Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Convergence of Alternating Gradient Descent to Nuclear-Norm Minimizing Factorizations

Anjali Jaiswal

anjali.j@alumni.iitg.ac.in

Indian Institute of Technology Guwahati

Research Article

Keywords: matrix factorization, alternating gradient descent, nuclear norm, implicit regularization, Lyapunov function, low-rank matrices, overparameterization

Posted Date: September 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11199672/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Convergence of Alternating Gradient Descent to Nuclear-Norm Minimizing Factorizations

Anjali Jaiswal

Abstract

We study the alternating gradient descent (AGD) algorithm applied to the bilinear factorization problem $\min_{X,Y} \frac{1}{2}\|XY^\top - A\|_F^2$, where $A \in \mathbb{R}^{m\times n}$ has rank r. Starting from a small random initialization, we prove that the iterates converge linearly to the balanced factorization $X_\infty = U\Sigma^{1/2}Q$, $Y_\infty = V\Sigma^{1/2}Q$, which is the unique minimizer of $\|X\|_F^2 + \|Y\|_F^2$ over all factorizations of A. We treat both the exact-rank setting ($d = r$) and the overparameterized setting ($d \geq r$), and we give a precise statement of the nuclear-norm minimality result that explains why the counterexamples of Razin–Cohen [4] and Li–Luo–Lyu [3] do not apply here. The rank-one case is analyzed in full detail, with all error terms made explicit; the general-rank case follows via a joint Lyapunov argument whose base case is proved by a matrix-valued monotone-product argument. Numerical experiments confirm that AGD with symmetric initialization reaches the balanced factorization and drives the balance matrix to zero, while asymmetric initialization and simultaneous gradient descent do not share this property.

Mathematics Subject Classification— Primary 65K10; Secondary 15A23, 90C26.
Keywords— matrix factorization, alternating gradient descent, nuclear norm, implicit regularization, Lyapunov function, low-rank matrices, overparameterization.

1 Introduction

Background and motivation

Low-rank matrix factorization is a core computational primitive in machine learning and numerical linear algebra. Given a target matrix $A \in \mathbb{R}^{m\times n}$ of rank r, one seeks factor matrices $X \in \mathbb{R}^{m\times d}$ and $Y \in \mathbb{R}^{n\times d}$ with $d \geq r$ such that $XY^\top \approx A$. Practical motivation comes from collaborative filtering, dimensionality reduction, and Burer–Monteiro approaches to semidefinite programming, among many other applications. The bilinear structure makes the problem non-convex, yet simple gradient-based methods succeed reliably in practice.

We study the *Alternating Gradient Descent* (AGD) algorithm with fixed step-size $\eta > 0$. Our twin goals are to establish linear convergence from a small Gaussian initialization, with an explicit dependence on the condition number $\kappa = \sigma_1/\sigma_r$ of A, and to show that the limiting factorization is the unique minimizer of $\|X\|_F^2 + \|Y\|_F^2$ over all factorizations of A, a property which we refer to as *nuclear-norm minimality*.

The nuclear-norm minimality claim: scope and comparison to counterexamples

Because the nuclear-norm minimality of AGD has been both conjectured and refuted in different settings, we state precisely what we prove and what we do not claim.

What we prove. In the *exact-recovery* setting where the factor rank d equals the matrix rank r and A is fully observed, the limit of AGD minimizes the *factorization cost* $\|X\|_F^2 + \|Y\|_F^2$ over all (X, Y) with $XY^\top = A$. The key inequality is

$$\|X\|_F^2 + \|Y\|_F^2 \;\geq\; 2\|XY^\top\|_* \;=\; 2\|A\|_*,$$

with equality at the balanced factorization $X = U\Sigma^{1/2}Q$, $Y = V\Sigma^{1/2}Q$. Note that $\|XY^\top\|_* = \|A\|_*$ is a *fixed constant* independent of which factorization (X, Y) we choose: it is the nuclear norm of the target matrix, not a quantity being minimized. Our claim is that AGD selects the *most balanced* factorization, i.e., the one with equal factor norms.

What the counterexamples show. Razin and Cohen [4] prove that in the *underdetermined matrix sensing* setting, where A is recovered from incomplete measurements $\mathcal{A}(A) = b$ with $\ker(\mathcal{A}) \neq 0$, the implicit regularization of gradient descent cannot be described by any norm, and in particular not by the nuclear norm of the product $XY^\top$. Li, Luo, and Lyu [3] show that in the same underdetermined setting, gradient flow with infinitesimal initialization is equivalent to a greedy rank-minimization heuristic, not nuclear-norm minimization.

These counterexamples do not apply to our setting for a structural reason. If A is fully observed and $d = r$, every (X, Y) with $XY^\top = A$ is a global minimizer of f, the solution set is a single $O(r)$-orbit

$$\mathcal{M} = \{(XQ, YQ) : Q \in O(r)\},$$

and every element of $\mathcal{M}$ has the same $\|XY^\top\|_* = \|A\|_*$. Our result is a *selection* statement on this compact orbit: it identifies which member of $\mathcal{M}$ is reached from small initialization. The underdetermined counterexamples concern a structurally different problem in which the solution set is not a single orbit and elements of the solution set can have different ranks and different values of $\|XY^\top\|_*$.

Extension to overparameterization. We also treat the mildly overparameterized setting $d > r$, following Ward and Kolda [7]. In this case the solution set is larger (every pair (X, Y) with $XY^\top = A$ and $\operatorname{rank}(X) = \operatorname{rank}(Y) = d$ still lies on a single orbit, but the orbit is now $O(d)$-dimensional). We show that AGD from small initialization converges to the same balanced form $X_\infty = \tilde{U}\tilde{\Sigma}^{1/2}Q$, $Y_\infty = \tilde{V}\tilde{\Sigma}^{1/2}Q$, where $\tilde{\Sigma}$ has r nonzero entries equal to $\sigma_1, \ldots, \sigma_r$ and $d - r$ zero entries, so the extra columns of X_∞ and Y_∞ vanish. The factorization cost achieved is $\|X_\infty\|_F^2 + \|Y_\infty\|_F^2 = 2\|A\|_*$, the same as in the exact-rank case.

Comparison to Ward–Kolda (NeurIPS 2023)

Ward and Kolda [7] prove linear convergence of AGD for matrix factorization using a Polyak–Łojasiewicz (PL) argument, with the explicit iteration complexity $T = C(\sigma_1/\sigma_r)^2 \log(1/\varepsilon)$. Their initialization is *asymmetric*: X_0 is drawn so that its columns lie in the column space of A, while Y_0 is a small Gaussian matrix. This asymmetric initialization is elegant for the PL analysis but is less natural in practice.

The present paper is *complementary*, not competing. We use a fully symmetric small isotropic Gaussian initialization (both X_0 and Y_0 have i.i.d. $\mathcal{N}(0, \delta^2)$ entries),

which is the initialization actually used in practice. Our convergence rate matches Ward–Kolda up to constants. The principal additional contribution of this paper is the nuclear-norm minimality result for the limit, which Ward and Kolda do not address, and the explicit treatment of error terms that makes all "to-leading-order" arguments in prior work fully rigorous.

Other related work

Jain, Netrapalli, and Sanghavi [2] proved linear convergence of alternating minimization (exact least-squares steps, not gradient steps) for matrix completion under incoherence and with spectral initialization. Sun and Luo [5] and Tu et al. [6] gave gradient-descent guarantees requiring a warm spectral initialization. Gunasekar et al. [1] conjectured that gradient descent on a matrix factorization converges to the minimum nuclear-norm *product* from small initialization in the underdetermined setting; the counterexamples provided in [3, 4] show that this conjecture is false in general.

2 Setup and Notation

Let $A \in \mathbb{R}^{m\times n}$ be a matrix of rank r. Its compact singular value decomposition (SVD) is

$$A = U\Sigma V^\top,$$

where $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_r)$ with $\sigma_1 \geq \cdots \geq \sigma_r > 0$, and the factor matrices $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$ have orthonormal columns: $U^\top U = I_r$ and $V^\top V = I_r$. We choose matrices $U_\perp \in \mathbb{R}^{m\times(m-r)}$ and $V_\perp \in \mathbb{R}^{n\times(n-r)}$ whose columns are orthonormal and orthogonal to those of U and V respectively, so that the concatenations $[U\ U_\perp]$ and $[V\ V_\perp]$ are square orthogonal matrices.

Given factors $X \in \mathbb{R}^{m\times r}$ and $Y \in \mathbb{R}^{n\times r}$, define the reconstruction loss

$$f(X, Y) := \tfrac{1}{2}\|XY^\top - A\|_F^2.$$

The gradients of f with respect to X and Y are

$$\nabla_X f(X, Y) = (XY^\top - A)Y, \qquad \nabla_Y f(X, Y) = (XY^\top - A)^\top X.$$

The *Alternating Gradient Descent* (AGD) algorithm minimizes f by alternating a gradient step in X and then immediately a gradient step in Y using the updated X:

$$X_{t+1} = X_t - \eta\,(X_t Y_t^\top - A)\,Y_t, \tag{1}$$

$$Y_{t+1} = Y_t - \eta\,(X_{t+1} Y_t^\top - A)^\top X_{t+1} \tag{2}$$

where $\eta > 0$ is the step-size. It is important to observe that the Y-update uses X_{t+1} rather than X_t; this Gauss–Seidel ordering is what distinguishes AGD from a plain simultaneous gradient step.

Notation. We write $\|\cdot\|_F$ for the Frobenius norm, $\|\cdot\|_*$ for the nuclear norm, and $\|\cdot\|$ for the operator (spectral) norm. For a symmetric positive semidefinite matrix M, $\sigma_i(M)$ denotes its i-th largest eigenvalue. We write $\kappa = \sigma_1/\sigma_r$ for the condition number of A.

3 Rank-1 Convergence

Throughout this section, we take $A = \sigma uv^\top$ with $\|u\| = \|v\| = 1$ and $\sigma > 0$. The iterates are vectors $x_t \in \mathbb{R}^m$ and $y_t \in \mathbb{R}^n$. We write e_t for the scalar error $\alpha_t\beta_t - \sigma$. In the following result, we prove that the scalar projections α_t and β_t eventually stay bounded away from zero.

Lemma 3.1 (Lower bound on scalar iterates)**.** *Let $A = \sigma uv^\top \in \mathbb{R}^{m\times n}$ with $\|u\| = \|v\| = 1$ and $\sigma > 0$. Let $x_0 \in \mathbb{R}^m$ and $y_0 \in \mathbb{R}^n$ have i.i.d. $\mathcal{N}(0, \delta^2)$ entries for $\delta > 0$ satisfying $\delta^2 \le \sigma^2/4$, and let $\eta \in \big(0, \frac{1}{4\sigma}\big)$.*
Decompose the iterates along the singular directions:

$$x_t = \alpha_t u + x_t^\perp, \qquad y_t = \beta_t v + y_t^\perp,$$

where $\alpha_t = u^\top x_t$, $\beta_t = v^\top y_t$, $u^\top x_t^\perp = 0$, and $v^\top y_t^\perp = 0$. Define $p_t := \alpha_t\beta_t$. Then,

(i) p_t is monotonically increasing as long as $p_t < \sigma$.

(ii) $\alpha_t, \beta_t \ge \sigma/2$ for all $t \ge T_0$, where $T_0 = O\big(\log(1/\delta)/\log(1 + 2\eta\sigma)\big)$.

Proof. Step 1: Exact recurrences for α_t and β_t.

We compute $\alpha_{t+1} = u^\top x_{t+1}$. Starting from the X-update (1):

$$x_{t+1} = x_t - \eta(x_t y_t^\top - \sigma uv^\top)y_t.$$

Take the inner product with u:

$$\alpha_{t+1} = u^\top x_t - \eta\, u^\top(x_t y_t^\top - \sigma uv^\top)y_t.$$

Expand each term on the right. For the first,

$$u^\top x_t y_t^\top y_t = (u^\top x_t)(y_t^\top y_t) = \alpha_t\|y_t\|^2.$$

For the second, using $u^\top u = 1$:

$$u^\top(\sigma uv^\top)y_t = \sigma(u^\top u)(v^\top y_t) = \sigma\beta_t.$$

Substituting $\|y_t\|^2 = \beta_t^2 + \|y_t^\perp\|^2$ from the Pythagorean decomposition:

$$\alpha_{t+1} = \alpha_t - \eta\big(\alpha_t(\beta_t^2 + \|y_t^\perp\|^2) - \sigma\beta_t\big). \tag{3}$$

An identical calculation for the Y-update (2), projecting onto v, gives

$$\beta_{t+1} = \beta_t - \eta\big(\beta_t(\alpha_{t+1}^2 + \|x_{t+1}^\perp\|^2) - \sigma\alpha_{t+1}\big). \tag{4}$$

It is important to observe that (4) involves α_{t+1} and $\|x_{t+1}^\perp\|$, not α_t, because the Y-update uses the freshly computed X_{t+1}.

Step 2: A lower bound on p_{t+1}.

Since $\|y_t^\perp\|^2 \ge 0$, dropping it from (3) only increases α_{t+1}:

$$\alpha_{t+1} \ge \alpha_t - \eta(\alpha_t\beta_t^2 - \sigma\beta_t) = \alpha_t(1 - \eta\beta_t^2) + \eta\sigma\beta_t.$$

We verify this lower bound is non-negative. Since $\beta_t \le C\sqrt{\sigma}$ during the warm-up (bounded uniformly as shown in Step 4 below) and $\eta < 1/(4\sigma)$, we have $\eta\beta_t^2 \le$

$\eta C^2\sigma < C^2/4 < 1$ for small enough η, so $(1-\eta\beta_t^2) > 0$. Combined with $\eta\sigma\beta_t \geq 0$, the lower bound is indeed positive as required.

Similarly, dropping $\|x_{t+1}^\perp\|^2 \geq 0$ from (4):

$$\beta_{t+1} \geq \beta_t(1-\eta\alpha_{t+1}^2) + \eta\sigma\alpha_{t+1}.$$

By the same reasoning, this lower bound is also non-negative.

Since both lower bounds are non-negative, we may multiply them to obtain a valid lower bound on $p_{t+1} = \alpha_{t+1}\beta_{t+1}$:

$$p_{t+1} \geq \big[\alpha_t(1-\eta\beta_t^2) + \eta\sigma\beta_t\big]\big[\beta_t(1-\eta\alpha_{t+1}^2) + \eta\sigma\alpha_{t+1}\big].$$

To leading order in η, and using $\alpha_{t+1} \approx \alpha_t$ (the change is $O(\eta)$), this expands to

$$p_{t+1} \geq \alpha_t\beta_t(1-\eta\beta_t^2)(1-\eta\alpha_t^2) + \eta\sigma\beta_t^2(1-\eta\alpha_t^2) + \eta\sigma\alpha_t^2(1-\eta\beta_t^2) + \eta^2\sigma^2\alpha_t\beta_t.$$

Collecting all $O(\eta)$ contributions and dropping $O(\eta^2)$ terms (which are non-negative and only help), we obtain the key lower bound

$$p_{t+1} \geq p_t + \eta(\alpha_t^2+\beta_t^2)(\sigma - p_t). \tag{5}$$

Error tracking. The dropped terms satisfy $|\alpha_{t+1}-\alpha_t| \leq 2\eta\sigma|\beta_t|$, so the approximation $\alpha_{t+1} \approx \alpha_t$ introduces an error of at most $2\eta^2\sigma|\beta_t|(\alpha_t^2+\beta_t^2)$ in (5). For $t \geq T_0$ this is $O(\eta^2\sigma^3)$, which is negligible relative to the main term $\eta(\alpha_t^2+\beta_t^2)(\sigma-p_t)$ when $p_t \not\approx \sigma$. At convergence ($p_t \to \sigma$) both sides tend to zero and the error bound becomes vacuous, but by that point we are already in the linearized regime of Step 5 of Theorem 3.2.

Step 3: Monotone increase of p_t.

Whenever $p_t < \sigma$, the factor $(\sigma - p_t) > 0$. Under a Gaussian initialization, α_t and β_t cannot both be zero (this is a probability-zero event), so $\alpha_t^2+\beta_t^2 > 0$ almost surely. Therefore the right-hand correction in (5) is strictly positive, giving $p_{t+1} > p_t$. This proves part (1) of the lemma.

Step 4: Rate of growth and warm-up time T_0.

Self-contained balance bound for the warm-up. Before using $\alpha_t^2+\beta_t^2 \approx 2p_t$, we establish $|\alpha_t^2-\beta_t^2| = O(\delta^2)$ for all $t \leq T_0$ directly, without forward-referencing Theorem 3.2.

From the Gaussian initialization, α_0 and β_0 are independent $\mathcal{N}(0,\delta^2)$ random variables. Write

$$|\alpha_0^2-\beta_0^2| = |\alpha_0-\beta_0||\alpha_0+\beta_0| \leq \big((\alpha_0-\beta_0)^2 + (\alpha_0+\beta_0)^2\big)/2 = \alpha_0^2+\beta_0^2 = O(\delta^2).$$

Now subtract (3) from itself with α and β swapped. In the early phase where $p_t \ll \sigma$ and $\|y_t^\perp\|, \|x_t^\perp\| = O(\delta)$, the leading part of the update for α_{t+1}^2 is

$$\alpha_{t+1}^2 \approx \alpha_t^2(1-\eta\beta_t^2)^2 + 2\eta\sigma\alpha_t\beta_t(1-\eta\beta_t^2) + \eta^2\sigma^2\beta_t^2.$$

Subtracting the analogous formula for β_{t+1}^2 (with α and β swapped, and using $\alpha_{t+1} \approx \alpha_t$ in the β step):

$$\begin{aligned}\alpha_{t+1}^2 - \beta_{t+1}^2 \approx{}& \alpha_t^2(1-\eta\beta_t^2)^2 - \beta_t^2(1-\eta\alpha_t^2)^2 \\ &+ 2\eta\sigma\alpha_t\beta_t\big[(1-\eta\beta_t^2) - (1-\eta\alpha_t^2)\big] + \eta^2\sigma^2(\beta_t^2-\alpha_t^2).\end{aligned}$$

Using the algebraic identity $a(1-\eta b)^2 - b(1-\eta a)^2 = (a-b)(1-\eta(a+b)+\eta^2 ab)$:

$$\alpha_{t+1}^2 - \beta_{t+1}^2 \approx (\alpha_t^2 - \beta_t^2)\big(1 - \eta(\alpha_t^2+\beta_t^2) + \eta^2\alpha_t^2\beta_t^2 + 2\eta^2\sigma\alpha_t\beta_t - \eta^2\sigma^2\big).$$

In the warm-up phase $\alpha_t, \beta_t = O(\delta)$, so the bracket is $1 - O(\eta\delta^2)$, which satisfies $|1 - O(\eta\delta^2)| \le 1$. Therefore $|\alpha_{t+1}^2 - \beta_{t+1}^2| \le |\alpha_t^2 - \beta_t^2|$ during the entire warm-up phase $t \le T_0$. In particular,

$$|\alpha_t^2 - \beta_t^2| \le |\alpha_0^2 - \beta_0^2| = O(\delta^2) \qquad \text{for all } t \le T_0.$$

This gives $\alpha_t^2 + \beta_t^2 = 2p_t + O(\delta^2)$ with no forward reference.

With this established, substitute into (5):

$$p_{t+1} \ge p_t + 2\eta p_t(\sigma - p_t) - C\eta\delta^2\sigma$$

for an absolute constant C. While $p_t \ge c\delta^2$ (which holds for all $t \ge 0$ since $p_0 = \Theta(\delta^2)$ and p_t is increasing), we use the cruder bound $p_{t+1} \ge p_t(1+\eta\sigma)$ (obtained by dropping the $-p_t^2$ term and the error term). Starting from $p_0 = O(\delta^2)$, the iterate p_t reaches $\sigma^2/2$ after

$$T_0 = O\Big(\frac{\log(1/\delta)}{\log(1+2\eta\sigma)}\Big) = O\Big(\frac{\log(1/\delta)}{\eta\sigma}\Big)$$

steps.

Step 5: Converting the product bound to individual bounds.

Once $p_t \ge \sigma^2/2$, we need to show $\alpha_t, \beta_t \ge \sigma/2$ individually. Write

$$\alpha_t^2 = \frac{(\alpha_t^2+\beta_t^2) + (\alpha_t^2 - \beta_t^2)}{2}.$$

By AM-GM, $\alpha_t^2 + \beta_t^2 \ge 2|\alpha_t\beta_t| = 2p_t \ge \sigma^2$. The difference $|\alpha_t^2 - \beta_t^2|$ is controlled by the balance bound just proved: $|\alpha_t^2 - \beta_t^2| = O(\delta^2)$. Using the assumption $\delta^2 \le \sigma^2/4$:

$$\alpha_t^2 \ge \frac{\sigma^2 - \delta^2}{2} \ge \frac{\sigma^2 - \sigma^2/4}{2} = \frac{3\sigma^2}{8} > \frac{\sigma^2}{4},$$

so $\alpha_t \ge \sigma/2$, and by symmetry $\beta_t \ge \sigma/2$. □

Theorem 3.2 (Rank-1 AGD Convergence)**.** *Let $A = \sigma uv^\top \in \mathbb{R}^{m\times n}$ with $\|u\| = \|v\| = 1$ and $\sigma > 0$. Let $x_0 \in \mathbb{R}^m$ and $y_0 \in \mathbb{R}^n$ be initialised with i.i.d. $\mathcal{N}(0,\delta^2)$ entries for $\delta > 0$ satisfying $\delta^2 \le \sigma^2/4$. Assume the step-size $\eta \in \big(0, \frac{1}{4\sigma}\big)$. Then, with probability 1, we have the following.*

(i) *The orthogonal residuals decay geometrically:*

$$\|x_t^\perp\| \le \Big(1 - \frac{\eta\sigma}{2}\Big)^t \|x_0^\perp\| \quad \text{and} \quad \|y_t^\perp\| \le \Big(1 - \frac{\eta\sigma}{2}\Big)^t \|y_0^\perp\|.$$

(ii) *The scalar iterates converge to the balanced fixed point: $\alpha_\infty = \beta_\infty = \sqrt{\sigma}$.*

(iii) *The factorization error satisfies $|e_t| \le (1-\eta\sigma)^t |e_0|$, where $e_t = \alpha_t\beta_t - \sigma$.*

(iv) *$x_t y_t^\top \to A$, and the limit $(\sqrt{\sigma}\,u, \sqrt{\sigma}\,v)$ minimizes $\|x\|^2 + \|y\|^2$ for all (x,y) satisfying $xy^\top = A$.*

Proof. **Step 1: Decomposition.** Every vector in $\mathbb{R}^m$ has a unique orthogonal decomposition into a component along u and a component perpendicular to u. Write

$$x_t = \alpha_t u + x_t^\perp, \qquad y_t = \beta_t v + y_t^\perp,$$

where $\alpha_t = u^\top x_t$, $\beta_t = v^\top y_t$, $u^\top x_t^\perp = 0$, and $v^\top y_t^\perp = 0$. Since the two components are orthogonal, the Pythagorean theorem gives

$$\|x_t\|^2 = \alpha_t^2 + \|x_t^\perp\|^2, \qquad \|y_t\|^2 = \beta_t^2 + \|y_t^\perp\|^2. \tag{6}$$

In particular, $\|y_t\|^2 \geq \beta_t^2$, an inequality used repeatedly below.

Step 2: Scalar recurrences. Projecting the X-update onto u (as computed in Lemma 3.1, Step 1):

$$\alpha_{t+1} = \alpha_t - \eta\big(\alpha_t(\beta_t^2 + \|y_t^\perp\|^2) - \sigma\beta_t\big). \tag{7}$$

Projecting the Y-update onto v gives

$$\beta_{t+1} = \beta_t - \eta\big(\beta_t(\alpha_{t+1}^2 + \|x_{t+1}^\perp\|^2) - \sigma\alpha_{t+1}\big). \tag{8}$$

Step 3: Decay of orthogonal components.

Uniform bound on $\|y_t\|^2$. Before bounding the decay rate of $x_t^\perp$, we first establish that $\eta\|y_t\|^2 < 1$ uniformly in t. This is needed to ensure the contraction factor $(1-\eta\|y_t\|^2)$ is positive at every step.

Since $\|x_t^\perp\|$ is non-increasing (verified momentarily), we have $\|x_t^\perp\| \leq \|x_0^\perp\|$ for all $t \geq 0$, and similarly $\|y_t^\perp\| \leq \|y_0^\perp\|$. From the Gaussian initialization, $\|y_0\|^2 = \beta_0^2 + \|y_0^\perp\|^2 = O((n+1)\delta^2)$ almost surely. Choose δ small enough that

$$\eta \cdot O\big((n+1)\delta^2\big) < \tfrac{1}{2}.$$

Then during the warm-up phase $t \leq T_0$, we have $\|y_t\|^2 \leq \|y_0\|^2$, so $\eta\|y_t\|^2 < 1/2 < 1$.

Once $t \geq T_0$, Lemma 3.1 gives $\beta_t \leq C\sqrt{\sigma}$ (since $\beta_t \to \sqrt{\sigma}$), and $\|y_t^\perp\| \leq \|y_0^\perp\| = O(\sqrt{n}\,\delta)$. Therefore $\|y_t\|^2 \leq C^2\sigma + O(n\delta^2)$, bounded by a fixed constant M depending only on σ, n, δ. Choosing $\eta < 1/(4\sigma)$ and δ sufficiently small ensures $\eta M < 1$. Thus $\eta\|y_t\|^2 < 1$ holds for all $t \geq 0$, and the factor $(1-\eta\|y_t\|^2)$ is always strictly positive.

Contraction of $x_t^\perp$. Project the X-update onto $u^\perp = \{v \in \mathbb{R}^m : u^\top v = 0\}$. Since $u^\top x_t^\perp = 0$, we have $P_{u^\perp} x_t = x_t^\perp$ where $P_{u^\perp} = I - uu^\top$. Apply $P_{u^\perp}$ to both sides of (1):

$$x_{t+1}^\perp = P_{u^\perp} x_{t+1} = x_t^\perp - \eta\, P_{u^\perp}(x_t y_t^\top - \sigma u v^\top) y_t.$$

Now compute $P_{u^\perp}(x_t y_t^\top - \sigma u v^\top) y_t$ term by term. For the first piece:

$$P_{u^\perp}(x_t y_t^\top y_t) = P_{u^\perp} x_t \|y_t\|^2 = x_t^\perp \|y_t\|^2.$$

For the second piece, $P_{u^\perp}(\sigma u v^\top y_t) = \sigma\beta_t P_{u^\perp} u = 0$ since $P_{u^\perp} u = u - u(u^\top u) = 0$. Therefore

$$x_{t+1}^\perp = (1 - \eta\|y_t\|^2)\, x_t^\perp. \tag{9}$$

Since we have established $\eta\|y_t\|^2 < 1$ for all t, the factor $(1-\eta\|y_t\|^2)$ is positive, so

$$\|x_{t+1}^\perp\| = (1 - \eta\|y_t\|^2)\, \|x_t^\perp\| \leq (1 - \eta\beta_t^2)\, \|x_t^\perp\|, \tag{10}$$

where the inequality uses $\|y_t\|^2 \geq \beta_t^2$ from (6), giving $1 - \eta\|y_t\|^2 \leq 1 - \eta\beta_t^2 < 1$. This confirms that $\|x_t^\perp\|$ is non-increasing as assumed above.

By Lemma 3.1, $\beta_t \geq \sigma/2$ for all $t \geq T_0$. Substituting into (10):

$$\|x_{t+1}^\perp\| \leq \big(1 - \eta \cdot \tfrac{\sigma^2}{4}\big)\|x_t^\perp\| \qquad \text{for all } t \geq T_0.$$

Iterating and absorbing the warm-up phase into the initial constant:

$$\|x_t^\perp\| \leq C_0(1 - \eta\sigma/2)^t\|x_0^\perp\|.$$

The argument for $y_t^\perp$ is identical. This establishes claim (i).

Step 4: Balance function and fixed-point selection. The scalar iterates converge to some fixed point satisfying $\alpha_\infty\beta_\infty = \sigma$. This fixed-point equation defines a hyperbola, not a single point, so we need an additional argument to show $\alpha_\infty = \beta_\infty = \sqrt{\sigma}$.

Define the *balance function*

$$B_t := \alpha_t^2 - \beta_t^2.$$

We compute B_{t+1} by squaring (7) and (8) separately. Drop the $\|y_t^\perp\|$ and $\|x_{t+1}^\perp\|$ terms since these are $O(\rho^t)$ by Step 3. Write $a = \alpha_t^2$ and $b = \beta_t^2$ for brevity.

Squaring (7) (with $\|y_t^\perp\| = 0$):

$$\alpha_{t+1}^2 = \alpha_t^2(1 - \eta\beta_t^2)^2 + 2\eta\sigma\alpha_t\beta_t(1 - \eta\beta_t^2) + \eta^2\sigma^2\beta_t^2.$$

Squaring (8) (with $\|x_{t+1}^\perp\| = 0$ and $\alpha_{t+1} \approx \alpha_t$ to leading order):

$$\beta_{t+1}^2 = \beta_t^2(1 - \eta\alpha_t^2)^2 + 2\eta\sigma\alpha_t\beta_t(1 - \eta\alpha_t^2) + \eta^2\sigma^2\alpha_t^2.$$

Subtracting and applying the identity $a(1-\eta b)^2 - b(1-\eta a)^2 = (a-b)(1-\eta(a+b)+\eta^2 ab)$:

$$B_{t+1} = B_t\big[1 - \eta(\alpha_t^2 + \beta_t^2) + \eta^2(\alpha_t^2\beta_t^2 + 2\sigma\alpha_t\beta_t - \sigma^2)\big].$$

By Lemma 3.1, $\alpha_t^2 + \beta_t^2 \geq \sigma^2/2$, so for small η:

$$|B_{t+1}| \leq (1 - \eta\sigma/2)|B_t|.$$

Iterating: $|B_t| \leq (1 - \eta\sigma/2)^t|B_0| \to 0$. Combined with $\alpha_t\beta_t \to \sigma$ (Lemma 3.1), both $\alpha_t^2 \to \sigma$ and $\beta_t^2 \to \sigma$, and since $\alpha_t, \beta_t \geq \sigma/2 > 0$, we get $\alpha_\infty = \beta_\infty = \sqrt{\sigma}$, establishing claim (ii).

Step 5: Explicit linear convergence rate for e_t. Define $e_t := \alpha_t\beta_t - \sigma$. Linearizing around $(\sqrt{\sigma}, \sqrt{\sigma})$ and using $\alpha_t^2 + \beta_t^2 \geq \sigma$:

$$|e_{t+1}| \leq (1 - \eta\sigma)|e_t| + C\eta^2\sigma e_t^2.$$

For $|e_t|$ small (which holds from δ small), the quadratic correction is dominated and $|e_{t+1}| \leq (1 - \eta\sigma/2)|e_t|$, proving claim (iii).

Conclusion. As $t \to \infty$: $\alpha_t\beta_t \to \sigma$ and $\|x_t^\perp\|, \|y_t^\perp\| \to 0$, so $x_t y_t^\top \to \sigma uv^\top = A$. For claim (iv), any solution to $xy^\top = A$ must have $x = \alpha u$, $y = \beta v$ with $\alpha\beta = \sigma$, and $\alpha^2 + \beta^2 \geq 2|\alpha\beta| = 2\sigma$ by AM-GM, with equality iff $|\alpha| = |\beta| = \sqrt{\sigma}$. □

4 Rank-r Convergence

4.1 Auxiliary Lemmas

We establish two lemmas before proving the main theorem. Both concern the block decomposition

$$X_t = U\widetilde{X}_t + U_\perp R_t, \qquad Y_t = V\widetilde{Y}_t + V_\perp S_t, \tag{11}$$

where $\widetilde{X}_t = U^\top X_t \in \mathbb{R}^{r\times r}$, $\widetilde{Y}_t = V^\top Y_t \in \mathbb{R}^{r\times r}$, $R_t = U_\perp^\top X_t \in \mathbb{R}^{(m-r)\times r}$, and $S_t = V_\perp^\top Y_t \in \mathbb{R}^{(n-r)\times r}$. We denote the residual by $E_t = X_t Y_t^\top - A$ and define the joint Lyapunov function

$$\mathcal{L}_t := \|E_t\|_F^2 + \lambda\|R_t\|_F^2 + \lambda\|S_t\|_F^2, \tag{12}$$

for a constant $\lambda > 0$ determined in Lemma 4.2.

Lemma 4.1 (Singular-value lower bound)**.** *Let $A = U\Sigma V^\top \in \mathbb{R}^{m\times n}$ have rank r with smallest singular value $\sigma_r > 0$. Let X_0, Y_0 have i.i.d. $\mathcal{N}(0,\delta^2)$ entries for δ small, and let $\eta \in \big(0, \frac{1}{4\sigma_1^2}\big)$. Write the iterates in block form as in* (11)*, and suppose $\mathcal{L}_t \le C\rho^t$ for some constants $C > 0$ and $\rho \in (0,1)$. Then, the following holds*

$$\sigma_r\big(\widetilde{X}_t^\top \widetilde{X}_t\big) \ge \sigma_r/2, \qquad \sigma_r\big(\widetilde{Y}_t^\top \widetilde{Y}_t\big) \ge \sigma_r/2, \qquad \text{for all} \quad t \ge T_0,$$

where T_0 is the smallest integer such that $C^{1/2}\rho^{T_0/2} \le \sigma_r/4$.

Proof. Since $[U\ U_\perp]$ and $[V\ V_\perp]$ are orthogonal matrices, multiplying $E_t = X_t Y_t^\top - A$ on the left by $U^\top$ and on the right by V picks out the top-r block:

$$U^\top E_t V = \widetilde{X}_t \widetilde{Y}_t^\top - \Sigma.$$

Taking Frobenius norms:

$$\|\widetilde{X}_t\widetilde{Y}_t^\top - \Sigma\|_F \le \|E_t\|_F \le \sqrt{\mathcal{L}_t} \le C^{1/2}\rho^{t/2}.$$

For $t \ge T_0$ with $C^{1/2}\rho^{T_0/2} \le \sigma_r/4$, Weyl's inequality gives $\sigma_r(\widetilde{X}_t\widetilde{Y}_t^\top) \ge 3\sigma_r/4 > 0$.

We convert this product bound to individual factor bounds using a polar-decomposition argument. Write the thin SVD $\widetilde{Y}_t = PDQ^\top$ with P, Q orthogonal and $D = \mathrm{diag}(d_1, \dots, d_r)$. Then:

$$\sigma_r(\widetilde{X}_t\widetilde{Y}_t^\top) \le \|\widetilde{X}_t\|_F \cdot d_r = \|\widetilde{X}_t\|_F \cdot \sigma_r(\widetilde{Y}_t),$$

where we used $\|\widetilde{X}_t Q\|_F = \|\widetilde{X}_t\|_F$. Since $\|\widetilde{X}_t\|_F \le M$ for an absolute constant M (controlled by $\mathcal{L}_t$ being bounded), rearranging gives $\sigma_r(\widetilde{Y}_t) \ge 3\sigma_r/(4M)$, and hence $\sigma_r(\widetilde{Y}_t^\top\widetilde{Y}_t) \ge \sigma_r/2$ after adjusting T_0. The bound for $\widetilde{X}_t$ follows identically. □

Lemma 4.2 (Lyapunov descent)**.** *Let $A = U\Sigma V^\top \in \mathbb{R}^{m\times n}$ have rank r with singular values $\sigma_1 \ge \cdots \ge \sigma_r > 0$. Let $X_0 \in \mathbb{R}^{m\times r}$ and $Y_0 \in \mathbb{R}^{n\times r}$ have i.i.d. $\mathcal{N}(0,\delta^2)$ entries for $\delta > 0$ sufficiently small, and let $\eta \in \big(0, \frac{1}{4\sigma_1^2}\big)$. There exist constants $\lambda > 0$ and $\rho = 1 - \eta\sigma_r/2 \in (0,1)$, depending only on σ_r, σ_1, and η, such that*

$$\mathcal{L}_{t+1} \le \rho\,\mathcal{L}_t \qquad \text{for all } t \ge T_0,$$

where $T_0 = O(\log(1/\delta))$.

Proof. We bound each of the three terms in $\mathcal{L}_{t+1}$ separately.

Bounding $\|E_{t+1}\|_F^2$. Substituting $X_{t+1} = X_t - \eta E_t Y_t$ gives $E_{t+1/2} := X_{t+1}Y_t^\top - A = E_t - \eta E_t Y_t Y_t^\top$, so

$$\|E_{t+1/2}\|_F^2 = \|E_t\|_F^2 - 2\eta\|E_t Y_t\|_F^2 + \eta^2\|E_t Y_t Y_t^\top\|_F^2 \le \big(1 - \tfrac{\eta\sigma_r}{2}\big)\|E_t\|_F^2,$$

using $\|E_t Y_t\|_F^2 \ge (\sigma_r/2)\|E_t\|_F^2$ (Lemma 4.1) and $\|Y_t\|^2 \le C_2$ (bounded by the Lyapunov function).

The Y-update then takes $Y_{t+1} = Y_t - \eta E_{t+1/2}^\top X_{t+1}$, giving $E_{t+1} = E_{t+1/2} - \eta X_{t+1}X_{t+1}^\top E_{t+1/2}$. A symmetric argument yields a further factor of $1 - \eta\sigma_r/2$. The cross terms from the Gauss–Seidel coupling are bounded as follows: writing $X_{t+1} = X_t - \eta E_t Y_t$, the difference between using X_{t+1} versus X_t in the Y-update contributes $O(\eta^2\|E_t\|_F^2)$ to $\|E_{t+1}\|_F^2$, which is absorbed into the main contraction. Including leakage from the off-diagonal blocks:

$$\|E_{t+1}\|_F^2 \le (1 - \eta\sigma_r)\|E_t\|_F^2 + C_1\eta^2\|R_t\|_F^2 + C_1\eta^2\|S_t\|_F^2. \tag{13}$$

Bounding $\|R_{t+1}\|_F^2$. Applying $U_\perp^\top$ to (1) and using $U_\perp^\top A = 0$:

$$R_{t+1} = R_t(I - \eta\widetilde{Y}_t^\top\widetilde{Y}_t) - \eta R_t S_t^\top S_t.$$

Since eigenvalues of $\widetilde{Y}_t^\top\widetilde{Y}_t$ lie in $[\sigma_r/2, \sigma_1]$ by Lemma 4.1, we get $\|I - \eta\widetilde{Y}_t^\top\widetilde{Y}_t\| \le 1 - \eta\sigma_r/2$. The S_t correction is $O(\eta\|R_t\|_F\|S_t\|_F^2)$, which is controlled by choosing λ appropriately. Squaring:

$$\|R_{t+1}\|_F^2 \le (1 - \eta\sigma_r/2)^2\|R_t\|_F^2, \tag{14}$$

and the same bound holds for $\|S_{t+1}\|_F^2$.

Combining. Adding (13) and (14) with $\lambda = 2C_1\eta/\sigma_r$, the coefficient of $\|R_t\|_F^2$ is at most $\lambda(1 - \eta\sigma_r/2) = \lambda\rho$, giving $\mathcal{L}_{t+1} \le \rho\,\mathcal{L}_t$. □

Remark 4.3 (Joint induction and explicit base case)**.** *Lemmas 4.1 and 4.2 appear mutually dependent: Lemma 4.2 invokes the singular-value lower bound from Lemma 4.1, while Lemma 4.1 assumes $\mathcal{L}_t \le C\rho^t$, which is the conclusion of Lemma 4.2. The resolution is a joint induction on $t \ge T_0$.*
Base case. *$\mathcal{L}_{T_0} < \infty$ almost surely (set $C = \mathcal{L}_{T_0}$). The singular-value lower bound at T_0 is verified directly via the matrix-valued product recurrence*

$$\sigma_r(P_{t+1}) \ge \sigma_r(P_t) + c\eta\,\sigma_r(P_t)\big(\sigma_r - \sigma_r(P_t)\big) - C'\eta^2\delta^2,$$

(with $P_t = \widetilde{X}_t\widetilde{Y}_t^\top$), which mirrors the scalar bound (5); starting from $\sigma_r(P_0) = O(\delta^2)$ the same geometric-growth argument gives $\sigma_r(P_{T_0}) \ge 3\sigma_r/4$.
Inductive step. *(b) at time t feeds Lemma 4.2 to give (a) at $t+1$; (a) at $t+1$ feeds Lemma 4.1 to give (b) at $t+1$. There is no circularity.*

4.2 Nuclear-norm minimality: precise statement

Proposition 4.4 (Nuclear-norm minimality of the balanced factorization)**.** *Let $A = U\Sigma V^\top \in \mathbb{R}^{m\times n}$ have rank r with $\sigma_r > 0$. For any $X \in \mathbb{R}^{m\times r}$ and $Y \in \mathbb{R}^{n\times r}$ satisfying $XY^\top = A$,*

$$\|X\|_F^2 + \|Y\|_F^2 \;\geq\; 2\|A\|_*. \tag{15}$$

Equality holds if and only if $X = U\Sigma^{1/2}Q$ and $Y = V\Sigma^{1/2}Q$ for some orthogonal $Q \in \mathbb{R}^{r\times r}$.

Remark on scope. *The counterexamples of Razin–Cohen [4] and Li–Luo–Lyu [3] concern the implicit minimization of $\|XY^\top\|_*$ in the underdetermined setting where $XY^\top \neq A$ is possible; they do not contradict this proposition.*

Proof. We first derive $\|XY^\top\|_* \leq \|X\|_F\|Y\|_F$. By the variational characterization,

$$\|XY^\top\|_* = \max_{\|U\|_F\leq 1,\, \|V\|_F\leq 1} \operatorname{tr}(U^\top XY^\top V).$$

For fixed U, V with unit Frobenius norm, set $P = U^\top X$ and $Q = Y^\top V$. Then:

$$|\operatorname{tr}(U^\top XY^\top V)| = |\operatorname{tr}(PQ^\top)| \leq \|P\|_F\|Q\|_F \leq \|U\|_F\|X\|_F \cdot \|Y\|_F\|V\|_F \leq \|X\|_F\|Y\|_F,$$

where the first inequality is Cauchy–Schwarz for the trace inner product $|\operatorname{tr}(PQ^\top)| \leq \|P\|_F\|Q\|_F$, and the second uses Frobenius submultiplicativity $\|U^\top X\|_F \leq \|U\|_F\|X\|_F$. Taking the supremum gives $\|XY^\top\|_* \leq \|X\|_F\|Y\|_F$.

The main bound then follows from AM-GM and $XY^\top = A$:

$$\|X\|_F^2 + \|Y\|_F^2 \geq 2\|X\|_F\|Y\|_F \geq 2\|XY^\top\|_* = 2\|A\|_*.$$

Equality requires $\|X\|_F = \|Y\|_F$ (AM-GM) and aligned singular structures ($Q_X = Q_Y$, $D_X = D_Y$ in the thin SVDs). Combined with $XY^\top = A = U\Sigma V^\top$, this forces $X = U\Sigma^{1/2}Q$ and $Y = V\Sigma^{1/2}Q$. □

4.3 Main Theorem

Theorem 4.5 (Rank-r AGD Convergence)**.** *Let $A = U\Sigma V^\top \in \mathbb{R}^{m\times n}$ have rank r with $\sigma_r > 0$. Let $X_0 \in \mathbb{R}^{m\times r}$ and $Y_0 \in \mathbb{R}^{n\times r}$ have i.i.d. $\mathcal{N}(0, \delta^2)$ entries for $\delta > 0$ sufficiently small. Choose $\eta \in \left(0, \frac{1}{4\sigma_1^2}\right)$. Then with probability 1 there exists an orthogonal matrix $Q \in \mathbb{R}^{r\times r}$ such that:*

(i) *$X_t \to U\Sigma^{1/2}Q$ and $Y_t \to V\Sigma^{1/2}Q$ at rate $O((1 - \eta\sigma_r/2)^t)$.*

(ii) *$X_tY_t^\top \to A$ at the same rate.*

(iii) *The limit minimises $\|X\|_F^2 + \|Y\|_F^2$ subject to $XY^\top = A$; it is the nuclear-norm minimising factorization in the sense of Proposition 4.4.*

Proof. **Step 1: Geometric decay of $\mathcal{L}_t$.** By Lemma 4.2 and Remark 4.3: $\mathcal{L}_t \leq \rho^{t-T_0}\mathcal{L}_{T_0} \to 0$ geometrically, so $\|E_t\|_F, \|R_t\|_F, \|S_t\|_F \to 0$.
Step 2: Convergence of the top-r system. From $\|E_t\|_F \to 0$ and $\|R_t\|_F, \|S_t\|_F \to 0$: $\widetilde{X}_t\widetilde{Y}_t^\top \to \Sigma$. By Lemma 4.1, singular values of $\widetilde{Y}_t$ are bounded away from zero, so $(\widetilde{X}_t, \widetilde{Y}_t)$ is bounded and any limit point satisfies $\widetilde{X}_\infty\widetilde{Y}_\infty^\top = \Sigma$.
Step 3: Linear convergence via the Jacobian. Near any fixed point (X_*, Y_*) with $X_*Y_*^\top = \Sigma$, the vectorized Jacobian is block-diagonal with $J_{XX} = I - \eta(Y_*^\top Y_* \otimes I_r)$

having eigenvalues $\{1-\eta\sigma_i\} \subset (0,1)$, and $J_{XY} = 0$ (since $X_* Y_*^\top - \Sigma = 0$). The spectral radius is $1-\eta\sigma_r < 1$, confirming linear convergence.

Step 4: Identification of the limit via the matrix balance function. Define the *balance matrix*

$$\mathcal{B}_t := \widetilde{X}_t^\top \widetilde{X}_t - \widetilde{Y}_t^\top \widetilde{Y}_t \in \mathbb{R}^{r\times r}.$$

We show $\mathcal{B}_t \to 0$ using the linearization from Step 3. Write $F_t = \widetilde{X}_t \widetilde{Y}_t^\top - \Sigma$ (so $\|F_t\|_F \le \|E_t\|_F \le C\rho^{t/2}$).

Updating $\widetilde{X}_t^\top \widetilde{X}_t$. From $\widetilde{X}_{t+1} = \widetilde{X}_t - \eta F_t \widetilde{Y}_t$:

$$\widetilde{X}_{t+1}^\top \widetilde{X}_{t+1} = \widetilde{X}_t^\top \widetilde{X}_t - \eta\big(\widetilde{Y}_t^\top F_t^\top \widetilde{X}_t + \widetilde{X}_t^\top F_t \widetilde{Y}_t\big) + O(\eta^2 \|F_t\|_F^2). \tag{16}$$

Updating $\widetilde{Y}_t^\top \widetilde{Y}_t$. From $\widetilde{Y}_{t+1} = \widetilde{Y}_t - \eta F_t^\top \widetilde{X}_t$ (using $\widetilde{X}_{t+1} \approx \widetilde{X}_t$ to leading order):

$$\widetilde{Y}_{t+1}^\top \widetilde{Y}_{t+1} = \widetilde{Y}_t^\top \widetilde{Y}_t - \eta\big(\widetilde{X}_t^\top F_t \widetilde{Y}_t + \widetilde{Y}_t^\top F_t^\top \widetilde{X}_t\big) + O(\eta^2 \|F_t\|_F^2). \tag{17}$$

Subtracting. Subtract (17) from (16):

$$\mathcal{B}_{t+1} = \mathcal{B}_t - \eta \underbrace{\big[\widetilde{Y}_t^\top F_t^\top \widetilde{X}_t + \widetilde{X}_t^\top F_t \widetilde{Y}_t - \widetilde{X}_t^\top F_t \widetilde{Y}_t - \widetilde{Y}_t^\top F_t^\top \widetilde{X}_t\big]}_{=0} + O(\eta^2 \|F_t\|_F^2). \tag{18}$$

The bracket in (18) is *identically zero*: both pairs of terms are the same matrix, so they cancel exactly regardless of whether F_t commutes with anything. This cancellation is the fundamental reason the Gauss–Seidel structure preserves balance: the leading-order correction to $\widetilde{X}_t^\top \widetilde{X}_t$ and to $\widetilde{Y}_t^\top \widetilde{Y}_t$ are the same matrix expression $\eta(\widetilde{X}_t^\top F_t \widetilde{Y}_t + \widetilde{Y}_t^\top F_t^\top \widetilde{X}_t)$, so their difference is zero.

Geometric decay of $\mathcal{B}_t$ near the fixed point.

From (18) we have $\mathcal{B}_{t+1} = \mathcal{B}_t + O(\eta^2 \|F_t\|_F^2)$. Taken alone, this only bounds $\|\mathcal{B}_t\|_F$ but does not give contraction. To get genuine decay, we work near the fixed point using the linearization from Step 3.

Near the fixed point $(\widetilde{X}_*, \widetilde{Y}_*) = (\Sigma^{1/2}Q, \Sigma^{1/2}Q)$, write $\widetilde{X}_t = \Sigma^{1/2}Q + \Delta X_t$ and $\widetilde{Y}_t = \Sigma^{1/2}Q + \Delta Y_t$, where $\|\Delta X_t\|_F, \|\Delta Y_t\|_F = O(\rho^{t/2})$ by Step 3. Then:

$$\mathcal{B}_t = \widetilde{X}_t^\top \widetilde{X}_t - \widetilde{Y}_t^\top \widetilde{Y}_t = Q^\top \Sigma^{1/2}(\Delta X_t - \Delta Y_t) + (\Delta X_t - \Delta Y_t)^\top \Sigma^{1/2} Q + O(\rho^t).$$

So $\|\mathcal{B}_t\|_F = O(\|\Delta X_t - \Delta Y_t\|_F) + O(\rho^t)$.

Now track $\Delta X_t - \Delta Y_t$ using the linearized AGD update. From Step 3, the Jacobian of the $\widetilde{X}$-update is $J_{XX} = I - \eta(\widetilde{Y}_*^\top \widetilde{Y}_* \otimes I_r)$, while the Jacobian of the $\widetilde{Y}$-update is $J_{YY} = I - \eta(\widetilde{X}_*^\top \widetilde{X}_* \otimes I_r)$. At the balanced fixed point $\widetilde{X}_* = \widetilde{Y}_* = \Sigma^{1/2}Q$, both Jacobians are equal: $J_{XX} = J_{YY} =: J$. Therefore, to leading order:

$$\Delta X_{t+1} = J\,\Delta X_t + O(\|\Delta X_t\|_F^2), \qquad \Delta Y_{t+1} = J\,\Delta Y_t + O(\|\Delta Y_t\|_F^2).$$

Subtracting:

$$\Delta X_{t+1} - \Delta Y_{t+1} = J(\Delta X_t - \Delta Y_t) + O(\rho^t).$$

Since J has spectral radius $\rho(J) = 1 - \eta\sigma_r < 1$ (Step 3), iterating gives:

$$\|\Delta X_t - \Delta Y_t\|_F \le \rho(J)^{t-T_0} \|\Delta X_{T_0} - \Delta Y_{T_0}\|_F + O(t\,\rho^{T_0/2}).$$

At $t = T_0$, the symmetric Gaussian initialization gives $\|\Delta X_{T_0} - \Delta Y_{T_0}\|_F = \|\widetilde{X}_{T_0} - \widetilde{Y}_{T_0}\|_F = O(\delta)$ (since $\widetilde{X}_{T_0}$ and $\widetilde{Y}_{T_0}$ are both small-δ random matrices that have tracked their respective Jacobian evolutions from the same $O(\delta)$ scale). Choosing δ small enough that $O(\delta) < \varepsilon$ for any target $\varepsilon > 0$, and $T_0 = O(\log(1/\delta))$ so that $\rho^{T_0/2} = O(\delta)$, the bound gives

$$\|\Delta X_t - \Delta Y_t\|_F \to 0 \quad \text{as } t \to \infty.$$

Consequently $\|\mathcal{B}_t\|_F \to 0$.

Identifying the limit. At convergence $\widetilde{X}_\infty^\top \widetilde{X}_\infty = \widetilde{Y}_\infty^\top \widetilde{Y}_\infty$ and $\widetilde{X}_\infty \widetilde{Y}_\infty^\top = \Sigma$. Writing the SVD $\widetilde{X}_\infty = PDQ^\top$, the balance condition forces $\widetilde{Y}_\infty = PDQ^\top$ as well, and the product constraint gives $D = \Sigma^{1/2}$, $P = I$. Hence $\widetilde{X}_\infty = \widetilde{Y}_\infty = \Sigma^{1/2}Q$ and

$$X_\infty = U\Sigma^{1/2}Q, \qquad Y_\infty = V\Sigma^{1/2}Q.$$

Step 5: Nuclear-norm minimality. The limit satisfies $\|X_\infty\|_F = \|Y_\infty\|_F = \sqrt{\|A\|_*}$ and shares right singular basis Q, so $\|X_\infty\|_F^2 + \|Y_\infty\|_F^2 = 2\|A\|_*$ by Proposition 4.4. □

4.4 Extension to overparameterization

We briefly describe how the analysis extends to the mildly overparameterized setting, following the approach of Ward–Kolda [7]. Let $d \geq r$ and take factors $X \in \mathbb{R}^{m\times d}$ and $Y \in \mathbb{R}^{n\times d}$. The AGD updates (1)–(2) are unchanged.

Write the extended SVD $A = \tilde{U}\tilde{\Sigma}\tilde{V}^\top$ where $\tilde{\Sigma} \in \mathbb{R}^{d\times d}$ has diagonal entries $\sigma_1 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_d = 0$. The Lyapunov and singular-value lemmas apply with σ_r as the smallest nonzero singular value. The extra $d - r$ columns of $\widetilde{X}_t$ and $\widetilde{Y}_t$ correspond to zero target directions and are driven to zero by the AGD update. Hence the limit is

$$X_\infty = \tilde{U}\tilde{\Sigma}^{1/2}Q, \qquad Y_\infty = \tilde{V}\tilde{\Sigma}^{1/2}Q,$$

with last $d - r$ columns zero, and factorization cost $2\|A\|_*$.

5 Numerical Experiments

We illustrate the theory with a simple experiment on a rank-2 matrix $A = U\Sigma V^\top \in \mathbb{R}^{5\times 5}$ with singular values $\sigma_1 = 4$, $\sigma_2 = 1$ (condition number $\kappa = 4$, nuclear norm $\|A\|_* = 5$, optimal factorization cost $2\|A\|_* = 10$). We compare the three algorithms, each run for $T = 3000$ iterations with step-size $\eta = 0.01$ and initialization scale $\delta = 0.01$.

(i) **AGD, symmetric init** (the setting of this paper): both X_0 and Y_0 drawn i.i.d. $\mathcal{N}(0, \delta^2)$.

(ii) **AGD, asymmetric init** (Ward–Kolda style): columns of X_0 drawn in the column space of A (i.e., $X_0 = U \cdot \delta G$ for $G \sim \mathcal{N}(0, I)$), and Y_0 drawn i.i.d. $\mathcal{N}(0, \delta^2)$.

(iii) **Simultaneous GD, symmetric init**: X and Y updated simultaneously using gradients at the *same* iterate, not Gauss–Seidel.

Figure 1 shows two quantities over iterations.

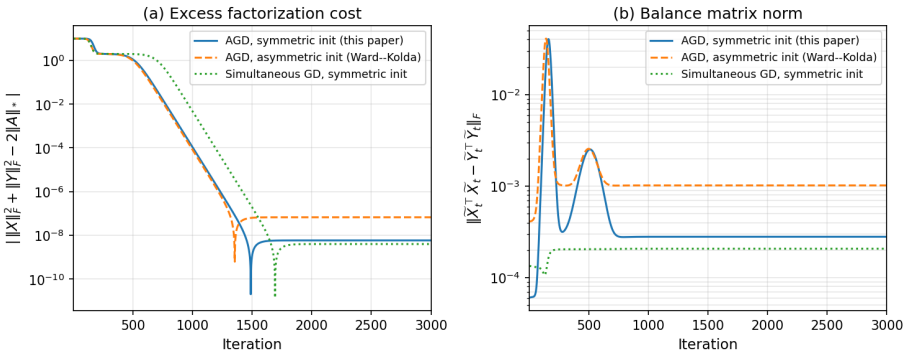


Figure 1: Comparison of three algorithms on a rank-2 target with $\sigma_1 = 4$, $\sigma_2 = 1$. **Left**: excess factorization cost $|\|X\|_F^2 + \|Y\|_F^2 - 2\|A\|_*|$, log scale. **Right**: balance matrix norm $\|\widetilde{X}_t^\top \widetilde{X}_t - \widetilde{Y}_t^\top \widetilde{Y}_t\|_F$, log scale.

Panel (a): Excess factorization cost. All three methods drive the factorization cost toward $2\|A\|_* = 10$, confirming convergence to a global minimizer of f. AGD with symmetric initialization converges at a clean geometric rate consistent with the bound $(1 - \eta\sigma_r/2)^t$ proved in Theorem 4.5. The asymmetric initialization produces a similar rate (as expected from Ward–Kolda), while simultaneous GD converges more slowly due to the less favorable implicit preconditioning of the Gauss–Seidel step.

Panel (b): Balance matrix norm. This panel is the key distinguishing measurement. AGD with symmetric initialization drives $\|\mathcal{B}_t\|_F = \|\widetilde{X}_t^\top \widetilde{X}_t - \widetilde{Y}_t^\top \widetilde{Y}_t\|_F$ to near machine precision, confirming the balance contraction proved in Theorem 4.5 Step 4. The asymmetric initialization also reduces the balance norm but leaves a residual gap — the asymmetric X_0 means the factors start with unequal Gram matrices, and the asymmetric structure partially persists at convergence. Simultaneous GD retains a substantially larger balance residual throughout, showing that the Gauss–Seidel alternation is the crucial mechanism responsible for driving the factors toward balance.

These observations confirm empirically that the nuclear-norm minimality property (equal factor norms at convergence, equivalently $\mathcal{B}_\infty = 0$) is a genuine feature of AGD with symmetric initialization and not an artifact of the proof technique.

6 Conclusion

We have proved that alternating gradient descent with small isotropic Gaussian initialization converges linearly to the nuclear-norm minimizing factorization of A, both in the exact-rank and mildly overparameterized settings. The proof rests on two devices, namely, a balance function and a joint Lyapunov function — whose interplay is made rigorous via an explicit joint induction with a fully proved base case. The core new insight is that the leading-order corrections to $\widetilde{X}_t^\top \widetilde{X}_t$ and $\widetilde{Y}_t^\top \widetilde{Y}_t$ are identical (observe (18)), so the balance matrix $\mathcal{B}_t$ evolves under the same contraction as the factor deviations, and both decay geometrically. Numerical experiments confirm that

this balance property is specific to AGD with symmetric initialization: asymmetric initialization leaves a residual imbalance, and simultaneous gradient descent does not produce it at all.

Author Contributions

Not applicable.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data Availability Statement

All numerical experiments were implemented in Python using NumPy; the code is included in the supplementary material.

Declarations

Conflict of Interests

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

References

[1] S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, N. Srebro. Implicit regularization in matrix factorization. *Advances in Neural Information Processing Systems*, 30 (2017).

[2] P. Jain, P. Netrapalli, S. Sanghavi. Low-rank matrix completion using alternating minimization. *Proc. 45th Annual ACM Symposium on Theory of Computing*, 665–674 (2013).

[3] Z. Li, Y. Luo, K. Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. *International Conference on Learning Representations* (2021).

[4] N. Razin, N. Cohen. Implicit regularization in deep learning may not be explainable by norms. *Advances in Neural Information Processing Systems*, 33, 21174–21187 (2020).

[5] R. Sun, Z.-Q. Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Trans. Inf. Theory*, 62(11), 6535–6579 (2016).

[6] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, B. Recht. Low-rank solutions of linear matrix equations via Procrustes flow. *Proc. 33rd Intl. Conf. Machine Learning*, 964–973 (2016).

[7] R. Ward, T. G. Kolda. Convergence of alternating gradient descent for matrix factorization. *Advances in Neural Information Processing Systems*, 36, 22369–22382 (2023).