SILVER STEPS BEYOND CONVEXITY: ACCELERATED CONVERGENCE FOR LOW-RANK MATRIX RECOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Low-rank matrix recovery problems are often solved through a low-rank factorization, which turns the rank constraint into a nonconvex optimization problem. Recent work has shown that carefully designed Silver stepsize schedules can accelerate gradient descent for convex optimization. In this paper, we study whether this acceleration extends to nonconvex low-rank factorization for matrix recovery from linear measurements. We show that gradient descent with a fixed-depth Silver stepsize schedule achieves an accelerated convergence rate when initialized near a constrained minimizer, while all intermediate iterates remain in its neighborhood despite the occasional large steps in the schedule. We also show that, even in a simple rank-one problem, the same fixed-depth Silver schedule diverges with nonvanishing probability under arbitrarily small Gaussian initialization, demonstrating that the local initialization condition cannot in general be removed. We apply our results to Gaussian matrix sensing and low-rank multivariate regression, and provide numerical experiments that support our theoretical findings.

1 INTRODUCTION

Low-rank matrix recovery problems are often solved by representing the unknown matrix as a product $X = UV^\top$ and optimizing over the two factors (Burer & Monteiro, 2003; Tu et al., 2016; Ma et al., 2021). This parameterization enforces the rank constraint but leads to a nonconvex optimization problem over the factors. Despite this nonconvexity, gradient descent on the factorized problem converges under suitable problem and initialization conditions, including initializations sufficiently close to a solution (Chen & Wainwright, 2015; Tu et al., 2016; Zheng & Lafferty, 2016) and small initializations in several low-rank recovery models (Li et al., 2018; Stöger & Soltanolkotabi, 2021; Soltanolkotabi et al., 2023; Kim & Chung, 2023).

Recently, Altschuler & Parrilo (2025a;b) introduced the Silver stepsize schedule, showing that, surprisingly, gradient descent on smooth convex problems can be accelerated simply by varying its stepsizes. A distinctive feature of the Silver schedule is that it contains occasional large stepsize spikes, which play a key role in its accelerated convergence. This stepsize-only acceleration has since been extended to proximal and projected gradient methods (Bok & Altschuler, 2025) and adapted to Riemannian optimization (Park et al., 2025).

However, existing results for Silver stepsizes rely on convexity or related geometric conditions and do not directly apply to the nonconvex factorized objectives arising in low-rank matrix recovery. Moreover, the large stepsize spikes in the Silver schedule may even prevent convergence in nonconvex problems. This motivates the following question:

> *Can a Silver stepsize schedule accelerate gradient descent for nonconvex low-rank factorization despite these large stepsize spikes?*

In this work, we investigate this question in the setting of low-rank recovery from linear measurements, where the goal is to solve

$$\min_{\mathrm{rank}(X)\le r} f(X), \text{ where } f(X) = \frac{1}{2}\|\mathcal{A}(X) - y\|_2^2.$$

Specifically, we factorize $X = UV^\top$, add a standard balancing penalty to the factorized objective, and run gradient descent with Silver stepsizes. The key challenge is that the large stepsize spikes in

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.

the Silver schedule may push the iterates outside the neighborhood of a solution where the factorized objective has favorable local curvature. To address this issue, we repeat a fixed finite block of Silver stepsizes throughout optimization and analyze the entire trajectory within each block, including both its intermediate iterates and its endpoint.

Our main result shows that, under suitable restricted curvature and initialization conditions, Silver stepsizes locally accelerate gradient descent despite the nonconvexity of low-rank factorization. In particular, a fixed block reduces the factor error by a constant factor using $\mathcal{O}(K^\theta)$ gradient evaluations, while all intermediate iterates remain near the solution, where K is the relevant condition number of the problem and $\theta = \log_{1+\sqrt{2}} 2 \simeq 0.786$ is the Silver exponent. Thus, locally, the method achieves the same dependence on the condition number as Silver acceleration for strongly convex problems (Altschuler & Parrilo, 2025b; Bok & Altschuler, 2025), improving the usual $\mathcal{O}(K)$ dependence of constant-step gradient descent. Moreover, we show that the local initialization condition cannot in general be removed: even for a simple rank-one problem, the same fixed Silver schedule diverges with nonvanishing probability under arbitrarily small Gaussian initialization. Finally, in statistical settings, we derive recovery guarantees for matrix sensing and low-rank multivariate regression using warm starts.

Our main contributions are summarized as follows:

- We show that Silver stepsizes accelerate gradient descent for nonconvex low-rank factorization, reducing the factor error by a constant factor in $\mathcal{O}(K^\theta)$ gradient evaluations.
- We give a rank-one counterexample showing that the local initialization condition cannot in general be removed, even under arbitrarily small Gaussian initialization.
- Under statistical observation models, we derive recovery guarantees for the accelerated method in matrix sensing and low-rank multivariate regression.

2 Related work

Silver stepsize acceleration. Altschuler & Parrilo (2025a;b) recently introduced the Silver stepsize schedule, which achieves acceleration for smooth convex and strongly convex optimization by modifying only the stepsizes. Subsequent work studies composition rules for constructing longer schedules (Grimmer et al., 2025) and optimal recursive splits within such constructions (Liu et al., 2026), while Ye & Liu (2026) establish worst-case lower bounds matching the Silver exponent up to subpolynomial factors for gradient descent with predetermined nonnegative stepsizes. Silver stepsizes have also been extended to proximal and projected gradient descent (Bok & Altschuler, 2025) and Riemannian optimization (Park et al., 2025). These results do not directly apply to the Euclidean factor-gradient setting considered here, where the objective is nonconvex and large Silver steps may leave the region of favorable local curvature.

Local convergence of low-rank factorizations. Low-rank factorization is a classical approach to reducing the computational cost of semidefinite optimization (Burer & Monteiro, 2003). For matrix recovery, gradient methods on the factorized problem are known to converge geometrically under suitable curvature and initialization conditions (Zheng & Lafferty, 2015; Candes et al., 2015; Chen & Wainwright, 2015; Tu et al., 2016; Barber & Ha, 2018), with related guarantees for matrix completion (Sun & Luo, 2016; Zheng & Lafferty, 2016). Ma et al. (2021) show that an explicit balancing penalty is unnecessary for asymmetric matrix sensing with spectral initialization, while more recent work improves the rank dependence of the sample complexity for spectrally initialized factorized gradient descent in positive-semidefinite Gaussian sensing (Stöger & Zhu, 2025). Motivated by these local convergence results, we investigate whether the same factorized gradient descent method can be accelerated by changing only its stepsizes.

Global geometry and small random initialization. Landscape and critical-point analyses establish the absence of spurious local minima and strict-saddle properties for suitable low-rank recovery problems (Bhojanapalli et al., 2016; Ge et al., 2016; 2017; Zhu et al., 2018; Ha et al., 2020). Such favorable geometry can support global optimization with appropriate algorithms, but does not by itself guarantee convergence under a Silver stepsize schedule. A separate line of work studies implicit regularization in matrix factorization (Arora et al., 2019), including recovery guarantees for

gradient descent with small initialization in overparameterized sensing (Li et al., 2018). Stöger & Soltanolkotabi (2021) show how small random initialization induces an initial spectral-alignment phase, with related analyses for asymmetric overparameterized sensing (Soltanolkotabi et al., 2023), symmetric rank-one matrix completion (Kim & Chung, 2023), and perturbed gradient-flow analyses for asymmetric matrix sensing (Wind, 2026). Our result gives a complementary counterexample: even in rank-one matrix sensing with full observation, the fixed Silver schedule diverges with probability bounded away from zero as the Gaussian initialization scale tends to zero. Thus, small random initialization alone does not guarantee convergence of the Silver schedule considered here.

Momentum and preconditioning. Other approaches accelerate low-rank optimization by modifying the factor dynamics. Zhou et al. (2020) develop a Nesterov-inspired factored method using momentum, while scaled gradient descent (Tong et al., 2021) preconditions each factor gradient using the Gram matrix of the other factor, removing the condition number of the solution matrix from the local iteration rate in suitable settings. Related developments include damped preconditioning for overparameterized sensing with small random initialization (Xu et al., 2026), global convergence from broader random initializations (Jia et al., 2023), improved sampling guarantees for ScaledGD (Li & Huang, 2026), and weight normalization with Riemannian optimization (Wei et al., 2026). In contrast to these works, we preserve the factor-gradient direction and vary only the stepsizes. Our goal is to understand how much acceleration can be achieved through stepsize scheduling alone.

3 Preliminaries and problem setup

3.1 Low-rank matrix sensing

Given a linear operator $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ and observations $y \in \mathbb{R}^m$, we consider the following rank-constrained optimization problem

$$\min_{\mathrm{rank}(X) \le r} f(X), \text{ where } f(X) = \frac{1}{2}\|\mathcal{A}(X) - y\|_2^2.$$

Let $\widehat{X} \in \mathrm{argmin}\{f(X) : \mathrm{rank}(X) \le r\}$ be a constrained minimizer, and assume throughout that $\mathrm{rank}(\widehat{X}) = r$. To optimize over low-rank matrices, we parameterize $X = UV^\top$, where $U \in \mathbb{R}^{d_1 \times r}$ and $V \in \mathbb{R}^{d_2 \times r}$, and write $W = (U, V)$ for the pair of factors. For a factor pair $W = (U, V)$, we write $\|W\|^2 := \|U\|_F^2 + \|V\|_F^2$ and $WQ := (UQ, VQ)$ for $Q \in \mathbb{R}^{r \times r}$. Following Tu et al. (2016); Zheng & Lafferty (2016), we add a standard balancing penalty to the factorized objective and consider

$$g_\lambda(U, V) = f(UV^\top) + \frac{\lambda}{2}\|U^\top U - V^\top V\|_{\mathrm{F}}^2, \qquad \lambda = \frac{\alpha}{4},$$

where α is the restricted curvature constant of $\mathcal{A}$ defined later (cf. Definition 3.1). Throughout, we apply gradient descent to g_λ with respect to the pair of factors $W = (U, V)$.

Fix a balanced factorization of $\widehat{X}$, denoted by $\widehat{W} = (\widehat{U}, \widehat{V})$, satisfying $\widehat{U}\widehat{V}^\top = \widehat{X}$ and $\widehat{U}^\top\widehat{U} = \widehat{V}^\top\widehat{V}$. Since applying the same orthogonal transformation to both factors leaves their product unchanged, we measure the factor error after optimal alignment by

$$\mathrm{dist}(W, \widehat{W}) := \min_{Q \in O(r)} \left(\|UQ - \widehat{U}\|_{\mathrm{F}}^2 + \|VQ - \widehat{V}\|_{\mathrm{F}}^2\right)^{1/2},$$

where $O(r) = \{Q \in \mathbb{R}^{r \times r} : Q^\top Q = I_r\}$ is the set of orthogonal matrices. Further details on balanced factorizations and the alignment distance are provided in Appendix B.5.

3.2 Restricted curvature and local geometry

We characterize the curvature of f along low-rank directions through the following property of the sensing operator (Negahban & Wainwright, 2012; Tu et al., 2016):

Definition 3.1 (Restricted strong convexity and smoothness)**.** The operator $\mathcal{A}$ satisfies rank-$2r$ restricted strong convexity and smoothness with constants $0 < \alpha \le \beta$ if

$$\alpha\|H\|_{\mathrm{F}}^2 \le \|\mathcal{A}(H)\|_2^2 \le \beta\|H\|_{\mathrm{F}}^2, \text{ for every } H \text{ with } \mathrm{rank}(H) \le 2r.$$

The corresponding restricted condition number is $\kappa := \beta/\alpha$.

At the minimizer $\widehat{X}$, the first-order optimality condition along the rank-r manifold gives $\nabla f(\widehat{X})\widehat{V} = 0, \nabla f(\widehat{X})^\top \widehat{U} = 0$ (Rockafellar & Wets, 1998). The gradient $\nabla f(\widehat{X})$ may nevertheless have a nonzero component, since constrained optimality only requires stationarity along tangent directions. We control its component normal to the rank-r manifold through the following assumption:

Assumption 3.2. For some fixed $0 \le \gamma \le 1/4$,

$$\|\nabla f(\widehat{X})\|_{\mathrm{op}} \le \gamma \alpha \sigma_r(\widehat{X}).$$

Assumption 3.2 is a deterministic condition at the minimizer $\widehat{X}$. In many statistical settings, the gradient at the true parameter consists of observation or sampling noise, which is typically required to be small relative to the signal strength for accurate recovery. Thus, when $\widehat{X}$ is close to the true parameter, it is natural to expect Assumption 3.2 to hold under a sufficiently small noise level. Section 5 makes this connection explicit by giving a sufficient condition on the noise for Assumption 3.2.

Next, define the condition number of $\widehat{X}$ by $\kappa_X := \sigma_1(\widehat{X})/\sigma_r(\widehat{X})$ and let $K := \kappa\kappa_X$. Let $\widetilde{Q} \in O(r)$ be an optimal alignment in the factor error defined in Section 3.1:

$$\widetilde{Q} \in \arg\min_{Q \in O(r)} \left\{ \|UQ - \widehat{U}\|_{\mathrm{F}}^2 + \|VQ - \widehat{V}\|_{\mathrm{F}}^2 \right\}^{1/2}.$$

The aligned error $(U\widetilde{Q} - \widehat{U}, V\widetilde{Q} - \widehat{V})$ is orthogonal to the directions that rotate the two factors together, which are flat directions of g_λ. Along the remaining directions, Assumption 3.2 ensures that the curvature of g_λ at $\widehat{W}$ is, up to constant factors, between $\alpha\sigma_r(\widehat{X})$ and $\beta\sigma_1(\widehat{X})$. Consequently, the Hessian $\nabla^2 g_\lambda(\widehat{W})$, restricted to the orthogonal complement of the flat directions, has condition number $\mathcal{O}(K)$. The precise statement and proof are given in Appendix B.5.

3.3 Silver blocks and their quadratic behavior

We choose the stepsizes of gradient descent by repeatedly applying the same depth-k Silver schedule. Specifically, let $\rho = 1 + \sqrt{2}$ be the silver ratio and set $\theta = \log_\rho 2$. Following Altschuler & Parrilo (2025a), the depth-k Silver schedule is defined recursively by

$$\alpha^{(0)} = [\,], \qquad \alpha^{(k)} = \big[\alpha^{(k-1)},\, 1 + \rho^{k-2},\, \alpha^{(k-1)}\big].$$

Here, the depth k denotes the number of recursive levels, and the resulting block contains $n_k = 2^k - 1$ stepsizes. Thus, each block consists of two copies of the depth-$(k-1)$ block with a larger central step between them. We keep k fixed throughout optimization and repeatedly apply the same block, restarting the schedule at each block boundary while continuing from the current iterate. We refer to this as the fixed-depth Silver schedule.

For a quadratic objective with condition number of order K, the analysis of Altschuler & Parrilo (2025b) motivates choosing the depth so that $\rho^k \asymp K$. With this choice, one Silver block gives a constant-factor reduction at its endpoint using $\mathcal{O}(K^\theta)$ gradient evaluations. For the nonconvex factorized objective, however, endpoint contraction alone is not sufficient to establish local convergence, since the large steps within a block may move intermediate iterates outside the local region where the quadratic approximation is valid. Appendix B.6 provides the quadratic contraction and the associated polynomial bounds used in our analysis.

4 Main results: local Silver acceleration

4.1 Local Silver acceleration for low-rank matrix sensing

We now study whether Silver stepsizes accelerate gradient descent on the factorized problem when initialized sufficiently close to a solution. Our main result gives an affirmative answer under the assumptions introduced above. Before stating the theorem, we first specify the block depth, stepsize normalization, and update rule used in our result.

Choose a block depth k satisfying $192K \le \rho^k < 192\rho K$, and set $\overline{\beta} := 4\beta\sigma_1(\widehat{X})$, $n_k := 2^k - 1$. Using these fixed parameters, run gradient descent on g_λ by repeating the depth-k Silver schedule every n_k steps, i.e.,

$$W_{t+1} = W_t - \frac{\alpha^{(k)}_{t \bmod n_k}}{\overline{\beta}} \nabla g_\lambda(W_t).$$

The following theorem gives our local acceleration guarantee for the fixed-depth Silver schedule.

Theorem 4.1. *Suppose Definition 3.1 and Assumption 3.2 hold. Let $\widehat{X}$ be a rank-r constrained minimizer with balanced factors $\widehat{W}$. Then there exists a universal constant $c > 0$ such that, if the initialization $W_0 = (U_0, V_0)$ satisfies*

$$D_0 := \operatorname{dist}(W_0, \widehat{W}) \le c\frac{\sqrt{\sigma_1(\widehat{X})}}{K},$$

then each block halves the factor error at its endpoint while keeping all intermediate iterates in the local neighborhood, i.e., for every $b \ge 0$,

$$\operatorname{dist}(W_{bn_k}, \widehat{W}) \le 2^{-b} D_0, \qquad \max_{0 \le s \le n_k} \operatorname{dist}(W_{bn_k+s}, \widehat{W}) \le 2\operatorname{dist}(W_{bn_k}, \widehat{W}).$$

Hence, for $0 < \varepsilon_W < D_0$, factor error at most ε_W is attained within

$$T = \mathcal{O}\left(K^\theta \left[1 + \log \frac{D_0}{\varepsilon_W}\right]\right),$$

gradient evaluations, with constants independent of the rank and the ambient dimensions.

The proof of Theorem 4.1 is given in Appendix A. The theorem shows that a fixed-depth Silver schedule achieves local acceleration while keeping all intermediate iterates controlled. The same block depth-k and normalization $\overline{\beta}$ can be used at every restart, without any adjustment as the iterates approach the solution. Despite the nonconvex factorization and its rotational symmetry, the resulting complexity matches the K^θ-type dependence of the Silver rate for strongly convex objectives (Altschuler & Parrilo, 2025b; Bok & Altschuler, 2025), with K playing the role of the relevant local condition number.

To relate factor error to matrix reconstruction accuracy, we apply Lemma A.1, which gives

$$\|U_T V_T^\top - \widehat{X}\|_{\mathrm{F}} \le 3\sqrt{\sigma_1(\widehat{X})}\operatorname{dist}(W_T, \widehat{W}), \tag{1}$$

whenever $\operatorname{dist}(W_T, \widehat{W}) \le \sqrt{\sigma_r(\widehat{X})}$. Therefore, the convergence in factor error in Theorem 4.1 also implies geometric convergence of the reconstructed matrix $U_T V_T^\top$ to the solution $\widehat{X}$.

Proof sketch. The key idea of the proof is to compare gradient descent on the factorized objective with the Silver iteration for its local quadratic approximation around the minimizer. At the beginning of each block, we optimally align the factors and let E_t denote the error from $\widehat{W}$. For a sufficiently small factor perturbation E, the gradient can be written as

$$\nabla g_\lambda(\widehat{W} + E) = \nabla^2 g_\lambda(\widehat{W})E + \mathcal{O}\left(\beta\sqrt{\sigma_1(\widehat{X})}\,\|E\|^2\right).$$

Thus, as long as the aligned errors E_t remain small, each factor update is the corresponding quadratic Silver update with an additional error that is second order in the aligned factor error.

Now the main difficulty is to verify this locality despite the large steps in the Silver schedule. The rotational symmetry of the factorization creates flat directions, so standard local strong convexity arguments do not apply directly. We overcome this by introducing a small quadratic correction along the rotation directions, which yields a local convexity inequality with an error that is second order in the factor distance. Combined with a finite-trajectory bound for the Silver schedule, this controls every intermediate iterate within a block, including those after the large central step, and a continuity argument closes the local bound. With this within-block stability established, the endpoint analysis becomes simpler: the quadratic part contracts according to the usual Silver polynomial bound, while the accumulated nonlinear errors remain small. Consequently, each block reduces the factor error by a constant factor, and repeating the same block gives geometric convergence.

4.2 Failure of fixed-depth Silver schedules

The local convergence guarantee in Theorem 4.1 raises the question of whether the fixed-depth Silver schedule can converge from arbitrary initializations. The following counterexample shows that this need not hold, even in a rank-one problem with full observation.

To see this, let $d_1 = d_2 = r = 1$, let $\mathcal{A}$ be the identity, and take $y = 1$. Then $f(X) = \frac{1}{2}(X-1)^2$ with $\widehat{X} = 1$, and the factorized objective becomes

$$g_\lambda(U, V) = \frac{1}{2}(UV - 1)^2 + \frac{1}{8}(U^2 - V^2)^2.$$

For this setting, $K = \kappa\kappa_X = 1$, so the parameters specified by Theorem 4.1 give the fixed normalization $\overline{\beta} = 4$ and depth $k = 6$.

Theorem 4.2. *Under the setting above, let* $U_0, V_0 \overset{\text{i.i.d.}}{\sim} N(0, \varepsilon_{\text{init}}^2)$. *Then there exist constants* $c > 0$ *and* $\varepsilon_0 > 0$ *such that, for every* $0 < \varepsilon_{\text{init}} \le \varepsilon_0$,

$$\Pr\left(\lim_{t\to\infty} U_t V_t = +\infty\right) \ge c.$$

The proof of Theorem 4.2 is given in Appendix D. Small random initialization is known to yield global recovery for constant-step factorized gradient descent in several low-rank matrix sensing problems (Stöger & Soltanolkotabi, 2021; Soltanolkotabi et al., 2023). In contrast, our result shows that this behavior need not carry over to the fixed Silver schedule considered in this work. In particular, the fixed Silver schedule diverges with a probability bounded away from zero even under arbitrarily small Gaussian initialization. Thus, even in this simple rank-one setting, small random initialization alone does not guarantee global convergence of the fixed Silver schedule, showing that a local initialization requirement cannot in general be avoided.

4.3 Choosing the schedule parameters from the initialization

The depth and normalization in Theorem 4.1 depend on the singular values of the unknown minimizer $\widehat{X}$. However, the proof of Theorem 4.1 in fact allows a broader range of parameters. Specifically, for a block starting at factor distance $D > 0$ from the solution, it suffices that the parameters k and $\overline{\beta}$ satisfy (see Appendix A.5)

$$3\beta\sigma_1(\widehat{X}) \le \overline{\beta} \le 5\beta\sigma_1(\widehat{X}), \qquad 192K \le \rho^k \le \frac{\sqrt{\sigma_1(\widehat{X})}}{100D}. \tag{2}$$

This allows us to choose the parameters using the initialization X_0. Under the local initialization condition of Theorem 4.1, the extreme nonzero singular values of X_0 approximate those of $\widehat{X}$ within factors of $3/4$ and $5/4$, yielding $\kappa_X \le 2\sigma_1(X_0)/\sigma_r(X_0) \le (10/3)\kappa_X$ (see Proposition A.3). We therefore set $\widehat{\kappa}_X := 2\sigma_1(X_0)/\sigma_r(X_0)$, $\overline{\beta}_0 := 4\beta\sigma_1(X_0)$, and let k_0 be the smallest integer satisfying $\rho^{k_0} \ge 192\kappa\widehat{\kappa}_X$. These bounds directly give

$$3\beta\sigma_1(\widehat{X}) \le \overline{\beta}_0 \le 5\beta\sigma_1(\widehat{X}), \qquad 192K \le \rho^{k_0} < 640\rho K.$$

Together with $D_0 \le c\sqrt{\sigma_1(\widehat{X})}/K$, this yields $\rho^{k_0} D_0/\sqrt{\sigma_1(\widehat{X})} \le 640\rho c \le 1/100$ for sufficiently small universal c, so $(k_0, \overline{\beta}_0)$ satisfies (2) for the first block. Since the factor error decreases after each block, the upper bound in (2) only becomes easier to satisfy. Hence, the same depth and normalization can be kept fixed throughout optimization, and the conclusions of Theorem 4.1 continue to hold with block length $\mathcal{O}(K^\theta)$.

5 Statistical recovery

5.1 Statistical model and recovery guarantee

In the previous sections, the observation vector y is treated as fixed, and the theoretical results provide purely deterministic convergence guarantees. We now specialize this setting to a statistical

observation model. Suppose that $y = \mathcal{A}(X_\star) + \varepsilon$, where $X_\star \in \mathbb{R}^{d_1 \times d_2}$ is the unknown true matrix with $\mathrm{rank}(X_\star) \le r$, and $\varepsilon \in \mathbb{R}^m$ is the measurement noise. Equivalently, if the linear operator $\mathcal{A}$ is represented by sensing matrices $A_1, \dots, A_m$, then $y_i = \langle A_i, X_\star \rangle + \varepsilon_i$ for $i = 1, \dots, m$.

Under this setting, the minimizer $\widehat{X}$ need not coincide with the true matrix $X_\star$. Since $X_\star$ is feasible for the rank-constrained problem, the optimality of $\widehat{X}$ gives $f(\widehat{X}) \le f(X_\star)$. Together with restricted strong convexity, this yields $\|\widehat{X} - X_\star\|_{\mathrm{F}} \le (2\sqrt{2r})/\alpha \|\mathcal{A}^*\varepsilon\|_{\mathrm{op}}$, where $\mathcal{A}^*$ denotes the adjoint of $\mathcal{A}$. Using restricted smoothness and the bound above, we can further show that

$$\|\nabla f(\widehat{X})\|_{\mathrm{op}} \le \beta \|\widehat{X} - X_\star\|_{\mathrm{F}} + \|\mathcal{A}^*\varepsilon\|_{\mathrm{op}} \le \left(1 + 2\sqrt{2}\,\kappa\sqrt{r}\right) \|\mathcal{A}^*\varepsilon\|_{\mathrm{op}}.$$

Hence, whenever

$$\left(1 + 2\sqrt{2}\,\kappa\sqrt{r}\right) \|\mathcal{A}^*\varepsilon\|_{\mathrm{op}} \le \gamma\alpha\sigma_r(\widehat{X}), \tag{3}$$

Assumption 3.2 follows. The derivation of these bounds is given in Appendix C.1. Together with Definition 3.1, Theorem 4.1 applies, showing that the fixed-depth Silver iteration converges geometrically to $\widehat{W}$ from the local initialization.

Now, combining the inequality (1) with the statistical error bound above gives the following recovery guarantee for gradient descent with Silver stepsizes (see Appendix C.1 for details):

$$\|U_T V_T^\top - X_\star\|_{\mathrm{F}} \le \underbrace{3\sqrt{\sigma_1(\widehat{X})}\,\mathrm{dist}(W_T, \widehat{W})}_{\text{optimization error}} + \underbrace{(2\sqrt{2r})/\alpha\,\|\mathcal{A}^*\varepsilon\|_{\mathrm{op}}}_{\text{statistical error}}. \tag{4}$$

At block endpoints $T = bn_k$, Theorem 4.1 gives $\mathrm{dist}(W_T, \widehat{W}) \le 2^{-b} D_0$, so the optimization error is at most $3\sqrt{\sigma_1(\widehat{X})}\,2^{-b} D_0$. Thus, the optimization error decreases geometrically, while the statistical error term does not decrease with further optimization toward $\widehat{X}$. We next illustrate these guarantees with Gaussian matrix sensing and low-rank multivariate regression, each with model-specific warm starts.

5.2 Gaussian matrix sensing and spectral initialization

We first specialize the general recovery result above to Gaussian matrix sensing.

Isotropic Gaussian sensing. The sensing matrices $A_1, \dots, A_m$ introduced in Section 5.1 have i.i.d. $\mathcal{N}(0, 1/m)$ entries. With sufficiently many measurements, $\mathcal{A}$ satisfies rank-$3r$ RIP with constant δ with high probability (see Appendix C) (Recht et al., 2010; Tu et al., 2016), so we may take $\alpha = 1 - \delta$ and $\beta = 1 + \delta$. For spectral initialization, form $Y_0 = \mathcal{A}^* y$ and let $X_0 = P_0 \Sigma_0 Q_0^\top$ be its best rank-r approximation. Set $U_0 = P_0 \Sigma_0^{1/2}$ and $V_0 = Q_0 \Sigma_0^{1/2}$ to obtain balanced factors (Tong et al., 2021). Combining the spectral initialization bound in Appendix C.2 with the statistical error bound in Section 5.1 gives

$$\|X_0 - \widehat{X}\|_F \le 2\delta \|X_\star\|_F + 2\sqrt{2r} \left(1 + \frac{1}{1-\delta}\right) \|\mathcal{A}^*\varepsilon\|_{\mathrm{op}}.$$

For $\mathrm{rank}(\widehat{X}) = r$, a sufficiently small right-hand side places the balanced factors of X_0 in the local basin around $\widehat{W}$. A precise sufficient condition and the noiseless Gaussian sample-size guarantee are given in Appendix C.2. Under this condition, Theorem 4.1 applies from the spectral initialization, and (4), with $\alpha = 1 - \delta$, gives the corresponding recovery guarantee for $X_\star$.

Anisotropic Gaussian sensing. More generally, let $S \in \mathbb{R}^{d_2 \times d_2}$ be known and positive definite, and let $A_i = G_i S^{1/2}$, where the G_i have i.i.d. $\mathcal{N}(0, 1/m)$ entries. Writing $[\mathcal{G}(H)]_i = \langle G_i, H \rangle$, we have $\mathcal{A}(H) = \mathcal{G}(HS^{1/2})$ and $\mathcal{A}^* y = (\mathcal{G}^* y) S^{1/2}$. Since multiplication by $S^{1/2}$ preserves rank, rank-$3r$ RIP for $\mathcal{G}$ yields

$$\alpha = (1 - \delta)\lambda_{\min}(S), \qquad \beta = (1 + \delta)\lambda_{\max}(S), \qquad \kappa = \frac{1 + \delta}{1 - \delta}\,\mathrm{cond}(S).$$

These bounds need not be sharp; restricted curvature remains applicable even when ill-conditioning precludes a small RIP constant in the original Frobenius norm.

The initializer must account for the covariance, since $\mathbb{E}[\mathcal{A}^* y] = X_\star S$ in the noiseless model. Set $Y_0 = \mathcal{A}^* y S^{-1/2} = \mathcal{G}^* y$ and choose X_0 so that $X_0 S^{1/2}$ is a best rank-r approximation of Y_0. Balanced factors are then obtained from an SVD of X_0. Combining the initialization and statistical error bounds in Appendix C.3 gives

$$\|X_0 - \widehat{X}\|_F \le \frac{2\delta \|X_\star S^{1/2}\|_F + 2\sqrt{2r}\,\|\mathcal{G}^*\varepsilon\|_{\mathrm{op}}}{\sqrt{\lambda_{\min}(S)}} + \frac{2\sqrt{2r}\,\|\mathcal{G}^*\varepsilon\|_{\mathrm{op}}}{(1-\delta)\sqrt{\lambda_{\min}(S)}}.$$

Proposition C.5 translates this covariance-dependent bound into a local initialization condition. In the noiseless case, entry into the local basin is ensured by $\delta \le c/[\sqrt{r}\,\mathrm{cond}(S)^{3/2}\kappa_X^{3/2}]$; a sufficient, not necessarily sharp, sample-size bound is given in Corollary C.6. With noise, Proposition C.4 also gives a sufficient condition for Assumption 3.2. Under these conditions and $\mathrm{rank}(\widehat{X}) = r$, Theorem 4.1 applies with $K = \kappa\kappa_X$ and the warm-start schedule of Section 4.3. Together with (4), this yields the corresponding recovery guarantee for $X_\star$.

5.3 Low-rank multivariate regression

We next consider low-rank multivariate regression, where $Y = Z\Theta_\star^\top + E$ with $\mathrm{rank}(\Theta_\star) \le r$, and the loss is $f(\Theta) = \frac{1}{2n}\|\Theta Z^\top - Y^\top\|_F^2$. Writing $\widehat{\Sigma}_Z := n^{-1}Z^\top Z \succ 0$, we have $\alpha = \lambda_{\min}(\widehat{\Sigma}_Z)$, $\beta = \lambda_{\max}(\widehat{\Sigma}_Z)$, and $\kappa = \mathrm{cond}(\widehat{\Sigma}_Z)$; see Appendix C.4. For initialization, let Θ_0 be the best rank-r approximation of the ordinary least-squares estimator $\Theta_{\mathrm{ols}} := n^{-1}Y^\top Z\widehat{\Sigma}_Z^{-1}$, and let W_0 be its balanced factors. For a rank-r constrained minimizer $\widehat{\Theta}$, Proposition C.9 gives a sufficient noise condition under which W_0 lies in the local basin.

Together with Assumption 3.2, this allows the warm-start schedule of Section 4.3 to reach factor error ε_W within $\mathcal{O}\big(K_{\mathrm{reg}}^\theta[1 + \log(D_0/\varepsilon_W)]\big)$ gradient evaluations, where $K_{\mathrm{reg}} := \kappa\,\sigma_1(\widehat{\Theta})/\sigma_r(\widehat{\Theta})$. Applying (4) then gives the corresponding recovery guarantee for $\Theta_\star$; explicit statistical error bounds, including the Gaussian noise case, are given in Appendix C.4.

6 Experiments

In this section, we empirically evaluate the fixed-depth Silver schedule, focusing on its dependence on conditioning and its performance under noisy Gaussian sensing. Additional experiments for low-rank multivariate regression are provided in Appendix E.3. Details on data generation, parameter choices, and implementation are given in Appendix E.

6.1 Conditioning dependence of iteration complexity

We compare constant-step gradient descent (Constant GD) and fixed-depth Silver stepsize gradient descent (Silver GD) on noiseless 64×48 rank-4 matrix sensing instances to examine their dependence on $K = \kappa\kappa_X$. We use the operator $\mathcal{A}(X) = \mathrm{vec}(XS^{1/2})$, where $\mathrm{vec}(\cdot)$ denotes vectorization. For this operator, $\kappa = \mathrm{cond}(S)$. We vary κ with $\kappa_X = 4$ and κ_X with $\kappa = 4$. In both cases, K ranges over $8, 16, 32, 64, 128, 256, 512$.

Both methods start from the same perturbed factors of $\widehat{X}$, with perturbation magnitude proportional to $1/K$, and use the same normalization. Constant GD uses stepsize $1/\overline{\beta}$, while Silver GD uses the depth prescribed by Theorem 4.1, fixed throughout each run. We measure the number of iterations T required to reach $\|U_T V_T^\top - \widehat{X}\|_F / \|U_0 V_0^\top - \widehat{X}\|_F \le 10^{-10}$.

Figure 1 plots $\log T$ against $\log K$ with least-squares linear fits. The fitted slopes are close to 1 for Constant GD and θ for Silver GD, consistent with the respective K and K^θ scalings. This accelerated scaling is observed whether K increases through κ or κ_X.

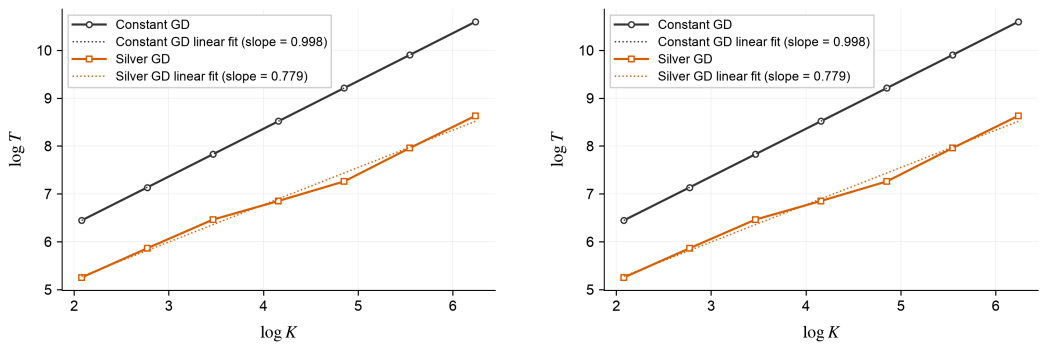

Figure 1: Gradient evaluations required for a 10^{-10} reduction in matrix error versus K on logarithmic axes. The left and right panels vary κ and κ_X, respectively. Least-squares fits are compared with reference exponents 1 for Constant GD and $\theta = \log_\rho 2 \approx 0.786$ for Silver GD.

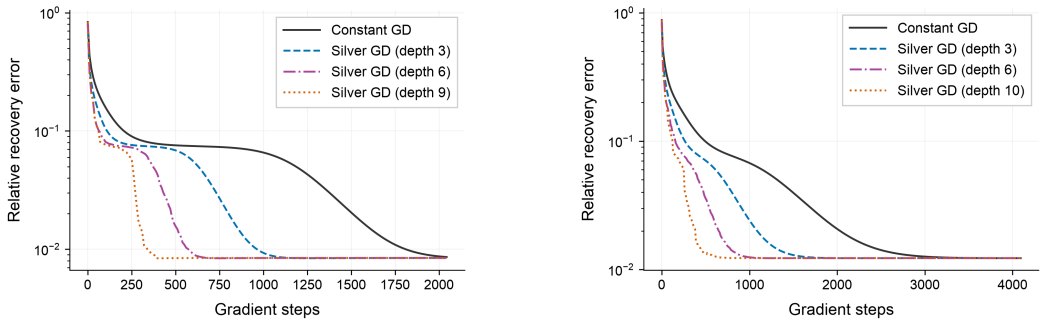

Figure 2: Relative recovery error $\|U_t V_t^\top - X_\star\|_F / \|X_\star\|_F$ versus gradient evaluations for isotropic (left) and anisotropic (right) sensing. Each panel compares Constant GD with the reference Silver depth and two smaller depths.

6.2 Noisy Gaussian sensing and depth sensitivity

We compare Constant GD and Silver GD on the isotropic and anisotropic Gaussian sensing models of Section 5.2, using their spectral and covariance-corrected spectral initializers, respectively. We use 64×48 rank-4 instances with $m = 1000$, $\kappa_X = 16$, and noise level $\nu = 0.01$; the anisotropic covariance has condition number 4. All methods share the same initialization, balancing penalty, and stepsize normalization.

The restricted curvature is estimated empirically from random low-rank probes: rank-$2r$ probes for isotropic sensing and rank-$3r$ probes of the underlying isotropic operator for anisotropic sensing, combined with the known covariance eigenvalues. Together with the initializer's singular values, these estimates determine a reference depth through the warm-start rule in Section 4.3. We compare this depth with two smaller depths and with Constant GD.

Figure 2 shows that Silver GD reaches the common recovery-error plateau faster than Constant GD in both models. The largest tested depth converges fastest, while demonstrating stable behavior at the tested depths on these instances.

7 Conclusion

We show that a fixed-depth Silver schedule accelerates factorized gradient descent for low-rank matrix recovery under a local regularity regime, and give a rank-one counterexample showing that a local initialization requirement cannot in general be avoided. We apply the theory to Gaussian sensing and multivariate regression. Numerical experiments support the theoretical K^θ iteration-complexity scaling and illustrate acceleration in noisy Gaussian sensing.

AI USE STATEMENT

We used generative AI tools as an interactive aid in theoretical development, implementation, and manuscript preparation. The authors formulated the research questions and mathematical results. Generative AI assisted in refining mathematical claims, examining arguments for possible gaps or errors, and organizing the exposition. The authors critically assessed these suggestions and verified all proofs and final theoretical results.

For the numerical experiments, the authors determined the experimental objectives and settings. Generative AI assisted in generating and organizing the code, implementing the experiments, and preparing figures. All AI-assisted code and the resulting experimental outputs were reviewed, tested, and verified by the authors.

For the literature review, the authors identified and evaluated the main relevant prior work. Generative AI was used to assist with literature search and discovery, including finding additional potentially relevant related work. All cited works were checked by the authors against the original sources, and the final selection and interpretation of the literature were determined by the authors. Generative AI also assisted in revising and polishing the manuscript for clarity and readability. The authors reviewed and approved the final text and take full responsibility for the content of this work.

REPRODUCIBILITY STATEMENT

The assumptions underlying our theoretical results are stated in the main text, with complete proofs provided in Appendices A–D. Appendix E describes the experimental setups, including data generation, initialization, parameter choices, random seeds, and stopping criteria. The code and instructions for reproducing the experiments and regenerating the figures are available at this anonymous repository. The repository includes the experiment configurations, dependency requirements, and a notebook for running all three experiments.

REFERENCES

Jason M. Altschuler and Pablo A. Parrilo. Acceleration by stepsize hedging: Silver Stepsize Schedule for smooth convex optimization. *Mathematical Programming*, 213(1):1105–1118, 2025a. doi: 10.1007/s10107-024-02164-2.

Jason M. Altschuler and Pablo A. Parrilo. Acceleration by stepsize hedging: Multi-step descent and the silver stepsize schedule. *J. ACM*, 72(2), 2025b. doi: 10.1145/3708502.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, 2019.

Rina Foygel Barber and Wooseok Ha. Gradient descent with non-convex constraints: local concavity determines convergence. *Information and Inference: A Journal of the IMA*, 7(4):755–806, 2018. doi: 10.1093/imaiai/iay002.

Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Global optimality of local search for low rank matrix recovery. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 3880–3888, 2016.

Jinho Bok and Jason M. Altschuler. Accelerating proximal gradient descent via silver stepsizes. In *Proceedings of Thirty Eighth Conference on Learning Theory*, pp. 421–453, 2025.

Samuel Burer and Renato D.C. Monteiro. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003. doi: 10.1007/s10107-002-0352-8.

Emmanuel J Candes, Xiaodong Li, and Mahdi Soltanolkotabi. Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

Yudong Chen and Martin J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees, 2015.

Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Advances in Neural Information Processing Systems*, 2016.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1233–1242, 2017.

Benjamin Grimmer, Kevin Shu, and Alex L. Wang. Composing optimized stepsize schedules for gradient descent. *Mathematics of Operations Research*, 2025. doi: 10.1287/moor.2024.0764.

Wooseok Ha, Haoyang Liu, and Rina Foygel Barber. An equivalence between critical points for rank constraints versus low-rank factorizations. *SIAM Journal on Optimization*, 30(4):2927–2955, 2020. doi: 10.1137/18M1231675.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975. doi: 10.1016/0047-259X(75)90042-1.

Xixi Jia, Hailin Wang, Jiangjun Peng, Xiangchu Feng, and Deyu Meng. Preconditioning matters: Fast global convergence of non-convex matrix factorization via scaled gradient descent. In *Advances in Neural Information Processing Systems*, pp. 76202–76213, 2023.

Daesung Kim and Hye Won Chung. Rank-1 matrix completion with gradient descent and small random initialization. In *Advances in Neural Information Processing Systems*, pp. 10530–10566, 2023.

Yuanzhi Li, Tengyu Ma, and Hongyang Zhang. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Proceedings of the 31st Conference On Learning Theory*, pp. 2–47, 2018.

Zhenxuan Li and Meng Huang. Scaled gradient descent for ill-conditioned low-rank matrix recovery with optimal sampling complexity, 2026.

Yu Liu, Kang Chen, Rujun Jiang, and Tianyu Wang. Optimal recursive composition and dyadic phase laws for gradient descent with predetermined stepsizes, 2026.

Cong Ma, Yuanxin Li, and Yuejie Chi. Beyond procrustes: Balancing-free gradient descent for asymmetric low-rank matrix sensing. *IEEE Transactions on Signal Processing*, 69:867–877, 2021. doi: 10.1109/TSP.2021.3051425.

Sahand Negahban and Martin J. Wainwright. Restricted strong convexity and weighted matrix completion: optimal bounds with noise. *J. Mach. Learn. Res.*, 13(1):1665–1697, 2012.

Jiyoung Park, Abhishek Roy, Jonathan W. Siegel, and Anirban Bhattacharya. Acceleration via silver step-size on riemannian manifolds with applications to wasserstein space. In *Advances in Neural Information Processing Systems*, pp. 36828–36884, 2025.

Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010. doi: 10.1137/070697835.

R Tyrrell Rockafellar and Roger JB Wets. *Variational analysis*. Springer, 1998.

Mahdi Soltanolkotabi, Dominik Stöger, and Changzhi Xie. Implicit balancing and regularization: Generalization and convergence guarantees for overparameterized asymmetric matrix sensing. In *Proceedings of Thirty Sixth Conference on Learning Theory*, pp. 5140–5142, 2023.

Dominik Stöger and Mahdi Soltanolkotabi. Small random initialization is akin to spectral learning: Optimization and generalization guarantees for overparameterized low-rank matrix reconstruction. In *Advances in Neural Information Processing Systems*, pp. 23831–23843, 2021.

Dominik Stöger and Yizhe Zhu. Non-convex matrix sensing: Breaking the quadratic rank barrier in the sample complexity extended abstract. In *Proceedings of Thirty Eighth Conference on Learning Theory*, pp. 5354–5355, 2025.

Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 62(11):6535–6579, 2016. doi: 10.1109/TIT.2016.2598574.

Tian Tong, Cong Ma, and Yuejie Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *Journal of Machine Learning Research*, 22(150):1–63, 2021.

Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Ben Recht. Low-rank solutions of linear matrix equations via procrustes flow. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 964–973, 2016.

Yudong Wei, Liang Zhang, Bingcong Li, and Niao He. On the benefits of weight normalization for overparameterized matrix sensing. In *International Conference on Learning Representations*, pp. 38818–38866, 2026.

Johan Sokrates Wind. Implicit regularization makes overparameterized asymmetric matrix sensing robust to perturbations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 384(2324):20250120, 2026. doi: 10.1098/rsta.2025.0120.

Xingyu Xu, Yandi Shen, Yuejie Chi, and Cong Ma. The power of preconditioning in overparameterized low-rank matrix sensing. *Information and Inference: A Journal of the IMA*, 15(1):iaaf033, 2026. doi: 10.1093/imaiai/iaaf033.

Yuhan Ye and Kaizhao Liu. Silver rate is (almost) optimal for gradient descent, 2026.

Qinqing Zheng and John Lafferty. A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *Advances in Neural Information Processing Systems*, 2015.

Qinqing Zheng and John Lafferty. Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

Dongruo Zhou, Yuan Cao, and Quanquan Gu. Accelerated factored gradient descent for low-rank matrix factorization. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 4430–4440, 2020.

Zhihui Zhu, Qiuwei Li, Gongguo Tang, and Michael B Wakin. Global optimality in low-rank matrix optimization. *IEEE Transactions on Signal Processing*, 66(13):3614–3628, 2018.

A LOCAL SILVER ACCELERATION: PROOF AND WARM-START PARAMETERS

The proof uses three technical ingredients collected in Appendix B: a finite-trajectory bound, local geometric estimates for balanced factorizations, and Silver polynomial bounds.

A.1 FIXED NORMALIZATION AND A BLOCK-START ALIGNMENT

We prove Theorem 4.1 under its stated assumptions and initialization condition, recalling the depth and normalization from Section 4.1:

$$192K \le \rho^k < 192\rho K, \qquad \overline{\beta} = 4\beta\sigma_1(\widehat{X}).$$

Recall that for a factor pair $Z = (Z_U, Z_V)$, write $\|Z\|^2 := \|Z_U\|_F^2 + \|Z_V\|_F^2$. Define the common-rotation subspace and its orthogonal complement by

$$\mathcal{V} := \{(\widehat{U}\Lambda, \widehat{V}\Lambda) : \Lambda^\top = -\Lambda\}, \qquad \mathcal{H} := \mathcal{V}^\perp,$$

and let $H_* := \nabla^2 g_\lambda(\widehat{W})$ denote the target Hessian. Further details on these subspaces, their geometric characterization, and the properties of the target Hessian are provided in Appendix B.5.

It suffices to establish stability and contraction over one block starting at distance $0 < D \le D_0$ from $\widehat{W}$. Reset the time index to zero and choose the best rotation Q at the block start, as in Appendix B.5. Define $E_t := W_t Q - \widehat{W}$. By Lemma B.2, $E_0 \in \mathcal{H}$ and $\|E_0\| = D$. For this proof, write $\tau := \sqrt{\sigma_1(\widehat{X})}$ and $R := \rho^k$. The depth and initialization conditions give

$$\frac{RD}{\tau} \le \frac{RD_0}{\tau} < 192\rho c \le \frac{1}{100},$$

after choosing the universal constant c sufficiently small.

Normalize the objective by

$$\varphi(E) := \frac{g_\lambda(\widehat{W} + E) - g_\lambda(\widehat{W})}{\overline{\beta}}, \qquad H := H_*/\overline{\beta}.$$

The aligned iteration then takes the form

$$E_{t+1} = E_t - \alpha_t^{(k)} \nabla\varphi(E_t),$$

with $\varphi(0) = 0$ and $\nabla\varphi(0) = 0$.

Let Π be the orthogonal projection onto the rotation space $\mathcal{V}$. Proposition B.4 yields

$$H\Pi = 0, \qquad 0 \preceq H \preceq \frac{3}{4} I, \qquad H|_{\mathcal{H}} \succeq \frac{3}{20K} I. \tag{5}$$

Dividing the bounds in Lemma B.5 by $\overline{\beta} \ge 3\beta\tau^2$ gives, for $\|E\| \le \tau$,

$$\begin{aligned} \|\nabla^2\varphi(E) - H\|_{\mathrm{op}} &\le \frac{6\|E\|}{\tau}, \\ \|\nabla\varphi(E) - HE\| &\le \frac{3\|E\|^2}{\tau}, \\ \left|\varphi(E) - \frac{1}{2}\langle E, HE\rangle\right| &\le \frac{\|E\|^3}{\tau}. \end{aligned} \tag{6}$$

A.2 LOCAL FUNCTION VALUES AND THE TWO-POINT COMPARISON

For $\|E\| \le 2D$, choose a minimizing rotation of $\widehat{W} + E$ and denote the resulting error by E'. Lemma B.2 and rotation invariance give

$$E' \in \mathcal{H}, \qquad \|E'\| \le \|E\|, \qquad \varphi(E') = \varphi(E).$$

This realignment is used only to bound the function value. By (5) and (6),

$$\varphi(E) \geq \left(\frac{3}{40K} - \frac{\|E'\|}{\tau}\right)\|E'\|^2 \geq \frac{3}{80K}\|E'\|^2 \geq 0.$$

The second inequality follows from $2D/\tau \leq 3/(80K)$, which is implied by $R \geq 192K$ and $RD/\tau \leq 1/100$. Also, $\Pi H = 0$, so the gradient remainder bound gives

$$\|\Pi\nabla\varphi(E)\| = \|\Pi(\nabla\varphi(E) - HE)\| \leq \frac{3\|E\|^2}{\tau} \leq B, \qquad B := \frac{12D^2}{\tau}. \tag{7}$$

To obtain a convex comparison on a neighborhood of the block, define

$$\mu := \frac{72D}{\tau}, \qquad F(E) := \varphi(E) + \frac{\mu}{2}\|\Pi E\|^2.$$

The iteration still uses φ; F is used only in the proof. Since $K \geq 1$, the bound on D/τ implies

$$6D \leq \tau, \qquad \mu \leq \frac{3}{800K} \leq \min\left\{\frac{3}{20K}, \frac{1}{12}\right\}.$$

For $\|E\| \leq 6D$, (6) yields

$$\|\nabla^2\varphi(E) - H\|_{\mathrm{op}} \leq \frac{36D}{\tau} = \frac{\mu}{2}.$$

Since H vanishes on $\mathcal{V}$ and $H|_{\mathcal{H}} \succeq \frac{3}{20K}I \succeq \mu I$, we have $H + \mu\Pi \succeq \mu I$. Moreover, $\nabla^2 F(E) = \nabla^2\varphi(E) + \mu\Pi$ because Π is an orthogonal projection. Combining these observations with the Hessian remainder bound gives

$$\frac{\mu}{2}I \preceq \nabla^2 F(E) \preceq \left(\frac{3}{4} + \frac{3\mu}{2}\right)I \preceq \frac{7}{8}I.$$

Thus F is convex and has an L-Lipschitz gradient, with $L = 7/8$, on the radius-$6D$ ball.

We now use the local convexity and smoothness of F to derive a two-point inequality for φ. For $\|x\|, \|y\| \leq 2D$, let

$$u := \nabla F(x) - \nabla F(y), \qquad z := x - \frac{u}{L}.$$

The gradient bound along the segment joining x and y gives

$$\|z\| \leq \|x\| + \frac{\|u\|}{L} \leq 2D + \|x - y\| \leq 6D.$$

All segments joining x, y, and z lie in the radius-$6D$ ball. Define $G(v) := F(v) - \langle\nabla F(y), v\rangle$. Convexity at y and L-smoothness along the segment from x to z give

$$G(y) \leq G(z) \leq G(x) + \langle u, z - x\rangle + \frac{L}{2}\|z - x\|^2 = G(x) - \frac{\|u\|^2}{2L},$$

since $\nabla G(y) = 0$, $\nabla G(x) = u$, and $z - x = -u/L$. Rearranging yields

$$2\big(F(x) - F(y) - \langle\nabla F(y), x - y\rangle\big) \geq L^{-1}\|\nabla F(x) - \nabla F(y)\|^2.$$

We now remove the added quadratic term. Write $v := \nabla\varphi(x) - \nabla\varphi(y)$ and $w := \Pi(x - y)$. Then

$$\begin{aligned}
&2\big(\varphi(x) - \varphi(y) - \langle\nabla\varphi(y), x - y\rangle\big) - \|\nabla\varphi(x) - \nabla\varphi(y)\|^2 \\
&\quad\geq L^{-1}\|v + \mu w\|^2 - \|v\|^2 - \mu\|w\|^2 \\
&\quad= \frac{1 - L}{L}\left\|v + \frac{\mu}{1 - L}w\right\|^2 - \left(\mu + \frac{\mu^2}{1 - L}\right)\|w\|^2 \\
&\quad\geq -2\mu\|w\|^2,
\end{aligned}$$

where the last step uses $L = 7/8$ and $\mu \leq 1/12$.

Therefore, for $\|x\|, \|y\| \leq 2D$,

$$2\big(\varphi(x) - \varphi(y) - \langle\nabla\varphi(y), x - y\rangle\big) - \|\nabla\varphi(x) - \nabla\varphi(y)\|^2 \geq -2\mu\|\Pi(x - y)\|^2. \tag{8}$$

The error depends only on the difference in rotation coordinates. Together with $\varphi \geq 0$ and (7), this inequality will yield the intermediate-iterate bound in the next subsection.

A.3 Intermediate stability by continuity in the initial point

We now prove that all iterates within the block remain in the radius-$2D$ ball. Suppose first that $\|E_t\| \le 2D$ for every $0 \le t \le n_k$. The function values are nonnegative, and (7) and (8) hold along the trajectory, including pairs involving the reference point 0. Since $\Pi E_0 = 0$, Lemma B.1 gives

$$\begin{aligned}\max_{0\le t\le n_k} \|E_t\|^2 &\le \|E_0\|^2 + 8\mu R^3 B^2 \\ &\le D^2 + 82944\left(\frac{RD}{\tau}\right)^3 D^2 \le \frac{9}{8}D^2, \end{aligned} \tag{9}$$

where we used $\mu = 72D/\tau$, $B = 12D^2/\tau$, and $RD/\tau \le 1/100$.

To show that $\|E_t\| \le 2D$ holds throughout the block, consider the same aligned iteration initialized at $E_0(s) := sE_0$, for $s \in [0,1]$. Keep k, $\overline{\beta}$, and the proof parameters D, μ, B fixed. Since the update is polynomial, each finite iterate $E_t(s)$ depends continuously on s. Hence

$$M(s) := \max_{0\le t\le n_k} \|E_t(s)\|$$

is continuous. Moreover, $M(0) = 0$ because $\nabla\varphi(0) = 0$. If $M(s) \ge 2D$ for some s, continuity gives an s_* with $M(s_*) = 2D$. At this value, all iterates lie in the closed radius-$2D$ ball, so the nonnegativity of φ, (7), and (8) apply along the trajectory. Since $\Pi E_0(s_*) = 0$ and $\|E_0(s_*)\| \le D$, (9) gives

$$M(s_*)^2 \le \|E_0(s_*)\|^2 + 8\mu R^3 B^2 \le \frac{9}{8}D^2 < 4D^2,$$

a contradiction.

Thus $M(s) < 2D$ throughout $[0,1]$. Applying (9) at $s = 1$ yields

$$\max_{0\le t\le n_k} \|E_t\|^2 \le \frac{9}{8}D^2 < 4D^2. \tag{10}$$

Thus, intermediate iterates remain stable despite the large Silver steps. We next prove endpoint contraction.

A.4 Endpoint contraction, restarting, and gradient complexity

We separate the endpoint error into its quadratic and nonlinear contributions. Write the aligned update as

$$E_{t+1} = (I - \alpha_t^{(k)} H)E_t + e_t, \qquad e_t := -\alpha_t^{(k)}\big(\nabla\varphi(E_t) - HE_t\big). \tag{11}$$

By (10), $\|E_t\| \le 2D \le \tau$, so (6) gives

$$\|e_t\| \le \frac{3\alpha_t^{(k)}}{\tau}\|E_t\|^2 \le \frac{12\alpha_t^{(k)} D^2}{\tau}. \tag{12}$$

Define the prefix polynomials by

$$p_t^{(k)}(\xi) := \prod_{j=0}^{t-1}(1 - \alpha_j^{(k)}\xi), \qquad 0 \le t \le n_k.$$

Here, empty products are equal to one.

Unrolling (11) over one block yields

$$E_{n_k} = p_{n_k}^{(k)}(H)E_0 + \sum_{t=0}^{n_k-1}\left[\prod_{j=t+1}^{n_k-1}(I - \alpha_j^{(k)}H)\right]e_t,$$

where an empty product is the identity. Since $E_0 \in \mathcal{H}$, the quadratic contraction in Appendix B.6 gives

$$\|p_{n_k}^{(k)}(H)E_0\| \le \sqrt{\frac{20K}{3R}}D \le \frac{\sqrt{5}}{12}D.$$

By (5) and Lemma B.7, each endpoint tail has operator norm at most one on the whole factor space. Thus no restriction on the rotation components of e_t is needed. Using (12) and the stepsize sum in the same lemma,

$$\sum_{t=0}^{n_k-1} \|e_t\| \le \frac{12D^2}{\tau}(R-1) \le 12\frac{RD}{\tau}D \le \frac{12}{100}D.$$

Combining the two contributions gives

$$\operatorname{dist}(W_{n_k}, \widehat{W}) \le \|E_{n_k}\| \le \left(\frac{\sqrt{5}}{12} + \frac{12}{100}\right) D < \frac{1}{2}D.$$

At each block endpoint, choose a new minimizing rotation for the analysis and repeat the argument. The starting distance decreases, so the condition $RD/\tau \le 1/100$ remains valid with the same depth and normalization. Induction therefore gives

$$\operatorname{dist}(W_{bn_k}, \widehat{W}) \le 2^{-b}D_0, \qquad b \ge 0.$$

Moreover, (10) implies

$$\max_{0\le s\le n_k} \operatorname{dist}(W_{bn_k+s}, \widehat{W}) \le 2\operatorname{dist}(W_{bn_k}, \widehat{W}).$$

If a block starts at distance zero, all subsequent iterates remain stationary; this also covers $D_0 = 0$.

For $0 < \varepsilon_W < D_0$, the blockwise contraction reduces the factor distance to ε_W within $\lceil \log_2(D_0/\varepsilon_W) \rceil$ blocks. Each block uses

$$n_k = 2^k - 1 < R^\theta < (192\rho K)^\theta$$

gradient evaluations, giving a total complexity of

$$\mathcal{O}\left(K^\theta \left[1 + \log \frac{D_0}{\varepsilon_W}\right]\right).$$

This completes the proof of Theorem 4.1.

A.5 Admissible depths

The proof allows a broader range of depths and normalizers than those chosen in Theorem 4.1. In particular, (5) and (6) remain valid whenever

$$3\beta\sigma_1(\widehat{X}) \le \overline{\beta} \le 5\beta\sigma_1(\widehat{X}).$$

For a block starting at distance $D > 0$, the stability and contraction arguments in Appendices A.3 and A.4 therefore apply provided

$$192K \le \rho^k \le \frac{\tau}{100D}.$$

The lower bound ensures quadratic contraction on $\mathcal{H}$, while the upper bound controls intermediate iterates and the accumulated nonlinear error.

An admissible integer depth exists when $D \le c\tau/K$ for a sufficiently small universal constant c. Indeed, the smallest integer k with $\rho^k \ge 192K$ satisfies

$$192K \le \rho^k < 192\rho K \le \frac{\tau}{100D}$$

whenever $c \le (19200\rho)^{-1}$. This choice gives $n_k = \mathcal{O}(K^\theta)$ and hence the complexity in Theorem 4.1.

As the block-start distance decreases, the upper depth bound increases, so the same depth and normalization remain admissible. These conditions are sufficient and need not describe the largest convergence basin. Appendix A.6 shows how to choose fixed parameters from the warm start.

A.6 Deterministic Error Conversions and Warm-Start Parameters

This subsection collects the deterministic error conversions and fixed-parameter arguments used in Sections 4.1 and 4.3.

We first establish a deterministic bound that converts factor distance into matrix error. This bound translates convergence of the factors into a matrix optimization-error bound and later allows us to compare the singular values of the initialization with those of the minimizer $\widehat{X}$.

Lemma A.1 (Factor error implies matrix error)**.** *Suppose* $\operatorname{dist}(W, \widehat{W}) \le \sqrt{\sigma_r(\widehat{X})}$. *Then*

$$\|UV^\top - \widehat{X}\|_F \le 3\sqrt{\sigma_1(\widehat{X})}\operatorname{dist}(W, \widehat{W}). \tag{13}$$

Proof. Choose $Q \in O(r)$ attaining $\operatorname{dist}(W, \widehat{W})$, and write $UQ = \widehat{U} + A, VQ = \widehat{V} + B$, so that $\|(A, B)\| = \operatorname{dist}(W, \widehat{W})$. Since $UV^\top = (UQ)(VQ)^\top$, we have $UV^\top - \widehat{X} = A\widehat{V}^\top + \widehat{U}B^\top + AB^\top$. Because $\widehat{U}$ and $\widehat{V}$ are balanced factors of $\widehat{X}$, $\|\widehat{U}\|_{\mathrm{op}} = \|\widehat{V}\|_{\mathrm{op}} = \sqrt{\sigma_1(\widehat{X})}$. Hence

$$\|A\widehat{V}^\top + \widehat{U}B^\top\|_F \le \sqrt{\sigma_1(\widehat{X})}\,(\|A\|_F + \|B\|_F) \le \sqrt{2\sigma_1(\widehat{X})}\,\|(A, B)\|.$$

Moreover,

$$\|AB^\top\|_F \le \|A\|_F\|B\|_F \le \frac{1}{2}\left(\|A\|_F^2 + \|B\|_F^2\right) = \frac{1}{2}\|(A, B)\|^2.$$

Using the assumption, $\|(A, B)\| = \operatorname{dist}(W, \widehat{W}) \le \sqrt{\sigma_r(\widehat{X})} \le \sqrt{\sigma_1(\widehat{X})}$, we conclude that

$$\|UV^\top - \widehat{X}\|_F \le \sqrt{2\sigma_1(\widehat{X})}\operatorname{dist}(W, \widehat{W}) + \frac{1}{2}\operatorname{dist}(W, \widehat{W})^2 \le 3\sqrt{\sigma_1(\widehat{X})}\operatorname{dist}(W, \widehat{W}).$$

□

As an immediate consequence, the factor contraction in Theorem 4.1 yields a matrix optimization-error bound at every block endpoint. Indeed, for $T = bn_k$, $\operatorname{dist}(W_T, \widehat{W}) \le 2^{-b}D_0$, and the local initialization condition ensures that the hypothesis of Lemma A.1 is satisfied. Therefore

$$\|U_T V_T^\top - \widehat{X}\|_F \le 3\sqrt{\sigma_1(\widehat{X})}\,2^{-b}D_0.$$

Thus the blockwise contraction in factor distance directly gives geometric decay of the matrix optimization error. The same deterministic bound will also be used below to compare the singular values of $X_0 = U_0V_0^\top$ with those of $\widehat{X}$.

We next establish a perturbation bound that converts matrix initialization error into factor distance. Let X_0 be a rank-r matrix with singular value decomposition $X_0 = P_0\Sigma_0Q_0^\top$, and define its balanced factors by $U_0 = P_0\Sigma_0^{1/2}$, $V_0 = Q_0\Sigma_0^{1/2}$.

Lemma A.2 (Matrix error ensures local initialization)**.** *Let X_0 be a rank-r matrix with balanced factors $W_0 = (U_0, V_0)$. There is a universal constant $c_0 > 0$ such that, if $\|X_0 - \widehat{X}\|_F \le c_0\sigma_r(\widehat{X})/(\kappa\sqrt{\kappa_X})$, then*

$$\operatorname{dist}(W_0, \widehat{W}) \le c\frac{\sqrt{\sigma_1(\widehat{X})}}{\kappa\kappa_X},$$

where c is the constant in Theorem 4.1. Thus, W_0 satisfies its local initialization condition.

Proof. Since $\kappa, \kappa_X \ge 1$, choosing $c_0 \le 1/2$ ensures $\|X_0 - \widehat{X}\|_{\mathrm{op}} \le \|X_0 - \widehat{X}\|_F \le \sigma_r(\widehat{X})/2$. Lemma 5.14 of Tu et al. (2016), applied to the balanced factorizations of X_0 and $\widehat{X}$, therefore gives

$$\operatorname{dist}(W_0, \widehat{W}) \le C\frac{\|X_0 - \widehat{X}\|_F}{\sqrt{\sigma_r(\widehat{X})}} \le Cc_0\frac{\sqrt{\sigma_1(\widehat{X})}}{\kappa\kappa_X},$$

where $C = \sqrt{2/(\sqrt{2} - 1)}$. Choosing $c_0 \le c/C$ completes the proof. □

We now show that the singular values of X_0 approximate those of $\widehat{X}$ within constant factors, allowing the depth and normalization to be fixed using quantities available at initialization.

Proposition A.3 (Singular values of the initial matrix)**.** *Suppose the initialization satisfies the local condition of Theorem 4.1, with its universal constant chosen sufficiently small. Then*

$$\|X_0 - \widehat{X}\|_{\mathrm{op}} \le \frac{1}{4}\sigma_r(\widehat{X}).$$

Consequently,

$$\frac{3}{4}\sigma_1(\widehat{X}) \le \sigma_1(X_0) \le \frac{5}{4}\sigma_1(\widehat{X}), \qquad \frac{3}{4}\sigma_r(\widehat{X}) \le \sigma_r(X_0) \le \frac{5}{4}\sigma_r(\widehat{X}),$$

and

$$\kappa_X \le \frac{2\sigma_1(X_0)}{\sigma_r(X_0)} \le \frac{10}{3}\kappa_X. \tag{14}$$

Proof. By the matrix-error bound in Lemma A.1,

$$\|X_0 - \widehat{X}\|_{\mathrm{op}} \le \|X_0 - \widehat{X}\|_F \le 3\sqrt{\sigma_1(\widehat{X})}\,\mathrm{dist}(W_0, \widehat{W}).$$

The local initialization condition ensures $\mathrm{dist}(W_0, \widehat{W}) \le c\frac{\sqrt{\sigma_1(\widehat{X})}}{\kappa\kappa_X}$. Thus

$$\|X_0 - \widehat{X}\|_{\mathrm{op}} \le \frac{3c}{\kappa}\sigma_r(\widehat{X}) \le \frac{1}{4}\sigma_r(\widehat{X}),$$

where the last inequality uses $\kappa \ge 1$ and $c \le 1/12$. Weyl's singular-value perturbation inequality now yields

$$|\sigma_i(X_0) - \sigma_i(\widehat{X})| \le \|X_0 - \widehat{X}\|_{\mathrm{op}} \le \frac{1}{4}\sigma_r(\widehat{X}), \qquad i = 1, \dots, r.$$

Taking $i = 1, r$ and using $\sigma_r(\widehat{X}) \le \sigma_1(\widehat{X})$ gives the stated singular-value bounds. Their ratios satisfy

$$\frac{3}{5}\kappa_X \le \frac{\sigma_1(X_0)}{\sigma_r(X_0)} \le \frac{5}{3}\kappa_X.$$

Multiplying by two proves (14). □

These estimates motivate the observable choices

$$\widehat{\kappa}_X := \frac{2\sigma_1(X_0)}{\sigma_r(X_0)}, \qquad \overline{\beta}_0 := 4\beta\sigma_1(X_0),$$

and let k_0 be the smallest integer satisfying $\rho^{k_0} \ge 192\kappa\widehat{\kappa}_X$.

Proposition A.4 (Depth and normalization from the initial factors)**.** *Under the assumptions and local initialization condition of Theorem 4.1, with a sufficiently small universal constant, the observable choices $(k_0, \overline{\beta}_0)$ lie in the broader admissible range established in Appendix A.5. The resulting iteration retains the endpoint contraction and intermediate-iterate bounds of Theorem 4.1, with $\mathcal{O}(K^\theta)$ gradient evaluations per block. The same parameters remain valid at every restart.*

Proof. Proposition A.3 gives $3\beta\sigma_1(\widehat{X}) \le \overline{\beta}_0 \le 5\beta\sigma_1(\widehat{X})$, verifying the normalization condition. By the minimality of k_0 and (14),

$$192K \le 192\kappa\widehat{\kappa}_X \le \rho^{k_0} < 192\rho\kappa\widehat{\kappa}_X \le 640\rho K.$$

The local initialization condition therefore implies $\rho^{k_0} D_0/\sqrt{\sigma_1(\widehat{X})} \le 640\rho c \le 1/100$ for sufficiently small c. These bounds verify the admissible conditions in Appendix A.5, so the proof of Theorem 4.1 applies with $(k_0, \overline{\beta}_0)$. Moreover, $\rho^{k_0} = \mathcal{O}(K)$ implies that each block requires $2^{k_0} - 1 = \mathcal{O}(K^\theta)$ gradient evaluations. As the block-start distance decreases after each restart, the same depth and normalization remain admissible. □

More precisely, the observable depth satisfies $k \le k_0 \le k+2$, where k is the depth chosen in Theorem 4.1. This follows from $\kappa_X \le \widehat{\kappa}_X \le (10/3)\kappa_X$, $\rho^2 > 10/3$, and the minimality of the two depths.

Thus, the depth and normalization can be selected before the first iteration using X_0 and valid restricted-curvature bounds α, β. These curvature bounds remain inputs and must satisfy the assumptions of Theorem 4.1.

We conclude with a standalone specification of the Silver method using the observable warm-start parameters developed above. Unlike the target-based parameter display in Theorem 4.1, the choices below are computable from the initial factors and valid restricted- curvature bounds. Their validity is guaranteed by Proposition A.4.

Algorithm 1 Silver acceleration from a warm start

Require: Balanced factors $W_0 = (U_0, V_0)$ satisfying the local initialization condition; valid restricted-curvature bounds $0 < \alpha \le \beta$

1: Set
$$X_0 = U_0 V_0^\top, \qquad \kappa = \frac{\beta}{\alpha}, \qquad \lambda = \frac{\alpha}{4}.$$
2: Compute once
$$\widehat{\kappa}_X = \frac{2\sigma_1(X_0)}{\sigma_r(X_0)}, \qquad \overline{\beta}_0 = 4\beta\sigma_1(X_0).$$
3: Choose the smallest integer k such that
$$\rho^k \ge 192\kappa\widehat{\kappa}_X, \qquad n_k = 2^k - 1.$$
4: **for** $b = 0, 1, 2, \ldots$ **do**
5: **for** $j = 0, \ldots, n_k - 1$ **do**
6: $t = bn_k + j$
7:
$$W_{t+1} = W_t - \frac{\alpha_j^{(k)}}{\overline{\beta}_0} \nabla g_\lambda(W_t).$$
8: **end for**
9: Check the desired stopping rule at the block endpoint $W_{(b+1)n_k}$.
10: **end for**

All quantities defining the schedule are computed once from W_0 before the first optimization step and remain fixed thereafter. At each restart, only the Silver schedule index is reset. In particular, the algorithm does not recompute curvature estimates, rebalance the factors, or perform any of the projections used only in the convergence proof.

B Supporting Lemmas for Local Silver Acceleration

This appendix collects the three technical ingredients used in Appendix A: the finite-trajectory stability argument, the local geometry of balanced factorizations, and the Silver polynomial product bounds.

B.1 Finite-trajectory assumptions

We prove the finite-trajectory bound used in Appendix A.3 to control all intermediate iterates of a Silver block. The lemma below allows an arbitrary orthogonal projection Π. Let φ be differentiable, with $\varphi(0) = 0$ and $\nabla\varphi(0) = 0$, and write

$$f_p := \varphi(x_p), \qquad g_p := \nabla\varphi(x_p), \qquad x_{p+1} = x_p - h_p g_p,$$

where $h_p = \alpha_p^{(k)}$. For ordered pairs of points, define

$$Q_{p,q} := 2(f_p - f_q) - 2\langle g_q, x_p - x_q\rangle - \|g_p - g_q\|^2.$$

These quantities are local to the proof. We use $\star$ for the reference point 0, so $x_\star = 0$.

Lemma B.1 (A finite Silver trajectory)**.** *Let Π be an orthogonal projection and let $\mu, B \geq 0$. For a depth-k Silver trajectory, assume $\Pi x_0 = 0$. Suppose that, at every trajectory point,*

$$f_i \geq 0, \qquad \|\Pi g_i\| \leq B,$$

and that, for all required ordered pairs of trajectory points and the reference point 0,

$$Q_{p,q} \geq -2\mu\|\Pi(x_p - x_q)\|^2.$$

Then

$$\max_{0\leq t\leq n_k} \|x_t\|^2 \leq \|x_0\|^2 + 8\mu\rho^{3k}B^2.$$

The assumptions concern only finitely many points and pairs; no global convexity or derivative bound on φ is required. The notation introduced above is used only for the proof and does not constitute additional input to the Silver algorithm. The proof proceeds by decomposing a block quantity along the Silver recursion, bounding the accumulated two-point error, and then controlling every prefix.

B.2 Joining two blocks: the exact identity

For a depth-j Silver block, let $T := \rho^j, n := 2^j - 1$, and define

$$\mathcal{S}_j := 2\sum_{i=0}^{n-1} h_i(f_i - f_n) - \sum_{i=0}^{n-1} h_i(h_i - 1)\|g_i\|^2 - T(T-1)\|g_n\|^2.$$

For the empty block, set $\mathcal{S}_0 := 0$. Now suppose $j \geq 1$, and write

$$S := \rho^{j-1}, \qquad \ell := 2^{j-1} - 1, \qquad a := 1 + \frac{S}{\rho} = h_\ell.$$

The left depth-$(j-1)$ block ends at x_ℓ, the central step takes x_ℓ to $x_{\ell+1}$, and the right block ends at x_n. Let $\mathcal{S}^L_{j-1}$ and $\mathcal{S}^R_{j-1}$ denote the corresponding quantities, each defined relative to its own endpoint. Then

$$\mathcal{S}_j = \mathcal{S}^L_{j-1} + \rho^2\mathcal{S}^R_{j-1} + \rho\sum_{i=\ell+1}^{n-1} h_i\big(Q_{\ell,i} + Q_{n,i}\big) + \rho Q_{\ell,n} + \rho(a-1)Q_{n,\ell}. \tag{15}$$

To prove this decomposition, define the displacement across the right block by $z := \sum_{i=\ell+1}^{n-1} h_i g_i = x_{\ell+1} - x_n$. The updates give $x_\ell - x_n = ag_\ell + z$, and $x_\ell + x_n - 2x_i = ag_\ell - z + 2\sum_{p=\ell+1}^{i-1} h_p g_p$. Expanding $\|z\|^2$ then gives

$$\sum_{i=\ell+1}^{n-1} h_i\langle g_i, x_\ell + x_n - 2x_i\rangle = a\langle z, g_\ell\rangle - \sum_{i=\ell+1}^{n-1} h_i^2\|g_i\|^2,$$

Using the definition of $Q_{p,q}$ and the identity $\sum_{i=\ell+1}^{n-1} h_i = S - 1$, we obtain

$$\begin{aligned}\sum_{i=\ell+1}^{n-1} h_i\big(Q_{\ell,i} + Q_{n,i}\big) =& 2(S-1)(f_\ell + f_n) - 4\sum_{i=\ell+1}^{n-1} h_i f_i + 2\sum_{i=\ell+1}^{n-1} h_i(h_i-1)\|g_i\|^2 \\ &+ 2\langle(1-a)g_\ell + g_n, z\rangle - (S-1)\big(\|g_\ell\|^2 + \|g_n\|^2\big),\end{aligned}$$

while the remaining term is

$$Q_{\ell,n} + (a-1)Q_{n,\ell} = 2(2-a)(f_\ell - f_n) + (2a^2 - 3a)\|g_\ell\|^2 - a\|g_n\|^2 + 2\langle(a-1)g_\ell - g_n, z\rangle.$$

The z-terms cancel. After multiplying by ρ and adding the two smaller-block terms, the right-interior coefficients match $\mathcal{S}_j$ since $\rho^2 - 2\rho = 1$. The four remaining coefficients, corresponding to $2f_\ell$, $2f_n$, $\|g_\ell\|^2$, and $\|g_n\|^2$, also agree. For example, the coefficient of $2f_\ell$ satisfies $-(S-1) + \rho(S+1-a) = a$. The other three coefficients are checked similarly using $a = 1 + S/\rho,\ \rho^2 = 2\rho + 1$.

The left-interior coefficients already coincide. Hence every coefficient agrees, proving the block identity for arbitrary $S = \rho^{j-1}$, including $j = 1$, where the two smaller blocks and the right-interior sums are empty.

B.3 Rotation displacement and accumulated two-point error

For any two trajectory indices $i \geq \ell$, the updates give $x_i - x_\ell = -\sum_{p=\ell}^{i-1} h_p g_p$. Hence, using $\|\Pi g_p\| \leq B$,

$$\|\Pi(x_i - x_\ell)\| \leq B \sum_{p=\ell}^{i-1} h_p.$$

For $i < \ell$, the same bound holds with the indices reversed. Now consider one join in the depth-j recursion. Write $S := \rho^{j-1}$, $T := \rho^j = \rho S$. The total stepsize from the left-block endpoint x_ℓ to any point in the right block is bounded by $a + (S-1) = \sqrt{2}\, S$. Therefore every new pair appearing in the join satisfies

$$\|\Pi(x_i - x_\ell)\| \leq \sqrt{2}\, SB, \qquad \|\Pi(x_n - x_i)\| \leq \sqrt{2}\, SB,$$

and hence the two-point defect assumption gives $Q_{p,q} \geq -4\mu S^2 B^2$ for each new $Q_{p,q}$ term created at the join. The total coefficient mass of these new terms in (15) is

$$\rho\,\{2(S-1) + 1 + (a-1)\} = \rho(\rho S - 1) \leq \rho^2 S.$$

We now prove by induction on j that $\mathcal{S}_j \geq -4\mu T^3 B^2,\ T = \rho^j$. The claim is immediate for the empty block, since $\mathcal{S}_0 = 0$.

Suppose the bound holds for depth $j-1$. Since each smaller block has parameter $S = \rho^{j-1}$, the two smaller-block contributions in (15) are bounded below by $-4\mu S^3 B^2 - 4\mu\rho^2 S^3 B^2$. The new join terms contribute at least $-4\mu\rho^2 S^3 B^2$. Therefore since $\rho^3 - (1 + 2\rho^2) = \sqrt{2} > 0$,

$$\mathcal{S}_j \geq -4\mu\big(1 + 2\rho^2\big) S^3 B^2 \geq -4\mu(\rho S)^3 B^2 = -4\mu T^3 B^2.$$

This closes the induction uniformly in the depth.

Notice that the argument only uses bounds on differences of rotation coordinates along each subblock. In particular, an interior subblock need not start with zero rotation projection.

B.4 Endpoint energies, central steps, and all prefixes

We next derive an endpoint energy inequality for each recursively occurring subblock. For a depth-j block, let $T = \rho^j$ and let x_0 and x_n denote its starting and ending points. Unrolling the updates gives

$$\|x_0\|^2 - \|x_n\|^2 = 2\sum_{i=0}^{n-1} h_i \langle g_i, x_i\rangle - \sum_{i=0}^{n-1} h_i^2 \|g_i\|^2.$$

Using $Q_{\star,i} = -2f_i + 2\langle g_i, x_i\rangle - \|g_i\|^2$ and $\sum_{i=0}^{n-1} h_i = T - 1$, we obtain

$$\begin{aligned}\mathcal{S}_j + \sum_{i=0}^{n-1} h_i Q_{\star,i} &= 2\sum_{i=0}^{n-1} h_i \langle g_i, x_i\rangle - \sum_{i=0}^{n-1} h_i^2 \|g_i\|^2 - 2(T-1)f_n - T(T-1)\|g_n\|^2 \\ &= \|x_0\|^2 - \|x_n\|^2 - 2(T-1)f_n - T(T-1)\|g_n\|^2. \end{aligned} \tag{16}$$

The last equality follows by telescoping from

$$\|x_i\|^2 - \|x_{i+1}\|^2 = 2h_i\langle g_i, x_i\rangle - h_i^2\|g_i\|^2.$$

Set $V := \rho^k B$ for the entire depth-k trajectory. Since $\Pi x_0 = 0$, the displacement bound from the previous subsection gives $\|\Pi x_i\| \leq V$ for every trajectory point. Moreover, every recursively occurring depth-j subblock satisfies $TB \leq V$. Therefore, using $\mathcal{S}_j \geq -4\mu T^3 B^2$ and $Q_{\star,i} \geq -2\mu\|\Pi x_i\|^2 \geq -2\mu V^2$, equation (16) yields

$$\|x_n\|^2 + T(T-1)\|g_n\|^2 + 2(T-1)f_n \leq \|x_0\|^2 + 6\mu T V^2. \tag{17}$$

This bound applies to every recursively occurring subblock, since it only uses the global estimate $\|\Pi x_i\| \leq V$, not zero rotation projection at the subblock start.

We now control the large central steps and all prefixes. For a depth-one step, $a = \sqrt{2}$. The two-point defect bounds for $(\star, x)$ and $(x, \star)$ give

$$\langle g, x\rangle \geq f(x) + \frac{1}{2}\|g\|^2 - \mu\|\Pi x\|^2, \qquad f(x) \geq \frac{1}{2}\|g\|^2 - \mu\|\Pi x\|^2.$$

Combining these inequalities yields $\langle g, x\rangle \geq \|g\|^2 - 2\mu\|\Pi x\|^2 \geq \|g\|^2 - 2\mu V^2$. Hence

$$\|x - ag\|^2 = \|x\|^2 - 2a\langle g, x\rangle + a^2\|g\|^2 \leq \|x\|^2 + 4\sqrt{2}\,\mu V^2.$$

Now consider a depth-j block with $j \geq 2$. Write $S := \rho^{j-1}$, $a := 1 + \frac{S}{\rho}$. Then $1 \leq a \leq S$. Let x be the endpoint of the left depth-$(j-1)$ subblock, and let $g = \nabla\varphi(x)$. The defect inequality for $Q_{\star,x}$ gives

$$\langle g, x\rangle \geq f(x) + \frac{1}{2}\|g\|^2 - \mu\|\Pi x\|^2 \geq f(x) + \frac{1}{2}\|g\|^2 - \mu V^2.$$

Combining this with (17) for the left subblock, we obtain

$$\|x - ag\|^2 \leq \|x\|^2 + (a^2 - a)\|g\|^2 - 2af(x) + 2a\mu V^2 \leq \|x_{\mathrm{start}}\|^2 + 8\mu S V^2.$$

Here we used $a^2 - a \leq S(S-1)$ and the nonnegativity $f(x) \geq 0$.

Let $A_j \mu V^2$ denote an upper bound on the excess squared norm of every prefix of any recursively occurring depth-j block relative to the squared norm at the start of that block. The depth-one estimate gives $A_1 = 4\sqrt{2}$. For $j \geq 2$, the first depth-$(j-1)$ block is controlled by the induction hypothesis, the central step by the preceding estimate, and the second depth-$(j-1)$ block contributes at most another $A_{j-1}\mu V^2$ relative to its own starting point. Thus we may take $A_j = A_{j-1} + 8\rho^{j-1}$. Summing the recurrence gives

$$A_j = 4\sqrt{2}\,\rho^j - 8 \leq 8\rho^j.$$

Therefore, at depth k,

$$\max_{0 \leq t \leq n_k} \|x_t\|^2 \leq \|x_0\|^2 + 8\rho^k \mu V^2 = \|x_0\|^2 + 8\mu\rho^{3k} B^2,$$

which proves Lemma B.1.

B.5 Local Geometry and Taylor Expansions of Balanced Factorizations

Let $\widehat{X} = \widehat{P}\Sigma\widehat{Q}^\top$ be a compact singular value decomposition of the rank-r target. Define its balanced factors

$$\widehat{U} = \widehat{P}\Sigma^{1/2}, \qquad \widehat{V} = \widehat{Q}\Sigma^{1/2},$$

and write $\widehat{W} = (\widehat{U}, \widehat{V})$. Then $\widehat{X} = \widehat{U}\widehat{V}^\top$ and $\widehat{U}^\top\widehat{U} = \widehat{V}^\top\widehat{V} = \Sigma$.

The local geometry uses stationarity along first-order rank-r changes,

$$\nabla f(\widehat{X})\widehat{V} = 0, \qquad \nabla f(\widehat{X})^\top\widehat{U} = 0,$$

together with the normal-gradient bound

$$\|\nabla f(\widehat{X})\|_{\mathrm{op}} \leq \gamma\alpha\sigma_r(\widehat{X}), \qquad 0 \leq \gamma \leq \frac{1}{4}.$$

For the constrained minimizer in the main text, stationarity follows from optimality, while the normal-gradient bound is imposed separately in Assumption 3.2.

Common orthogonal rotations preserve both the represented matrix and the objective: $(UQ)(VQ)^\top = UV^\top$ and $g_\lambda(WQ) = g_\lambda(W)$ for $Q \in O(r)$. Let $Q(t) \in O(r)$ be a smooth curve with $Q(0) = I$. Differentiating $Q(t)^\top Q(t) = I$ at $t = 0$ gives $Q'(0)^\top + Q'(0) = 0$. Writing $\Lambda := Q'(0)$, we recover the rotation subspaces introduced in Appendix A.1:

$$\mathcal{V} := \{(\widehat{U}\Lambda, \widehat{V}\Lambda) : \Lambda^\top = -\Lambda\}, \qquad \mathcal{H} := \mathcal{V}^\perp.$$

The space $\mathcal{V}$ therefore consists exactly of infinitesimal symmetry directions that rotate the two factors simultaneously while leaving their product unchanged. These directions are distinct from balance-changing directions, which alter $U^\top U - V^\top V$ and are penalized by the balancing term.

A pair (A, B) lies in $\mathcal{H}$ precisely when $\widehat{U}^\top A + \widehat{V}^\top B$ is symmetric. Indeed, for every skew-symmetric Λ, $\langle (A, B), (\widehat{U}\Lambda, \widehat{V}\Lambda)\rangle = \langle \widehat{U}^\top A + \widehat{V}^\top B, \Lambda\rangle$, and symmetric matrices are the orthogonal complement of skew-symmetric matrices. The following lemma shows that minimizing over rotations places the factor error in $\mathcal{H}$.

Lemma B.2 (Error after the best rotation)**.** *For $W = (U, V)$, choose any $Q \in O(r)$ attaining* $\mathrm{dist}(W, \widehat{W})$. *Then*

$$E := WQ - \widehat{W} \in \mathcal{H}, \qquad \|E\| = \mathrm{dist}(W, \widehat{W}).$$

Proof. A minimizing rotation exists by compactness of $O(r)$. Let $Q \in O(r)$ be any minimizer. For every skew-symmetric Λ, $Q\exp(t\Lambda) \in O(r)$, so optimality gives

$$0 = \frac{d}{dt}\|WQe^{t\Lambda} - \widehat{W}\|^2\Big|_{t=0} = 2\langle E, WQ\Lambda\rangle.$$

Writing $E = (A, B)$, we have $\langle E, E\Lambda\rangle = \mathrm{tr}((A^\top A + B^\top B)\Lambda) = 0$, since $A^\top A + B^\top B$ is symmetric and Λ is skew-symmetric. As $WQ = \widehat{W} + E$, it follows that $\langle E, \widehat{W}\Lambda\rangle = 0$ for every skew-symmetric Λ. Hence $E \in \mathcal{V}^\perp = \mathcal{H}$. Finally, $\|E\| = \mathrm{dist}(W, \widehat{W})$ by the choice of Q. □

For the block analysis, choose a rotation attaining the factor distance at the block start and keep it fixed throughout the block. Since $\nabla g_\lambda(WQ) = \nabla g_\lambda(W)Q$, the aligned iterates satisfy the same gradient update. This alignment is used only in the analysis; the algorithm does not rotate or rebalance the factors.

For a direction $\Delta = (A, B) \in \mathcal{H}$, define the first-order changes in the represented matrix and balancing residual by

$$D_\Delta := A\widehat{V}^\top + \widehat{U}B^\top, \qquad S_\Delta := \widehat{U}^\top A + A^\top\widehat{U} - \widehat{V}^\top B - B^\top\widehat{V}.$$

Together, these variations control the factor perturbation on $\mathcal{H}$.

Lemma B.3 (Balanced-factor coercivity)**.** *For every $\Delta = (A, B) \in \mathcal{H}$,*

$$\sigma_r(\widehat{X})\|\Delta\|^2 \le \|D_\Delta\|_F^2 + \frac{1}{4}\|S_\Delta\|_F^2 \le 2\sigma_1(\widehat{X})\|\Delta\|^2.$$

Proof. Decompose A and B into their core and orthogonal components:

$$A = \widehat{P}A_0 + A_\perp, \qquad B = \widehat{Q}B_0 + B_\perp,$$

where $\widehat{P}^\top A_\perp = 0, \widehat{Q}^\top B_\perp = 0$.

The orthogonality conditions give

$$S_\Delta = \Sigma^{1/2}A_0 + A_0^\top\Sigma^{1/2} - \Sigma^{1/2}B_0 - B_0^\top\Sigma^{1/2},$$

so only the core components contribute to S_Δ. The three components of D_Δ are mutually orthogonal, yielding

$$\|D_\Delta\|_F^2 = \|A_0\Sigma^{1/2} + \Sigma^{1/2}B_0^\top\|_F^2 + \|A_\perp\Sigma^{1/2}\|_F^2 + \|B_\perp\Sigma^{1/2}\|_F^2.$$

The last two terms lie between $\sigma_r(\widehat{X})$ and $\sigma_1(\widehat{X})$ times $\|A_\perp\|_F^2 + \|B_\perp\|_F^2$.

We now compute the core contribution to $\|D_\Delta\|_F^2 + \frac{1}{4}\|S_\Delta\|_F^2$. Write $\sigma_i := \sigma_i(\widehat{X})$ and $\tau_i := \sqrt{\sigma_i}$. Since $\Delta \in \mathcal{H}$, the matrix $\Sigma^{1/2}(A_0 + B_0)$ is symmetric. For each diagonal index i, the entries $a = (A_0)_{ii}$ and $b = (B_0)_{ii}$ contribute $\sigma_i(a + b)^2 + \sigma_i(a - b)^2 = 2\sigma_i(a^2 + b^2)$.

For each non-diagonal pair (i, j) with $i < j$, set $a = (A_0)_{ij}$, $b = (B_0)_{ij}$, $c = (A_0)_{ji}$, and $d = (B_0)_{ji}$. With $x = a + b$, $y = a - b$, $u = c + d$, and $v = c - d$, the symmetry condition becomes $\tau_i x = \tau_j u$. These entries contribute

$$\begin{aligned}(a\tau_j + \tau_i d)^2 + (c\tau_i + \tau_j b)^2 + \frac{1}{2}(\tau_i y + \tau_j v)^2 &= \frac{1}{2}(\tau_j x + \tau_i u)^2 + \frac{1}{2}(\sigma_i + \sigma_j)(y^2 + v^2) \\ &= (\sigma_i + \sigma_j)(a^2 + b^2 + c^2 + d^2).\end{aligned}$$

The last equality uses $(\tau_j x + \tau_i u)^2 = (\sigma_i + \sigma_j)(x^2 + u^2)$, which follows from $\tau_i x = \tau_j u$.

Thus the diagonal weights $2\sigma_i$ and non-diagonal weights $\sigma_i + \sigma_j$ all lie between $2\sigma_r(\widehat{X})$ and $2\sigma_1(\widehat{X})$. Combining this with the orthogonal-component bounds and $\|\Delta\|^2 = \|A_0\|_F^2 + \|B_0\|_F^2 + \|A_\perp\|_F^2 + \|B_\perp\|_F^2$ proves the claim. □

The balanced-factor coercivity bound now yields curvature bounds for the target Hessian $H_* = \nabla^2 g_\lambda(\widehat{W})$. Since g_λ is invariant under common orthogonal rotations and $\nabla g_\lambda(\widehat{W}) = 0$, the Hessian vanishes on $\mathcal{V}$. Thus, it remains to characterize its restriction to $\mathcal{H} = \mathcal{V}^\perp$.

Proposition B.4 (Curvature after removing rotations)**.** *Suppose Definition 3.1 and Assumption 3.2 hold. Then,*

$$(1 - \gamma)\alpha\sigma_r(\widehat{X})I \preceq H_*|_{\mathcal{H}} \preceq (2 + \gamma)\beta\sigma_1(\widehat{X})I.$$

Consequently, the condition number of the positive part of H_* is at most $\frac{2+\gamma}{1-\gamma}K \le 3K$.

Proof. For $\Delta = (A, B) \in \mathcal{H}$, direct differentiation at the balanced target gives

$$\langle \Delta, H_*\Delta\rangle = \|\mathcal{A}(D_\Delta)\|_2^2 + \frac{\alpha}{4}\|S_\Delta\|_F^2 + 2\langle \nabla f(\widehat{X}), AB^\top\rangle.$$

Since $\mathrm{rank}(D_\Delta) \le 2r$, the restricted curvature bounds, Lemma B.3, and $\alpha \le \beta$ imply

$$\alpha\sigma_r(\widehat{X})\|\Delta\|^2 \le \|\mathcal{A}(D_\Delta)\|_2^2 + \frac{\alpha}{4}\|S_\Delta\|_F^2 \le 2\beta\sigma_1(\widehat{X})\|\Delta\|^2.$$

The remaining term is controlled by the normal-gradient assumption:

$$\begin{aligned}2\left|\langle \nabla f(\widehat{X}), AB^\top\rangle\right| &\le 2\|\nabla f(\widehat{X})\|_{\mathrm{op}}\|AB^\top\|_* \\ &\le \|\nabla f(\widehat{X})\|_{\mathrm{op}}\left(\|A\|_F^2 + \|B\|_F^2\right) \\ &\le \gamma\alpha\sigma_r(\widehat{X})\|\Delta\|^2.\end{aligned}$$

where $\|\cdot\|_*$ is the nuclear norm. Combining these estimates and using $\alpha\sigma_r(\widehat{X}) \le \beta\sigma_1(\widehat{X})$ gives

$$(1 - \gamma)\alpha\sigma_r(\widehat{X})I \preceq H_*|_{\mathcal{H}} \preceq (2 + \gamma)\beta\sigma_1(\widehat{X})I.$$

Because H_* is symmetric and vanishes on $\mathcal{V}$, its positive spectrum is precisely that of its restriction to $\mathcal{H}$. Therefore

$$\mathrm{cond}(H_*|_{\mathcal{H}}) \le \frac{2 + \gamma}{1 - \gamma}\frac{\beta}{\alpha}\frac{\sigma_1(\widehat{X})}{\sigma_r(\widehat{X})} = \frac{2 + \gamma}{1 - \gamma}\kappa\kappa_X \le 3K,$$

where the last inequality uses $0 \le \gamma \le 1/4$. □

These bounds characterize the quadratic approximation at the target.

The curvature bounds at the target describe the quadratic approximation of g_λ. To use this approximation in the local convergence analysis, we also need to control the errors in the function value, gradient, and Hessian near $\widehat{W}$.

Lemma B.5 (Local expansions). *For every factor perturbation E with $\|E\| \le \sqrt{\sigma_1(\widehat{X})}$,*

$$\left\|\nabla g_\lambda(\widehat{W}+E) - H_* E\right\| \le 9\beta\sqrt{\sigma_1(\widehat{X})}\,\|E\|^2.$$

On the same ball,

$$\left\|\nabla^2 g_\lambda(\widehat{W}+E) - H_*\right\|_{\mathrm{op}} \le 18\beta\sqrt{\sigma_1(\widehat{X})}\,\|E\|,$$
$$\left| g_\lambda(\widehat{W}+E) - g_\lambda(\widehat{W}) - \frac{1}{2}\langle E, H_* E\rangle \right| \le 3\beta\sqrt{\sigma_1(\widehat{X})}\,\|E\|^3.$$

Proof. Throughout the ball $\|W - \widehat{W}\| \le \sqrt{\sigma_1(\widehat{X})}$, the operator norms of U and V are both at most $2\sqrt{\sigma_1(\widehat{X})}$, by the triangle inequality and $\|\widehat{U}\|_{\mathrm{op}} = \|\widehat{V}\|_{\mathrm{op}} = \sqrt{\sigma_1(\widehat{X})}$. We bound the third derivatives of the loss and balancing terms separately.

For $X(U,V) := UV^\top$ and unit factor directions $\Delta_i = (A_i, B_i)$, differentiation gives

$$DX[\Delta_i] = A_i V^\top + U B_i^\top, \qquad D^2X[\Delta_i,\Delta_j] = A_i B_j^\top + A_j B_i^\top,$$

with $D^3X = 0$. The factor norm bounds and the Cauchy–Schwarz inequality yield

$$\|DX[\Delta_i]\|_F \le 2\sqrt{2}\,\sqrt{\sigma_1(\widehat{X})}, \qquad \|D^2X[\Delta_i,\Delta_j]\|_F \le 1.$$

Both matrix derivatives have rank at most $2r$. Thus, restricted smoothness implies

$$\left|\left\langle \mathcal{A}\big(D^2X[\Delta_i,\Delta_j]\big), \mathcal{A}\big(DX[\Delta_k]\big)\right\rangle\right| \le 2\sqrt{2}\,\beta\sqrt{\sigma_1(\widehat{X})}.$$

Since f is quadratic, every third derivative of $f(X(U,V))$ is the sum of three such pairings and is therefore bounded in absolute value by $6\sqrt{2}\,\beta\sqrt{\sigma_1(\widehat{X})}$.

For the balancing residual $C(U,V) := U^\top U - V^\top V$,

$$DC[\Delta_i] = U^\top A_i + A_i^\top U - V^\top B_i - B_i^\top V,$$
$$D^2C[\Delta_i,\Delta_j] = A_i^\top A_j + A_j^\top A_i - B_i^\top B_j - B_j^\top B_i.$$

The same norm inequalities give

$$\|DC[\Delta_i]\|_F \le 4\sqrt{2}\,\sqrt{\sigma_1(\widehat{X})}, \qquad \|D^2C[\Delta_i,\Delta_j]\|_F \le 2.$$

The third derivative of $\frac{\lambda}{2}\|C(U,V)\|_F^2$ is the sum of three pairings of the form $\lambda\langle D^2C[\Delta_i,\Delta_j], DC[\Delta_k]\rangle$. Using $\lambda = \frac{\alpha}{4} \le \frac{\beta}{4}$, its absolute value is therefore at most

$$24\sqrt{2}\,\lambda\sqrt{\sigma_1(\widehat{X})} \le 6\sqrt{2}\,\beta\sqrt{\sigma_1(\widehat{X})},$$

Combining the two contributions yields

$$\|\nabla^3 g_\lambda(W)\|_{\mathrm{op}} \le 12\sqrt{2}\,\beta\sqrt{\sigma_1(\widehat{X})} < 18\beta\sqrt{\sigma_1(\widehat{X})}$$

throughout the ball. All numerical factors in these bounds are independent of the rank and ambient dimensions.

Integrating this bound along the segment from $\widehat{W}$ to $\widehat{W}+E$ gives the Hessian estimate. Two further integrations, using $\nabla g_\lambda(\widehat{W}) = 0$ and $\nabla^2 g_\lambda(\widehat{W}) = H_*$, yield the gradient and function-value estimates, with constants $18/2 = 9$ and $18/6 = 3$, respectively.

□

B.6 Quadratic Contraction and Polynomial Bounds for Silver Blocks

We first establish the quadratic contraction underlying the local convergence analysis. Recall that $\rho = 1 + \sqrt{2}$ and that a depth-k Silver block contains $n_k = 2^k - 1$ stepsizes:

$$\alpha^{(0)} = [\,], \qquad \alpha^{(k)} = \big[\alpha^{(k-1)},\, 1 + \rho^{k-2},\, \alpha^{(k-1)}\big].$$

Recall the prefix polynomials introduced in Appendix A.4:

$$p_t^{(k)}(\xi) := \prod_{j=0}^{t-1}(1 - \alpha_j^{(k)}\xi), \qquad 0 \le t \le n_k.$$

with empty products equal to one. For a positive semidefinite matrix H, gradient descent on the quadratic $\frac{1}{2}\langle z, Hz\rangle$ with stepsizes $\alpha_j^{(k)}/L$ then satisfies $z_t = p_t^{(k)}(H/L)z_0$.

The endpoint estimate follows from the Silver convergence guarantee.

Theorem B.6 (Theorem 1.1 in Altschuler & Parrilo (2025a))**.** *Let F be convex and L-smooth, and let $x_\star$ be a minimizer. Gradient descent with stepsizes $\alpha_j^{(k)}/L$ satisfies*

$$F(x_{n_k}) - F(x_\star) \le \frac{L\|x_0 - x_\star\|^2}{1 + \sqrt{4\rho^{2k} - 3}} \le \frac{L\|x_0 - x_\star\|^2}{2\rho^k}.$$

Applying this result with $L = 1$ to $F(x) = \xi x^2/2$ for $0 \le \xi \le 1$ and $x_0 = 1$ gives

$$\sup_{0\le\xi\le 1} \xi\big(p_{n_k}^{(k)}(\xi)\big)^2 \le \rho^{-k}.$$

Consequently, if the positive eigenvalues of H lie in $[m, L]$ with $m > 0$, the spectral theorem yields

$$\|p_{n_k}^{(k)}(H/L)z\| \le \sqrt{\frac{L}{m\rho^k}}\,\|z\|, \qquad z \perp \ker H.$$

For the factor objective, set $\overline{\beta} = 4\beta\sigma_1(\widehat{X})$. By Proposition B.4, $0 \preceq H_*/\overline{\beta} \preceq \frac{3}{4}I, (H_*/\overline{\beta})|_{\mathcal{H}} \succeq \frac{3}{20K}I$. Thus, for an aligned initial error $E_0 \in \mathcal{H}$ and $\rho^k \ge 192K$,

$$\|p_{n_k}^{(k)}(H_*/\overline{\beta})E_0\| \le \sqrt{\frac{20K}{3\rho^k}}\,\|E_0\| \le \frac{\sqrt{5}}{12}\|E_0\|.$$

We next bound prefixes, endpoint tails, and internal products of a Silver block. These bounds control how errors propagate within the block.

Lemma B.7 (Silver product bounds)**.** *For every $k \ge 0$ and $\xi \in [0, 1]$, all prefixes and endpoint tails satisfy*

$$|p_t^{(k)}(\xi)| \le 1, \qquad \left|\prod_{j=t}^{n_k-1}(1 - \alpha_j^{(k)}\xi)\right| \le 1, \qquad 0 \le t \le n_k.$$

For $k \ge 2$, every internal product satisfies

$$\left|\prod_{j=s}^{t-1}(1 - \alpha_j^{(k)}\xi)\right| \le \rho^{k-2}, \qquad 0 \le s \le t \le n_k.$$

Moreover,

$$\sum_{j=0}^{n_k-1}\alpha_j^{(k)} = \rho^k - 1.$$

The product bounds also hold in operator norm when ξ is replaced by a symmetric matrix with spectrum in $[0, 1]$. Empty products are understood to equal one.

Proof. Write $a_k := 1 + \rho^{k-2}$ for the central stepsize and $P_k := p_{n_k}^{(k)}$ for the endpoint polynomial. We first establish the auxiliary bound

$$|R_k(\xi)| \le 1, \qquad R_k(\xi) := P_{k-1}(\xi)(1 - a_{k+1}\xi), \qquad k \ge 1.$$

The cases $k = 1, 2$ follow from

$$R_1(\xi) = 1 - 2\xi, \qquad R_2(\xi) = 1 - (2 + 2\sqrt{2})\xi(1 - \xi).$$

For $k \ge 3$, the palindromic recursion gives

$$R_k(\xi) = R_{k-1}(\xi)^2 - 2\rho^{k-3}\xi(1 - \xi)P_{k-2}(\xi)^2.$$

The subtracted term is nonnegative and, by the scalar endpoint bound in Appendix B.6, satisfies

$$2\rho^{k-3}\xi(1 - \xi)P_{k-2}(\xi)^2 \le \frac{2}{\rho}(1 - \xi) \le \frac{2}{\rho} < 1.$$

Hence $-2/\rho \le R_k(\xi) \le 1$ by induction, proving the auxiliary bound.

We now prove the prefix bound by induction on k. The cases $k = 0, 1$ are immediate. For $k \ge 2$, prefixes contained in the first depth-$(k-1)$ block satisfy the bound by the induction hypothesis. The prefix ending immediately after the central step satisfies

$$P_{k-1}(\xi)(1 - a_k\xi) = \big[P_{k-2}(\xi)(1 - a_{k-1}\xi)\big] R_{k-1}(\xi).$$

The bracketed term is a prefix of the smaller block, so both factors have magnitude at most one. Every remaining prefix is its product with a prefix of the second depth-$(k-1)$ block, so the same bound applies. Since the schedule is palindromic, each endpoint tail equals a prefix product of the same length.

For the internal-product bound, a product either lies within one smaller block or contains the central step. In the latter case, it consists of a tail of the left block, the central factor, and a prefix of the right block. The two side products have magnitude at most one, while

$$\sup_{\xi \in [0,1]} |1 - a_k\xi| = \rho^{k-2}, \qquad k \ge 2.$$

Products contained in a smaller block are controlled by induction, starting from the depth-one bound 1.

Finally, if T_k denotes the sum of the depth-k steps, then $T_0 = 0$ and $T_k = 2T_{k-1} + a_k$. Using $\rho^2 = 2\rho + 1$, induction gives $T_k = \rho^k - 1$. The matrix product bounds follow by applying the scalar bounds to each eigenvalue in an orthonormal eigenbasis. □

In particular, the endpoint tail bound applies on the whole factor space, including the zero eigenspace.

C Model Specializations, Statistical Recovery, and Initialization

We first connect our restricted-curvature framework to the restricted isometry property (RIP). We then derive model-independent statistical recovery bounds and a sufficient condition for the normal-gradient assumption, before specializing the analysis to isotropic and anisotropic Gaussian sensing and multivariate regression, together with model-specific initialization guarantees.

Definition C.1 (Restricted Isometry Property)**.** The operator $\mathcal{A}$ satisfies the rank-$2r$ restricted isometry property (RIP) with constant $\delta \in [0, 1)$ if

$$(1 - \delta)\|H\|_{\mathrm{F}}^2 \le \|\mathcal{A}(H)\|_2^2 \le (1 + \delta)\|H\|_{\mathrm{F}}^2 \qquad \text{for every } H \text{ with } \operatorname{rank}(H) \le 2r.$$

Hence, if $\mathcal{A}$ satisfies the rank-$2r$ RIP with constant δ, then Definition 3.1 holds with

$$\alpha = 1 - \delta, \qquad \beta = 1 + \delta, \qquad \kappa = \frac{1 + \delta}{1 - \delta}.$$

Under the remaining assumptions of Theorem 4.1, the theorem therefore applies without any distributional assumption on $\mathcal{A}$ or statistical model for the observations.

C.1 Statistical Recovery and Normal-Gradient Control

We now introduce a generative observation model and relate the optimization target $\widehat{X}$ to the true matrix $X_\star$. The results in this subsection are model-independent: they give a deterministic statistical-error bound and a sufficient noise condition for the normal-gradient assumption.

We now specialize to the statistical observation model

$$y = \mathcal{A}(X_\star) + \varepsilon, \qquad \operatorname{rank}(X_\star) \le r. \tag{18}$$

Since $X_\star$ is feasible for the rank-constrained problem, the constrained-minimizer formulation of $\widehat{X}$ automatically gives

$$f(\widehat{X}) \le f(X_\star). \tag{19}$$

The following deterministic bound relates the optimization target $\widehat{X}$ to the true matrix $X_\star$.

Lemma C.2 (Statistical error of the constrained minimizer)**.** *Suppose rank-$2r$ RSC holds with constant α. Then*

$$\|\widehat{X} - X_\star\|_F \le \frac{2\sqrt{2r}}{\alpha}\|\mathcal{A}^*\varepsilon\|_{\mathrm{op}}. \tag{20}$$

Proof. By (18) and (19), $\frac{1}{2}\|\mathcal{A}(\widehat{X} - X_\star) - \varepsilon\|_2^2 \le \frac{1}{2}\|\varepsilon\|_2^2$. Expanding and cancelling the noise term gives

$$\frac{1}{2}\|\mathcal{A}(\widehat{X} - X_\star)\|_2^2 \le \langle \varepsilon, \mathcal{A}(\widehat{X} - X_\star)\rangle = \langle \mathcal{A}^*\varepsilon, \widehat{X} - X_\star\rangle.$$

Since $\operatorname{rank}(\widehat{X} - X_\star) \le 2r$, by the duality between the operator and nuclear norms,

$$\langle \mathcal{A}^*\varepsilon, \widehat{X} - X_\star\rangle \le \|\mathcal{A}^*\varepsilon\|_{\mathrm{op}}\|\widehat{X} - X_\star\|_* \le \sqrt{2r}\,\|\mathcal{A}^*\varepsilon\|_{\mathrm{op}}\|\widehat{X} - X_\star\|_F.$$

Rank-$2r$ RSC gives

$$\frac{\alpha}{2}\|\widehat{X} - X_\star\|_F^2 \le \sqrt{2r}\,\|\mathcal{A}^*\varepsilon\|_{\mathrm{op}}\|\widehat{X} - X_\star\|_F,$$

and cancelling one factor of $\|\widehat{X} - X_\star\|_F$ proves (20). □

Combining Lemma C.2 and Lemma A.1 gives, at every block endpoint T,

$$\|U_T V_T^\top - X_\star\|_F \le 3\sqrt{\sigma_1(\widehat{X})}\,\operatorname{dist}(W_T, \widehat{W}) + \frac{2\sqrt{2r}}{\alpha}\|\mathcal{A}^*\varepsilon\|_{\mathrm{op}}.$$

In particular, for $T = bn_k$, the block-endpoint contraction yields

$$\|U_T V_T^\top - X_\star\|_F \le 3\sqrt{\sigma_1(\widehat{X})}\,2^{-b}D_0 + \frac{2\sqrt{2r}}{\alpha}\|\mathcal{A}^*\varepsilon\|_{\mathrm{op}}.$$

Only the first term is reduced by further optimization toward $\widehat{X}$; the second term controls the statistical discrepancy between $\widehat{X}$ and $X_\star$.

We next give a sufficient condition under which the small-normal-gradient assumption at $\widehat{X}$ follows from the noise level. Since $\nabla f(\widehat{X}) = \mathcal{A}^*\mathcal{A}(\widehat{X} - X_\star) - \mathcal{A}^*\varepsilon$, for unit vectors u and v,

$$\begin{aligned} |u^\top \nabla f(\widehat{X}) v| &= \left|\langle \mathcal{A}(uv^\top), \mathcal{A}(\widehat{X} - X_\star)\rangle - \langle \mathcal{A}^*\varepsilon, uv^\top\rangle\right| \\ &\le \beta\|\widehat{X} - X_\star\|_F + \|\mathcal{A}^*\varepsilon\|_{\mathrm{op}}. \end{aligned}$$

Taking the supremum over unit u, v and using Lemma C.2 gives

$$\|\nabla f(\widehat{X})\|_{\mathrm{op}} \le \left(1 + 2\sqrt{2}\,\kappa\sqrt{r}\right)\|\mathcal{A}^*\varepsilon\|_{\mathrm{op}},$$

where $\kappa = \beta/\alpha$. Therefore, a sufficient condition for $\|\nabla f(\widehat{X})\|_{\mathrm{op}} \le \gamma\alpha\sigma_r(\widehat{X})$ is

$$\left(1 + 2\sqrt{2}\,\kappa\sqrt{r}\right)\|\mathcal{A}^*\varepsilon\|_{\mathrm{op}} \le \gamma\alpha\sigma_r(\widehat{X}).$$

All bounds above are deterministic statements for the realized operator $\mathcal{A}$ and noise vector ε. Distributional assumptions on the noise are needed only when these realized quantities are subsequently bounded with high probability. Moreover, (20) is an upper bound on the statistical error of $\widehat{X}$; it does not assert a matching lower bound or that this error level is unavoidable.

C.2 Isotropic Gaussian Sensing and Spectral Initialization

We apply the preceding statistical bounds to isotropic Gaussian sensing and justify the spectral initializer in Section 5.2. We first consider the noiseless case. Then $\widehat{X} = X_\star$, and we assume $\mathrm{rank}(X_\star) = r,\ \sigma_r(X_\star) > 0$. Suppose that $\mathcal{A}$ satisfies rank-$3r$ RIP with constant δ:

$$(1-\delta)\|H\|_F^2 \le \|\mathcal{A}(H)\|_2^2 \le (1+\delta)\|H\|_F^2$$

for every matrix H with $\mathrm{rank}(H) \le 3r$. In particular, the rank-$2r$ restricted-curvature constants required for acceleration may be taken as

$$\alpha = 1-\delta, \qquad \beta = 1+\delta, \qquad \kappa = \frac{1+\delta}{1-\delta}.$$

Recall the spectral initializer

$$Y_0 = \mathcal{A}^* y = \mathcal{A}^*\mathcal{A}(X_\star),$$

and let X_0 be the best rank-r approximation of Y_0, with balanced factors W_0.

Proposition C.3 (Spectral initialization for Gaussian sensing)**.** *There is a universal constant $c > 0$ such that, if*

$$\delta \le \frac{c}{\sqrt{r}\,\kappa_X^{3/2}}, \tag{21}$$

then the balanced factors W_0 satisfy the local initialization condition of Theorem 4.1. Moreover, for normalized Gaussian measurements, it is sufficient to have

$$m \ge Cr^2\kappa_X^3(d_1+d_2)\log(d_1+d_2) \tag{22}$$

measurements.

Proof. Since X_0 is the best rank-r approximation of Y_0 and $\mathrm{rank}(X_\star) = r$,

$$\|Y_0 - X_0\|_F^2 \le \|Y_0 - X_\star\|_F^2.$$

Expanding both sides gives

$$\|X_0 - X_\star\|_F^2 \le 2\langle Y_0 - X_\star, X_0 - X_\star\rangle.$$

Since

$$Y_0 - X_\star = (\mathcal{A}^*\mathcal{A} - I)X_\star,$$

rank-$3r$ RIP and the polarization identity imply

$$|\langle(\mathcal{A}^*\mathcal{A} - I)X_\star, H\rangle| \le \delta\|X_\star\|_F\|H\|_F$$

for every H with $\mathrm{rank}(H) \le 2r$. Taking $H = X_0 - X_\star$ therefore gives

$$\|X_0 - X_\star\|_F \le 2\delta\|X_\star\|_F \le 2\delta\sqrt{r}\,\sigma_1(X_\star).$$

Choose the constant in (21) sufficiently small that $\delta \le 1/2$, and hence $\kappa \le 3$. Then

$$\|X_0 - X_\star\|_F \le \frac{2c\,\sigma_r(X_\star)}{\sqrt{\kappa_X}} \le \frac{6c\,\sigma_r(X_\star)}{\kappa\sqrt{\kappa_X}}.$$

For $6c$ no larger than the constant in Lemma A.2, this is its required matrix-error bound. In particular, the error is at most $\sigma_r(X_\star)/2$, so Weyl's inequality and $\mathrm{rank}(X_0) \le r$ give $\mathrm{rank}(X_0) = r$. The lemma therefore gives the required local initialization condition. Decrease c further, if necessary, to accommodate the warm-start parameter choice in Proposition A.4.

For normalized Gaussian measurements, standard low-rank RIP results imply that rank-$3r$ RIP with constant δ holds with high probability provided

$$m \ge C\delta^{-2}r(d_1+d_2)\log(d_1+d_2).$$

The sample-size condition in (21) therefore implies (22). □

We next return to the noisy model $y = \mathcal{A}(X_\star) + \varepsilon$. The same initialization argument gives

$$\|X_0 - X_\star\|_F \le 2\delta\|X_\star\|_F + 2\sqrt{2r}\,\|\mathcal{A}^*\varepsilon\|_{\mathrm{op}}.$$

Combining this estimate with Lemma C.2 and the triangle inequality yields

$$\|X_0 - \widehat{X}\|_F \le 2\delta\|X_\star\|_F + 2\sqrt{2r}\left(1 + \frac{1}{\alpha}\right)\|\mathcal{A}^*\varepsilon\|_{\mathrm{op}}.$$

Suppose $\mathrm{rank}(\widehat{X}) = r$ and the right-hand side is at most $c_0\sigma_r(\widehat{X})/(\kappa\sqrt{\kappa_X})$ for a sufficiently small universal $c_0 > 0$. Weyl's inequality then gives $\mathrm{rank}(X_0) = r$, and Lemma A.2 places its balanced factors in the certified local basin. If Assumption 3.2 also holds, Theorem 4.1 and Proposition A.4 apply.

The rank-$3r$ RIP condition above is used only to certify the spectral initializer; the acceleration theorem itself requires only rank-$2r$ restricted curvature. The conditions (21)–(22) and the noisy initialization bound above are sufficient conditions for entering the local basin; they are not recovery lower bounds.

C.3 Anisotropic Gaussian Sensing and Covariance-Corrected Initialization

We now extend the isotropic analysis to anisotropic Gaussian sensing, examining how the sensing covariance affects restricted curvature and spectral initialization. Recall from Section 5.2 that $A_i = G_i S^{1/2}$, where $S \in \mathbb{R}^{d_2\times d_2}$ is known and positive definite and the entries of the G_i are mutually independent $\mathcal{N}(0, 1/m)$ variables. Writing $[\mathcal{G}(H)]_i = \langle G_i, H\rangle$, we have

$$\mathcal{A}(H) = \mathcal{G}(HS^{1/2}), \qquad \mathcal{A}^*v = (\mathcal{G}^*v)S^{1/2}.$$

We first work on an event where $\mathcal{G}$ satisfies rank-$3r$ RIP with constant $0 \le \delta < 1$. Only the measurement ensemble and the initializer are changed. The Silver iteration is still gradient descent on g_λ in the original factor coordinates.

We first establish the restricted curvature bounds and examine the limitation of small RIP in the original Frobenius norm. Since $S^{1/2}$ is invertible, $\mathrm{rank}(HS^{1/2}) = \mathrm{rank}(H)$. Consequently, for every $\mathrm{rank}(H) \le 3r$,

$$(1-\delta)\|HS^{1/2}\|_F^2 \le \|\mathcal{A}(H)\|_2^2 \le (1+\delta)\|HS^{1/2}\|_F^2. \tag{23}$$

The eigenvalue bounds $\lambda_{\min}(S)\|H\|_F^2 \le \|HS^{1/2}\|_F^2 \le \lambda_{\max}(S)\|H\|_F^2$ give the valid restricted curvature constants

$$\alpha = (1-\delta)\lambda_{\min}(S), \quad \beta = (1+\delta)\lambda_{\max}(S), \quad \kappa = \frac{1+\delta}{1-\delta}\,\mathrm{cond}(S). \tag{24}$$

These need not be the sharp realized constants. The same lower bound also makes f coercive on the closed set $\{X : \mathrm{rank}(X) \le r\}$, ensuring existence of a minimizer.

For completeness, anisotropy can rule out a small RIP constant in the original Frobenius norm even after a scalar rescaling. If $a > 0$ and $0 \le \delta_0 < 1$ satisfy $(1-\delta_0)\|H\|_F^2 \le a\|\mathcal{A}(H)\|_2^2 \le (1+\delta_0)\|H\|_F^2$ for every rank-one H, then

$$\delta_0 \ge \frac{(1-\delta)\,\mathrm{cond}(S) - (1+\delta)}{(1-\delta)\,\mathrm{cond}(S) + (1+\delta)}. \tag{25}$$

Indeed, take $H = uv^\top$ with u, v unit vectors and v an eigenvector of S. For the smallest eigenvalue, $1 - \delta_0 \le a(1+\delta)\lambda_{\min}(S)$; for the largest, $a(1-\delta)\lambda_{\max}(S) \le 1 + \delta_0$. Eliminating a proves the claim. The bound is informative when its numerator is positive. For fixed $\delta < 1$, the lower bound approaches one as $\mathrm{cond}(S)$ increases.

We next use the covariance structure to bound the statistical and initialization errors. The rank-preserving change of variables permits a sharper statistical bound than direct substitution into Lemma C.2.

Proposition C.4 (Statistical error under anisotropic sensing). *Suppose* $y = \mathcal{A}(X_\star) + \varepsilon$, *with* $\operatorname{rank}(X_\star) \le r$, *and let* $\widehat{X} \in \arg\min_{\operatorname{rank}(X)\le r} f(X)$. *Under* (23),

$$\|(\widehat{X} - X_\star)S^{1/2}\|_F \le \frac{2\sqrt{2r}}{1-\delta}\|\mathcal{G}^*\varepsilon\|_{\mathrm{op}}, \tag{26}$$

and consequently

$$\|\widehat{X} - X_\star\|_F \le \frac{2\sqrt{2r}}{(1-\delta)\sqrt{\lambda_{\min}(S)}}\|\mathcal{G}^*\varepsilon\|_{\mathrm{op}}. \tag{27}$$

Moreover,

$$\|\nabla f(\widehat{X})\|_{\mathrm{op}} \le \sqrt{\lambda_{\max}(S)}\left(1 + \frac{2\sqrt{2r}(1+\delta)}{1-\delta}\right)\|\mathcal{G}^*\varepsilon\|_{\mathrm{op}}. \tag{28}$$

Thus, when $\operatorname{rank}(\widehat{X}) = r$, *an explicit sufficient condition for Assumption 3.2 is that the right-hand side of* (28) *be at most* $\gamma(1-\delta)\lambda_{\min}(S)\sigma_r(\widehat{X})$.

Proof. Apply the basic inequality from Lemma C.2 to $H = (\widehat{X} - X_\star)S^{1/2}$, which has rank at most $2r$. Since $f(\widehat{X}) \le f(X_\star)$,

$$\frac{1}{2}\|\mathcal{G}(H)\|_2^2 \le \langle \mathcal{G}^*\varepsilon, H\rangle \le \sqrt{2r}\,\|\mathcal{G}^*\varepsilon\|_{\mathrm{op}}\|H\|_F.$$

The RIP lower bound and cancellation of $\|H\|_F$ prove (26); the case $H = 0$ is immediate. Multiplication by $S^{-1/2}$ proves (27). For unit vectors u, v, the RIP upper bound and Cauchy–Schwarz give

$$|\langle uv^\top, \mathcal{G}^*\mathcal{G}(H)\rangle| = |\langle \mathcal{G}(uv^\top), \mathcal{G}(H)\rangle| \le (1+\delta)\|H\|_F.$$

Taking the supremum yields $\|\mathcal{G}^*\mathcal{G}(H)\|_{\mathrm{op}} \le (1+\delta)\|H\|_F$. Finally,

$$\nabla f(\widehat{X}) = \big(\mathcal{G}^*\mathcal{G}(H) - \mathcal{G}^*\varepsilon\big)S^{1/2}.$$

The triangle inequality and (26) prove (28). □

For initialization, form $Y_0 = \mathcal{A}^*y\,S^{-1/2} = \mathcal{G}^*y$ and choose X_0 so that $X_0S^{1/2}$ is a best rank-r approximation of Y_0. Compute the balanced factors W_0 from an SVD of X_0; transforming the balanced factors of the best approximation of Y_0 need not preserve balance. The initializer uses only the observations, sensing matrices, known S, and rank r. It does not use $X_\star$ or $\widehat{X}$.

$Y_0 = \mathcal{A}^*y\,S^{-1/2} = \mathcal{G}^*y$

Proposition C.5 (Initialization error and entry into the local basin). *Under rank-*$3r$ *RIP for* $\mathcal{G}$ *and* $\operatorname{rank}(X_\star) \le r$,

$$\|X_0 - X_\star\|_F \le \frac{2\delta\|X_\star S^{1/2}\|_F + 2\sqrt{2r}\,\|\mathcal{G}^*\varepsilon\|_{\mathrm{op}}}{\sqrt{\lambda_{\min}(S)}}, \tag{29}$$

$$\|X_0 - \widehat{X}\|_F \le \frac{2\delta\|X_\star S^{1/2}\|_F + 2\sqrt{2r}\big(1 + (1-\delta)^{-1}\big)\|\mathcal{G}^*\varepsilon\|_{\mathrm{op}}}{\sqrt{\lambda_{\min}(S)}}. \tag{30}$$

Suppose $\operatorname{rank}(\widehat{X}) = r$, *and retain* $\kappa_X = \sigma_1(\widehat{X})/\sigma_r(\widehat{X})$ *and* (24). *If the right-hand side of* (30) *is at most* $c_0\sigma_r(\widehat{X})/(\kappa\sqrt{\kappa_X})$, *with a sufficiently small universal* $c_0 > 0$, *then* $\operatorname{rank}(X_0) = r$ *and its balanced factors satisfy the local initialization condition of Theorem 4.1. If the normal-gradient condition in Proposition C.4 also holds, the same fixed Silver iteration and the warm-start parameter choice in Proposition A.4 apply.*

Proof. The best-approximation property gives

$$\|Y_0 - X_0S^{1/2}\|_F^2 \le \|Y_0 - X_\star S^{1/2}\|_F^2.$$

With $H = (X_0 - X_\star)S^{1/2}$, expansion gives

$$\begin{aligned}\|H\|_F^2 &\le 2\langle Y_0 - X_\star S^{1/2}, H\rangle \\ &= 2\langle(\mathcal{G}^*\mathcal{G} - I)(X_\star S^{1/2}), H\rangle + 2\langle\mathcal{G}^*\varepsilon, H\rangle \\ &\le 2\delta\|X_\star S^{1/2}\|_F\|H\|_F + 2\sqrt{2r}\,\|\mathcal{G}^*\varepsilon\|_{\mathrm{op}}\|H\|_F.\end{aligned}$$

Here $\mathrm{rank}(H) \le 2r$ and $\mathrm{rank}(X_\star S^{1/2}) \le r$. Every matrix in their linear span has rank at most $3r$. Restricting $\mathcal{G}^*\mathcal{G} - I$ to this span, RIP and the spectral theorem bound its bilinear form by δ times the product of Frobenius norms. This justifies the last inequality without an ambient operator-norm bound. Cancellation and $\|X_0 - X_\star\|_F \le \|H\|_F/\sqrt{\lambda_{\min}(S)}$ prove (29). The triangle inequality and (27) prove (30).

Choose $c_0 \le 1/2$ and no larger than the constant in Lemma A.2. The assumed error bound is then at most $\sigma_r(\widehat{X})/2$. Since $\mathrm{rank}(X_0) \le r$, Weyl's inequality implies $\sigma_r(X_0) \ge \sigma_r(\widehat{X})/2 > 0$. Lemma A.2 therefore gives the required factor-distance bound. If necessary, decrease c_0 once more to meet the smaller universal constant in Proposition A.4. The additional gradient condition verifies Assumption 3.2; all other curvature and rank hypotheses have already been checked. □

Corollary C.6 (Noiseless Gaussian sample bound)**.** *Suppose $\varepsilon = 0$ and $\mathrm{rank}(X_\star) = r$. For a sufficiently small universal $c > 0$, the initializer enters the local basin whenever*

$$\delta \le \frac{c}{\sqrt{r}\ \mathrm{cond}(S)^{3/2}\kappa_X^{3/2}}. \tag{31}$$

Consequently, for the normalized Gaussian ensemble, the initializer enters the local basin with high probability provided

$$m \ge Cr^2\,\mathrm{cond}(S)^3\kappa_X^3(d_1 + d_2)\log(d_1 + d_2). \tag{32}$$

Here $\widehat{X} = X_\star$ and $\kappa_X = \sigma_1(X_\star)/\sigma_r(X_\star)$. This is a sufficient bound for the specified one-step spectral initializer to reach the Silver basin, not a statistical recovery lower bound or a claim of optimal dependence on conditioning.

Proof. The rank-$2r$ lower curvature bound implies $\widehat{X} = X_\star$ uniquely in the noiseless problem, and $\nabla f(\widehat{X}) = 0$. Take c in (31) small enough that $\delta \le 1/2$. Then $\kappa \le 3\,\mathrm{cond}(S)$ and

$$\begin{aligned}\|X_0 - X_\star\|_F &\le 2\delta\frac{\|X_\star S^{1/2}\|_F}{\sqrt{\lambda_{\min}(S)}} \\ &\le 2\delta\sqrt{r}\,\sigma_1(X_\star)\sqrt{\mathrm{cond}(S)} \\ &\le \frac{2c\,\sigma_r(X_\star)}{\mathrm{cond}(S)\sqrt{\kappa_X}} \le \frac{6c\,\sigma_r(X_\star)}{\kappa\sqrt{\kappa_X}}.\end{aligned}$$

Choose $6c \le c_0$ from Proposition C.5. This proves entry into the basin, including the balanced-factor and warm-start parameter requirements.

The standard Gaussian RIP estimate gives rank-$3r$ RIP with constant δ and probability at least $1 - 2\exp(-c\delta^2 m)$ when $m \ge C\delta^{-2}r(d_1 + d_2)\log(d_1 + d_2)$ (Recht et al., 2010; Tu et al., 2016). Choosing δ equal to the right-hand side of (31), with a fixed sufficiently small constant, yields (32) and failure probability at most $2\exp\{-cm/[r\,\mathrm{cond}(S)^3\kappa_X^3]\}$. No extra factor of $\mathrm{cond}(S)$ is needed to obtain RIP for $\mathcal{G}$ itself: the conditioning factors above enter when converting the initialization error to the original Silver basin. □

We now specialize the deterministic noise bounds to Gaussian noise and obtain a recovery guarantee for the true matrix. Suppose additionally that the ε_i are independent $\mathcal{N}(0, \nu^2/m)$ variables, independent of the sensing matrices. This convention matches the normalized $\mathcal{N}(0, 1/m)$ measurements: multiplying each observation by $\sqrt{m}$ gives unit-variance Gaussian sensing entries and noise variance ν^2. Conditionally on ε, the entries of $\mathcal{G}^*\varepsilon$ are independent $\mathcal{N}(0, \|\varepsilon\|_2^2/m)$ variables. Equivalently,

$$\mathcal{G}^*\varepsilon \,\big|\, \varepsilon \overset{d}{=} \frac{\|\varepsilon\|_2}{\sqrt{m}}\,G,$$

where G has independent standard Gaussian entries. The Gaussian operator-norm bound used in Appendix C.4, together with $\Pr\{\|\varepsilon\|_2 > \nu(1 + t/\sqrt{m})\} \le e^{-t^2/2}$, gives, for every $t \ge 0$, with probability at least $1 - 2e^{-t^2/2}$,

$$\|\mathcal{G}^*\varepsilon\|_{\mathrm{op}} \le \frac{\nu}{\sqrt{m}} \left(1 + \frac{t}{\sqrt{m}}\right) \left(\sqrt{d_1} + \sqrt{d_2} + t\right). \tag{33}$$

The vector-norm tail follows from Gaussian concentration of the Euclidean norm and its expectation bound $\mathbb{E}\|N(0, I_m)\|_2 \le \sqrt{m}$. The conditional matrix-norm estimate is taken before restricting to the RIP event. A union bound, rather than an independence claim, combines (33) with that event.

On the intersection, Proposition C.4 yields

$$\|\widehat{X} - X_\star\|_F \le \frac{2\sqrt{2r}\,\nu}{(1-\delta)\sqrt{m\lambda_{\min}(S)}} \left(1 + \frac{t}{\sqrt{m}}\right) \left(\sqrt{d_1} + \sqrt{d_2} + t\right). \tag{34}$$

Its probability is at least $1 - 2\exp(-c\delta^2 m) - 2\exp(-t^2/2)$ under the preceding Gaussian RIP sample requirement. Substitution of (33) into Propositions C.4 and C.5 also gives explicit sufficient noise conditions for the normal gradient and initialization. The rank-exactness condition at $\widehat{X}$ remains as in the main theorem; when $\mathrm{rank}(X_\star) = r$, it follows, for example, if the right-hand side of (34) is less than $\sigma_r(X_\star)$.

Under these local conditions, Theorem 4.1 and Lemma A.1 give at every block endpoint $T = bn_k$

$$\begin{aligned}\|U_T V_T^\top - X_\star\|_F \le{}& 3\sqrt{\sigma_1(\widehat{X})}\, 2^{-b} D_0 \\ &+ \frac{2\sqrt{2r}\,\nu}{(1-\delta)\sqrt{m\lambda_{\min}(S)}} \left(1 + \frac{t}{\sqrt{m}}\right) \left(\sqrt{d_1} + \sqrt{d_2} + t\right).\end{aligned} \tag{35}$$

The first term is the optimization error; the second is an upper bound on the statistical error, not an asserted unavoidable floor. The conditioning factor $\mathrm{cond}(S)^3$ in (32) concerns entry of this initializer into the local basin and does not appear in (34). Setting $S = I$ recovers the initialization bounds in Appendix C.2 and gives the corresponding Gaussian noise recovery guarantee.

C.4 Multivariate Regression

We now turn to multivariate regression, deriving its restricted curvature, statistical error bounds, and truncated-OLS initialization guarantee.

Restricted curvature and the limitation of RIP. We consider the low-rank multivariate regression problem introduced in Subsection 5.3:

$$\min_{\mathrm{rank}(\Theta) \le r} \frac{1}{2n} \|\Theta Z^\top - Y^\top\|_{\mathrm{F}}^2.$$

For fixed data Z and Y, this has sensing operator $\mathcal{A}_Z(H) = n^{-1/2}\,\mathrm{vec}(HZ^\top)$ and observation vector $n^{-1/2}\,\mathrm{vec}(Y^\top)$. Write $\widehat{\Sigma}_Z := n^{-1} Z^\top Z$.

Proposition C.7 (Exact curvature for multivariate regression)**.** *Suppose that $\widehat{\Sigma}_Z \succ 0$. Then, for every matrix H,*

$$\lambda_{\min}(\widehat{\Sigma}_Z)\|H\|_{\mathrm{F}}^2 \le \|\mathcal{A}_Z(H)\|_2^2 \le \lambda_{\max}(\widehat{\Sigma}_Z)\|H\|_{\mathrm{F}}^2.$$

Consequently, Definition 3.1 holds with

$$\alpha = \lambda_{\min}(\widehat{\Sigma}_Z), \qquad \beta = \lambda_{\max}(\widehat{\Sigma}_Z), \qquad \kappa = \mathrm{cond}(\widehat{\Sigma}_Z).$$

Proof. Writing the rows of H as $h_i^\top$, we have

$$\|\mathcal{A}_Z(H)\|_2^2 = \mathrm{tr}(H\widehat{\Sigma}_Z H^\top) = \sum_i h_i^\top \widehat{\Sigma}_Z h_i.$$

Since $\widehat{\Sigma}_Z \succ 0$,

$$\lambda_{\min}(\widehat{\Sigma}_Z)\|h_i\|_2^2 \le h_i^\top \widehat{\Sigma}_Z h_i \le \lambda_{\max}(\widehat{\Sigma}_Z)\|h_i\|_2^2.$$

Summing over the rows gives the result. □

The positive-definiteness assumption is necessary for a positive restricted strong convexity constant: if $\widehat{\Sigma}_Z$ is singular, any nonzero $b \in \ker(\widehat{\Sigma}_Z)$ and nonzero a give a rank-one matrix $H = ab^\top$ with $\mathcal{A}_Z(H) = 0$. Even when $\widehat{\Sigma}_Z \succ 0$, ill-conditioning can preclude a small RIP constant, as the following proposition shows.

Proposition C.8 (Lower bound on the RIP constant)**.** *Suppose $\widehat{\Sigma}_Z \succ 0$ and write $\kappa_Z := \mathrm{cond}(\widehat{\Sigma}_Z)$. If some $c > 0$ and $\delta_0 \in [0, 1)$ satisfy*

$$(1 - \delta_0)\|H\|_{\mathrm{F}}^2 \le c\|\mathcal{A}_Z(H)\|_2^2 \le (1 + \delta_0)\|H\|_{\mathrm{F}}^2$$

for every rank-one matrix H, then

$$\delta_0 \ge \frac{\kappa_Z - 1}{\kappa_Z + 1}.$$

Proof. Take $H = ab^\top$, where a is a unit vector and b is a unit eigenvector of $\widehat{\Sigma}_Z$. Then $\|H\|_F = 1$ and $\|\mathcal{A}_Z(H)\|_2^2 = b^\top \widehat{\Sigma}_Z b$. Choosing eigenvectors for the smallest and largest eigenvalues gives $c\lambda_{\min}(\widehat{\Sigma}_Z) \ge 1 - \delta_0$ and $c\lambda_{\max}(\widehat{\Sigma}_Z) \le 1 + \delta_0$. Hence $\kappa_Z \le (1 + \delta_0)/(1 - \delta_0)$, which yields the stated bound. □

Thus, a small RIP constant may be unattainable even on rank-one directions, whereas the restricted-curvature condition remains valid for ill-conditioned designs. For a rank-exactly-r constrained minimizer $\widehat{\Theta}$, set $\kappa_\Theta := \frac{\sigma_1(\widehat{\Theta})}{\sigma_r(\widehat{\Theta})}$ and $K_{\mathrm{reg}} := \kappa_Z \kappa_\Theta$. Under the remaining target and initialization assumptions, Theorem 4.1 yields the fixed-depth Silver convergence guarantee in Subsection 5.3 with effective condition number K_{reg}.

Statistical recovery. Under the model and notation of Section 5.3, we derive the statistical error bound and its conditional Gaussian specialization. We use the operator identification and restricted-curvature constants established above. The fixed-full-block optimization rate remains

$$\mathcal{O}\left(K_{\mathrm{reg}}^{\theta}\left[1 + \log\frac{D_0}{\varepsilon_W}\right]\right),$$

as stated in Section 5.3; we therefore focus here on the statistical term.

For multivariate regression, the least-squares loss takes the general form $\frac{1}{2}\|\mathcal{A}_Z(\Theta) - y\|_2^2$. Since $Y^\top = \Theta_\star Z^\top + E^\top$, the normalized noise is $\varepsilon = n^{-1/2}\,\mathrm{vec}(E^\top)$, and because $\mathcal{A}_Z^*(\mathrm{vec}(M)) = n^{-1/2}MZ$, we have $\mathcal{A}_Z^*\varepsilon = n^{-1}E^\top Z$. Lemma C.2 with $\alpha = \lambda_{\min}(\widehat{\Sigma}_Z)$, therefore gives

$$\|\widehat{\Theta} - \Theta_\star\|_F \le \frac{2\sqrt{2r}}{\lambda_{\min}(\widehat{\Sigma}_Z)}\left\|\frac{1}{n}E^\top Z\right\|_{\mathrm{op}}.$$

To express this bound in terms of the noise variance, suppose that, conditionally on Z, the entries of E are independent $\mathcal{N}(0, \nu^2)$ variables. Then the rows of $n^{-1}E^\top Z$ are independent centered Gaussian vectors with covariance $\frac{\nu^2}{n}\widehat{\Sigma}_Z$. Equivalently, conditionally on Z,

$$\frac{1}{n}E^\top Z \overset{d}{=} \frac{\nu}{\sqrt{n}}G\widehat{\Sigma}_Z^{1/2},$$

where $G \in \mathbb{R}^{d_1 \times d_2}$ has independent standard Gaussian entries. For every $t \ge 0$, the Gaussian operator-norm bound $\|G\|_{\mathrm{op}} \le \sqrt{d_1} + \sqrt{d_2} + t$ holds with probability at least $1 - e^{-t^2/2}$. Consequently, with the same conditional probability,

$$\left\|\frac{1}{n}E^\top Z\right\|_{\mathrm{op}} \le \frac{\nu\sqrt{\lambda_{\max}(\widehat{\Sigma}_Z)}}{\sqrt{n}}\left(\sqrt{d_1} + \sqrt{d_2} + t\right),$$

and hence

$$\|\widehat{\Theta} - \Theta_\star\|_F \le \frac{2\sqrt{2r}\,\nu}{\sqrt{n}}\frac{\sqrt{\lambda_{\max}(\widehat{\Sigma}_Z)}}{\lambda_{\min}(\widehat{\Sigma}_Z)}\left(\sqrt{d_1} + \sqrt{d_2} + t\right).$$

Truncated-OLS initialization. We now justify the truncated-OLS warm start used in Section 5.3. Define $\Theta_{\mathrm{ols}} := \frac{1}{n} Y^\top Z \widehat{\Sigma}_Z^{-1}$, let Θ_0 be its best rank-r approximation, and let W_0 be the balanced factors of Θ_0.

Proposition C.9 (Truncated OLS enters the local basin)**.** *Let $\widehat{\Theta}$ be a rank-exactly-r constrained minimizer, suppose $\mathrm{rank}(\Theta_\star) \le r$ and write $\kappa_\Theta = \frac{\sigma_1(\widehat{\Theta})}{\sigma_r(\widehat{\Theta})}$. There exists a universal constant $c > 0$ such that, if*

$$\left\| \frac{1}{n} E^\top Z \right\|_{\mathrm{op}} \le c \frac{\alpha \sigma_r(\widehat{\Theta})}{\kappa \sqrt{r \kappa_\Theta}},$$

then the balanced factors W_0 satisfy the local initialization condition of Theorem 4.1.

Proof. The observation model $Y^\top = \Theta_\star Z^\top + E^\top$ gives $\Theta_{\mathrm{ols}} - \Theta_\star = \frac{1}{n} E^\top Z \widehat{\Sigma}_Z^{-1}$. Since $\|\widehat{\Sigma}_Z^{-1}\|_{\mathrm{op}} = \frac{1}{\alpha}$, we have

$$\|\Theta_{\mathrm{ols}} - \Theta_\star\|_{\mathrm{op}} \le \frac{1}{\alpha} \left\| \frac{1}{n} E^\top Z \right\|_{\mathrm{op}}.$$

Because $\mathrm{rank}(\Theta_\star) \le r$ and Θ_0 is the best rank-r approximation of Θ_{ols}, the Eckart–Young theorem gives $\|\Theta_0 - \Theta_{\mathrm{ols}}\|_{\mathrm{op}} \le \|\Theta_\star - \Theta_{\mathrm{ols}}\|_{\mathrm{op}}$ and the triangle inequality gives $\|\Theta_0 - \Theta_\star\|_{\mathrm{op}} \le 2\|\Theta_{\mathrm{ols}} - \Theta_\star\|_{\mathrm{op}}$. Using $\mathrm{rank}(\Theta_0 - \Theta_\star) \le 2r$, we obtain

$$\|\Theta_0 - \Theta_\star\|_F \le \frac{2\sqrt{2r}}{\alpha} \left\| \frac{1}{n} E^\top Z \right\|_{\mathrm{op}}.$$

On the other hand, the statistical error bound in Appendix C.1, specialized to multivariate regression, gives

$$\|\widehat{\Theta} - \Theta_\star\|_F \le \frac{2\sqrt{2r}}{\alpha} \left\| \frac{1}{n} E^\top Z \right\|_{\mathrm{op}}.$$

Thus, the triangle inequality and the assumed noise bound give

$$\|\Theta_0 - \widehat{\Theta}\|_F \le \frac{4\sqrt{2r}}{\alpha} \left\| \frac{1}{n} E^\top Z \right\|_{\mathrm{op}} \le \frac{4\sqrt{2}\, c\, \sigma_r(\widehat{\Theta})}{\kappa \sqrt{\kappa_\Theta}}.$$

For sufficiently small c, this error is at most $\sigma_r(\widehat{\Theta})/2$ and satisfies the matrix-error condition in Lemma A.2. Since $\mathrm{rank}(\Theta_0) \le r$, Weyl's inequality gives $\mathrm{rank}(\Theta_0) = r$. The lemma then places its balanced factors in the local basin of Theorem 4.1. □

Provided $\widehat{\Theta}$ also satisfies the normal-gradient condition in Theorem 4.1, the Silver iteration can be initialized from the balanced factors of Θ_0. The depth and normalization are computed once from this warm start using Proposition A.4 and remain fixed throughout optimization.

The truncated OLS matrix Θ_0 is used only as a convenient warm start and generally differs from the exact reduced-rank least-squares solution obtained after whitening (Izenman, 1975).

D A Scalar Counterexample for Fixed-Depth Silver Schedules

This appendix proves Theorem 4.2. We first derive the two-factor dynamics and an escape region. We then study the balanced inverse dynamics, construct escaping balanced intervals, and extend them to two-dimensional sets of initializations. Finally, we show that these sets have probability bounded away from zero under arbitrarily small Gaussian initialization.

D.1 Two-factor dynamics and an escape region

We begin with the full two-factor dynamics, without imposing the balancing condition $U = V$. In the scalar full-observation setting of Theorem 4.2,

$$g_\lambda(U, V) = \frac{1}{2}(UV - 1)^2 + \frac{1}{8}(U^2 - V^2)^2,$$

where $\lambda = 1/4$ and the fixed normalizer is $\overline{\beta} = 4$. Introduce the orthogonal coordinates

$$p := \frac{U+V}{\sqrt{2}}, \qquad z := \frac{U-V}{\sqrt{2}}, \qquad R^2 := p^2 + z^2 = U^2 + V^2, \qquad UV = \frac{p^2 - z^2}{2}.$$

The change of variables $(U, V) \mapsto (p, z)$ is orthogonal. Hence, if $U_0, V_0 \overset{\text{i.i.d.}}{\sim} N(0, \varepsilon_{\text{init}}^2)$, then $p_0, z_0 \overset{\text{i.i.d.}}{\sim} N(0, \varepsilon_{\text{init}}^2)$ as well.

For a Silver coefficient a, the factor update with normalizer $\overline{\beta} = 4$ becomes

$$p^+ = \left(1 + \frac{a}{4} - \frac{a}{8}R^2\right) p, \qquad z^+ = \left(1 - \frac{a}{4} - \frac{a}{8}R^2\right) z. \tag{36}$$

These are simply the original factor-gradient updates written in the (p, z) coordinates.

We first identify a region from which escape is irreversible.

Lemma D.1 (Irreversible escape region)**.** *Suppose that at some iteration t,*

$$R_t^2 \geq 16, \qquad \frac{|z_t|}{|p_t|} \leq \frac{1}{8}.$$

Then

$$U_{t+j}V_{t+j} \longrightarrow +\infty \qquad \text{as } j \to \infty.$$

Proof. Every Silver coefficient satisfies $a \geq \sqrt{2}$. If $R^2 \geq 16$, both multipliers in (36) are negative, and their absolute values are at least $7\sqrt{2}/4 - 1 > \sqrt{2}$. Consequently, $R_{t+j}^2 \geq 2^j R_t^2$ for $j \geq 0$.

Moreover, whenever $p \neq 0$,

$$\frac{|z^+/p^+|}{|z/p|} = 1 + \frac{4}{R^2 - 2 - 8/a} \leq 1 + \frac{8}{R^2},$$

where the last inequality follows from $2 + 8/a \leq 2 + 4\sqrt{2} < 8 \leq R^2/2$. Hence, using $1 + x \leq e^x$

$$\frac{|z_{t+j}|}{|p_{t+j}|} \leq \frac{1}{8}\prod_{\ell=0}^{j-1}\left(1 + \frac{8}{R_{t+\ell}^2}\right) \leq \frac{1}{8}\exp\left(\sum_{\ell=0}^{\infty}\frac{8}{R_{t+\ell}^2}\right) \leq \frac{1}{8}\exp\left(\frac{8}{R_t^2}\sum_{\ell=0}^{\infty}2^{-\ell}\right) \leq \frac{e}{8} < \frac{1}{2}.$$

Therefore, using $|z_{t+j}| < |p_{t+j}|/2$ and hence $R_{t+j}^2 = p_{t+j}^2 + z_{t+j}^2 \leq \frac{5}{4}p_{t+j}^2$,

$$U_{t+j}V_{t+j} = \frac{p_{t+j}^2 - z_{t+j}^2}{2} \geq \frac{3}{8}p_{t+j}^2 \geq \frac{3}{10}R_{t+j}^2 \longrightarrow +\infty \qquad \text{as } j \to \infty.$$

□

Thus it is enough to construct a set of initial conditions whose trajectories reach the region

$$R^2 > 16, \qquad |z| < \frac{|p|}{8}. \tag{37}$$

In the next subsection, we first construct such trajectories on the balanced invariant line $z = 0$, and then use them to build a two-dimensional set of escaping initializations.

D.2 Balanced inverse dynamics

We next restrict the dynamics to the balanced invariant line $z = 0$. By (36), if $z = 0$ at one iteration, then it remains zero at every subsequent iteration. On this line, $U = V = \frac{p}{\sqrt{2}}$. Define the squared factor amplitude

$$s := U^2 = V^2 = \frac{p^2}{2}.$$

Then a Silver coefficient a induces the scalar update

$$s^+ = T_a(s) := s\left[1 + \frac{a}{4}(1 - s)\right]^2. \tag{38}$$

To construct small escaping intervals, we next establish the required inverse properties of this map.

Lemma D.2 (Inverse branch of the scalar update)**.** *For every $a > 0$, the map T_a is strictly increasing and strictly concave on $[0, 1/3]$. For every $z \in [0, 1/3]$, there is therefore a unique $q_a(z) \in [0, 1/3]$ satisfying*

$$T_a(q_a(z)) = z.$$

The inverse map q_a is increasing and convex, with

$$q_a(0) = 0, \qquad 0 < q_a(z) < z$$

for $z > 0$. Moreover, if $a \geq \sqrt{2}$, then

$$q_a(z) \leq \eta^{-1} z, \qquad \eta := \left(1 + \frac{\sqrt{2}}{6}\right)^2 > 1.$$

Proof. Differentiating (38) gives

$$T_a'(s) = \left[1 + \frac{a}{4}(1-s)\right]\left[1 + \frac{a}{4}(1-3s)\right], \qquad T_a''(s) = -a - \frac{a^2}{4} + \frac{3a^2 s}{8}.$$

For $0 \leq s \leq 1/3$, both factors in $T_a'(s)$ are strictly positive, while $T_a''(s) \leq -a - \frac{a^2}{8} < 0$. Hence T_a is strictly increasing and strictly concave on $[0, 1/3]$. Also, $T_a(0) = 0$, $T_a(1/3) > 1/3$. Thus $q_a : [0, 1/3] \to [0, 1/3]$ is well defined. Since T_a is increasing and concave, q_a is also increasing and convex. Finally, if $a \geq \sqrt{2}$ and $0 \leq s \leq 1/3$, then

$$T_a(s) = s\left[1 + \frac{a}{4}(1-s)\right]^2 \geq s\left(1 + \frac{a}{6}\right)^2 \geq \eta s.$$

Applying this inequality at $s = q_a(z)$ yields $q_a(z) \leq \eta^{-1} z$. □

For a depth-k Silver block with coefficients $a_0, \ldots, a_{n_k - 1}$, define the forward block map B_k and its inverse branch J_k by

$$B_k := T_{a_{n_k-1}} \circ \cdots \circ T_{a_0}, \qquad J_k := q_{a_0} \circ \cdots \circ q_{a_{n_k-1}}.$$

Then

$$B_k(J_k(z)) = z, \qquad 0 < J_k(z) \leq \eta^{-n_k} z, \qquad 0 < z \leq \frac{1}{3}. \tag{39}$$

Moreover, every intermediate squared amplitude in the forward block starting from $J_k(z)$ lies in $[0, z]$.

Since each q_a is increasing and convex, J_k is also increasing and convex. The same is therefore true for any composition of the maps J_k. These inverse maps will be used in the next subsection to construct balanced escaping intervals arbitrarily close to the origin.

D.3 A ROBUST ESCAPING RECTANGLE

We now use the inverse maps from the previous subsection to construct two-dimensional sets of escaping initializations arbitrarily close to the origin.

Lemma D.3 (A robust escaping rectangle)**.** *There exist arbitrarily small $b > 0$ for which one can find $\delta > 0$ such that every initialization satisfying*

$$(p_0, z_0) \in [b, 3b/2] \times [-\delta, \delta]$$

obeys

$$U_t V_t \longrightarrow +\infty.$$

Proof. For $j \geq 0$, define $F_j := J_6^j \circ J_5$, $b_j := \sqrt{2F_j(1/8)}$, where J_6^j denotes j repeated compositions of J_6. By (39), we have $b_j \longrightarrow 0$ as $j \to \infty$. Since J_5 and J_6 are increasing and convex, so is F_j. Moreover, $F_j(0) = 0$. For any convex function F with $F(0) = 0$ and $0 < x < y$, $F(x) \leq \frac{x}{y} F(y)$. Applying this to F_j with $x = 1/8$ and $y = 1/3$ gives $F_j(1/3) \geq \frac{8}{3} F_j(1/8)$. Hence

$$\sqrt{2F_j(1/3)} \geq \sqrt{\frac{8}{3}}\, b_j > \frac{3}{2} b_j.$$

It follows that every balanced starting point satisfying $b_j \le p \le \frac{3}{2}b_j,\ z = 0$, has squared amplitude

$$s = \frac{p^2}{2} \in [F_j(1/8), F_j(1/3)].$$

Since $F_j = J_6^j \circ J_5$ is increasing, after j complete depth-6 blocks followed by the first $n_5 = 31$ steps of the next block, the squared amplitude lies in $s \in [1/8, 1/3]$.

The next coefficient is the central depth-6 coefficient $a_{31} = 1 + \rho^4 = 18 + 12\sqrt{2} > 32$. Since $T_{a_{31}}$ is increasing on $[0, 1/3]$,

$$T_{a_{31}}(s) \ge T_{a_{31}}(1/8) = \frac{1}{8}\left(1 + \frac{7a_{31}}{32}\right)^2 > 8.$$

Thus, since $z = 0$ implies $R^2 = p^2 = 2s$, after one additional step we have $R^2 > 16$. Therefore the trajectory enters the escape region (37), and Lemma D.1 implies $U_t V_t \longrightarrow +\infty$.

We now thicken this balanced interval into a two-dimensional set. Fix j and write $b := b_j$. The finite-step map after $jn_6 + 32$ iterations is continuous, while the escape region (37) is open. Since

$$[b, 3b/2] \times \{0\}$$

is compact and every point in this set is mapped into the escape region, there exists $\delta > 0$ such that every initialization in

$$[b, 3b/2] \times [-\delta, \delta]$$

also reaches the escape region after the same number of steps. Finally, since $b_j \to 0$, such values of b can be chosen arbitrarily small. □

D.4 Scaled rectangles and Gaussian probability

We next propagate the escaping rectangle from Lemma D.3 to smaller scales.

Lemma D.4 (Scaled divergent rectangles)**.** *Fix an escaping rectangle*

$$[b, 3b/2] \times [-\delta, \delta]$$

from Lemma D.3, with $b > 0$ chosen sufficiently small. Define

$$\Lambda := \prod_{i=0}^{n_6-1}\left(1 + \frac{a_i}{4}\right) > 1, \qquad \ell_m := b\Lambda^{-m}.$$

Then there exists an integer m_0 such that, for every $m \ge m_0$, every initialization in

$$\mathcal{B}_m := \left\{(p, z) : \frac{5}{4}\ell_m \le |p| \le \frac{11}{8}\ell_m, \quad |z| \le \ell_m\right\}$$

satisfies

$$U_t V_t \longrightarrow +\infty.$$

Proof. Set $h := 1 + \frac{\sqrt{2}}{4}$, $a_{\max} := \max_{0 \le i < n_6} a_i = 18 + 12\sqrt{2}$. Since Lemma D.3 allows b to be chosen arbitrarily small, choose it so that

$$2(11b/8)^2 \le \min\left\{\frac{4}{a_{\max}}, \frac{h^2 - 1}{4}\right\}. \tag{40}$$

When $R^2 \le 4/a_{\max}$, write $\chi := 1 - aR^2/8 \in [1/2, 1]$. Then (36) becomes

$$p^+ = (\chi + a/4)p, \qquad z^+ = (\chi - a/4)z.$$

Since $\chi > 0$ and $a > 0$, $\frac{|\chi - a/4|}{\chi + a/4} < 1$. The Silver block contains finitely many coefficients, and $\chi \in [1/2, 1]$. Hence there exists $0 < \zeta < 1$ such that

$$0 < p^+ \le \left(1 + \frac{a}{4}\right)p, \qquad \frac{|z^+|}{p^+} \le \zeta\frac{|z|}{p} \tag{41}$$

whenever $p > 0$ and $R^2 \le 4/a_{\max}$. Consider a point in $\mathcal{B}_m$ with $p_0 > 0$, and set $N := mn_6$, $L_t := \prod_{i=0}^{t-1}\left(1 + \frac{a_i}{4}\right)$, where the coefficients are extended periodically with period n_6. Then $L_N = \Lambda^m$. We claim that, for $0 \le t \le N$,

$$0 < p_t \le L_t p_0, \qquad \frac{|z_t|}{p_t} \le \zeta^t, \qquad R_t^2 \le 2(L_t p_0)^2 \le 2(11b/8)^2. \tag{42}$$

At $t = 0$, $\frac{|z_0|}{p_0} \le \frac{4}{5} \le 1$, $L_N p_0 \le \frac{11}{8} b$. Suppose (42) holds at time $t < N$. Then (40) gives $R_t^2 \le 4/a_{\max}$, so (41) applies. The first two bounds therefore hold at time $t+1$, and they imply the third bound as well. This proves (42) by induction. To obtain a lower bound on p_N, write

$$p_{t+1} = \left(1 + \frac{a_t}{4}\right) p_t (1 - \xi_t), \qquad \xi_t := \frac{a_t R_t^2}{8(1 + a_t/4)}.$$

Then $0 \le \xi_t \le \frac{R_t^2}{2}$. Since every Silver coefficient is at least $\sqrt{2}$, $\frac{L_t}{L_N} \le h^{-(N-t)}$. Therefore,

$$\sum_{t=0}^{N-1} \xi_t \le (L_N p_0)^2 \sum_{t=0}^{N-1} \left(\frac{L_t}{L_N}\right)^2 \le (11b/8)^2 \sum_{j=1}^{\infty} h^{-2j} = \frac{(11b/8)^2}{h^2 - 1} \le \frac{1}{8}.$$

Using $\prod_{t=0}^{N-1}(1-\xi_t) \ge 1 - \sum_{t=0}^{N-1} \xi_t$, together with $\frac{5}{4} b \le L_N p_0 \le \frac{11}{8} b$, we obtain $\frac{35}{32} b \le p_N \le \frac{11}{8} b$. Moreover, (42) gives $|z_N| \le \frac{11}{8} b \zeta^{mn_6}$. Choose m_0 sufficiently large so that $\frac{11}{8} b \zeta^{mn_6} \le \delta$ for every $m \ge m_0$. Since $[35b/32, 11b/8] \subset [b, 3b/2]$, we have

$$(p_N, z_N) \in [b, 3b/2] \times [-\delta, \delta].$$

Lemma D.3 therefore implies that the trajectory diverges.

Finally, the update (36) is symmetric under $p \mapsto -p$, while $UV = (p^2 - z^2)/2$ is unchanged. Hence the same conclusion holds when $p_0 < 0$. Thus every initialization in $\mathcal{B}_m$ diverges for every $m \ge m_0$. □

Proof of Theorem 4.2. Let b, Λ, and m_0 be as in Lemma D.4, and set $\varepsilon_0 := b\Lambda^{-m_0}$. For any $0 < \varepsilon_{\text{init}} \le \varepsilon_0$, choose $m \ge m_0$ such that $\ell_m \le \varepsilon_{\text{init}} < \Lambda \ell_m$ and define $r_0 := \frac{\ell_m}{\varepsilon_{\text{init}}} \in [1/\Lambda, 1]$.

By the orthogonal change of variables introduced in Section D.1, p_0 and z_0 are independent $N(0, \varepsilon_{\text{init}}^2)$ random variables. Therefore, if Φ denotes the standard normal distribution function,

$$\Pr\big((p_0, z_0) \in \mathcal{B}_m\big) = 2\left[\Phi\left(\frac{11 r_0}{8}\right) - \Phi\left(\frac{5 r_0}{4}\right)\right][2\Phi(r_0) - 1].$$

The right-hand side is continuous and strictly positive on $[1/\Lambda, 1]$. Hence

$$c := \min_{r \in [1/\Lambda, 1]} 2\left[\Phi\left(\frac{11 r}{8}\right) - \Phi\left(\frac{5 r}{4}\right)\right][2\Phi(r) - 1] > 0.$$

By Lemma D.4, every point in $\mathcal{B}_m$ is a divergent initialization. Therefore

$$\Pr(U_t V_t \longrightarrow +\infty) \ge c, \qquad 0 < \varepsilon_{\text{init}} \le \varepsilon_0.$$

□

E Experimental Details and Additional Results

Across experiments, Constant GD and Silver GD minimize the balanced factor objective g_λ defined in Section 3.1, starting from identical initial factors within each instance. Writing $C(U, V) = U^\top U - V^\top V$, the gradients are

$$\nabla_U g_\lambda = \nabla f(UV^\top)V + 2\lambda U C(U, V), \qquad \nabla_V g_\lambda = \nabla f(UV^\top)^\top U - 2\lambda V C(U, V).$$

Both factors are updated simultaneously using gradients evaluated at the current iterate. With the depth-k schedule $\alpha^{(k)}$ and block length $n_k = 2^k - 1$ defined in Section 3.3, Constant GD and

Silver GD use stepsizes $1/\overline{\beta}$ and $\alpha^{(k)}_{t \bmod n_k}/\overline{\beta}$, respectively; for the initialization-based choice in Section 4.3, $\overline{\beta}$ is replaced by $\overline{\beta}_0$. At each block boundary, only the schedule index is reset: optimization continues from the current factors with fixed depth, normalizer, and balancing penalty. Iteration counts measure joint gradient updates, excluding initialization, curvature probes, and terminal diagnostic evaluations. Numerical histories and settings are saved separately from figures, allowing plots to be regenerated without rerunning optimization.

E.1 Experiment 1: conditioning dependence of iteration complexity

Problem generation. For the experiments in Section 6.1, set $d_1 = 64$, $d_2 = 48$, and $r = 4$. We generate orthonormal matrices $P \in \mathbb{R}^{d_1 \times r}$ and $Q \in \mathbb{R}^{d_2 \times r}$ by reduced QR factorization of independent standard Gaussian matrices, with diagonal signs standardized. The noiseless target is $X_\star = \widehat{X} = P\Sigma Q^\top$, where $\Sigma_{jj} = \kappa_X^{-(j-1)/(r-1)}$ for $j = 1, \dots, r$. The sensing model is $\mathcal{A}(X) = \mathrm{vec}(XS^{1/2})$, with

$$S = O\,\mathrm{diag}(s_1, \dots, s_{d_2})O^\top, \qquad s_j = \kappa^{-(j-1)/(d_2-1)},$$

where O is an orthogonal matrix obtained by Gaussian QR factorization. Thus $\sigma_1(\widehat{X}) = \beta = 1$, $\alpha = 1/\kappa$, and $\lambda = \alpha/4$. The two sweeps vary either κ with $\kappa_X = 4$, or κ_X with $\kappa = 4$, over $\{2, 4, 8, 16, 32, 64, 128\}$. Within each sweep, the singular-vector and covariance-eigenvector orientations remain fixed; only the specified spectrum changes.

Initialization and schedule. Let $\widehat{U} = P\Sigma^{1/2}$ and $\widehat{V} = Q\Sigma^{1/2}$. We draw a Gaussian factor perturbation (A, B), normalize it so that $\|A\|_F^2 + \|B\|_F^2 = 1$, and set

$$U_0 = \widehat{U} + \frac{2}{K}A, \qquad V_0 = \widehat{V} + \frac{2}{K}B, \qquad K = \kappa\kappa_X.$$

The perturbation direction is shared across cases within each sweep. The multiplier 2 is an experimental choice, not a certified value of the small universal constant in Theorem 4.1. We set $\overline{\beta} = 4\beta\sigma_1(\widehat{X}) = 4$ and choose the smallest $k \geq 1$ satisfying $\rho^k \geq 192K$. For $K = 8, 16, 32, 64, 128, 256, 512$, the depths are $9, 10, 10, 11, 12, 13, 14$, respectively. Both methods receive at most $20n_k$ updates per case.

Stopping and fitting. We record every update and stop at the first iteration T satisfying

$$\frac{\|U_T V_T^\top - \widehat{X}\|_F}{\|U_0 V_0^\top - \widehat{X}\|_F} \leq 10^{-10}.$$

This criterion is checked at every update, including intermediate Silver iterates rather than only block endpoints. Runs stop on nonfinite values or factor entries exceeding 10^{50} in absolute value. We fit $\log T = a + p \log K$ by ordinary least squares over successful cases, where a and p denote the fitted intercept and slope, without substituting the update budget for unreached thresholds. All seven cases succeeded for both methods in each sweep. The fitted slopes are 0.998 and 0.779 for Constant GD and Silver GD when varying κ, and 0.983 and 0.770 when varying κ_X.

E.2 Experiment 2: isotropic and anisotropic Gaussian sensing

Data and initialization. For Section 6.2, set $d_1 = 64$, $d_2 = 48$, $r = 4$, and $m = 1000$. The true target $X_\star$ has random Gaussian-QR singular vectors and singular values geometrically spaced from 1 to $1/16$. We generate $y = \mathcal{A}(X_\star) + \varepsilon$ with independent $\varepsilon_i \sim \mathcal{N}(0, \nu^2/m)$ and $\nu = 0.01$. Here the prescribed singular values belong to $X_\star$; the constrained empirical minimizer $\widehat{X}$ generally differs from $X_\star$ because of noise.

Let $\mathcal{P}_r$ denote a best rank-r approximation obtained by truncated SVD. In the isotropic model, the entries of each A_i are independent $\mathcal{N}(0, 1/m)$ variables, and $X_0 = \mathcal{P}_r(\mathcal{A}^* y)$. In the anisotropic model, $A_i = G_i S^{1/2}$, where the entries of G_i have the same isotropic distribution and S has a random Gaussian-QR eigenbasis with eigenvalues geometrically spaced from 1 to $1/4$. Using the operator $\mathcal{G}$ from Section 5.2, we set

$$X_0 = \mathcal{P}_r(\mathcal{G}^* y)S^{-1/2}.$$

In both models, an SVD $X_0 = P_0\Sigma_0 Q_0^\top$ gives the balanced factors $U_0 = P_0\Sigma_0^{1/2}$ and $V_0 = Q_0\Sigma_0^{1/2}$. NumPy `SeedSequence` provides separate streams for sensing, targets, noise, and probes. The anisotropic run preserves the stream indices for sensing, targets, and noise and uses an additional covariance stream, matching these random components across the two models.

Empirical curvature. Each probe has the form $H = P\,\mathrm{diag}(s/\|s\|_2)Q^\top$, where P and Q are independent Gaussian-QR orthonormal factors with the prescribed number of columns; thus $\|H\|_F = 1$. The entries of s cycle through three choices: all ones, geometric values from 1 to 0.01, and independent values $\exp(z_j)$ with $z_j \sim \mathrm{Unif}[-4, 0]$. For isotropic sensing, 1000 rank-$2r$ probes determine

$$\widehat{\alpha} = \min_H \|\mathcal{A}(H)\|_2^2, \qquad \widehat{\beta} = \max_H \|\mathcal{A}(H)\|_2^2,$$

where the extrema are taken only over the sampled probes. An independent set of 1000 rank-$3r$ probes supplies diagnostics without changing the parameters. For anisotropic sensing, 1000 rank-$3r$ probes of $\mathcal{G}$ determine

$$\widehat{\delta} = \max_H \big|\|\mathcal{G}(H)\|_2^2 - 1\big|, \qquad \widehat{\alpha} = (1-\widehat{\delta})\lambda_{\min}(S), \qquad \widehat{\beta} = (1+\widehat{\delta})\lambda_{\max}(S).$$

Here $\lambda_{\min}(S) = 1/4$ and $\lambda_{\max}(S) = 1$. Probes are evaluated in batches of 32. The hats distinguish these sampled estimates from the certified curvature and RIP constants in the theory. Neither the curvature estimates nor the entry of the spectral initializations into the local basin is certified for these numerical instances.

Parameter, budget, and error metric. Using the initialization-based notation of Section 4.3, set

$$\widehat{\kappa}_X = \frac{2\sigma_1(X_0)}{\sigma_r(X_0)}, \qquad \overline{\beta}_0 = 4\widehat{\beta}\sigma_1(X_0), \qquad \lambda = \widehat{\alpha}/4.$$

The reference depth is the smallest $k_0 \geq 1$ satisfying $\rho^{k_0} \geq 192\frac{\widehat{\beta}}{\widehat{\alpha}}\widehat{\kappa}_X$. Thus the warm-start formulas use the sampled curvature estimates in place of α and β. We compare Constant GD with Silver depths $\lfloor k_0/3 \rfloor$, $\lfloor 2k_0/3 \rfloor$, and k_0; the anisotropic implementation clamps depths to at least one and removes duplicates. The realized depths are $(3, 6, 9)$ for isotropic sensing and $(3, 6, 10)$ for anisotropic sensing. All methods receive the same budget $4n_{k_0}$: 2044 and 4092 updates, respectively. The plotted quantity is relative recovery error, $\frac{\|U_t V_t^\top - X_\star\|_F}{\|X_\star\|_F}$. Histories retain the initial state, every tenth update, all Silver block endpoints, and terminal states. Tolerance stopping is disabled for these noisy runs, and all trajectories complete their budgets. The divergence guard stops a run if values become nonfinite or relative recovery error exceeds 10^8 times the larger of one and its initial value. For optional noiseless runs, termination requires a 10^{-12} reduction in matrix error relative to its initial value; isotropic Silver checks termination at block endpoints, whereas anisotropic Silver can stop at any qualifying update.

E.3 Experiment 3: low-rank multivariate regression

Model and data. This additional experiment evaluates the regression model of Section 5.3, which is not plotted in the main text. We generate $Y = Z\Theta_\star^\top + E$ with $n = 384$ observations, $d_1 = 64$, $d_2 = 48$, and $r = 4$. The true coefficient matrix $\Theta_\star \in \mathbb{R}^{d_1 \times d_2}$ has rank r, random Gaussian-QR singular vectors, and singular values geometrically spaced from 1 to $1/4$. Rows of $Z \in \mathbb{R}^{n \times d_2}$ are independent Gaussian vectors with population covariance C_Z. Its random orthogonal eigenbasis remains fixed, while its eigenvalues range geometrically from 1 to $1/q$, for $q \in \{4, 16, 64\}$. Here $q = \mathrm{cond}(C_Z)$ denotes population design conditioning, distinct from the empirical quantity $\kappa = \mathrm{cond}(\widehat{\Sigma}_Z)$ used in the theory. Noise entries are independent $\mathcal{N}(0, \nu^2)$ variables with $\nu = 0.01$, as in Appendix C.4; unlike Experiment 2, their variance is not divided by the measurement count. Separate streams generate the target, design, and noise, and their base Gaussian draws are reused across the three covariance conditions.

Empirical minimizer and initialization. Following Section 5.3, write

$$f(\Theta) = \frac{1}{2n}\|\Theta Z^\top - Y^\top\|_F^2, \qquad \widehat{\Sigma}_Z = \frac{1}{n}Z^\top Z, \qquad \Theta_{\mathrm{ols}} = \frac{1}{n}Y^\top Z\widehat{\Sigma}_Z^{-1}.$$

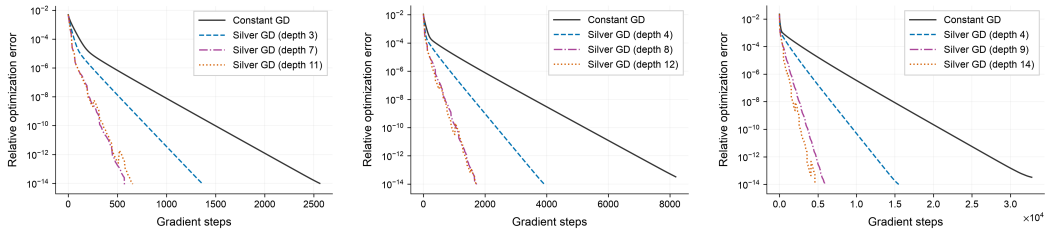

Figure 3: **Low-rank multivariate regression.** Relative optimization error $\|U_t V_t^\top - \widehat{\Theta}\|_F / \|\widehat{\Theta}\|_F$ versus gradient evaluations for population design condition numbers $q = 4, 16, 64$ (left to right). All instances use $n = 384$, $d_1 = 64$, $d_2 = 48$, $r = 4$, and noise level $\nu = 0.01$. Within each panel, Constant GD and three Silver depths share the same initialization and stepsize normalization. Curves terminate at the optimization tolerance 10^{-14} or the common update budget.

We compute Θ_{ols} using a reduced QR factorization of Z, avoiding an explicit inverse of $Z^\top Z$. With $\mathcal{P}_r$ denoting the rank-r truncation operator defined above, the initialization and rank-constrained empirical minimizer are

$$\Theta_0 = \mathcal{P}_r(\Theta_{\mathrm{ols}}), \qquad \widehat{\Theta} = \mathcal{P}_r(\Theta_{\mathrm{ols}} \widehat{\Sigma}_Z^{1/2}) \widehat{\Sigma}_Z^{-1/2}.$$

These matrices generally differ because the loss is weighted by the empirical design covariance. All methods start from the balanced SVD factors of Θ_0, as in Appendix C.4. The reference $\widehat{\Theta}$ is computed directly by the weighted truncation formula rather than approximated by a long optimization run. The prescribed target condition number 4 concerns $\Theta_\star$, whereas κ_Θ in the theory is defined using $\widehat{\Theta}$.

Parameters and stopping. As in Appendix C.4, use

$$\alpha = \lambda_{\min}(\widehat{\Sigma}_Z), \qquad \beta = \lambda_{\max}(\widehat{\Sigma}_Z), \qquad \kappa = \frac{\beta}{\alpha}.$$

We set $\lambda = \alpha/4$ and $\overline{\beta}_0 = 4\beta\sigma_1(\Theta_0)$. Applying the warm-start rule to Θ_0, choose the smallest $k_0 \geq 1$ satisfying

$$\rho^{k_0} \geq 192\kappa \frac{2\sigma_1(\Theta_0)}{\sigma_r(\Theta_0)}.$$

The three Silver depths are $\lfloor k_0/3 \rfloor$, $\lfloor 2k_0/3 \rfloor$, and k_0. For $q = 4, 16, 64$, the realized empirical condition numbers κ are approximately 7.32, 23.61, and 87.74, and the depth triples are $(3, 7, 11)$, $(4, 8, 12)$, and $(4, 9, 14)$, respectively. The common per-case budgets are $2n_{k_0}$, namely 4094, 8190, and 32766 updates.

We plot relative optimization error $\frac{\|U_t V_t^\top - \widehat{\Theta}\|_F}{\|\widehat{\Theta}\|_F}$ and stop at its first value at most 10^{-14}. This criterion is checked at every update for all methods, including intermediate Silver iterates. Histories retain the initial state, every tenth update, all Silver block endpoints, and terminal states. The divergence guard uses relative recovery error with respect to $\Theta_\star$, with the same rule as in Appendix E.2. The histories also store this recovery error, which differs from the plotted optimization error. For $q = 4, 16, 64$, the relative statistical errors $\frac{\|\widehat{\Theta} - \Theta_\star\|_F}{\|\Theta_\star\|_F}$ are approximately 0.0114, 0.0163, and 0.0246, respectively.

Additional results. Figure 3 shows that all tested Silver depths reach the optimization tolerance sooner than Constant GD in the instances. Constant GD reaches the tolerance in 2560 updates for $q = 4$ and exhausts the budgets for $q = 16, 64$, ending at relative optimization errors approximately 3.2×10^{-14} and 3.3×10^{-14}. For $q = 64$, Silver depths 4, 9, and 14 reach the tolerance in 15499, 5877, and 4614 updates. The largest depth is not uniformly fastest: for $q = 4$, depth 7 reaches the tolerance in 572 updates, compared with 656 for depth 11. The very small stopping tolerance makes exact first-hit counts sensitive to floating-point arithmetic; these single-instance comparisons illustrate optimization behavior rather than establish a universal ordering of depths.