bound the ReLU remainder. Since u and q are orthonormal,

$$\|\Psi_0(\mathbf{w})\|^2 = \frac{b^2}{4\pi^2}\left[(\rho\theta - \sin\theta)^2 + \theta^2 s^2\right].$$

Also

$$\|\mathbf{w} - \mathbf{w}^\star\|^2 = (a - b\rho)^2 + b^2 s^2 \geq b^2 s^2.$$

The case $\theta = 0$ gives $\Psi_0(\mathbf{w}) = 0$ by continuity. For $\theta \in (0, \pi/2]$,

$$\frac{\|\Psi_0(\mathbf{w})\|^2}{\|\mathbf{w} - \mathbf{w}^\star\|^2} \leq \frac{1}{4\pi^2}\left[(\theta\cot\theta - 1)^2 + \theta^2\right].$$

It remains to use a scalar inequality. Let $g(\theta) = (1 - \theta\cot\theta)/\theta = 1/\theta - \cot\theta$. Then $g'(\theta) = \csc^2\theta - 1/\theta^2 \geq 0$ on $(0, \pi/2]$ because $\sin\theta \leq \theta$, and $g(0+) = 0$, $g(\pi/2) = 2/\pi$. Hence $0 \leq 1 - \theta\cot\theta \leq 2\theta/\pi$. Therefore

$$\|\Psi_0(\mathbf{w})\| \leq \frac{\theta}{2\pi}\sqrt{1 + \frac{4}{\pi^2}}\,\|\mathbf{w} - \mathbf{w}^\star\| = C_0\theta\,\|\mathbf{w} - \mathbf{w}^\star\|.$$

Multiplying by $(1-\alpha)^2$ and using $\theta \leq \theta_0$ gives the column bound. Squaring, summing over columns, and taking square roots proves

$$\|\Psi_\alpha(\mathbf{W})\|_{\mathrm{F}} \leq (1-\alpha)^2 C_0\theta_0\left\|\mathbf{W} - \widetilde{\mathbf{W}}^\star\right\|_{\mathrm{F}}.$$

A.4 Proof of Lemma 4.4

The proof is for a fixed path independent of the sample. The fresh-batch stage statement follows by conditioning on the previous stages; after conditioning, the current iterate is fixed and the new batch is independent.

Fix t and j, write $h_{t,j} = w_{t,j} - \widetilde{w}_j^\star$, and define the vector summand

$$Z_i^{t,j} = \big(\sigma_\alpha(\mathbf{x}_i^\top w_{t,j}) - \sigma_\alpha(\mathbf{x}_i^\top \widetilde{w}_j^\star)\big)\sigma'_\alpha(\mathbf{x}_i^\top w_{t,j})\mathbf{x}_i.$$

Then

$$\nabla_{w_j}\mathcal{L}_N(\mathbf{W}_t) - \nabla_{w_j}\mathcal{L}_{\mathrm{pop}}(\mathbf{W}_t) = \frac{1}{N}\sum_{i=1}^N \left(Z_i^{t,j} - \mathbb{E}Z_i^{t,j}\right).$$

At the nondifferentiable point of σ_α, choose any derivative in $[\alpha, 1]$; all bounds below use only $|\sigma'_\alpha| \le 1$.

Let $u \in \mathbb{S}^{d-1}$. The one-dimensional projection is bounded in absolute value by a product of Gaussian linear forms:

$$\begin{aligned}\left|u^\top Z_i^{t,j}\right| &\le \left|u^\top \mathbf{x}_i\right| \left|\sigma_\alpha(\mathbf{x}_i^\top w_{t,j}) - \sigma_\alpha(\mathbf{x}_i^\top \widetilde{w}_j^\star)\right| \left|\sigma'_\alpha(\mathbf{x}_i^\top w_{t,j})\right| \\ &\le \left|u^\top \mathbf{x}_i\right| \left|\mathbf{x}_i^\top h_{t,j}\right|,\end{aligned}$$

because σ_α is 1-Lipschitz for $\alpha \in [0,1]$. Hence

$$\left\| \left|u^\top \mathbf{x}\right| \left|\mathbf{x}^\top h_{t,j}\right| \right\|_{\psi_1} \le C \left\|u^\top \mathbf{x}\right\|_{\psi_2} \left\|\mathbf{x}^\top h_{t,j}\right\|_{\psi_2} \le C \left\|h_{t,j}\right\|.$$

Centering increases the ψ_1 norm by at most a universal factor, so

$$\left\|u^\top \left(Z_i^{t,j} - \mathbb{E}Z_i^{t,j}\right)\right\|_{\psi_1} \le C \left\|h_{t,j}\right\|.$$

For the variance, centering can only decrease the second moment:

$$\begin{aligned}\mathrm{Var}\left(u^\top Z_i^{t,j}\right) &\le \mathbb{E}\left[(u^\top Z_i^{t,j})^2\right] \\ &\le \mathbb{E}\left[(u^\top \mathbf{x}_i)^2 (\mathbf{x}_i^\top h_{t,j})^2\right] \le 3\left\|h_{t,j}\right\|^2,\end{aligned}$$

where the last inequality is Isserlis' theorem.

Bernstein's inequality for centered sub-exponential variables now gives, for any $\tau > 0$, with probability at least $1 - 2e^{-\tau}$,

$$\left|\frac{1}{N}\sum_{i=1}^N u^\top \left(Z_i^{t,j} - \mathbb{E}Z_i^{t,j}\right)\right| \le C\left[\left\|h_{t,j}\right\|\sqrt{\frac{\tau}{N}} + \left\|h_{t,j}\right\|\frac{\tau}{N}\right].$$

If $N \ge C_1\tau$, the linear Bernstein term is absorbed into the square-root term:

$$\left|\frac{1}{N}\sum_{i=1}^N u^\top \left(Z_i^{t,j} - \mathbb{E}Z_i^{t,j}\right)\right| \le C\left\|h_{t,j}\right\|\sqrt{\frac{\tau}{N}}.$$

Let $\mathcal{N}$ be a $1/2$-net of $\mathbb{S}^{d-1}$ with $|\mathcal{N}| \le 5^d$, and set

$$\tau = \log\frac{2\cdot 5^d m(T+1)}{\delta}.$$

Union bounding over $u \in \mathcal{N}$, all columns j, and all path points $t = 0, \dots, T$ gives the preceding scalar bound simultaneously. For any $z \in \mathbb{R}^d$,

$$\|z\|_2 \le 2\sup_{u\in\mathcal{N}} \left|u^\top z\right|,$$

so, on the same event,

$$\left\|\nabla_{w_j}\mathcal{L}_N(\mathbf{W}_t) - \nabla_{w_j}\mathcal{L}_{\mathrm{pop}}(\mathbf{W}_t)\right\|_2 \le C\left\|h_{t,j}\right\|\sqrt{\frac{\log(2\cdot 5^d m(T+1)/\delta)}{N}}.$$

Summing the column bounds in quadrature yields, for every t,

$$\|\nabla\mathcal{L}_N(\mathbf{W}_t) - \nabla\mathcal{L}_{\mathrm{pop}}(\mathbf{W}_t)\|_{\mathrm{F}} \le C \left\|\mathbf{W}_t - \widetilde{\mathbf{W}}^\star\right\|_{\mathrm{F}} \sqrt{\frac{\log(2\cdot 5^d m(T+1)/\delta)}{N}}.$$

Finally,

$$\log\frac{2\cdot 5^d m(T+1)}{\delta} \le Cd\log\left(\frac{10dm(T+1)}{\delta}\right)$$

after enlarging the universal constant. Taking the maximum over t proves the stated bound.

A.5 Proof of Proposition 4.8 and a near-best bound

Proof. Since $D = uv^\top$ is a leading singular dyad with unit singular vectors, $\|D\|_{\mathrm{F}} = 1$ and $\langle A, D\rangle = \sigma$. Using $R = A - E$ and expanding the square gives

$$\begin{aligned}\|R\|_{\mathrm{F}}^2 - \|R - sD\|_{\mathrm{F}}^2 &= 2s\,\langle R, D\rangle - s^2\\ &= s\big(2\sigma - s - 2\,\langle E, D\rangle\big)\\ &\ge s\big(2\sigma - s - 2\,\|E\|_2\big),\end{aligned}$$

where $|\langle E, D\rangle| = \left|u^\top E v\right| \le \|E\|_2$ and $s \ge 0$. The right-hand side is positive under the stated strict inequality. Since $\|E\|_2 \le \|E\|_{\mathrm{F}}$, the same conclusion holds under $0 < s < 2(\sigma - \|E\|_{\mathrm{F}})$. □

The result is deterministic and permits a scalar fitted on the same batch as A. In the Leaky-ReLU setting, the proof of Proposition 4.5 gives $\left\|A_k - R_k^W\right\|_{\mathrm{F}} \le \eta_\alpha \left\|R_k^W\right\|_{\mathrm{F}}$ under sector control and the fresh-batch concentration event. Using that bound retains those assumptions; the deterministic progress inequality itself does not require them.

Near-best canonical update. Set $\delta = \|E\|_2$, $P_1(A) = \sigma D$, and $\mathrm{tail}_1(R) = (\sum_{j>1}\sigma_j(R)^2)^{1/2}$. For the canonical amplitude $s = \sigma$,

$$0 \le \|R - P_1(A)\|_{\mathrm{F}}^2 - \mathrm{tail}_1(R)^2 \le 4\sigma\delta + \delta^2.$$

Indeed, the optimal rank-one correction reduces squared Frobenius error by $\sigma_1(R)^2$. If $I = \|R\|_{\mathrm{F}}^2 - \|R - P_1(A)\|_{\mathrm{F}}^2$ is the achieved decrease, Proposition 4.8 gives $I \ge \sigma^2 - 2\sigma\delta$, while Weyl's inequality gives $\sigma_1(R) \le \sigma + \delta$. Hence

$$0 \le \sigma_1(R)^2 - I \le (\sigma+\delta)^2 - (\sigma^2 - 2\sigma\delta) = 4\sigma\delta + \delta^2,$$

which is the claimed bound. This comparison concerns squared parameter-residual error for $s = \sigma$, not arbitrary scalar refits or prediction loss.

A.6 Proof of Proposition 4.5 and statement and proof of Corollary A.1

We first prove Proposition 4.5, then state and prove Corollary A.1, which records the deterministic accounting one gets when a free-factor solver is compared to its direction-locked counterpart.

Proof of Proposition 4.5. Set

$$A_k = -\nabla\mathcal{L}_{\mathcal{B}_k}(\mathbf{W}^{\langle k\rangle})/\mu_\alpha.$$

Because $\widetilde{\mathbf{W}}^\star = \mathbf{W}_0 + \Delta$ and $\mathbf{W}^{\langle k\rangle} = \mathbf{W}_0 + \sum_{j<k} U_j$, we have

$$\mathbf{W}^{\langle k\rangle} - \widetilde{\mathbf{W}}^\star = -R_k^W.$$

Lemma 4.3 gives

$$-\nabla\mathcal{L}_{\mathrm{pop}}(\mathbf{W}^{\langle k\rangle})/\mu_\alpha = R_k^W - \Psi_\alpha(\mathbf{W}^{\langle k\rangle})/\mu_\alpha.$$

Since $\mu_\alpha > 0$, dividing the negative gradient by μ_α preserves its leading singular dyads. This normalization gives the residual unit coefficient in the population decomposition, calibrating A_k and its initial spectral amplitude $\sigma_1(A_k)$ to the parameter-residual scale without eliminating nonlinear

bias or sampling error. Theorem 4.1 applies to any matrix surrogate satisfying its oracle-relative error condition. Proposition 4.5 instantiates this principle with the normalized gradient: the displayed coefficient η_α and the value of the amplitude reference $\sigma_1(A_k)$ depend on this normalization. Other gradient-based surrogates require their own approximation bounds and, for update-error guarantees, amplitude-slack accounting relative to their leading singular value.

On the concentration event of Lemma 4.4,

$$\begin{aligned}\left\|A_k - R_k^W\right\|_{\mathrm{F}} &\le \frac{\left\|\Psi_\alpha(\mathbf{W}^{\langle k\rangle})\right\|_{\mathrm{F}}}{\mu_\alpha} + \frac{\left\|\nabla\mathcal{L}_{\mathcal{B}_k}(\mathbf{W}^{\langle k\rangle}) - \nabla\mathcal{L}_{\mathrm{pop}}(\mathbf{W}^{\langle k\rangle})\right\|_{\mathrm{F}}}{\mu_\alpha} \\ &\le \frac{(1-\alpha)^2 C_0\theta_0 + \gamma_N}{\mu_\alpha}\left\|R_k^W\right\|_{\mathrm{F}} = \eta_\alpha\left\|R_k^W\right\|_{\mathrm{F}}.\end{aligned}$$

Therefore, with

$$\varepsilon_k := \eta_\alpha\left\|R_k^W\right\|_{\mathrm{F}} + \left\|R_k^W - \Delta_k^\star\right\|_{\mathrm{F}},$$

the hypothesis gives

$$\|A_k - \Delta_k^\star\|_{\mathrm{F}} \le \varepsilon_k \le \mathcal{T}_k^\star/4.$$

Write $E_k = A_k - \Delta_k^\star$ and let $\sigma_{1,k} \ge \sigma_{2,k} \ge \cdots$ be the singular values of $\Delta_k^\star$. Let $\hat\sigma_{1,k} \ge \hat\sigma_{2,k} \ge \cdots$ be those of A_k. Weyl's inequality gives, for every ℓ,

$$|\hat\sigma_{\ell,k} - \sigma_{\ell,k}| \le \|E_k\|_2 \le \|E_k\|_{\mathrm{F}}.$$

In particular,

$$\hat\sigma_{1,k} - \hat\sigma_{2,k} \ge (\sigma_{1,k} - \sigma_{2,k}) - 2\|E_k\|_2 \ge \mathcal{T}_k^\star/2,$$

and $\hat\sigma_{1,k} \ge \sigma_{1,k} - \|E_k\|_2 \ge 3\sigma_{1,k}/4$ since $\mathcal{T}_k^\star \le \sigma_{1,k}$. Thus the leading singular direction remains simple and separated.

We now apply the rank-one form of Wedin's theorem. Equivalently, apply the Davis-Kahan theorem to the symmetric dilation

$$\mathcal{D}(M) = \begin{bmatrix} 0 & M \\ M^\top & 0 \end{bmatrix}.$$

The perturbation satisfies $\|\mathcal{D}(E_k)\|_2 = \|E_k\|_2 \le \|E_k\|_{\mathrm{F}}$, while the relevant eigenvalue gap of $\mathcal{D}(\Delta_k^\star)$ around its leading positive eigenvalue is at least $\mathcal{T}_k^\star = \sigma_{1,k} - \sigma_{2,k}$. Since $\|E_k\|_2 \le \mathcal{T}_k^\star/4$, the one-dimensional left and right singular subspaces obey

$$\sin\angle(\hat u_k, u_{1,k}) + \sin\angle(\hat v_k, v_{1,k}) \le C\frac{\|E_k\|_{\mathrm{F}}}{\mathcal{T}_k^\star}.$$

Choosing the joint signs so that the products are compared consistently,

$$\|\hat u_k - u_{1,k}\|_2 \le 2\sin\angle(\hat u_k, u_{1,k}), \qquad \|\hat v_k - v_{1,k}\|_2 \le 2\sin\angle(\hat v_k, v_{1,k}).$$

Therefore

$$\begin{aligned}\left\|\hat u_k\hat v_k^\top - u_{1,k}v_{1,k}^\top\right\|_{\mathrm{F}} &\le \left\|(\hat u_k - u_{1,k})\hat v_k^\top\right\|_{\mathrm{F}} + \left\|u_{1,k}(\hat v_k - v_{1,k})^\top\right\|_{\mathrm{F}} \\ &\le C\frac{\|E_k\|_{\mathrm{F}}}{\mathcal{T}_k^\star} \le C\frac{\varepsilon_k}{\mathcal{T}_k^\star}.\end{aligned}$$

This proves the stated dyad bound.

It remains to convert dyad error into update error. The algorithm returns a scalar s_k satisfying

$$\varepsilon_{k,\mathrm{amp}} = |s_k - \sigma_1(A_k)| = |s_k - \hat\sigma_{1,k}|.$$

Weyl's inequality also gives

$$|\hat\sigma_{1,k} - \sigma_{1,k}| \le \|E_k\|_2 \le \varepsilon_k,$$

so

$$|s_k - \sigma_{1,k}| \le \varepsilon_{k,\mathrm{amp}} + \varepsilon_k.$$

Using $U_k^\star = \sigma_{1,k}u_{1,k}v_{1,k}^\top$,

$$\begin{aligned}\|U_k - U_k^\star\|_{\mathrm{F}} &= \left\|s_k\hat u_k\hat v_k^\top - \sigma_{1,k}u_{1,k}v_{1,k}^\top\right\|_{\mathrm{F}} \\ &\le |s_k - \sigma_{1,k}| + \sigma_{1,k}\left\|\hat u_k\hat v_k^\top - u_{1,k}v_{1,k}^\top\right\|_{\mathrm{F}} \\ &\le \varepsilon_{k,\mathrm{amp}} + \varepsilon_k + C\sigma_{1,k}\frac{\varepsilon_k}{\mathcal{T}_k^\star}.\end{aligned}$$

Since $\mathcal{T}_k^\star \le \sigma_{1,k}$, the additive ε_k term is absorbed into the last term after increasing the universal constant. This yields

$$\|U_k - U_k^\star\|_{\mathrm{F}} \le \varepsilon_{k,\mathrm{amp}} + C' \frac{\varepsilon_k}{\mathcal{T}_k^\star} \sigma_1(\Delta_k^\star),$$

which is the claimed stage bound.

Scalar refinement toward the oracle amplitude. With E_k and ε_k as above, if s_k lies between $\sigma_1(A_k)$ and $\sigma_1(\Delta_k^\star)$, Weyl's inequality gives

$$\varepsilon_{k,\mathrm{amp}} \le |\sigma_1(A_k) - \sigma_1(\Delta_k^\star)| \le \|E_k\|_2 \le \varepsilon_k.$$

Since $\mathcal{T}_k^\star \le \sigma_1(\Delta_k^\star)$, the stage bound then yields

$$\|U_k - U_k^\star\|_{\mathrm{F}} \le (C' + 1) \frac{\varepsilon_k}{\mathcal{T}_k^\star} \sigma_1(\Delta_k^\star).$$

Under this interval hypothesis, the amplitude deviation is controlled by the same surrogate-error bound. It can nevertheless increase from zero while empirical loss decreases. Minimizing the nonlinear empirical loss does not enforce the interval hypothesis, so arbitrary refinements retain the separate amplitude-slack term.

Corollary A.1 (Accounting for a rotated rank-one stage solver)**.** *At stage k of an actual free-factor path, construct the reference direction-locked update $\bar{U}_k = s_k \hat{u}_k \hat{v}_k^\top$ from the same current iterate $\mathbf{W}^{\langle k \rangle}$ and the same stage batch $\mathcal{B}_k$ used by the free-factor solver. Suppose Proposition 4.5 applies at this iterate and batch. If the free-factor solver returns $\widetilde{U}_k = b_k a_k^\top$, define $\varepsilon_{k,\mathrm{rot}} = \left\|\widetilde{U}_k - \bar{U}_k\right\|_{\mathrm{F}}$. Then*

$$\left\|\widetilde{U}_k - U_k^\star\right\|_{\mathrm{F}} \le \varepsilon_{k,\mathrm{rot}} + \left\|\bar{U}_k - U_k^\star\right\|_{\mathrm{F}},$$

where the second term is bounded by Proposition 4.5. Consequently, if Proposition 4.5 holds at every stage along the actual free-factor path, Proposition 4.6 applies to the free-factor sequence with the additional sum $\sum_{k=1}^r \varepsilon_{k,\mathrm{rot}}$. This does not prove that $\varepsilon_{k,\mathrm{rot}}$ is small; it only makes the missing solver-stability term explicit. The hypothesis that Proposition 4.5 applies along the actual free-factor path is itself a precondition: free-factor iterates differ from direction-locked iterates, and we do not separately analyze when the sector and perturbation-budget conditions continue to hold along that path.

Proof of Corollary A.1. The result is deterministic. At a single stage, insert the direction-locked reference update $\bar{U}_k$ and use the triangle inequality:

$$\left\|\widetilde{U}_k - U_k^\star\right\|_{\mathrm{F}} \le \left\|\widetilde{U}_k - \bar{U}_k\right\|_{\mathrm{F}} + \left\|\bar{U}_k - U_k^\star\right\|_{\mathrm{F}}.$$

The first term is the definition of $\varepsilon_{k,\mathrm{rot}}$. The second term is exactly the direction-locked stage error controlled by Proposition 4.5. Repeating the same insertion for each stage and applying Proposition 4.6 gives the same additive end-to-end bound with the added term

$$\sum_{k=1}^r \varepsilon_{k,\mathrm{rot}}.$$

No probabilistic or optimization claim about the size of $\varepsilon_{k,\mathrm{rot}}$ is used.

A.7 Statement and proof of Proposition A.2

The same surrogate-to-dyad principle applies to smooth matched activations through oracle Hessian preconditioning. This is a local, analysis-level analogue: it requires oracle Hessian preconditioning, is not used in the all-stage theorem, and is included to scope the surrogate principle beyond Leaky-ReLU.

Proposition A.2 (Smooth matched Hessian-preconditioned surrogate)**.** *Consider the one-matrix teacher-student model*

$$y_j = \phi(x^\top \widetilde{w}_j^\star), \qquad \mathcal{L}(\mathbf{W}) = \frac{1}{2} \sum_{j=1}^m \mathbb{E}\left[\left(\phi(x^\top w_j) - \phi(x^\top \widetilde{w}_j^\star)\right)^2\right],$$

with $\mathcal{L}_{\mathcal{B}}$ denoting the corresponding fresh-batch empirical loss. Assume the matched activation satisfies $|\phi'| \le B_1$ and ϕ' is B_2-Lipschitz. Let $R = \widetilde{\mathbf{W}}^\star - \mathbf{W}$, $r_j = \widetilde{w}_j^\star - w_j$, and assume

$$\sup_{\|a\|=\|b\|=\|u\|=1} \mathbb{E}\left|(a^\top x)(b^\top x)(u^\top x)\right| \le M_3.$$

Define

$$H_j^\star = \mathbb{E}[\phi'(x^\top \widetilde{w}_j^\star)^2 x x^\top], \qquad \mathcal{P}_\star(G) = [(H_1^\star)^{-1} g_1, \dots, (H_m^\star)^{-1} g_m],$$

and suppose $H_j^\star \succeq \lambda_{\min} I$ for all j, with $\lambda_{\min} > 0$. If a fresh stage batch is drawn after $\mathbf{W}$ is fixed and $\left\|x^\top a\right\|_{\psi_2} \le K \|a\|$, then with probability at least $1-\delta$,

$$\|\mathcal{P}_\star(-\nabla \mathcal{L}_{\mathcal{B}}(\mathbf{W})) - R\|_{\mathrm{F}} \le \left[\frac{2B_1 B_2 M_3}{\lambda_{\min}} \max_j \|r_j\| + \frac{\gamma_{\mathrm{sm}}}{\lambda_{\min}}\right] \|R\|_{\mathrm{F}},$$

where

$$\gamma_{\mathrm{sm}} = C B_1^2 K^2 \left[\sqrt{\frac{dm + \log(2/\delta)}{|\mathcal{B}|}} + \frac{dm + \log(2/\delta)}{|\mathcal{B}|}\right].$$

Proof of Proposition A.2. For a single column, write $w = w_j$, $\widetilde{w}^\star = \widetilde{w}_j^\star$, and $e = w - \widetilde{w}^\star = -r_j$. Since $|\phi'| \le B_1$, the loss is differentiable under the expectation by dominated convergence. With $z_t = x^\top(\widetilde{w}^\star + te)$, the fundamental theorem of calculus gives

$$\phi(x^\top w) - \phi(x^\top \widetilde{w}^\star) = \int_0^1 \phi'(z_t) x^\top e \, dt.$$

Hence

$$\nabla_{w_j} \mathcal{L}(\mathbf{W}) = H_j(\mathbf{W}, \widetilde{\mathbf{W}}^\star)(w_j - \widetilde{w}_j^\star),$$

where

$$H_j(\mathbf{W}, \widetilde{\mathbf{W}}^\star) = \mathbb{E}\left[x x^\top \int_0^1 \phi'(x^\top(\widetilde{w}_j^\star + t(w_j - \widetilde{w}_j^\star)))\phi'(x^\top w_j) dt\right].$$

For unit vectors a, b, Lipschitzness of ϕ' gives

$$\begin{aligned}
&\left|a^\top (H_j(\mathbf{W}, \widetilde{\mathbf{W}}^\star) - H_j^\star) b\right| \\
&\le \mathbb{E}\left[\left|a^\top x\right| \left|b^\top x\right| \int_0^1 \left|\phi'(z_t)\phi'(z_1) - \phi'(z_0)^2\right| dt\right] \\
&\le 2B_1 B_2 \mathbb{E}\left|(a^\top x)(b^\top x)(x^\top e)\right| \le 2B_1 B_2 M_3 \|e\|.
\end{aligned}$$

Thus

$$-\nabla_{w_j} \mathcal{L}(\mathbf{W}) = H_j^\star r_j - q_j, \qquad \|q_j\| \le 2B_1 B_2 M_3 \|r_j\|^2.$$

Applying $(H_j^\star)^{-1}$ columnwise and summing over j gives the population bound

$$\|\mathcal{P}_\star(-\nabla \mathcal{L}(\mathbf{W})) - R\|_{\mathrm{F}} \le \frac{2B_1 B_2 M_3}{\lambda_{\min}} \left(\max_j \|r_j\|\right) \|R\|_{\mathrm{F}}.$$

It remains to control the empirical gradient. For any fixed matrix $U = [u_1, \dots, u_m]$ with $\|U\|_{\mathrm{F}} = 1$, the single-sample directional gradient satisfies

$$|\langle \nabla \ell_i(\mathbf{W}), U\rangle| \le B_1^2 \sum_{j=1}^m \left|x_i^\top r_j\right| \left|x_i^\top u_j\right|.$$

The product of two sub-Gaussian random variables is sub-exponential, so

$$\left\|(x_i^\top r_j)(x_i^\top u_j)\right\|_{\psi_1} \le C K^2 \|r_j\| \|u_j\|.$$

By the triangle inequality for ψ_1 norms and Cauchy–Schwarz over columns,

$$\|\langle \nabla \ell_i(\mathbf{W}) - \nabla \mathcal{L}(\mathbf{W}), U\rangle\|_{\psi_1} \le C B_1^2 K^2 \|R\|_{\mathrm{F}}.$$

Bernstein's inequality for each fixed U, followed by a $1/2$-net of the Frobenius unit sphere in $\mathbb{R}^{d\times m}$ with cardinality at most 5^{dm}, gives with probability at least $1-\delta$,

$$\|\nabla \mathcal{L}_{\mathcal{B}}(\mathbf{W}) - \nabla \mathcal{L}(\mathbf{W})\|_{\mathrm{F}} \le \gamma_{\mathrm{sm}} \|R\|_{\mathrm{F}}.$$

Applying $\mathcal{P}_\star$ to this deviation costs a factor $\lambda_{\min}^{-1}$. Combining the empirical deviation with the population bound proves Proposition A.2. □

A.8 Proof of Proposition 4.6 and Proposition 4.7

Let

$$D_k = R_k^W - \Delta_k^\star.$$

Using

$$R_k^W = \Delta - \sum_{j<k} U_j, \qquad \Delta_k^\star = \Delta - \sum_{j<k} U_j^\star,$$

we get

$$D_k = -\sum_{j<k} (U_j - U_j^\star).$$

After r stages,

$$\Delta - \widehat{\Delta}^{(r)} = \Delta_{r+1}^\star + D_{r+1}.$$

The triangle inequality gives

$$\left\| \Delta - \widehat{\Delta}^{(r)} \right\|_{\mathrm{F}} \le \left\| \Delta_{r+1}^\star \right\|_{\mathrm{F}} + \sum_{k=1}^{r} \|U_k - U_k^\star\|_{\mathrm{F}}.$$

It remains to identify the oracle tail. If $\Delta = \sum_{\ell \ge 1} \sigma_\ell u_\ell v_\ell^\top$ is an SVD, then successive exact best rank-1 deflation removes $\sigma_1 u_1 v_1^\top, \dots, \sigma_r u_r v_r^\top$ (with an arbitrary choice inside tied singular subspaces). Hence $\Delta_{r+1}^\star = \sum_{\ell > r} \sigma_\ell u_\ell v_\ell^\top$, and $\left\| \Delta_{r+1}^\star \right\|_{\mathrm{F}} = (\sum_{j>r} \sigma_j(\Delta)^2)^{1/2}$.

For any input $\mathbf{x}$, the coordinatewise Leaky-ReLU map is 1-Lipschitz for $\alpha \in [0, 1]$ in Euclidean norm:

$$\left\| \sigma_\alpha(\mathbf{x}^\top \mathbf{W}) - \sigma_\alpha(\mathbf{x}^\top \widetilde{\mathbf{W}}^\star) \right\|_2 \le \left\| \mathbf{x}^\top (\mathbf{W} - \widetilde{\mathbf{W}}^\star) \right\|_2.$$

Taking expectation over $\mathbf{x} \sim \mathcal{N}(0, I_d)$ gives

$$\mathcal{L}_{\mathrm{pop}}(\mathbf{W}) \le \frac{1}{2} \mathbb{E} \left\| \mathbf{x}^\top (\mathbf{W} - \widetilde{\mathbf{W}}^\star) \right\|_2^2 = \frac{1}{2} \left\| \mathbf{W} - \widetilde{\mathbf{W}}^\star \right\|_{\mathrm{F}}^2.$$

The last equality follows from

$$\mathbb{E} \left\| \mathbf{x}^\top A \right\|_2^2 = \mathbb{E} \operatorname{tr}(A^\top \mathbf{x}\mathbf{x}^\top A) = \operatorname{tr}(A^\top A) = \|A\|_{\mathrm{F}}^2$$

with $A = \mathbf{W} - \widetilde{\mathbf{W}}^\star$.

A.9 Statement and proof of Corollary A.3

Corollary A.3 (Quantitative fixed-direction scalar refitting). *Fix dyads $\widehat{D}_k = \hat{u}_k \hat{v}_k^\top$, $k = 1, \dots, r$, and define $\mathbf{W}(c) = \mathbf{W}_0 + \sum_{k=1}^r c_k \widehat{D}_k$. Let $s_k^\star = \sigma_1(\Delta_k^\star)$ and $D_k^\star = U_k^\star / s_k^\star$ when $s_k^\star > 0$; terms with $s_k^\star = 0$ are omitted. If $\hat{c}$ is an ε-optimal population scalar refit,*

$$\mathcal{L}_{\mathrm{pop}}(\mathbf{W}(\hat{c})) \le \inf_{c \in \mathbb{R}^r} \mathcal{L}_{\mathrm{pop}}(\mathbf{W}(c)) + \varepsilon,$$

then

$$\mathcal{L}_{\mathrm{pop}}(\mathbf{W}(\hat{c})) \le \frac{1}{2} \left(\left(\sum_{j>r} \sigma_j(\Delta)^2 \right)^{1/2} + \sum_{k=1}^{r} s_k^\star \left\| \widehat{D}_k - D_k^\star \right\|_{\mathrm{F}} \right)^2 + \varepsilon.$$

An ε-optimal scalar refit therefore controls fitted loss for fixed directions; it does not imply better directions or smaller parameter error.

Proof of Corollary A.3. Let $s^\star = (s_1^\star, \dots, s_r^\star)$ be the oracle amplitude vector. By ε-optimality,

$$\mathcal{L}_{\mathrm{pop}}(\mathbf{W}(\hat{c})) \le \mathcal{L}_{\mathrm{pop}}(\mathbf{W}(s^\star)) + \varepsilon.$$

Using $\Delta = \Delta_{r+1}^\star + \sum_{k=1}^r s_k^\star D_k^\star$,

$$\begin{aligned}\left\|\mathbf{W}(s^\star) - \widetilde{\mathbf{W}}^\star\right\|_{\mathrm{F}} &= \left\|\sum_{k=1}^r s_k^\star \widehat{D}_k - \Delta\right\|_{\mathrm{F}} \\ &= \left\|-\Delta_{r+1}^\star + \sum_{k=1}^r s_k^\star(\widehat{D}_k - D_k^\star)\right\|_{\mathrm{F}} \\ &\le \left\|\Delta_{r+1}^\star\right\|_{\mathrm{F}} + \sum_{k=1}^r s_k^\star \left\|\widehat{D}_k - D_k^\star\right\|_{\mathrm{F}}.\end{aligned}$$

The first term is the oracle rank-r tail. Applying Proposition 4.7 to $\mathbf{W}(s^\star)$ gives the displayed population-loss bound. Finally, for any fixed scalar vector s, that vector is feasible for the same scalar-refit problem, so an ε-optimal refit for either $\mathcal{L}_{\mathrm{pop}}$ or an empirical refit-batch loss satisfies $\mathcal{L}(\mathbf{W}(\hat{c})) \le \mathcal{L}(\mathbf{W}(s)) + \varepsilon$ for that objective. The argument uses only feasibility of comparator scalars and does not assert that the fixed dyads themselves are accurate.

A.10 Statement and proof of Theorem A.4

The conservative multi-stage certificate combines the per-stage perturbation bound (Proposition 4.5) with the additive propagation step (Proposition 4.6) into a closed end-to-end statement for the Gaussian Leaky-ReLU instantiation.

Theorem A.4 (Conservative all-stage error certificate)**.** *Assume $\mathcal{T}_k^\star > 0$ for $k \le r$, and assume the sector condition of Lemma 4.3 holds at every iterate $\mathbf{W}^{\langle k\rangle}$ with the same angle bound θ_0. For each stage, let $\gamma_k \ge 0$ and define*

$$E_k = \left\{\left\|\nabla\mathcal{L}_{\mathcal{B}_k}(\mathbf{W}^{\langle k\rangle}) - \nabla\mathcal{L}_{\mathrm{pop}}(\mathbf{W}^{\langle k\rangle})\right\|_{\mathrm{F}} \le \gamma_k \left\|R_k^W\right\|_{\mathrm{F}}\right\}.$$

Set

$$\eta_k = \frac{(1-\alpha)^2 C_0\theta_0 + \gamma_k}{\mu_\alpha}, \quad s_k^\star = \sigma_1(\Delta_k^\star), \quad \rho_k^\star = \frac{s_k^\star}{\mathcal{T}_k^\star}, \quad \kappa_k^\star = \frac{\|\Delta_k^\star\|_{\mathrm{F}}}{\mathcal{T}_k^\star},$$

and let $\varepsilon_{k,\mathrm{amp}} = \left|s_k - \sigma_1(-\nabla\mathcal{L}_{\mathcal{B}_k}(\mathbf{W}^{\langle k\rangle})/\mu_\alpha)\right|$. Define $d_1 = 0$ and

$$d_{k+1} = \big(1 + C'\rho_k^\star(1+\eta_k)\big)d_k + \varepsilon_{k,\mathrm{amp}} + C'\eta_k\kappa_k^\star s_k^\star.$$

If, for every $k \le r$,

$$\eta_k\left(\kappa_k^\star + \frac{d_k}{\mathcal{T}_k^\star}\right) + \frac{d_k}{\mathcal{T}_k^\star} \le \frac{1}{4},$$

then on $E = \cap_{k=1}^r E_k$,

$$\left\|R_k^W - \Delta_k^\star\right\|_{\mathrm{F}} \le d_k, \qquad \|U_k - U_k^\star\|_{\mathrm{F}} \le d_{k+1} - d_k \quad (k \le r).$$

Consequently,

$$\left\|\Delta - \widehat{\Delta}^{(r)}\right\|_{\mathrm{F}} \le \left(\sum_{j>r}\sigma_j(\Delta)^2\right)^{1/2} + d_{r+1},$$

and

$$\mathcal{L}_{\mathrm{pop}}(\mathbf{W}_0 + \widehat{\Delta}^{(r)}) \le \frac{1}{2}\left[\left(\sum_{j>r}\sigma_j(\Delta)^2\right)^{1/2} + d_{r+1}\right]^2.$$

Reading the recurrence. The amplification factor $1 + C'\rho_k^\star(1+\eta_k)$ in front of d_k records how much past parameter error the current stage inherits, while the additive terms $\varepsilon_{k,\mathrm{amp}} + C'\eta_k\kappa_k^\star s_k^\star$ record what this stage contributes from scalar slack and nonlinearity bias. The sufficient condition $\eta_k(\kappa_k^\star + d_k/\mathcal{T}_k^\star) + d_k/\mathcal{T}_k^\star \le 1/4$ requires the combined nonlinearity-plus-sampling slack η_k and the relative past mismatch $d_k/\mathcal{T}_k^\star$ to remain a constant fraction of the current residual gap. Because past mismatch enters in Frobenius norm, a stage that consumes a sizable fraction of its gap budget tightens every later stage's margin: this is why the certificate is conservative and not a generic convergence theorem.

Status of $\varepsilon_{k,\mathrm{amp}}$. The scalar slack $\varepsilon_{k,\mathrm{amp}} = |s_k - \sigma_1(A_k)|$ enters the recurrence as an additive per-stage term but is not bounded by Algorithm 1 in general. When $t_k = 0$, the algorithm sets $s_k = \sigma_1(A_k)$ and $\varepsilon_{k,\mathrm{amp}} = 0$ exactly, so the theorem applies with the slack removed. When $t_k > 0$, scalar refinement on $\mathcal{L}_{\mathcal{B}_k}(\mathbf{W}^{\langle k\rangle} + s_k\hat{u}_k\hat{v}_k^\top)$ may move s_k away from $\sigma_1(A_k)$; the theorem treats $\varepsilon_{k,\mathrm{amp}}$ as a per-stage measured quantity (reported by the diagnostic $|s_k - \sigma_1(A_k)|$ in Appendix B), and the budget β_k in Corollary A.6 accommodates any user-supplied tolerance. An analytical bound on $\varepsilon_{k,\mathrm{amp}}$ for the refined scalar is outside the scope of the local theorem and depends on the specific scalar solver.

Proof of Theorem A.4. For each stage let $D_k = R_k^W - \Delta_k^\star$. Since $R_1^W = \Delta = \Delta_1^\star$, we have $D_1 = 0 = d_1$. Work on $E = \cap_{k=1}^r E_k$ and suppose inductively that $\|D_k\|_{\mathrm{F}} \le d_k$. Then

$$\frac{\|R_k^W\|_{\mathrm{F}}}{\mathcal{T}_k^\star} \le \frac{\|\Delta_k^\star\|_{\mathrm{F}} + d_k}{\mathcal{T}_k^\star} = \kappa_k^\star + \frac{d_k}{\mathcal{T}_k^\star}, \qquad \frac{\|R_k^W - \Delta_k^\star\|_{\mathrm{F}}}{\mathcal{T}_k^\star} \le \frac{d_k}{\mathcal{T}_k^\star}.$$

The envelope condition therefore implies the sufficient condition in Proposition 4.5. Applying that proposition at stage k gives

$$\begin{aligned}\|U_k - U_k^\star\|_{\mathrm{F}} &\le \varepsilon_{k,\mathrm{amp}} + C'\left[\eta_k\left(\kappa_k^\star + \frac{d_k}{\mathcal{T}_k^\star}\right) + \frac{d_k}{\mathcal{T}_k^\star}\right] s_k^\star \\ &= \varepsilon_{k,\mathrm{amp}} + C'\eta_k\kappa_k^\star s_k^\star + C'\rho_k^\star(1+\eta_k)d_k = d_{k+1} - d_k.\end{aligned}$$

Moreover,

$$D_{k+1} = R_{k+1}^W - \Delta_{k+1}^\star = (R_k^W - U_k) - (\Delta_k^\star - U_k^\star) = D_k - (U_k - U_k^\star),$$

so

$$\|D_{k+1}\|_{\mathrm{F}} \le \|D_k\|_{\mathrm{F}} + \|U_k - U_k^\star\|_{\mathrm{F}} \le d_k + (d_{k+1} - d_k) = d_{k+1}.$$

This closes the induction and proves the simultaneous residual-mismatch and stage-error bounds. The final parameter bound follows from Proposition 4.6 and the telescoping estimate

$$\sum_{k=1}^r \|U_k - U_k^\star\|_{\mathrm{F}} \le \sum_{k=1}^r (d_{k+1} - d_k) = d_{r+1}.$$

The population-prediction bound is then Proposition 4.7.

A.11 Statements and proofs of Corollaries A.5 and A.6

The first corollary specializes Theorem A.4 to a uniform fresh-batch setting and certifies sector control along the path; the second converts the per-stage feasibility condition into an oracle certification margin.

Corollary A.5 (Uniform fresh-batch all-stage bound with sector certification)**.** *Assume the setting of Theorem A.4 except possibly its sector condition. Suppose that, for all $k \le r$,*

$$\kappa_k^\star \le \kappa_\star, \quad \rho_k^\star \le \rho_\star, \quad s_k^\star \le s_\star, \quad \mathcal{T}_k^\star \ge \tau_\star > 0, \quad \varepsilon_{k,\mathrm{amp}} \le \varepsilon_{\mathrm{amp}}.$$

Assume also

$$m_{\min} := \min_{j:\widetilde{w}_j^\star \ne 0} \left\|\widetilde{w}_j^\star\right\|_2 > 0.$$

Each stage uses an independent fresh batch of size N, with failure probability δ/r per stage. Let

$$\gamma = C_2\sqrt{\frac{d\log\left(\frac{10dmr}{\delta}\right)}{N}}, \qquad \eta = \frac{(1-\alpha)^2 C_0\theta_0 + \gamma}{\mu_\alpha},$$

and define $a = 1 + C'\rho_\star(1+\eta)$, $b = \varepsilon_{\mathrm{amp}} + C'\eta\kappa_\star s_\star$, *and*

$$\bar{d}_1 = 0, \qquad \bar{d}_{k+1} = a\bar{d}_k + b.$$

Since $\rho_\star \geq 1$ *and* $\eta \geq 0$, *we have* $a > 1$, *and hence*

$$\bar{d}_k = b\frac{a^{k-1}-1}{a-1}.$$

If $N \geq C_1 d \log\left(\frac{10dmr}{\delta}\right)$ *and*

$$\eta\left(\kappa_\star + \frac{\bar{d}_r}{\tau_\star}\right) + \frac{\bar{d}_r}{\tau_\star} \leq \frac{1}{4},$$

and

$$\max_{1\leq k\leq r}\left\{\|\Delta_k^\star\|_{\mathrm{F}} + \bar{d}_k\right\} \leq m_{\min}\sin\theta_0,$$

then with probability at least $1-\delta$,

$$\left\|R_k^W - \Delta_k^\star\right\|_{\mathrm{F}} \leq \bar{d}_k \quad (k \leq r), \qquad \left\|\Delta - \widehat{\Delta}^{(r)}\right\|_{\mathrm{F}} \leq \left(\sum_{j>r}\sigma_j(\Delta)^2\right)^{1/2} + \bar{d}_{r+1},$$

and the same squared right-hand side times $1/2$ *bounds* $\mathcal{L}_{\mathrm{pop}}(\mathbf{W}_0 + \widehat{\Delta}^{(r)})$. *The displayed maximum certifies the sector condition at every stage via Lemma C.1.*

Feasibility and certified horizon. One nonempty feasible regime follows directly from the displayed conditions. Fix the problem geometry and a finite horizon r with positive oracle gaps and $\tau_\star > 0$. Suppose the strict sufficient parameter-closeness condition $\|\Delta\|_{\mathrm{F}} < m_{\min}\sin\theta_0$ holds for $\theta_0 \in (0, \pi/2)$. With spectral amplitudes ($t_k = 0$), we have $\varepsilon_{\mathrm{amp}} = 0$. As $\alpha \to 1$ from below and N increases, η and b tend to zero while a remains bounded. For this fixed r, therefore, $\bar{d}_r \to 0$. Since $\|\Delta_k^\star\|_{\mathrm{F}} \leq \|\Delta\|_{\mathrm{F}}$, both the gap and sector checks above hold for sufficiently near-linear $\alpha < 1$ and sufficiently large finite N. Sufficiently small positive amplitude slack can also be accommodated. This gives a local, near-linear feasible regime for the original fresh-batch certificate. The induction bounds the next mismatch by adding the norms of the past mismatch and the current update error. This allows worst-case alignment and does not exploit cancellation between them. Conversely, with fixed $a > 1$ and $b > 0$, the envelope grows geometrically and eventually exceeds any fixed positive error allowance. Thus persistent error injection limits the certified horizon, although a prescribed finite horizon can remain feasible when the errors are small enough. Exhausting this allowance means that the certificate ceases to apply, not that the algorithm necessarily stops making useful residual progress.

Corollary A.6 (Stage-wise sample and sector certification budgets)**.** *Assume the setting of Theorem A.4. Let* $\delta_1, \ldots, \delta_r > 0$ *with* $\sum_k \delta_k \leq \delta$, *and fix scalar amplitude budgets* $\beta_k \geq 0$. *Define* $d_1 = 0$ *and, before stage* k, *set*

$$e_k = \frac{d_k}{\mathcal{T}_k^\star}, \qquad M_k = \frac{\mu_\alpha(1/4 - e_k)}{\kappa_k^\star + e_k} - (1-\alpha)^2 C_0\theta_0.$$

If $M_k \leq 0$, *the all-stage certificate cannot certify another greedy direction at stage* k *from sample size alone. If* $M_k > 0$ *and the fresh stage batch size satisfies*

$$N_k \geq C_1 d\log\left(\frac{10dm}{\delta_k}\right), \qquad N_k \geq C\frac{d\log\left(\frac{10dm}{\delta_k}\right)}{M_k^2},$$

then Lemma 4.4 gives the stage event E_k *with*

$$\gamma_k = C_2\sqrt{\frac{d\log\left(\frac{10dm}{\delta_k}\right)}{N_k}} \leq M_k.$$

If also $\varepsilon_{k,\mathrm{amp}} \leq \beta_k$, *update the envelope using*

$$\eta_k = \frac{(1-\alpha)^2 C_0\theta_0 + \gamma_k}{\mu_\alpha}, \qquad d_{k+1} = \left(1 + C'\rho_k^\star(1+\eta_k)\right)d_k + \beta_k + C'\eta_k\kappa_k^\star s_k^\star.$$

If the margin is positive at every stage, then with probability at least $1-\delta$ the conclusions of Theorem A.4 hold with this envelope. Moreover, if $m_{\min} = \min_{j:\widetilde{w}_j^\star \neq 0} \left\|\widetilde{w}_j^\star\right\|_2 > 0$ and $\|\Delta_k^\star\|_{\mathrm{F}} + d_k \leq m_{\min} \sin\theta_0$ for every stage, then the sector assumption in Theorem A.4 is certified by Lemma C.1. This is an oracle certification rule: it uses $\mathcal{T}_k^\star, \kappa_k^\star, \rho_k^\star$, and d_k, and is not an observable optimal stopping theorem.

Proofs. At each stage, apply Lemma 4.4 conditionally at the fixed current iterate with a fresh batch and failure probability δ/r. Since $N \geq C_1 d \log\left(\frac{10dmr}{\delta}\right)$, the concentration events

$$E_k = \left\{ \left\| \nabla \mathcal{L}_{\mathcal{B}_k}(\mathbf{W}^{\langle k \rangle}) - \nabla \mathcal{L}_{\mathrm{pop}}(\mathbf{W}^{\langle k \rangle}) \right\|_{\mathrm{F}} \leq \gamma \left\| R_k^W \right\|_{\mathrm{F}} \right\}$$

hold simultaneously with probability at least $1-\delta$, with

$$\gamma = C_2 \sqrt{\frac{d \log\left(\frac{10dmr}{\delta}\right)}{N}}.$$

Work on this event. We prove by induction that

$$\left\| R_k^W - \Delta_k^\star \right\|_{\mathrm{F}} \leq \bar{d}_k \qquad (k = 1, \ldots, r).$$

The base case holds because $R_1^W = \Delta = \Delta_1^\star$ and $\bar{d}_1 = 0$. Assume the claim holds at stage k. Then

$$\left\| R_k^W \right\|_{\mathrm{F}} \leq \|\Delta_k^\star\|_{\mathrm{F}} + \bar{d}_k \leq m_{\min} \sin\theta_0,$$

where the last inequality is the displayed sector-certification assumption. By Lemma C.1, the sector condition of Lemma 4.3 holds at $\mathbf{W}^{\langle k \rangle}$. Also

$$\frac{\left\| R_k^W \right\|_{\mathrm{F}}}{\mathcal{T}_k^\star} \leq \kappa_k^\star + \frac{\bar{d}_k}{\mathcal{T}_k^\star} \leq \kappa_\star + \frac{\bar{d}_r}{\tau_\star}, \qquad \frac{\left\| R_k^W - \Delta_k^\star \right\|_{\mathrm{F}}}{\mathcal{T}_k^\star} \leq \frac{\bar{d}_r}{\tau_\star}.$$

Thus the displayed uniform condition implies the sufficient condition in Proposition 4.5. Applying Proposition 4.5 with $\eta_k = \eta$ gives

$$\|U_k - U_k^\star\|_{\mathrm{F}} \leq \varepsilon_{\mathrm{amp}} + C'\eta\kappa_\star s_\star + C'\rho_\star(1+\eta)\bar{d}_k = \bar{d}_{k+1} - \bar{d}_k.$$

Since

$$R_{k+1}^W - \Delta_{k+1}^\star = (R_k^W - \Delta_k^\star) - (U_k - U_k^\star),$$

we get

$$\left\| R_{k+1}^W - \Delta_{k+1}^\star \right\|_{\mathrm{F}} \leq \bar{d}_k + (\bar{d}_{k+1} - \bar{d}_k) = \bar{d}_{k+1}.$$

This closes the simultaneous induction: before each application of Proposition 4.5, sector control has already been certified. The parameter-error bound follows from Proposition 4.6 and the telescoping sum

$$\sum_{k=1}^{r} \|U_k - U_k^\star\|_{\mathrm{F}} \leq \sum_{k=1}^{r} (\bar{d}_{k+1} - \bar{d}_k) = \bar{d}_{r+1}.$$

Finally, Proposition 4.7 gives the population-loss bound.

At stage k, Lemma 4.4 applied conditionally at the fixed current iterate with failure probability δ_k gives

$$\gamma_{N_k} \leq C_2 \sqrt{\frac{d \log\left(\frac{10dm}{\delta_k}\right)}{N_k}},$$

provided $N_k \geq C_1 d \log\left(\frac{10dm}{\delta_k}\right)$. The displayed lower bound on N_k and a large enough universal constant C ensure $\gamma_{N_k} \leq \gamma_k \leq M_k$. By the definitions of η_k, M_k, and $e_k = d_k/\mathcal{T}_k^\star$,

$$\begin{aligned} \eta_k(\kappa_k^\star + e_k) + e_k &\leq \frac{(1-\alpha)^2 C_0 \theta_0 + M_k}{\mu_\alpha}(\kappa_k^\star + e_k) + e_k \\ &= \frac{1}{4}. \end{aligned}$$

Hence the envelope condition in Theorem A.4 holds at stage k. If $\varepsilon_{k,\mathrm{amp}} \le \beta_k$, the displayed recurrence is an upper-envelope version of the recurrence in Theorem A.4. Repeating this argument stage by stage and union bounding $\sum_k \delta_k \le \delta$ gives the event $\cap_k E_k$ with probability at least $1-\delta$. Theorem A.4 then gives the claimed all-stage residual, parameter-error, and population-loss conclusions. If $M_k \le 0$, then even setting the sampling term γ_k to zero leaves no positive margin in the sufficient condition, so this proof cannot certify the next stage from more samples alone. The final sector-certification statement is exactly Lemma C.1 applied with $\left\|R_k^W\right\|_{\mathrm{F}} \le \|\Delta_k^\star\|_{\mathrm{F}} + d_k$ on the certified envelope.

B Synthetic Theory Diagnostics

B.1 Reproducibility details and synthetic setup

We run two synthetic suites. The quick suite uses $d = 40$, $m = 32$, $r = 5$, $N = 4096$, three seeds, and a total budget of 180 scalar-solver steps for the direction-locked method. The large suite uses $d = 128$, $m = 96$, $r = 8$, $N = 32768$, eight seeds, and a total budget of 1200 scalar-solver steps with at least 25 steps per stage. We use $\alpha = 0.05$ throughout, except in the nonlinearity sweep where we vary α explicitly. Unless stated otherwise, the sequential runs use fresh stage batches. We report training loss on a fixed reporting batch and prediction loss on a held-out test batch.

For scalar refinement we run full-batch Adam on a single amplitude per stage with learning rate 0.025, PyTorch default betas and epsilon, and a fresh optimizer state at every stage. Simultaneous baselines initialize random factors i.i.d. at scale 10^{-2} and train without weight decay. The scalar-solver ablation (Figure 9) uses the large mild-spectrum setting with the same seeds and stage batches as the main run.

We generate teacher spectra from three singular-value profiles: steep, $s_i \propto e^{-0.65(i-1)}$; mild, $s_i \propto e^{-0.22(i-1)}$; flat, $s_i \propto 1 - 0.025(i-1)$. We normalize so that $\|s\|_2 = 2$ in the large suite. In the rank-mismatch sweep we use learner ranks $\{2, 4, 6, 8, 10, 12\}$ against a rank-8 teacher. Ranks above 8 are over-specified and not covered by the local oracle-gap condition, since the teacher residual has been exhausted. In the nonlinearity sweep (Figure 1b) we use $N = 65536$ and vary $\alpha \in \{0, 0.3, 0.6, 0.9\}$ and $\|\Delta\|_{\mathrm{F}} \in \{0.5, 1.0, 1.5, 2.0, 2.5\}$, reusing the eight seeds of the large suite.

B.2 Certified regime and local mechanism diagnostics

The dashed $1/4$ line marks the direct surrogate test $\|A_k - \Delta_k^\star\|_{\mathrm{F}} / \mathcal{T}_k^\star \le 1/4$ of Theorem 4.1. In the synthetic plot labels, $\widetilde{A}_k$ denotes the same empirical surrogate A_k. Under the sector and concentration assumptions of Proposition 4.5,

$$\frac{\|A_k - \Delta_k^\star\|_{\mathrm{F}}}{\mathcal{T}_k^\star} \le \frac{\eta_\alpha \left\|R_k^W\right\|_{\mathrm{F}} + \left\|R_k^W - \Delta_k^\star\right\|_{\mathrm{F}}}{\mathcal{T}_k^\star}.$$

Passing the measured test directly certifies the dyad bound of Theorem 4.1, but does not by itself verify the more conservative analytical budget on the right. Conversely, a measured ratio above $1/4$ rules out satisfying all hypotheses of Proposition 4.5 simultaneously. These checks use the known synthetic teacher, not an observable stopping rule. We label teacher-spectrum profiles steep, mild, and flat, as in Section B.1, and flat is the hardest case because the top dyad is least stable when residual singular-value gaps are small.

Before stress-testing the rank-8 run, Figure 3 checks the direct surrogate condition in a deliberately favorable regime: a separated rank-3 teacher with N=131072. The measured ratio stays below $1/4$ at every stage, certifying the per-stage dyad bound of Theorem 4.1; the sector and scalar diagnostics remain small.

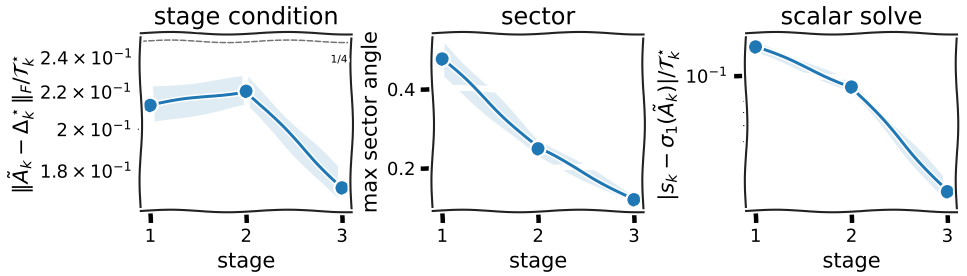


Figure 3: Calibration of the direct surrogate condition: separated rank-3 teacher with N=131072. Per-stage perturbation ratio (left), sector angle (middle), and gap-normalized scalar mismatch (right) on the steep profile. The empirical ratio stays below the $1/4$ sufficient threshold of Theorem 4.1 at every stage, certifying its per-stage dyad error bound. The rank-8 trajectories in Figure 2 fail the plotted $1/4$ top-dyad perturbation test at every tested stage, yet satisfy the gap-free parameter-residual progress condition throughout (Appendix B.3).

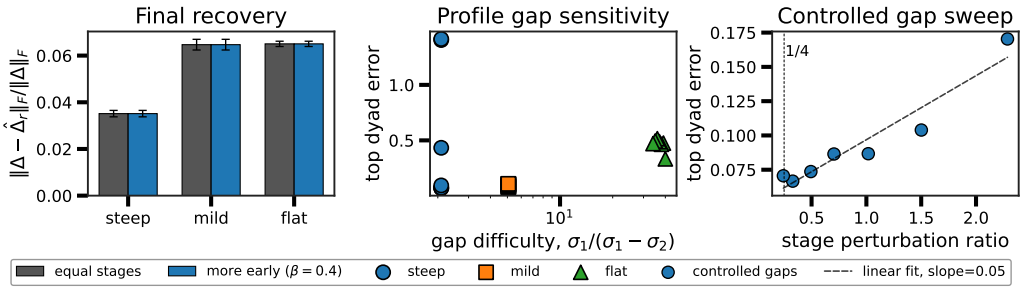


Figure 4: Spectral-gap sensitivity. Steep/mild/flat denote increasingly less separated teacher singular spectra. Left: final relative parameter error $\|\Delta - \widehat{\Delta}_r\|_{\mathrm{F}}/\|\Delta\|_{\mathrm{F}}$ on the rank-8 run is larger for flatter spectra, consistent with smaller residual gaps making each stage's top dyad less stable. Middle: top dyad error grows with the inverse residual relative gap $\sigma_1/(\sigma_1 - \sigma_2)$ across the three spectra. Right: in a controlled sweep that varies only the residual spectral gap (problem size, activation, sample size, and residual norm fixed), the top dyad error grows roughly linearly with the stage perturbation ratio $\|A - \Delta^\star\|_{\mathrm{F}}/\mathcal{T}^\star$; the dashed vertical line marks the $1/4$ sufficient threshold of Theorem 4.1. The figure is consistent with the gap-normalized perturbation prediction of Theorem 4.1/Proposition 4.5 and with the sharpness statement of Proposition 4.2; it is not a global convergence statement.

We then return to the rank-8 run and expand the measured diagnostics in Figure 5. Alongside the direct surrogate ratio, we report realized past-stage mismatch, the oracle residual-to-gap ratio, sector angle, dyad error, and scalar slack. Past-stage mismatch and scalar slack are normalized by $\mathcal{T}_k^\star$. These measurements illustrate ingredients of the theorem chain; the realized mismatch is not the recursive envelope d_k, and the figure does not compute or verify the all-stage certificate of Theorem A.4.

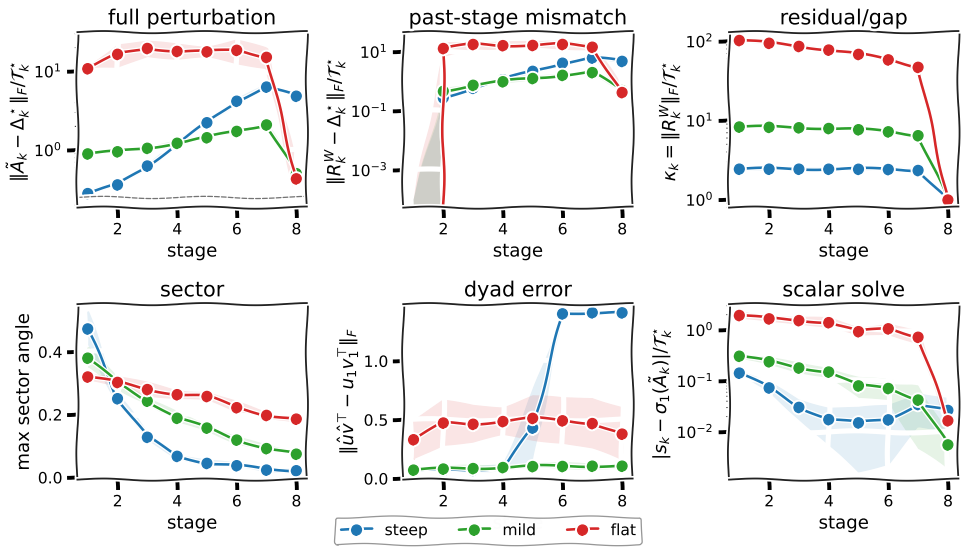


Figure 5: Expanded per-stage diagnostics on the rank-8 run. From top left to bottom right: direct surrogate ratio $\|A_k - \Delta_k^\star\|_{\mathrm{F}} / \mathcal{T}_k^\star$ with the $1/4$ threshold of Theorem 4.1; realized gap-normalized mismatch $\left\|R_k^W - \Delta_k^\star\right\|_{\mathrm{F}} / \mathcal{T}_k^\star$; oracle residual-to-gap ratio $\kappa_k^\star = \|\Delta_k^\star\|_{\mathrm{F}} / \mathcal{T}_k^\star$; maximum sector angle; observed top-dyad error $\left\|\hat{u}_k \hat{v}_k^\top - u_{1,k} v_{1,k}^\top\right\|_{\mathrm{F}}$; and gap-normalized scalar slack $\varepsilon_{k,\mathrm{amp}} / \mathcal{T}_k^\star$. Curves show seed means of these quantities, not the recursive envelope or its feasibility bound.

B.3 Gap-free progress on the rank-8 trajectories

We reanalyze the recorded trajectories underlying Figure 2 without rerunning training. The data file is `spectrum_sensitivity.csv` in `outputs/large/tables/`. We select `experiment=spectrum` rows with locked directions (`stage_rule=locked`), fresh batches (`fresh_batches=1`), and zero values of both `schedule_alpha` and `schedule_beta`. These are 24 runs: steep/mild/flat spectra, seeds 7–14, and eight stages per run, with 150 scalar-Adam steps per stage.

Let $p_k = \left\|\Delta - \widehat{\Delta}^{(k)}\right\|_{\mathrm{F}} / \|\Delta\|_{\mathrm{F}}$ be the post-update `param_rel`, with $p_0 = 1$. The pre-update relative surrogate error is $g_k = \left\|A_k - R_k^W\right\|_{\mathrm{F}} / \left\|R_k^W\right\|_{\mathrm{F}}$ (`grad_residual_rel`). The recorded `spectral_ref_scalar` gives $\sigma_k = \sigma_1(A_k)$ before scalar refinement, whereas `scalar_value` gives the applied amplitude s_k after refinement. We reconstruct

$$\delta_{F,k} = g_k p_{k-1} \|\Delta\|_{\mathrm{F}} = \left\|A_k - R_k^W\right\|_{\mathrm{F}}, \qquad \zeta_k = \frac{s_k + 2\delta_{F,k}}{2\sigma_k}.$$

Here $\|\Delta\|_{\mathrm{F}}$ is the first-stage `stage_residual_norm`; later values of that field are oracle-residual norms and are not used in this reconstruction. For $s_k > 0$ and $\sigma_k > 0$, the test $\zeta_k < 1$ is exactly the Frobenius sufficient condition of Proposition 4.8. Both signs and this test hold for all 192 recorded updates. Moreover, every measured squared-residual decrease satisfies

$$\|\Delta\|_{\mathrm{F}}^2 \left(p_{k-1}^2 - p_k^2\right) \geq s_k(2\sigma_k - s_k - 2\delta_{F,k}) > 0.$$

Table 1: Stagewise checks on the recorded Figure 2 trajectories. Oracle counts passes of the plotted test $\|A_k - \Delta_k^\star\|_{\mathrm{F}} / \mathcal{T}_k^\star \le 1/4$ (`stage_condition_ratio`); progress counts $s_k > 0$ and $\zeta_k < 1$; decrease counts $p_k < p_{k-1}$. Each row pools eight stages from each of eight seeds, not 64 independent runs.

Spectrum	Oracle	Progress	Decrease	Maximum ζ_k
Steep	0/64	64/64	64/64	0.699631
Mild	0/64	64/64	64/64	0.739105
Flat	0/64	64/64	64/64	0.829223

This is an a posteriori comparison using the known synthetic teacher. It establishes parameter-residual progress for these trajectories, not top-dyad recovery, monotonic prediction loss, or an observable stopping rule. Positivity after scalar refinement is verified here, not assumed for arbitrary runs.

B.4 Sample dependence and end-to-end accounting

Sample size and learner rank enter the theorem at different points. The fresh batch size N enters through the concentration rate γ_N of Lemma 4.4, and the learner rank r enters through the oracle rank-r tail in Proposition 4.6. We test the two effects separately.

In Figure 6 we sweep N on the rank-8 run. Larger N tightens the maximum-over-stages perturbation ratio and improves both final parameter recovery and held-out prediction loss. The observed log-log slope is in qualitative agreement with the $N^{-1/2}$ guide of Lemma 4.4, since final error also contains nonlinear bias, gap effects, scalar slack, and accumulated past-stage mismatch.

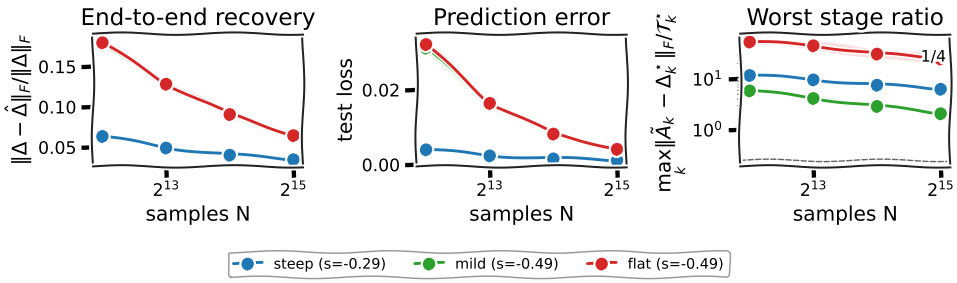


Figure 6: End-to-end sample scaling on the rank-8 run for steep/mild/flat teacher spectra. Increasing the fresh stage-batch size N improves final parameter recovery (left) and held-out prediction loss (middle); the right panel reports the maximum-over-stages perturbation ratio $\max_k \|A_k - \Delta_k^\star\|_{\mathrm{F}} / \mathcal{T}_k^\star$, with the dashed $1/4$ line of Theorem 4.1. Log-log slope annotations compare the observed scaling against the $N^{-1/2}$ sampling guide of Lemma 4.4; agreement is qualitative, since final error also includes nonlinear bias, gap effects, scalar slack, and accumulated past-stage mismatch (Theorem A.4). The panels are jointly consistent with the theory chain: larger N tightens gradient concentration (Lemma 4.4), which improves stage-wise dyad recovery (Proposition 4.5) and propagates to final parameter and prediction error through Proposition 4.6/Proposition 4.7.

Fresh versus reused stage batches. We compare the two protocols on the same rank-8 teachers using the large-suite configuration and eight seeds per spectrum. At each of eight stages, both protocols use a batch of $N = 32768$ samples. Fresh batches use $8N = 262144$ distinct training samples, whereas reuse uses the same fixed batch of $N = 32768$ samples at every stage. The mean final relative parameter errors are similar across the three tested spectra (Table 2). This comparison fixes per-stage batch size, not the total distinct-sample budget. It provides empirical evidence for batch reuse in these settings without extending the fresh-batch concentration guarantee.

Table 2: Final relative parameter-recovery error, mean ± sample standard deviation over eight seeds.

Spectrum	Fresh batches	Reused batch
Steep	0.0351 ± 0.0015	0.0354 ± 0.0022
Mild	0.0647 ± 0.0024	0.0638 ± 0.0018
Flat	0.0650 ± 0.0012	0.0649 ± 0.0015

In Figure 7 we sweep the learner rank r on the same teacher. We plot the learner error, the oracle rank-r tail, and the Proposition 4.6 accounting quantity (oracle tail plus the sum of measured per-stage errors) on a single axis. On the rank-≤ 8 side the learner sits at or below the accounting curve, as the additive bound predicts. Ranks $r > 8$ are over-specified: the oracle residual is exhausted, $\mathcal{T}_k^\star = 0$, and the local gap condition no longer certifies new dyads.

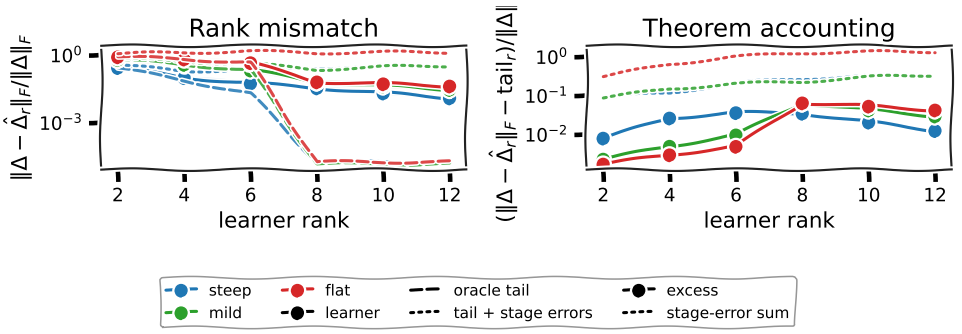


Figure 7: Rank mismatch against a rank-8 teacher (steep/mild/flat spectra). Left: final relative parameter error $\|\Delta - \widehat{\Delta}_r\|_{\mathrm{F}}/\|\Delta\|_{\mathrm{F}}$ versus learner rank r. Dashed: normalized oracle rank-r truncated-SVD tail $\mathrm{tail}_r/\|\Delta\|_{\mathrm{F}}$ with $\mathrm{tail}_r = (\sum_{j>r} \sigma_j(\Delta)^2)^{1/2}$. Solid: the direction-locked learner. Dotted: the Proposition 4.6 accounting quantity, oracle tail plus the sum of measured per-stage errors. Right: extra error above the oracle tail, $(\|\Delta - \widehat{\Delta}_r\|_{\mathrm{F}} - \mathrm{tail}_r)/\|\Delta\|_{\mathrm{F}}$, alongside the accumulated measured stage-error term. Under Proposition 4.6 the learner is bounded by the dotted curve, which is what the figure shows on the rank-≤ 8 side. Ranks > 8 are over-specified diagnostics: all nonzero oracle components have already been removed, $\mathcal{T}_k^\star = 0$, and the local gap condition no longer certifies new dyads.

B.5 Nonlinearity, ablations, and failure modes

The remaining synthetic figures probe how the local mechanism behaves when we vary parts of the setup that the theory holds fixed.

In Figure 8 we sweep the Leaky-ReLU slope α in the full sequential algorithm, not only in the stage-1 gradient check. Parameter recovery is fairly stable across slopes in this controlled setting, and prediction error increases as the teacher becomes nearly linear with fixed $\|\Delta\|_{\mathrm{F}}$. The mean perturbation ratio decreases with α, in line with the $(1-\alpha)^2$ scaling of the nonlinear remainder term in Lemma 4.3. In the rank-8 mild-spectrum configuration the ratio stays above the $1/4$ threshold, so the trend supports the predicted dependence without certifying the sufficient condition globally.

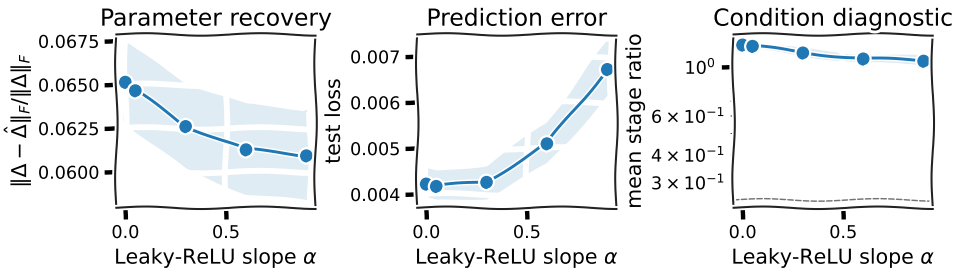


Figure 8: End-to-end recovery as the activation becomes more linear (rank-8 mild spectrum). Left/middle: final parameter and prediction error across $\alpha \in [0, 1)$. Right: empirical perturbation ratio averaged over stages; the dashed $1/4$ line is the sufficient threshold of Theorem 4.1. The mean ratio decreases with α, consistent with the $(1-\alpha)^2$ scaling of the nonlinear remainder term in Lemma 4.3, but stays above $1/4$ in this configuration. The trend is consistent with the predicted dependence; it does not certify the sufficient condition globally.

In Figure 9 we ablate the one-dimensional scalar solver under the direction lock. In the large mild-spectrum run, pure spectral amplitude gives relative parameter error 0.0806 ± 0.0029. Exact piecewise-quadratic line search, scalar Adam, and backfitting all previous scalars are nearly indistinguishable: 0.0654 ± 0.0015, 0.0654 ± 0.0015, and 0.0653 ± 0.0015. Orthogonalizing new directions slightly increases final error in this configuration (0.0667 ± 0.0017; protocol below). Line search, scalar Adam, and scalar backfitting improve final recovery over spectral amplitude even though the mean gap-normalized mismatch of the newly fitted scalar, $\varepsilon_{k,\mathrm{amp}}/\mathcal{T}_k^\star$, increases from zero to about 0.127. Mean per-stage dyad errors remain similar (0.0983–0.0993). Thus, in this configuration, a larger amplitude-deviation diagnostic accompanies better final recovery. The free-factor stage heuristic is more accurate (0.0106 ± 0.0004) but rotates away from the locked dyads with $\varepsilon_{\mathrm{rot}} \approx 0.068$ on average, and it is covered by the theorem only through Corollary A.1.

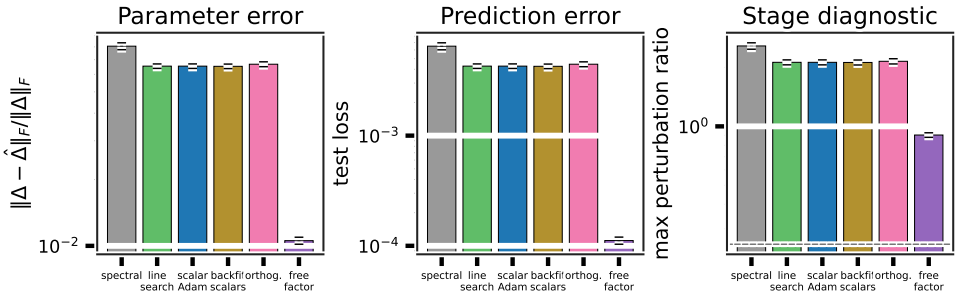


Figure 9: Scalar-solver ablation under the direction lock (large mild-spectrum run). Line search and scalar Adam improve on spectral amplitude and achieve similar final errors to scalar backfitting. Backfitting revises earlier amplitudes and is an empirical variant of the frozen-update procedure. The free-factor heuristic is shown for context: it is more accurate but rotates away from the locked dyads, and Corollary A.1 only records the resulting rotation term $\varepsilon_{k,\mathrm{rot}}$ as a deterministic diagnostic without bounding it. The "orthog." bar uses the orthogonalization control described below. Error bars show population standard deviations across eight seeds.

Orthogonalization control. The "orthog." variant in Figure 9 subtracts projections of each new left and right gradient singular vector onto the previously retained left and right vectors, respectively, and normalizes the results. If a remaining norm is below 10^{-8}, it retains that original unit singular vector. It then fits the new scalar with the same Adam learning rate, stage-wise step budgets, and fresh batches as the scalar-Adam control. In the rank-8 mild-spectrum run, final relative parameter error is higher in all eight paired seeds (7–14), by approximately 0.0007–0.0020, while mean per-stage dyad error decreases from 0.0983 to 0.0884. Thus this variant improves dyad alignment but not final recovery in the tested setting. It is an empirical modification of the top-gradient-dyad procedure, not a claim about the general value of orthogonality constraints.

In Figure 10 we keep the same direction-locked algorithm and vary five axes that the theorem holds fixed: dimension, target rank, output noise, activation mismatch, and correlated inputs. We use the figure as a failure-mode diagnostic, not as additional evidence for a stronger theorem. Recovery degrades when the perturbation ratio grows, when the residual gaps shrink, or when the data geometry moves away from the isotropic Gaussian setting, in agreement with the local analysis.

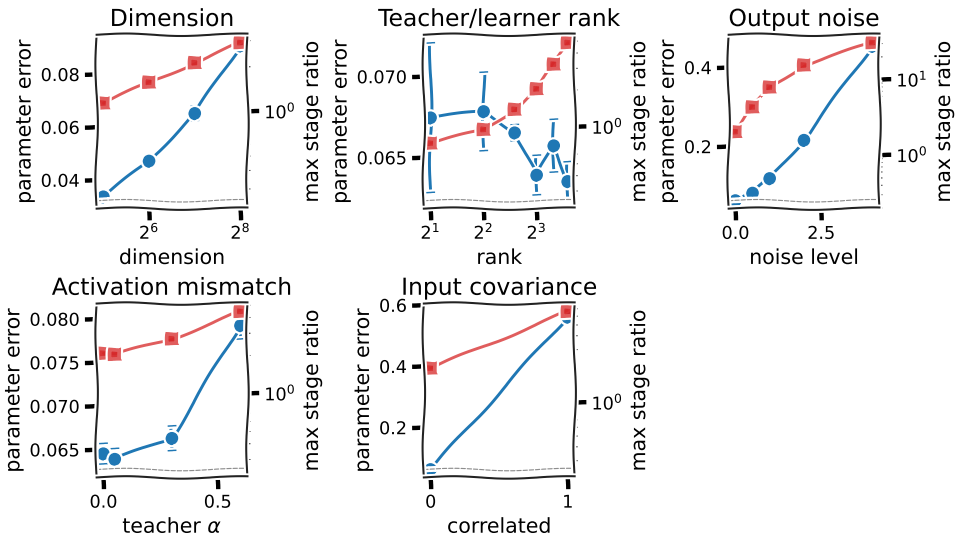


Figure 10: Synthetic robustness stress tests, outside the clean theorem assumptions. Final parameter error and the maximum empirical stage-condition ratio as we vary dimension, target rank, output noise, activation mismatch, and input covariance one axis at a time. The figure is a failure-mode diagnostic for the assumptions of Lemma 4.3/Proposition 4.5: recovery degrades when the perturbation ratio grows, residual gaps shrink, or the data geometry leaves the isotropic Gaussian setting. It is not additional evidence for a stronger theorem.

Finally, in Figure 11 we compare the analyzed direction-locked variants ($t_k{=}0$ spectral and scalar-refined) against three baselines. The free-factor sequential method shares the SVD initialization but then optimizes both factors without a direction lock, and is covered by the theorem only through the rotation term of Corollary A.1. The last two baselines train a simultaneous rank-r factorization $\Delta = BA^\top$, from spectral and from random initialization, and lie outside the per-stage theorem. The scalar-refined locked, free-factor sequential, and simultaneous methods each use 1200 Adam steps in total; the locked spectral variant uses no Adam steps. Both locked variants use eight independent stage batches of 32,768 examples, totaling 262,144 distinct training examples. The free-factor and simultaneous baselines each reuse one fixed batch of 32,768 examples. Matching optimization steps therefore does not match distinct training-data budgets or total computation. We include this comparison for empirical context, not as additional theorem evidence.

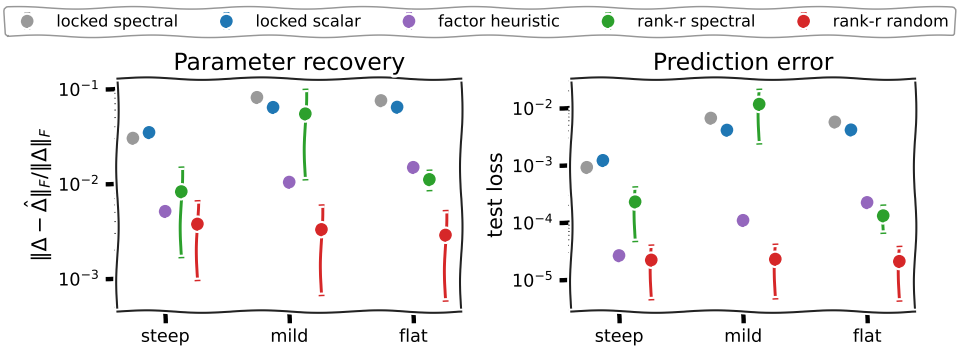


Figure 11: Baseline comparison on the large run. The two direction-locked variants (t_k=0 spectral and scalar-refined) are the methods analyzed by Proposition 4.5. The free-factor sequential baseline shares the same SVD initialization but optimizes both factors per stage; it is covered by the theorem only through the additive rotation term $\varepsilon_{k,\mathrm{rot}}$ of Corollary A.1. Simultaneous rank-r training, from spectral or random initialization, is shown for empirical context and is outside the stage perturbation theorem. The locked variants use 262,144 distinct training examples, whereas each free-factor or simultaneous baseline reuses 32,768. The Adam-trained variants each use 1200 steps; the locked spectral variant uses none. The comparison is not matched for distinct training data or total computation.

C Auxiliary Theory and Diagnostics

C.1 Deterministic sector check

The sector assumption in Lemma 4.3 is not enforced by Algorithm 1. The following elementary check records when it is implied by ordinary parameter closeness. It is sufficient, not necessary: it uses the smallest nonzero teacher-column norm, so it can be restrictive when that norm is small even if the columnwise angle condition holds.

Lemma C.1 (Parameter error implies sector control)**.** *Let $\widetilde{w}_j^\star$ denote the jth column of $\widetilde{\mathbf{W}}^\star$, and assume $m_{\min} = \min_{j:\widetilde{w}_j^\star \neq 0} \|\widetilde{w}_j^\star\|_2 > 0$. If, for some $\theta_0 \in (0, \pi/2)$,*

$$\left\|\mathbf{W} - \widetilde{\mathbf{W}}^\star\right\|_{\mathrm{F}} \leq m_{\min} \sin\theta_0,$$

then every learner column corresponding to a nonzero teacher column is nonzero and satisfies $\angle(w_j, \widetilde{w}_j^\star) \leq \theta_0$. Consequently, along a stage-wise envelope it is sufficient that

$$\|\Delta_k^\star\|_{\mathrm{F}} + d_k \leq m_{\min} \sin\theta_0$$

to certify the sector condition at stage k, since $\left\|\mathbf{W}^{\langle k\rangle} - \widetilde{\mathbf{W}}^\star\right\|_{\mathrm{F}} = \|R_k^W\|_{\mathrm{F}} \leq \|\Delta_k^\star\|_{\mathrm{F}} + d_k$.

Proof. Fix a column with $\widetilde{w}_j^\star \neq 0$. The Frobenius bound implies $\|w_j - \widetilde{w}_j^\star\|_2 \leq m_{\min} \sin\theta_0 \leq \|\widetilde{w}_j^\star\|_2 \sin\theta_0 < \|\widetilde{w}_j^\star\|_2$, so $w_j \neq 0$. The distance from $\widetilde{w}_j^\star$ to the line spanned by w_j is $\|\widetilde{w}_j^\star\|_2 \sin\angle(w_j, \widetilde{w}_j^\star)$, and this distance is at most $\|w_j - \widetilde{w}_j^\star\|_2$. Also $\|w_j - \widetilde{w}_j^\star\|_2 < \|\widetilde{w}_j^\star\|_2$ implies $w_j^\top \widetilde{w}_j^\star > 0$, so the angle lies in $[0, \pi/2)$. Therefore

$$\sin\angle(w_j, \widetilde{w}_j^\star) \leq \frac{\|w_j - \widetilde{w}_j^\star\|_2}{\|\widetilde{w}_j^\star\|_2} \leq \sin\theta_0.$$

Since both angles lie in $[0, \pi/2]$, the sector bound follows. The final claim uses $R_k^W = \widetilde{\mathbf{W}}^\star - \mathbf{W}^{\langle k\rangle}$ and $R_k^W = \Delta_k^\star + (R_k^W - \Delta_k^\star)$. □

C.2 Independent-batch dyad stability

The following diagnostic is not used as a main contribution, but it is a useful check on whether independent fresh batches agree on the next dyad.

Corollary C.2 (Independent-batch top-dyad stability). *Fix a stage k after $\mathbf{W}^{\langle k\rangle}$ has been fixed, and let $\mathcal{B}_k^{(1)}, \dots, \mathcal{B}_k^{(M)}$ be independent fresh micro-batches, with $M \geq 2$. Define*

$$A_k^{(m)} := -\nabla \mathcal{L}_{\mathcal{B}_k^{(m)}}(\mathbf{W}^{\langle k\rangle})/\mu_\alpha, \qquad D_k^{(m)} := \hat{u}_k^{(m)} \hat{v}_k^{(m)\top}.$$

Let $D_k^\star = u_{1,k} v_{1,k}^\top$ be the top singular dyad of $\Delta_k^\star$. Suppose that Proposition 4.5 applies to each micro-batch and gives $\left\| D_k^{(m)} - D_k^\star \right\|_{\mathrm{F}} \leq \delta_k^{(m)}$. Then, for all $m \neq \ell$,

$$\left\| D_k^{(m)} - D_k^{(\ell)} \right\|_{\mathrm{F}} \leq \delta_k^{(m)} + \delta_k^{(\ell)}.$$

Consequently,

$$I_k(M) := \frac{2}{M(M-1)} \sum_{1 \leq m < \ell \leq M} \left\| D_k^{(m)} - D_k^{(\ell)} \right\|_{\mathrm{F}} \leq \frac{2}{M} \sum_{m=1}^{M} \delta_k^{(m)}.$$

In particular, if $\delta_k^{(m)} \leq \bar{\delta}_k$ for all m, then $I_k(M) \leq 2\bar{\delta}_k$.

Proof. Insert the oracle dyad and use the triangle inequality:

$$\left\| D_k^{(m)} - D_k^{(\ell)} \right\|_{\mathrm{F}} \leq \left\| D_k^{(m)} - D_k^\star \right\|_{\mathrm{F}} + \left\| D_k^{(\ell)} - D_k^\star \right\|_{\mathrm{F}} \leq \delta_k^{(m)} + \delta_k^{(\ell)}.$$

Averaging over unordered pairs gives the displayed bound, because each $\delta_k^{(m)}$ appears in exactly $M-1$ pairs. □

C.3 Diagnostic explanation of the stage condition

For equal-size micro-batches of size n, Proposition 4.5 gives, on the corresponding fresh-batch concentration events,

$$\delta_k^{(m)} \leq C \frac{\eta_\alpha(\theta_0, \gamma_n) \left\| R_k^W \right\|_{\mathrm{F}} + \left\| R_k^W - \Delta_k^\star \right\|_{\mathrm{F}}}{\mathcal{T}_k^\star}.$$

Thus $I_k(M)$ inherits the same sampling, nonlinearity, past-error, and gap dependence as the one-stage perturbation term. The converse is not claimed: small $I_k(M)$ does not prove that Proposition 4.5 holds, since several batches can agree on the same wrong direction.

C.4 Sector-free progress in a near-linear regime

We retain the noiseless single-layer Leaky-ReLU teacher and squared loss, but replace Gaussianity by $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = I_d$. The following result controls the actual parameter residual rather than identifying an oracle singular dyad. In this subsection, $\nabla \mathcal{L}_{\mathrm{pop}}$ denotes the expectation of the selected sample gradient under the kink convention of Section 3, as in Lemma 4.4.

Lemma C.3 (Sector-free population surrogate bound). *Let $\alpha \in [0, 1)$ and $\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = I_d$. The jth column of the selected population gradient is*

$$\nabla_{w_j} \mathcal{L}_{\mathrm{pop}}(\mathbf{W}) = \mathbb{E}\left[\left(\sigma_\alpha(\mathbf{x}^\top w_j) - \sigma_\alpha(\mathbf{x}^\top \widetilde{w}_j^\star) \right) \sigma_\alpha'(\mathbf{x}^\top w_j) \mathbf{x} \right],$$

where $\sigma_\alpha'(0)$ uses the fixed kink convention. Set

$$R = \widetilde{\mathbf{W}}^\star - \mathbf{W}, \qquad \mu_\alpha = \frac{1+\alpha^2}{2}, \qquad \beta_\alpha = \frac{1-\alpha^2}{1+\alpha^2}, \qquad A = -\nabla \mathcal{L}_{\mathrm{pop}}(\mathbf{W})/\mu_\alpha.$$

For every $\mathbf{W}$, including zero learner columns,

$$\left\| \nabla \mathcal{L}_{\mathrm{pop}}(\mathbf{W}) + \mu_\alpha R \right\|_{\mathrm{F}} \leq \frac{1-\alpha^2}{2} \left\| R \right\|_{\mathrm{F}}, \qquad \left\| A - R \right\|_{\mathrm{F}} \leq \beta_\alpha \left\| R \right\|_{\mathrm{F}}.$$

No sector condition is required.

Proof. Fix a column and write $u = \mathbf{x}^\top w_j$, $v = \mathbf{x}^\top \widetilde{w}_j^\star$, and $r_j = \widetilde{w}_j^\star - w_j$. The secant slope

$$a_\alpha(u, v) = \frac{\sigma_\alpha(u) - \sigma_\alpha(v)}{u - v}$$

belongs to $[\alpha, 1]$ when $u \neq v$. Set $a_\alpha(u, u) = \alpha$, so that $\sigma_\alpha(u) - \sigma_\alpha(v) = a_\alpha(u, v)(u - v)$ in all cases. Consequently,

$$\nabla_{w_j} \mathcal{L}_{\text{pop}}(\mathbf{W}) = -K_j r_j, \qquad K_j = \mathbb{E}\left[a_\alpha(u, v)\sigma'_\alpha(u)\mathbf{x}\mathbf{x}^\top\right].$$

Both scalar factors lie in $[\alpha, 1]$, and isotropy gives

$$\alpha^2 I_d \preceq K_j \preceq I_d.$$

Since μ_α is the midpoint of $[\alpha^2, 1]$,

$$\left\|\nabla_{w_j} \mathcal{L}_{\text{pop}}(\mathbf{W}) + \mu_\alpha r_j\right\|_2 = \left\|(\mu_\alpha I_d - K_j) r_j\right\|_2 \leq \frac{1 - \alpha^2}{2} \left\|r_j\right\|_2.$$

Summing the squared column bounds proves the first inequality. Dividing by μ_α proves the second. □

Corollary C.4 (Near-linear population residual contraction)**.** *Under the assumptions of Lemma C.3, let D be a leading singular dyad of A with unit singular vectors, set $\sigma = \sigma_1(A)$, and update $\mathbf{W}^+ = \mathbf{W} + \sigma D$. If $R \neq 0$ and*

$$b_R := \beta_\alpha \frac{\|R\|_{\text{F}}}{\sigma_1(R)} < \frac{1}{3},$$

then

$$\|R\|_{\text{F}}^2 - \left\|\widetilde{\mathbf{W}}^\star - \mathbf{W}^+\right\|_{\text{F}}^2 \geq (1 - b_R)(1 - 3b_R)\sigma_1(R)^2 > 0.$$

In particular, let $p = \min\{d, m\}$ and suppose

$$\bar{b} := \beta_\alpha \sqrt{p} < \frac{1}{3}, \qquad c_{\alpha,p} := \frac{(1 - \bar{b})(1 - 3\bar{b})}{p}.$$

Start from $\mathbf{W}^{\langle 1 \rangle} = \mathbf{W}_0$, with no restriction on $\left\|\mathbf{W}_0 - \widetilde{\mathbf{W}}^\star\right\|_{\text{F}}$, and write $R_k^W = \widetilde{\mathbf{W}}^\star - \mathbf{W}^{\langle k \rangle}$. At stage k, set $A_k = -\nabla \mathcal{L}_{\text{pop}}(\mathbf{W}^{\langle k \rangle})/\mu_\alpha$ and let $\hat{u}_k \hat{v}_k^\top$ be a leading singular dyad of A_k with unit singular vectors. Repeat the population updates

$$U_k = \sigma_1(A_k)\hat{u}_k \hat{v}_k^\top, \qquad \mathbf{W}^{\langle k+1 \rangle} = \mathbf{W}^{\langle k \rangle} + U_k,$$

taking $U_k = 0$ when $A_k = 0$. Then, for every $k \geq 1$,

$$\left\|R_k^W\right\|_{\text{F}}^2 \leq (1 - c_{\alpha,p})^{k-1} \left\|R_1^W\right\|_{\text{F}}^2.$$

Thus the parameter residual converges geometrically to zero without a nearby initialization, sector control, or a spectral gap.

Proof. Write $E = A - R$ and $e = \|E\|_2 \leq \|E\|_{\text{F}}$. By Lemma C.3, $e \leq \beta_\alpha \|R\|_{\text{F}} = b_R \sigma_1(R) < \sigma_1(R)/3$. Weyl's inequality gives $\sigma \geq \sigma_1(R) - e > 2e$. Applying Proposition 4.8 with $s = \sigma$ yields

$$\begin{aligned} \|R\|_{\text{F}}^2 - \|R - \sigma D\|_{\text{F}}^2 &\geq \sigma(\sigma - 2e) \\ &\geq (\sigma_1(R) - e)(\sigma_1(R) - 3e) \\ &\geq (1 - b_R)(1 - 3b_R)\sigma_1(R)^2. \end{aligned}$$

The middle inequality uses that $z(z - 2e)$ is increasing for $z \geq e$.

For the uniform statement, $\|R\|_{\text{F}} \leq \sqrt{p}\,\sigma_1(R)$ implies $b_R \leq \bar{b}$. Therefore, at every nonzero residual,

$$\|R\|_{\text{F}}^2 - \|R - \sigma D\|_{\text{F}}^2 \geq (1 - \bar{b})(1 - 3\bar{b})\sigma_1(R)^2 \geq c_{\alpha,p} \|R\|_{\text{F}}^2.$$

If $A = 0$, the surrogate bound and $\beta_\alpha \leq \bar{b} < 1/3$ imply $R = 0$, which remains zero under a zero update. Iteration proves the claimed geometric bound. □

For $R \neq 0$, its stable rank is $\mathrm{sr}(R) := \|R\|_{\mathrm{F}}^2 / \sigma_1(R)^2$. Thus the one-step condition is $\beta_\alpha \sqrt{\mathrm{sr}(R)} < 1/3$ and depends on residual stable rank rather than ambient dimension. In particular, a rank-one residual satisfies it whenever $\alpha > 1/\sqrt{2}$. Using this condition throughout a run requires controlling the residual stable rank at every stage.

The uniform condition is equivalent to

$$\alpha > \sqrt{\frac{3\sqrt{p}-1}{3\sqrt{p}+1}}.$$

For the quick and large suites of Appendix B.1, $p = 32$ and $p = 96$, respectively, and these thresholds are approximately 0.943 and 0.967. The $\alpha = 0.05$ settings and the stated sweep through $\alpha = 0.9$ do not satisfy this uniform population-contraction condition. It replaces a local initialization requirement by a dimension-dependent near-linearity requirement. Here p is the ambient rank bound, not the teacher rank: imperfect updates can increase the residual rank. The result concerns exact population updates with spectral amplitudes. It does not imply recovery within $r^\star$ stages or monotonic prediction loss. This population corollary does not by itself guarantee progress for finite-batch runs of Algorithm 1, with or without scalar refinement.

D Frozen-Feature Squared-Loss Bridge

D.1 Setup and motivation for the bridge

This appendix is a secondary frozen-feature squared-loss bridge. It isolates a one-matrix, frozen-representation setting and is not a theorem for deep-network LoRA, cross-entropy, or representation drift.

D.2 Deterministic frozen-feature bridge proposition

Proposition D.1 (Frozen-feature squared-loss bridge)**.** *Let $z \in \mathbb{R}^d$ be a frozen feature vector with covariance $\Sigma = \mathbb{E}[zz^\top]$, and consider the noiseless linear-head teacher $y = z^\top(\mathbf{W}_0 + \Delta)$. For the squared loss*

$$\mathcal{L}_{\mathrm{ff}}(\mathbf{W}) = \frac{1}{2}\mathbb{E}\left\|z^\top \mathbf{W} - z^\top(\mathbf{W}_0 + \Delta)\right\|_2^2,$$

define $R_k^W = \Delta - \sum_{j<k} U_j$ as before. Then

$$-\nabla \mathcal{L}_{\mathrm{ff}}(\mathbf{W}^{\langle k \rangle}) = \Sigma R_k^W.$$

For a fresh empirical batch $\mathcal{B}$, the same identity holds with Σ replaced by $\widehat{\Sigma}_{\mathcal{B}} = |\mathcal{B}|^{-1} \sum_{z_i \in \mathcal{B}} z_i z_i^\top$. Consequently,

$$\left\| -\nabla \mathcal{L}_{\mathrm{ff}}(\mathbf{W}^{\langle k \rangle}) - R_k^W \right\|_{\mathrm{F}} \leq \|\Sigma - I\|_2 \left\| R_k^W \right\|_{\mathrm{F}},$$

and similarly with $\widehat{\Sigma}_{\mathcal{B}}$ for the empirical gradient.

In particular, if the features are whitened so that $\Sigma = I$, then the population negative gradient equals the current residual exactly. The population-gradient squared-loss analogue of Algorithm 1, with the same direction-locked extraction rule and exact scalar amplitudes $s_k = \sigma_1(R_k^W)$, recovers the rank-r truncated SVD of Δ:

$$\sum_{k=1}^{r} U_k = \Delta_{(r)}, \qquad \left\| \Delta - \sum_{k=1}^{r} U_k \right\|_{\mathrm{F}} = \left(\sum_{j>r} \sigma_j(\Delta)^2 \right)^{1/2}.$$

The same conclusion holds if the selected SVD dyads are kept fixed and their scalar amplitudes are jointly refit by least squares. More generally, for any fixed set of dyads, joint scalar refitting cannot increase the squared-loss objective relative to the stagewise scalar values on those same dyads.

D.3 FINITE-SAMPLE FROZEN-FEATURE BRIDGE PROPOSITION

Proposition D.2 (Finite-sample frozen-feature bridge)**.** *In the setting of Proposition D.1, assume* $\mathcal{T}_k^\star > 0$ *and assume the frozen features are population-whitened and sub-Gaussian:* $\mathbb{E}[zz^\top] = I$ *and* $\|z\|_{\psi_2} \le K$*. At a fixed stage* k*, draw a fresh batch* $\mathcal{B}_k$ *of size* N *after* $\mathbf{W}^{\langle k\rangle}$ *is fixed, and set*

$$\widetilde{A}_k = -\nabla \mathcal{L}_{\mathrm{ff},\mathcal{B}_k}(\mathbf{W}^{\langle k\rangle}) = \widehat{\Sigma}_{\mathcal{B}_k} R_k^W.$$

There is a constant C_K *depending only on* K *such that, with probability at least* $1-\delta$*,*

$$\left\|\widetilde{A}_k - R_k^W\right\|_{\mathrm{F}} \le \gamma_{\mathrm{ff}} \left\|R_k^W\right\|_{\mathrm{F}}, \qquad \gamma_{\mathrm{ff}} = C_K\left[\sqrt{\frac{d+\log(2/\delta)}{N}} + \frac{d+\log(2/\delta)}{N}\right].$$

Consequently, if

$$\gamma_{\mathrm{ff}} \left\|R_k^W\right\|_{\mathrm{F}} + \left\|R_k^W - \Delta_k^\star\right\|_{\mathrm{F}} \le \frac{\mathcal{T}_k^\star}{4},$$

then the top dyad $\hat{u}_k\hat{v}_k^\top$ *of* $\widetilde{A}_k$ *satisfies*

$$\left\|\hat{u}_k\hat{v}_k^\top - u_{1,k}v_{1,k}^\top\right\|_{\mathrm{F}} \le C\frac{\gamma_{\mathrm{ff}} \left\|R_k^W\right\|_{\mathrm{F}} + \left\|R_k^W - \Delta_k^\star\right\|_{\mathrm{F}}}{\mathcal{T}_k^\star}.$$

Moreover, if $U_k = s_k\hat{u}_k\hat{v}_k^\top$ *and* $\varepsilon_{k,\mathrm{amp}} = \left|s_k - \sigma_1(\widetilde{A}_k)\right|$*, then*

$$\|U_k - U_k^\star\|_{\mathrm{F}} \le \varepsilon_{k,\mathrm{amp}} + C'\frac{\gamma_{\mathrm{ff}} \left\|R_k^W\right\|_{\mathrm{F}} + \left\|R_k^W - \Delta_k^\star\right\|_{\mathrm{F}}}{\mathcal{T}_k^\star}\sigma_1(\Delta_k^\star).$$

D.4 PROOFS OF THE BRIDGE PROPOSITIONS

Proof of Proposition D.1. The squared-loss objective can be written as

$$\begin{aligned}\mathcal{L}_{\mathrm{ff}}(\mathbf{W}) &= \frac{1}{2}\mathbb{E}\,\mathrm{tr}\left[(\mathbf{W}-\mathbf{W}_0-\Delta)^\top zz^\top(\mathbf{W}-\mathbf{W}_0-\Delta)\right] \\ &= \frac{1}{2}\,\mathrm{tr}\left[(\mathbf{W}-\mathbf{W}_0-\Delta)^\top\Sigma(\mathbf{W}-\mathbf{W}_0-\Delta)\right].\end{aligned}$$

Differentiating gives

$$\nabla\mathcal{L}_{\mathrm{ff}}(\mathbf{W}) = \Sigma(\mathbf{W}-\mathbf{W}_0-\Delta).$$

At stage k, $\mathbf{W}^{\langle k\rangle} - \mathbf{W}_0 - \Delta = -R_k^W$, and hence $-\nabla\mathcal{L}_{\mathrm{ff}}(\mathbf{W}^{\langle k\rangle}) = \Sigma R_k^W$. The empirical identity is identical with $\widehat{\Sigma}_{\mathcal{B}} = |\mathcal{B}|^{-1}\sum_{z_i\in\mathcal{B}} z_i z_i^\top$ in place of Σ. Therefore

$$\left\|\Sigma R_k^W - R_k^W\right\|_{\mathrm{F}} = \left\|(\Sigma - I)R_k^W\right\|_{\mathrm{F}} \le \|\Sigma - I\|_2 \left\|R_k^W\right\|_{\mathrm{F}},$$

and the same deterministic bound holds with $\widehat{\Sigma}_{\mathcal{B}}$.

When $\Sigma = I$, the population negative gradient is exactly R_k^W. If $\Delta = \sum_{j\ge1}\sigma_j u_j v_j^\top$ is an SVD, then stage 1 extracts $u_1v_1^\top$ with amplitude σ_1. After subtracting this component, the residual is $\sum_{j\ge2}\sigma_j u_j v_j^\top$, so induction gives the first r SVD components. The final parameter error is therefore the oracle tail. For the same selected SVD dyads $D_j = u_jv_j^\top$, joint scalar refitting solves

$$\min_{c\in\mathbb{R}^r}\frac{1}{2}\left\|\Delta - \sum_{j=1}^r c_j D_j\right\|_{\mathrm{F}}^2.$$

The dyads are orthonormal in Frobenius inner product, so the unique minimizer on the selected span is $c_j = \langle\Delta, D_j\rangle = \sigma_j$ up to arbitrary choices inside tied singular subspaces. Finally, for any fixed dyads and any fixed stagewise scalar values, those values are feasible for the joint least-squares problem; the optimizer over the same fixed span cannot have larger squared loss.

Proof of Proposition D.2. For isotropic sub-Gaussian features, the standard sample-covariance bound (see, e.g., Vershynin, 2018) gives, with probability at least $1-\delta$,

$$\left\|\widehat{\Sigma}_{\mathcal{B}_k} - I\right\|_2 \le C_K \left[\sqrt{\frac{d+\log(2/\delta)}{N}} + \frac{d+\log(2/\delta)}{N}\right] = \gamma_{\mathrm{ff}},$$

where C_K depends only on the sub-Gaussian norm bound. Since $\widetilde{A}_k = \widehat{\Sigma}_{\mathcal{B}_k} R_k^W$,

$$\left\|\widetilde{A}_k - R_k^W\right\|_{\mathrm{F}} = \left\|(\widehat{\Sigma}_{\mathcal{B}_k} - I)R_k^W\right\|_{\mathrm{F}} \le \left\|\widehat{\Sigma}_{\mathcal{B}_k} - I\right\|_2 \left\|R_k^W\right\|_{\mathrm{F}} \le \gamma_{\mathrm{ff}} \left\|R_k^W\right\|_{\mathrm{F}}.$$

Therefore

$$\left\|\widetilde{A}_k - \Delta_k^\star\right\|_{\mathrm{F}} \le \gamma_{\mathrm{ff}} \left\|R_k^W\right\|_{\mathrm{F}} + \left\|R_k^W - \Delta_k^\star\right\|_{\mathrm{F}}.$$

If this quantity is at most $\mathcal{T}_k^\star/4$, Wedin's singular-vector perturbation bound applied to $\widetilde{A}_k$ and $\Delta_k^\star$ gives the displayed dyad estimate, up to a universal numerical constant. Weyl's inequality gives

$$\left|\sigma_1(\widetilde{A}_k) - \sigma_1(\Delta_k^\star)\right| \le \gamma_{\mathrm{ff}} \left\|R_k^W\right\|_{\mathrm{F}} + \left\|R_k^W - \Delta_k^\star\right\|_{\mathrm{F}}.$$

Combining this amplitude perturbation with the dyad perturbation exactly as in the proof of Proposition 4.5 yields the stated bound on $\|U_k - U_k^\star\|_{\mathrm{F}}$.

D.5 Frozen-feature bridge experiments

We use the frozen-feature bridge in two complementary ways. In a *synthetic-feature bridge* we replace real labels with a known low-rank Leaky-ReLU teacher and measure the same per-stage diagnostics as in the synthetic suite, since the teacher shift is known and we can compute oracle residuals, gaps, dyad error, sector angles, and the perturbation ratio of Theorem 4.1. In a *real-label bridge* we fit a single classifier head over the same frozen features and report the rank/active-state frontier that motivates the procedure. Proposition D.1 gives the exact identity $-\nabla\mathcal{L}_{\mathrm{ff}} = \Sigma R$ for a noiseless linear-head teacher under squared loss. This motivates testing covariance control through whitening. Our nonlinear synthetic teacher and real-label tasks do not satisfy that proposition's full hypotheses, so these experiments test the covariance-control motivation rather than verify its exact gradient identity.

Datasets and backbones. We use CIFAR-10 and CIFAR-100 (Krizhevsky, 2009) and Flowers102 (Nilsback & Zisserman, 2008). We pass each image through an ImageNet-pretrained ResNet-18 (He et al., 2016) or MobileNetV3-Small (Howard et al., 2019) and freeze the backbone. We sample at most 12000 training images and 3000 test images at a fixed feature seed. By default, we standardize coordinates using training means and standard deviations, project to dimension 256 with a fixed-seed, data-independent random orthonormal matrix, and rescale projected coordinates by their training standard deviations. We then whiten using the training covariance with eigenvalues floored at $5 \cdot 10^{-3}$ and rescale once more by training standard deviations. All data-dependent transformations are fitted once on the selected training features, held fixed across stages, and applied unchanged to test features. We do not update any backbone parameter.

Real-label adaptation. For the real-label bridge we first fit a frozen linear head on 4000 labeled examples for 500 Adam steps at learning rate $5 \cdot 10^{-2}$. We then adapt only the head matrix. Locked methods use fresh stage batches of size 4096. Scalar Adam takes 120 steps at learning rate $5 \cdot 10^{-2}$. The free-factor sequential baseline shares the SVD initialization but optimizes both factors for 160 Adam steps at learning rate $3 \cdot 10^{-2}$. The simultaneous baseline optimizes all factors jointly for 650 Adam steps at the same factor learning rate. We sweep final ranks $\{1, 2, 4, 6\}$, run each frontier with cross-entropy and with a one-vs-rest squared-loss head initialized by ridge regression, and use five seeds 11–15. Unless stated otherwise, every $\pm$ in this appendix and every error bar in Figures 16, 14, 17, and 18 denotes one sample standard deviation (divisor $n-1$) of the per-seed test accuracy across these five seeds, not a standard error. The squared-loss head is closer to the bridge proposition because the loss is quadratic in the head output.

We measure active Adam-state memory as a function of trainable adaptation parameters: zero for locked spectral extraction, one scalar for locked scalar refinement, r scalars for fixed-direction refitting,

one rank-1 factor pair for free-factor sequential training, and rank-r factors for simultaneous training. We also record active optimizer-state size and adaptation FLOPs. At rank 6 with cross-entropy, scalar refitting uses $4.8 \cdot 10^{-5}$ MB of active Adam state for the trainable scalars, whereas simultaneous training uses 0.0128, 0.0171, and 0.0172 MB on CIFAR-10, CIFAR-100, and Flowers102 respectively. On CIFAR-10 the one-run scalar-refit frontier reaches $75.1\% \pm 1.3\%$ at rank 6 after 22.2 head-adaptation GFLOPs, while separately training simultaneous ranks $1, 2, 4, 6$ reaches $77.6\% \pm 0.2\%$ at rank 6 after 53.9 cumulative head-adaptation GFLOPs.

Synthetic-feature bridge. We replace CIFAR labels with a known Leaky-ReLU teacher of 24 outputs, rank 4, $\alpha = 0.05$, and shift Frobenius norm 0.65. Because the teacher is known we report oracle residuals, residual gaps, dyad error, sector angles, and the perturbation ratio of Theorem 4.1. We run the diagnostic on raw and standardized features at the full backbone dimension (512 for ResNet-18 and 576 for MobileNetV3-Small), on features projected to dimension 256 with PCA, on the same PCA-projected features after whitening with additive ridge values $\{10^{-1}, 10^{-2}, 5 \cdot 10^{-3}\}$ and training-based coordinate rescaling, and on a randomly rotated whitened control. Unlike the default random projection, this ablation uses the leading 256 eigenvectors of the standardized training-feature covariance, with the same fitted transformations applied to test features. The PCA-projected variant provides a dimension-matched reference for the additional whitening and rescaling steps. The Gaussian theorem does not apply on these features. We use the diagnostic to measure how close the bridge is to the local geometry needed by the theorem before considering larger fine-tuning runs.

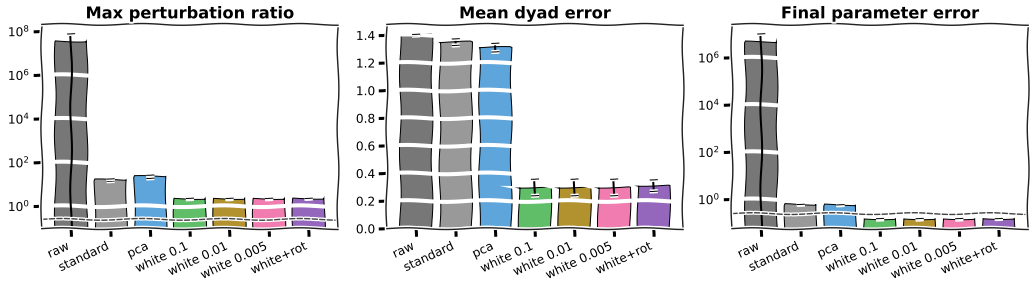


Figure 12: Frozen-feature bridge: effect of preprocessing on a known low-rank teacher. Two pretrained backbones (ResNet-18, MobileNetV3-Small) with raw, standardized, PCA-projected, and whitened features (three ridge values). Whitening is the preprocessing that brings $\widehat{\Sigma}$ closest to I; consistent with the $\|\Sigma - I\|_2 \|R\|_{\mathrm{F}}$ perturbation bound of Proposition D.1, it improves the perturbation ratio, top dyad error, and final parameter error compared with the alternatives. The direct quarter-gap threshold is not reached on either backbone, so the figure illustrates the bridge mechanism rather than verifying the Gaussian theorem on real features.

Whitening drives the bridge perturbation ratio down (Figure 12). Figure 12 reports the per-stage diagnostic across preprocessing choices on both backbones. Whitening is the only preprocessing that brings the perturbation ratio close to the direct quarter-gap threshold. Even after whitening this threshold is not reached on either backbone, so the figure illustrates the bridge mechanism but does not verify the Gaussian theorem on real features.

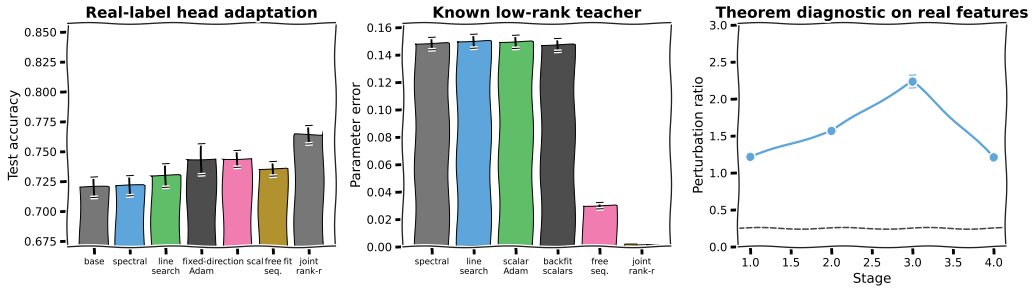


Figure 13: Frozen-feature CIFAR-10 bridges. Left: real-label adaptation of a single classifier head over frozen pretrained features. Middle: parameter recovery for a known low-rank Leaky-ReLU teacher on the same features. Right: empirical perturbation ratio for the locked scalar method on the synthetic-feature bridge; the dashed line is the direct $1/4$ threshold of Theorem 4.1. The threshold is not met on these features even after whitening, so the figure is bridge evidence for the residual-surrogate mechanism (Proposition D.1/Proposition D.2) rather than a verification of the Gaussian Leaky-ReLU theorem on real data.

Three views of the CIFAR-10 bridge (Figure 13). Figure 13 aggregates three views of the CIFAR-10 bridge: real-label head adaptation, parameter recovery against the known teacher, and the per-stage perturbation ratio from the synthetic-feature bridge. The perturbation-ratio panel does not meet the direct quarter-gap threshold, and we read the figure as bridge evidence for the residual-surrogate mechanism rather than as a verification of the Gaussian theorem.

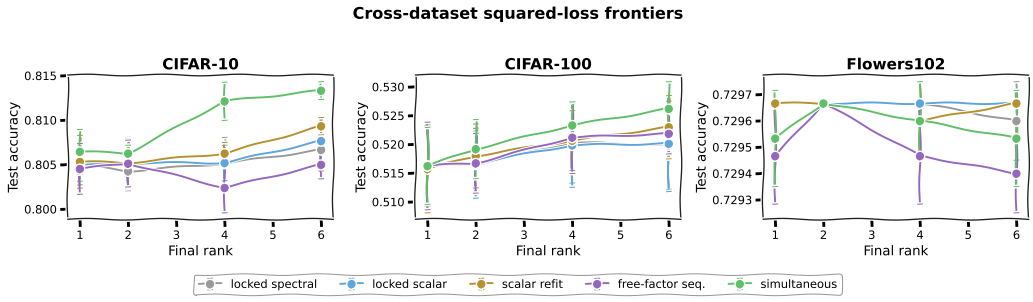


Figure 14: Squared-loss one-vs-rest head frontiers on frozen ResNet-18 features. This bridge is closer to the analyzed setting than cross-entropy: the adapted object is a single matrix and the loss is quadratic in the head outputs, matching Proposition D.1. At rank 6, fixed-direction scalar refitting nearly matches simultaneous rank-r training on all three datasets. The figure is empirical evidence for the active-state/rank-frontier tradeoff, not a verification of the Gaussian theorem in deep settings.

Frontiers under squared loss (Figure 14). Figure 14 reports rank/accuracy frontiers under the one-vs-rest squared-loss head, which is the loss most aligned with Proposition D.1. At rank 6, fixed-direction scalar refitting nearly matches simultaneous rank-r training on all three datasets. This is the closest the bridge gets to the analyzed setting and is empirical evidence for the active-state/rank-frontier tradeoff, not a verification of the Gaussian theorem.

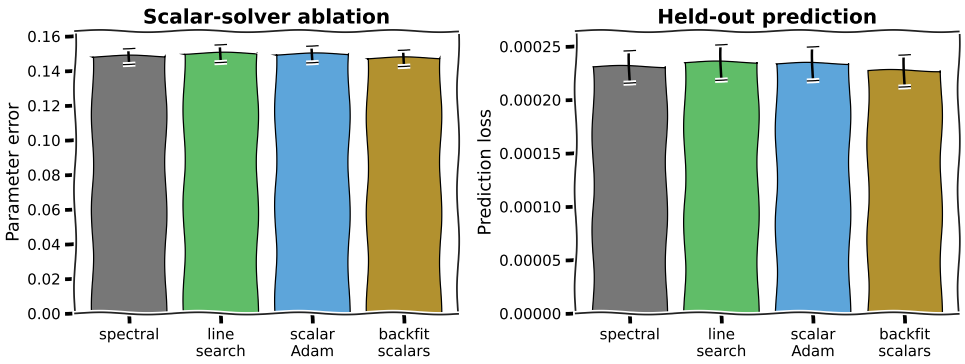


Figure 15: Scalar-solver ablation on the synthetic-feature bridge. Spectral amplitude, numerical line search, scalar Adam, and scalar backfitting yield similar final errors after whitening.

Scalar-solver ablation under the bridge (Figure 15). In Figure 15 we repeat the synthetic scalar-solver ablation on the synthetic-feature bridge. Spectral amplitude, line search, scalar Adam, and scalar backfitting give similar final errors after whitening.

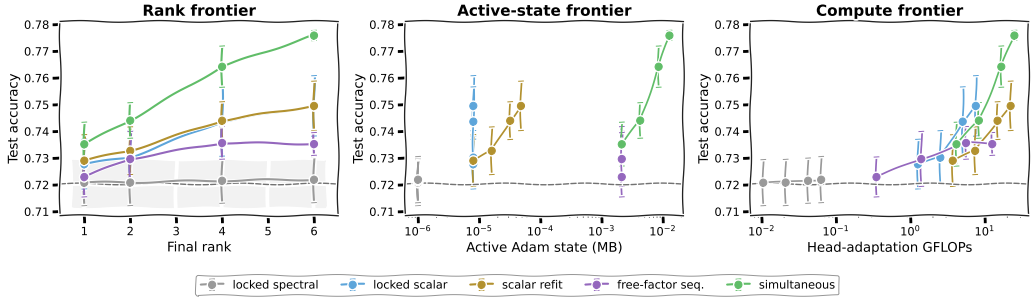


Figure 16: Frozen-feature CIFAR-10 rank/active-state frontier. Adapted object: one classifier head over frozen ImageNet-pretrained ResNet-18 features, with the base head fit on 4000 labels. Fixed-direction scalar refitting is a bridge for the analyzed setting (Proposition D.1/Proposition D.2) and produces an entire rank frontier in a single run with bounded active optimizer state. Simultaneous rank-r training is an empirical baseline outside the theorem: it is more accurate at each rank but uses rank-r active optimizer state and must be retrained per target rank.

Cross-entropy active-state frontier (Figure 16). Figure 16 reports the cross-entropy rank/active-state frontier on CIFAR-10 with the same 4000-example base head. Fixed-direction scalar refitting yields the entire rank frontier in a single run with bounded active optimizer state. Simultaneous rank-r training is more accurate at each rank but uses rank-r active state and must be retrained for every target rank. Cross-entropy is outside the squared-loss bridge of Proposition D.1, so we report this only for empirical context.

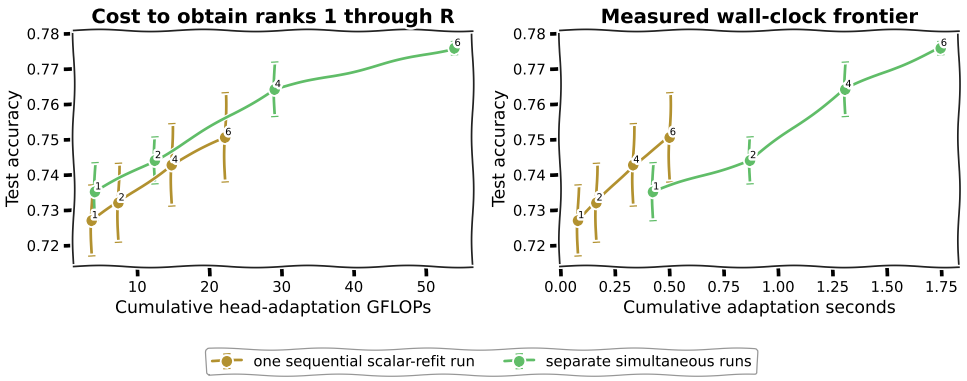


Figure 17: Anytime rank frontier on CIFAR-10 with the same 4000-example base head as Figure 16. A single fixed-direction scalar-refit run yields ranks $1, 2, 4, 6$ along one path; the simultaneous baseline is more accurate at each rank but requires separately training a model per target rank. The figure illustrates the active-state/anytime tradeoff implied by sequential rank-one extraction; the underlying Gaussian theorem is not exercised here.

Anytime frontier from a single run (Figure 17). Figure 17 highlights the anytime nature of the procedure: a single fixed-direction scalar-refit run produces ranks $1, 2, 4, 6$ along one path. The simultaneous baseline is more accurate at each rank but requires a separate model per target rank.

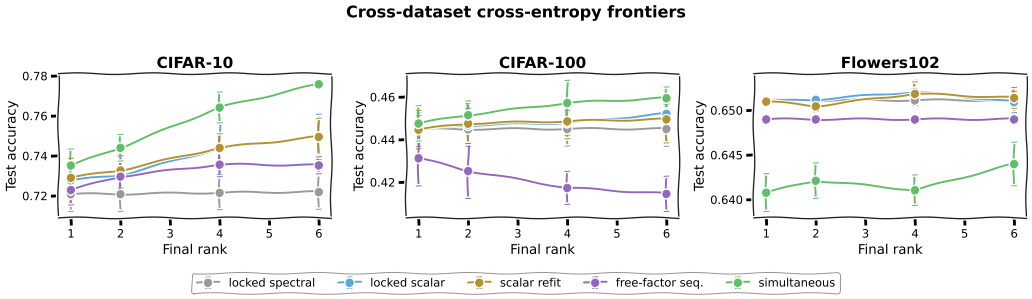


Figure 18: Cross-dataset cross-entropy head frontiers on frozen ResNet-18 features (base heads fit on 4000 labels). CIFAR-10 shows a useful active-state tradeoff for fixed-direction scalar refitting ($72.0\% \to 75.0\%$ from the base head to rank 6, against 77.6% for simultaneous rank-6 training); on CIFAR-100 the refit gain ($44.5\% \to 45.0\%$) is within one seed standard deviation (≈ 1 pp); Flowers102 is saturated by the frozen base head and is included as a control where sequential updates do not help. Cross-entropy is outside the squared-loss bridge of Proposition D.1, so this panel is empirical evidence for the active-state tradeoff rather than a verification of the bridge.

Cross-dataset frontiers (Figure 18). In Figure 18 we extend the cross-entropy frontier to CIFAR-10, CIFAR-100, and Flowers102. CIFAR-10 shows a useful active-state tradeoff for fixed-direction scalar refitting ($72.0\% \to 75.0\%$ from the base head to rank 6, against 77.6% for simultaneous rank-6 training). On CIFAR-100 the refit gain ($44.5\% \to 45.0\%$) is within one seed standard deviation (≈ 1 pp), while simultaneous rank-6 training reaches 46.0%. Flowers102 is saturated by the frozen base head, and we include it as a control where sequential updates do not help.

E Gradient-Selected versus Random Directions in a Transformer

We test the value of gradient-based direction selection at a fixed adapter-parameter budget in DeBERTa-v3-base. Both methods add eight frozen rank-one directions to each query, key, and value projection, giving 288 trainable adapter amplitudes in addition to the task head (pooler and classifier); the pretrained backbone weights remain frozen. The gradient-selected variant uses leading

singular vectors of the negative loss gradient on one mini-batch per extraction. The random control instead samples independent unit vectors. These correspond to the `m2` and `v9` run identifiers, respectively. Standard rank-1 LoRA on the same query, key, and value projections has $55{,}296$ trainable adapter parameters, compared with 288 scalar amplitudes here. The random-direction control therefore isolates direction selection at matched adapter capacity. The LoRA reference runs in Table 3 place rank-1 and rank-2 adapters on the same projections (A Kaiming-uniform, B zero-initialized, scale $\alpha_{\mathrm{LoRA}}/r = 2$, no LoRA dropout, no adapter weight decay) and use adapter learning rate 3×10^{-4} for both ranks, selected by the same three-point sweep on full-data CoLA and MRPC that fixed 3×10^{-3} for the amplitudes, and not re-tuned per cell; rank 1 was not swept. An additional rank-2 arm at 10^{-3} (ten seeds, all cells) raises SST-2 at $k = 16$ to 80.75 ± 3.79 and MNLI at $k = 16$ to 51.06 ± 5.31 and changes the other cells by less than 1.5 points. The head-only reference trains the pooler and classifier alone.

Protocol. After a common task-head warm-up, directions are extracted sequentially and frozen. New amplitudes start at zero, while all accumulated amplitudes and the task head are jointly trained between extractions. AdamW state persists across adaptation stages, and both methods reuse the finite training subset. Runs are paired by seed (42–51) and training subset, with $k = 16, 64, 256$ examples per class. The respective head warm-up budgets are $300, 400, 600$ steps, followed by $700, 1200, 2000$ adaptation steps, with $30, 40, 60$ steps between successive extractions. Remaining adaptation steps follow the eighth extraction. Both methods use batch size 32, maximum sequence length 128, head learning rate 10^{-4}, and adapter learning rate 3×10^{-3}. This matches adapter capacity and optimization-step budgets, not total compute. Each of the eight gradient-selected extractions adds one forward–backward pass on a mini-batch of 32 examples in fp32, outside the bf16 autocast used for training steps, and a dense SVD of each of the 36 query, key, and value projection gradients (768×768 each), retaining only the leading singular pair. In total, this adds eight gradient passes and 288 SVDs; the random control instead samples unit vectors.

We did not profile extraction separately. The recorded wall-clock times include model loading, data preparation, head warm-up, adaptation, and periodic and final development-set evaluation. Across 120 seed-matched pairs (four tasks, three training-set sizes, ten seeds), the mean recorded time was 441.3 s for gradient selection and 348.8 s for random directions, a mean paired difference of 92.5 s and a ratio of means of 1.27. Gradient-selected runs took longer in 116 of 120 pairs (Table 4). The campaign ran multiple workers on each H100 GPU, so these timings include resource contention. They describe whole-run times under that execution setup, not isolated extraction cost or a controlled hardware-speed comparison.

Table 3: Gradient-selected versus random frozen directions in DeBERTa-v3-base, with head-only and LoRA reference points. Final development metric, mean ± sample standard deviation over ten paired seeds (42–51); k is the number of training examples per class. The metric is accuracy (%) for SST-2 and MNLI (matched development set), Matthews correlation $\times 100$ for CoLA, and F1 $\times 100$ for MRPC. The second header row gives trainable adapter parameters: 288 amplitudes for each direction method, 0 for the head-only run, and 55,296 and 110,592 for LoRA with $r = 1$ and $r = 2$ on the query, key, and value projections of all twelve layers ($\alpha_{\mathrm{LoRA}} = 2r$). Every method additionally trains the same task head (pooler and classifier; 592,130 parameters, 592,899 for MNLI) at learning rate 10^{-4}, on the same seed-determined training subset, with the same head warm-up, adaptation-step budget, and final-step evaluation. LoRA uses one fixed adapter learning rate, 3×10^{-4}, for both ranks and all cells, without per-cell retuning. Gradient-selected and random directions form the matched-capacity control; the head-only run and LoRA are reference points at their respective adapter budgets, not budget-matched baselines.

Task	k	Gradient-selected 288	Random 288	Head only 0	LoRA r=1 55,296	LoRA r=2 110,592
SST-2	16	78.99 ± 4.28	68.13 ± 3.24	67.32 ± 3.74	78.61 ± 5.07	77.67 ± 5.37
	64	86.67 ± 3.46	70.45 ± 5.18	68.27 ± 5.87	87.31 ± 2.69	87.81 ± 2.84
	256	90.81 ± 0.91	80.01 ± 3.27	75.67 ± 3.28	91.87 ± 0.71	92.08 ± 1.02
MNLI	16	49.21 ± 4.47	41.92 ± 2.53	41.60 ± 2.51	48.38 ± 6.19	48.94 ± 5.25
	64	69.65 ± 1.55	51.59 ± 2.50	49.74 ± 3.00	70.08 ± 3.09	71.14 ± 2.54
	256	76.60 ± 1.16	55.58 ± 4.52	52.00 ± 5.00	80.56 ± 0.61	80.77 ± 0.84
CoLA	16	44.43 ± 4.99	41.05 ± 6.69	39.88 ± 7.16	45.97 ± 4.11	45.79 ± 4.03
	64	51.95 ± 3.04	51.81 ± 2.26	51.06 ± 2.27	52.62 ± 3.20	52.98 ± 2.37
	256	60.66 ± 2.13	54.65 ± 1.33	53.64 ± 0.95	60.05 ± 1.87	59.44 ± 1.78
MRPC	16	77.77 ± 5.39	77.09 ± 6.54	77.07 ± 7.52	77.94 ± 5.15	79.14 ± 3.17
	64	81.51 ± 3.18	80.91 ± 1.81	80.24 ± 1.83	80.99 ± 2.44	81.33 ± 1.95
	256	85.64 ± 2.08	82.56 ± 1.70	81.69 ± 2.66	86.94 ± 1.39	87.90 ± 1.49

Table 4: Recorded wall-clock time per run (s) for the paired runs of Table 3, mean ± sample standard deviation over ten seeds. Times include model loading, data preparation, head warm-up, adaptation, and periodic and final evaluation, with several runs sharing each GPU. "Extra" is the mean paired difference; "Ratio" is the ratio of means; "Slower" counts seeds in which the gradient-selected run took longer. Differences and ratios are computed from unrounded values.

Task	k	Gradient-selected	Random	Extra	Ratio	Slower
SST-2	16	415 ± 28	307 ± 8	$+108$	1.35	10/10
	64	377 ± 32	282 ± 14	$+95$	1.34	10/10
	256	436 ± 56	326 ± 34	$+110$	1.34	10/10
MNLI	16	471 ± 52	379 ± 43	$+92$	1.24	9/10
	64	591 ± 84	516 ± 65	$+75$	1.15	10/10
	256	604 ± 86	544 ± 75	$+60$	1.11	8/10
CoLA	16	452 ± 43	298 ± 13	$+154$	1.52	10/10
	64	376 ± 33	270 ± 16	$+106$	1.39	10/10
	256	416 ± 49	314 ± 33	$+102$	1.32	10/10
MRPC	16	387 ± 26	303 ± 8	$+83$	1.27	10/10
	64	345 ± 32	284 ± 20	$+61$	1.22	10/10
	256	424 ± 83	361 ± 35	$+63$	1.17	9/10

Results and scope. On SST-2 and MNLI, gradient-selected directions outperform random directions in all ten paired seeds at each tested training-set size (Table 3). Mean paired accuracy gains range from 7.29 to 21.02 percentage points. Benefits vary across tasks: at $k = 16, 64, 256$, the corresponding mean gains are $3.38, 0.15, 6.01$ points on CoLA (Matthews correlation $\times 100$) and $0.68, 0.61, 3.08$

points on MRPC (F1 $\times 100$), so the low-shot effects there are less pronounced. For $k = 16, 64, 256$, gradient-selected directions outperform the random control in 8, 6, and 10 of the ten paired CoLA runs, and in 8, 6, and 8 of the ten paired MRPC runs, respectively. Random directions add up to approximately 4.3 points over the head-only run. At $k \leq 64$ the 288-amplitude method's mean scores are within 1.6 points of both LoRA ranks in all eight cells; at $k = 256$ LoRA $r{=}1$ leads by 1.1 (SST-2), 4.0 (MNLI), and 1.3 (MRPC) points and trails by 0.6 on CoLA. These comparisons are at LoRA's fixed learning rate and carry no significance test. These results support the practical value of gradient-informed direction selection at a fixed adapter-parameter budget. They concern a multilayer, joint-amplitude-refit variant with reused data, rather than the freeze-each-update procedure in Algorithm 1, and do not extend its fresh-batch recovery guarantee to deep networks.