On Initializations for NMF

by

Sammy Shi

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Applied Mathematics

Waterloo, Ontario, Canada, 2026

© Sammy Shi 2026

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In Ward and Kolda (2023) and Xu et al. (2024), the authors proposed a sketching based initialization for unconstrained two block matrix factorization (**MF**) with provable guarantee. The purpose of this thesis is to examine the limits of extending these two proof frameworks in the context of Non-negative Matrix Factorization (**NMF**), as well as its numerical implications.

The **first chapter** deals with fundamentals of NMF: its formulation, the identifiability issue and guarantees, a survey of iterative methods to solve it, and finally some complexity results.

In the **second chapter**, I will quickly go over spectral initialization techniques for matrix optimization problems, and introduce sketching, a family of randomization techniques for numerical linear algebra, with a particular focus on its uses in iterative methods. The core focus of this chapter lies at the intersection of these two topics: a sketching-based initialization for the unconstrained two-block MF problem, using either non-alternating or alternating gradient descent (**GD**). We will go over the proof frameworks. In the **third chapter**, I will present my attempts to generalize the proof frameworks to NMF and the conclusions based on the partial progress.

The **fourth chapter** deals with numerics. We will compare the performance of different base methods in different data regimes using all of the initializations mentioned in this thesis.

Acknowledgements

First, I want to thank my supervisor, Giang Tran, for her guidance and feedback throughout my time at Waterloo. The regurgitation of relevant literatures at the early stage turned out to be pivotal for my research. I am also deeply grateful for her superb work ethic and dedication - she remained incredibly diligent during a demanding and special period for her, which I truly appreciate. She expects the absolute best from her students. Additionally, one of her technical contributions to this project that I would like to highlight is that extrapolation helps to contract the log-barrier as a non-linear component when using a constant barrier parameter in Xu et al. (2024). She also helps to suggest the untanglement between θ and ρ_* in Section 3.2.8.

Andersen Ang (Southampton) is my co-advisor de facto. We first met at ISMP24 in Montréal and worked more intensively together starting early 2025. He brings buckets of new ideas and optimism into this project. Before my memory fades, I want to note some of the technical concepts we discussed that I found fascinating, even if they ended up being less central to the final thesis. These included the use of various smooth reparameterization techniques (like the log-barrier and Hadamard product), the application of the Hoffman constant Burke and Tseng (1996) (although in hindsight, this is used almost exclusively for LPs), and various equivalent forms of the Polyak-Łojasiewicz (PŁ) condition Karimi et al. (2016) and other geometric inequalities, alongside many other NMF fundamentals.

Both of the two put enourmous efforts in crafting this thesis.

I also want to thank Aukosh Jagannath and Steve Vavasis for being excellent teachers.

I would also like to thank Steve Vavasis (again) and Henry Wolkowicz, for agreeing to sit on the committee, read the thesis, and provide feedback despite their tight schedules.

I am also grateful to all the AMATH administrative staff for their patience and for letting me bug them on a daily basis.

Finally, I acknowledge the financial support of Canadian taxpayers (via NSERC RGPIN 50503-10842), as well as the ongoing support of my family.

Table of Contents

Author's Declaration ii

Abstract iii

Acknowledgements iv

List of Figures viii

List of Tables xi

1 A quick tour of NMF 1

1.1 Notations . . . 1

1.2 What is NMF . . . 3

1.3 Geometric interpretation . . . 3

1.4 Identifiability issue . . . 4

1.4.1 Uniqueness guarantee . . . 5

1.5 Zoo of iterative algorithms . . . 8

1.5.1 MU . . . 8

1.5.2 HALS: Hierarchical Alternating Least Squares . . . 12

1.5.3 PGD, AltPGD, and PALM . . . 13

1.5.4 ADMM . . . 19

1.5.5 Extrapolation and accelerated versions . . . 22

1.6 Some complexity results . . . 25

2 Spectral techniques for initializations 26

2.1 A few canonical problems under the two-stage approach . . . 26

2.1.1 Why spectral initialization works well (under gradient methods)? . . . 26

2.1.2 Proximity to ground-truth via matrix perturbation . . . 28
2.1.3 Local curvature condition . . . 29
2.1.4 Putting it together: Orthogonal Procrustes Flow as a special case . . . 31
2.1.5 Implicit bias of (small) Gaussian initialization . . . 34
2.1.6 Some default options for initializations of NMF . . . 34
2.2 Sketching techniques in optimization . . . 38
2.2.1 Some examples of sketches . . . 39
2.2.2 Sketching techniques in iterative methods . . . 42
2.3 A sketching based initialization . . . 44
2.3.1 PŁ-based approach . . . 44
2.3.2 Contraction-based approach . . . 48
2.3.3 Detour: Nyström initialization works well on a different base method . . . 54

3 Do their frameworks fit for NMF? 56
3.1 The attempt for the PŁ framework . . . 56
3.1.1 The need for non-negativity at initialization . . . 56
3.1.2 The need (and failure) for a meaningful proximal PŁ inequality . . . 57
3.2 The attempt for the contraction-based framework . . . 59
3.2.1 Motivations and re-parametrization techniques . . . 59
3.2.2 Set-up . . . 60
3.2.3 Initialization and positivity with a Gaussian max bound . . . 60
3.2.4 Dynamics decomposition . . . 63
3.2.5 Contractivity via spectral analysis of $\mathbf{T}_{\text{bar}}$ . . . 68
3.2.6 Control on the nonlinear terms . . . 72
3.2.7 Convergence . . . 76
3.2.8 The neccessities (and pitfalls) of those conditional statements . . . 81

4 Numerics 85
4.1 Data matrices . . . 85
4.1.1 Simplex data via Dirichlet and logit-normal . . . 85
4.1.2 Reference matrices . . . 87
4.1.3 Data clouds and the existence of exact NMFs . . . 87
4.2 Base solvers . . . 88

4.3 Initialization schemes . . . 89
4.4 Hyperparameters . . . 91
4.5 Results . . . 91
4.5.1 MU . . . 91
4.5.2 HALS . . . 93
4.5.3 AltPGD . . . 94
4.5.4 PGD . . . 94
4.5.5 ADMM . . . 95
4.6 Discussion . . . 97
4.6.1 Which initialization wins, and when . . . 97
4.6.2 Solver sensitivity to initialization . . . 98
4.6.3 Regime-specific anomalies . . . 98

References **117**

List of Figures

4.1 Option 1 — *dense, scattered, well-conditioned*. Left: Dirichlet with $\boldsymbol{\alpha} = (1, 1, 1)$ (uniform on Δ_2). Right: logit-normal with $\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \mathbf{I}_3$. Columns of $\mathbf{A} \in \mathbb{R}^{3\times 500}$ are scattered directly in $\mathbb{R}^3_+$; the dashed triangle marks Δ_2. *Exact NMF exists at* $d = 3$ (*trivial factorization* $\mathbf{W} = \mathbf{I}_3, \mathbf{H} = \mathbf{A}$); *minimum rel. error* $= 0$. 89

4.2 Option 2 — *dense, concentrated, ill-conditioned*. $\boldsymbol{\alpha} = (50, 50, 50)$; $\boldsymbol{\Sigma} = 0.01\,\mathbf{I}_3$. The cloud collapses to a near-rank-1 cluster around the centroid $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ of Δ_2; the singular value ratio σ_3/σ_1 drops by an order of magnitude relative to Option 1. *Exact NMF exists at* $d = 3$; *minimum rel. error* $= 0$. 90

4.3 Option 3 — *sparse, scattered, well-conditioned*. $\boldsymbol{\alpha} = (0.3, 0.3, 0.3)$; $\boldsymbol{\Sigma} = 8\,\mathbf{I}_3$. Mass concentrates at the three vertices of Δ_2, approximating a scaled permutation of $\mathbf{I}_3$ when restricted to any three sample columns. *Exact NMF exists at* $d = 3$; *minimum rel. error* $= 0$. 90

4.4 Option 4 — *sparse, concentrated, ill-conditioned*. $\boldsymbol{\alpha} = (50, 0.3, 0.3)$ for Dirichlet; $\boldsymbol{\mu} = (4, 0, 0)^\top$, $\boldsymbol{\Sigma} = 0.05\,\mathbf{I}_3$ for logit-normal. A single dominant corner attracts essentially all probability mass; the cloud is the worst-conditioned of the four options. *Exact NMF exists at* $d = 3$; *minimum rel. error* $= 0$. 91

4.5 MU on Option 1 (dense, scattered, well-conditioned). Rows compare the two simplex distributions; columns compare the iteration- and time-axis metrics. *Exact NMF exists at* $d = 3$; *minimum rel. error* $= 0$. 92

4.6 MU on Option 2 (dense, concentrated, ill-conditioned). *Exact NMF exists at* $d = 3$; *minimum rel. error* $= 0$. 93

4.7 MU on Option 3 (sparse, scattered, well-conditioned). *Exact NMF exists at* $d = 3$; *minimum rel. error* $= 0$. 94

4.8 MU on Option 4 (sparse, concentrated, ill-conditioned). *Exact NMF exists at* $d = 3$; *minimum rel. error* $= 0$. 95

4.9 MU on the two reference matrices. The first row (`wardkolda`) uses $\sigma = 0$; the second row (`synthetic_noisy`) uses $\sigma = 0.05$ and visibly departs from the noiseless baseline. *Exact NMF exists at* $d = 10$ *for* `wardkolda` (*min. rel. error* $= 0$); `synthetic_noisy` *has no exact rank-*10 *NMF* (*min. rel. error strictly positive,* ≈ 0.033 *in the runs*). 96

4.10 HALS on Option 1 (dense, scattered, well-conditioned). *Exact NMF exists at* $d = 3$*; minimum rel. error* $= 0$. . . . 97
4.11 HALS on Option 2 (dense, concentrated, ill-conditioned). *Exact NMF exists at* $d = 3$*; minimum rel. error* $= 0$. . . . 98
4.12 HALS on Option 3 (sparse, scattered, well-conditioned). *Exact NMF exists at* $d = 3$*; minimum rel. error* $= 0$. . . . 99
4.13 HALS on Option 4 (sparse, concentrated, ill-conditioned). *Exact NMF exists at* $d = 3$*; minimum rel. error* $= 0$. . . . 100
4.14 HALS on the two reference matrices. *Exact NMF exists at* $d = 10$ *for* `wardkolda` *(min. rel. error* $= 0$*);* `synthetic_noisy` *has no exact rank-*10 *NMF (min. rel. error strictly positive).* . . . 101
4.15 AltPGD on Option 1 (dense, scattered, well-conditioned). *Exact NMF exists at* $d = 3$*; minimum rel. error* $= 0$. . . . 102
4.16 AltPGD on Option 2 (dense, concentrated, ill-conditioned). *Exact NMF exists at* $d = 3$*; minimum rel. error* $= 0$. . . . 103
4.17 AltPGD on Option 3 (sparse, scattered, well-conditioned). *Exact NMF exists at* $d = 3$*; minimum rel. error* $= 0$. . . . 104
4.18 AltPGD on Option 4 (sparse, concentrated, ill-conditioned). *Exact NMF exists at* $d = 3$*; minimum rel. error* $= 0$. . . . 105
4.19 AltPGD on the two reference matrices. *Exact NMF exists at* $d = 10$ *for* `wardkolda` *(min. rel. error* $= 0$*);* `synthetic_noisy` *has no exact rank-*10 *NMF (min. rel. error strictly positive).* . . . 106
4.20 PGD on Option 1 (dense, scattered, well-conditioned). *Exact NMF exists at* $d = 3$*; minimum rel. error* $= 0$. . . . 107
4.21 PGD on Option 2 (dense, concentrated, ill-conditioned). *Exact NMF exists at* $d = 3$*; minimum rel. error* $= 0$. . . . 108
4.22 PGD on Option 3 (sparse, scattered, well-conditioned). *Exact NMF exists at* $d = 3$*; minimum rel. error* $= 0$. . . . 109
4.23 PGD on Option 4 (sparse, concentrated, ill-conditioned). *Exact NMF exists at* $d = 3$*; minimum rel. error* $= 0$. . . . 110
4.24 PGD on the two reference matrices. *Exact NMF exists at* $d = 10$ *for* `wardkolda` *(min. rel. error* $= 0$*);* `synthetic_noisy` *has no exact rank-*10 *NMF (min. rel. error strictly positive).* . . . 111
4.25 ADMM on Option 1 (dense, scattered, well-conditioned). *Exact NMF exists at* $d = 3$*; minimum rel. error* $= 0$. . . . 112
4.26 ADMM on Option 2 (dense, concentrated, ill-conditioned). *Exact NMF exists at* $d = 3$*; minimum rel. error* $= 0$. . . . 113
4.27 ADMM on Option 3 (sparse, scattered, well-conditioned). *Exact NMF exists at* $d = 3$*; minimum rel. error* $= 0$. . . . 114

4.28 ADMM on Option 4 (sparse, concentrated, ill-conditioned). *Exact NMF exists at $d = 3$; minimum rel. error $= 0$.* 115

4.29 ADMM on the two reference matrices. *Exact NMF exists at $d = 10$ for* `wardkolda` *(min. rel. error $= 0$);* `synthetic_noisy` *has no exact rank-10 NMF (min. rel. error strictly positive).* 116

List of Tables

2.1 Common matrix recovery problems and their objectives. 27

2.2 Surrogate matrices, spectral initializations, and perturbation guarantees for the problems in Table 2.1. 29

4.1 Qualitative property tags and the parameter levers that realize them. Adapted from Ward and Kolda (2023)-style spectral-control intuitions; full discussion in the preceding chapters. 86

4.2 The four admissible simplex regimes and their parameter values, as resolved in the data layer. 86

4.3 The five base solvers under comparison. The active-set column indicates whether iterates expose a binary indicator $[\mathbf{W} \leq \tau\,;\, \mathbf{H} \leq \tau]$ that supports the Hamming-distance manifold-stability metric (not reported in the figures below; available in the raw run output for the four projection-style solvers). 88

4.4 Deterministic initializations. 89

4.5 Randomized initializations. The Ward–Kolda variants use half-normal sub-Gaussian sketches $\mathbf{\Phi}_1 \in \mathbb{R}^{n\times d}$, $\mathbf{\Phi}_2 \in \mathbb{R}^{n\times d}$, $\mathbf{\Phi}_{rX} \in \mathbb{R}^{m\times d}$, $\mathbf{\Phi}_{rY} \in \mathbb{R}^{n\times d}$ drawn once per seed. The constants $C = 4$, $\nu = 10^{-10}$, $D = C\nu/9$ follow the paper's defaults; σ_1 is the leading singular value of $\mathbf{A}$. The factor convention is the project's $(\mathbf{W}_0, \mathbf{H}_0)$; the paper's $(\mathbf{X}_0, \mathbf{Y}_0)$ map back via $\mathbf{W}_0 = \mathbf{X}_0$, $\mathbf{H}_0 = \mathbf{Y}_0^\top$. . . . 91

4.6 Experimental hyperparameters held fixed across the sweep. 92

Chapter 1

A quick tour of NMF

NMF has emerged as a fundamental computational tool in machine learning, data mining, and signal processing. By enforcing non-negativity constraints on the factor matrices, NMF produces highly interpretable, parts-based representations of complex data. This chapter provides a comprehensive overview of NMF, laying the theoretical and algorithmic groundwork for the subsequent developments in this thesis. We begin by formally defining the problem, followed by an exploration of its geometric properties and inherent identifiability challenges. Finally, we review a wide array of iterative algorithms and state-of-the-art optimization techniques used to solve NMF in practice.

1.1 Notations

Throughout this thesis, scalars are written in standard font, while vectors and matrices are denoted by bold lowercase and uppercase letters, respectively.

Sets and spaces

We use $\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{R}_{++}$ to denote the sets of real, non-negative real, and strictly positive real numbers. The space of n-dimensional real vectors is denoted by $\mathbb{R}^n$, and the space of $m \times n$ real matrices by $\mathbb{R}^{m\times n}$. The subsets of entrywise non-negative and strictly positive matrices are denoted by $\mathbb{R}_+^{m\times n}$ and $\mathbb{R}_{++}^{m\times n}$, respectively. Similarly, $\mathbb{S}^{n\times n}$, $\mathbb{S}_+^{n\times n}$, and $\mathbb{S}_{++}^{n\times n}$ represent the spaces of symmetric, positive semi-definite, and positive definite matrices.

Norms and indexing

The standard Euclidean norm (or spectral norm for matrices) is denoted by $\|\cdot\|$ or $\|\cdot\|_2$, while $\|\cdot\|_F$ represents the Frobenius norm. For a matrix $\mathbf{A}$, its (i,j)-th entry is denoted by $\mathbf{A}(i,j)$ or the more compact subscript form $\mathbf{A}_{ij}$. Its i-th row and j-th column are denoted by $\mathbf{A}(i,:)$ (or $\mathbf{A}_{i,:}$) and $\mathbf{A}(:,j)$ (or $\mathbf{A}_{:,j}$), respectively.

Matrix inequalities

When the dimension of $\mathbf{A}$ is clear from context, the entrywise inequality $\mathbf{A} \geq 0$ concisely indicates that $\mathbf{A} \in \mathbb{R}_+^{m \times n}$. More generally, $\mathbf{A} \geq C$ and $\mathbf{A} > C$ signify that all entries of $\mathbf{A}$ are greater than or equal to, or strictly greater than, a scalar constant C. Similarly, we write $\succeq 0$ and $\succ 0$ to denote positive semi-definiteness and positive definiteness when the dimension is clear.

Entrywise non-negative projection

For a given matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, the notations $[\mathbf{A}]_+$, $\max\{0, \mathbf{A}\}$, and $\delta_+(\mathbf{A})$ are used interchangeably to denote the entrywise projection onto the non-negative orthant. This operation sets all negative entries of the matrix to zero while leaving non-negative entries unchanged, defined elementwise as:

$$[\mathbf{A}]_+(i,j) = \max\{0, \mathbf{A}_{ij}\}.$$

Hadamard operations

For operations in NMF, $\odot$ represents the Hadamard (entrywise) product between two matrices $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{m \times n}$, defined as:

$$(\mathbf{P} \odot \mathbf{Q})_{ij} = \mathbf{P}_{ij}\mathbf{Q}_{ij}.$$

The entrywise inverse of a strictly positive matrix $\mathbf{A} \in \mathbb{R}_{++}^{m \times n}$ is denoted by $\mathbf{A}^{\odot(-1)}$, where:

$$[\mathbf{A}^{\odot(-1)}]_{ij} = \frac{1}{\mathbf{A}_{ij}}.$$

Factorization conventions

We consider low-rank factorizations where the target rank d (or r) satisfies $d \ll \min\{m, n\}$. For unconstrained two-block matrix factorizations, the objective is typically formulated as:

$$\min_{\mathbf{X},\mathbf{Y}} \frac{1}{2}\|\mathbf{A} - \mathbf{X}\mathbf{Y}^\top\|_F^2,$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$, and the factors $\mathbf{X} \in \mathbb{R}^{m \times d}$ and $\mathbf{Y} \in \mathbb{R}^{n \times d}$ are both "tall" matrices. However, following standard NMF conventions, we instead structure the right factor as a "flat" matrix $\mathbf{H} \in \mathbb{R}_+^{d \times n}$, yielding the objective:

$$\min_{\mathbf{W},\mathbf{H}} \frac{1}{2}\|\mathbf{A} - \mathbf{W}\mathbf{H}\|_F^2 \quad \text{s.t. } \mathbf{W} \in \mathbb{R}_+^{m \times d},\ \mathbf{H} \in \mathbb{R}_+^{d \times n}.$$

1.2 What is NMF

Definition 1.2.1: Exact NMF (Problem 2.1. in Gillis (2020))

Given a matrix $\mathbf{A} \in \mathbb{R}_+^{m\times n}$, and a target factorization rank r, compute, if possible, two matrices $\mathbf{W} \in \mathbb{R}_+^{m\times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r\times n}$ so that

$$\mathbf{A} = \mathbf{WH}. \tag{1.1}$$

We call $\mathbf{WH}$ an **Exact NMF** of $\mathbf{A}$ with size r. Equivalently, we sometimes call $(\mathbf{W}, \mathbf{H})$ a solution to the Exact size-r NMF problem of $\mathbf{A} \in \mathbb{R}_+^{m\times n}$ if the pair satisfies (1.1).

Sometimes we will use d to represent the inner dimension of the factorization in the case where we are operating in an overparametrized regime $d > r$. Also, we always assume $r \leq \min\{m, n\}$.

We focus on the optimization under the Frobenius and Euclidean norms [1].

Definition 1.2.2: (Approximate) NMF

Given a matrix $\mathbf{A} \in \mathbb{R}_+^{m\times n}$, and a target factorization rank r with $r \ll \min\{m, n\}$, we aim to find two matrices $\mathbf{W} \in \mathbb{R}_+^{m\times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r\times n}$ so that

$$\begin{aligned} \min_{\mathbf{W},\mathbf{H}} \quad & \frac{1}{2}\|\mathbf{A} - \mathbf{WH}\|_F^2 \\ \text{s.t.} \quad & \mathbf{W} \in \mathbb{R}_+^{m\times r}, \mathbf{H} \in \mathbb{R}_+^{r\times n}. \end{aligned} \tag{1.2}$$

A pair $(\mathbf{W}, \mathbf{H})$ that minimizes (1.2) is called an (**approximate**) **NMF** of $\mathbf{A}$ with size r.

Having established the algebraic formulation of both exact and approximate NMF, it is instructive to examine the problem from a geometric perspective. Understanding the geometric intuition behind NMF not only clarifies the role of the non-negativity constraints but also illuminates why finding a unique factorization can be so mathematically challenging.

1.3 Geometric interpretation

The geometric picture of NMF can be roughly thought of as the following:

- The matrix $\mathbf{A}$ represents the data cloud living in the non-negative orthant for which we would like to perform a low-rank factorization. If $\mathbf{A}$ is of rank r, then there exist r distinct rays which constitute the convex cone/hull generated by the columns of $\mathbf{A}$.
- The matrix $\mathbf{W}$ can be thought of as the basis for the approximation, and $\mathbf{H}$ can be treated as the non-negative scaling parameter for $\mathbf{W}$.

Rigorously, the convex cone generated by the columns of $\mathbf{A}$ is

$$\text{cone}(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^m \mid \mathbf{x} = \mathbf{Ay} \text{ for some } \mathbf{y} \in \mathbb{R}_+^n\}. \tag{1.3}$$

[1] Many applications and algorithms are formulated and can be derived under β-divergences or from an information geometry perspective Amari (2016). In the case when $\beta = 2$, we recover the Frobenius norm.

Elements in cone$(\mathbf{A})$ are non-negative linear combinations of the columns of $\mathbf{A}$. Note that for the j-th column of $\mathbf{A}$, we have $\mathbf{A}(:, j) = \mathbf{W}\mathbf{H}(:, j)$. Since $\mathbf{W} \geq \mathbf{0}$ and $\mathbf{H} \geq \mathbf{0}$, we have

$$\mathbf{A}(:, j) \in \text{cone}(\mathbf{W}) \subseteq \mathbb{R}_+^m, \tag{1.4}$$

and as a consequence

$$\text{cone}(\mathbf{A}) \subseteq \text{cone}(\mathbf{W}) \subseteq \mathbb{R}_+^m. \tag{1.5}$$

Therefore, the goal is to find a matrix $\mathbf{W}$ whose cone sits in between the two cones cone$(\mathbf{A})$ and $\mathbb{R}_+^m$, while best "enclosing" cone$(\mathbf{A})$.

This geometric intuition, fitting a cone within a larger orthant while successfully enclosing the data, naturally leads to a critical question: is there a single, unique cone that satisfies these conditions? As we will explore next, NMF generally suffers from severe non-uniqueness, making the study of its identifiability a deeply constrained problem.

1.4 Identifiability issue

In unconstrained two-block MF, the identifiability/non-uniqueness of the solutions arises from two places: rotational ambiguity and scaling ambiguity. Specifically, given $\mathbf{A} \in \mathbb{R}^{m \times n}$, we are looking for a factorization consisting of products of an unconstrained pair: $\mathbf{X} \in \mathbb{R}^{m \times r}$ and $\mathbf{Y} \in \mathbb{R}^{r \times n}$ so that

$$\mathbf{A} = \mathbf{X}\mathbf{Y}. \tag{1.6}$$

Suppose $\mathbf{X}_*$ and $\mathbf{Y}_*$ is an exact pair of factorizations, then:

- $(\mathbf{X}\mathbf{Q}, \mathbf{Q}^{-1}\mathbf{Y})$ is another valid solution pair for any invertible matrix $\mathbf{Q} \in \mathbb{R}^{r \times r}$.
- $(c\mathbf{X}_*, c^{-1}\mathbf{Y}_*)$ is another valid solution pair for any non-zero scalar $c \in \mathbb{R} \setminus \{0\}$.

Similar issues happen in NMF, where the ambiguity is given by permutation and scaling.

Definition 1.4.1: Equivalence of two exact NMFs

Let $(\mathbf{W}, \mathbf{H})$ and $(\mathbf{W}', \mathbf{H}')$ be two Exact NMFs of the same $\mathbf{A} \in \mathbb{R}_+^{m \times n}$ with size r, i.e.

$$\mathbf{A} = \mathbf{W}\mathbf{H} = \mathbf{W}'\mathbf{H}'. \tag{1.7}$$

We say they are **equivalent** if there exist a permutation matrix $\mathbf{\Pi} \in \{0, 1\}^{r \times r}$ and a diagonal scaling matrix $\mathbf{D} \in \mathbb{R}^{r \times r}$ with positive diagonal elements such that

$$\mathbf{W}' = \mathbf{W}\mathbf{\Pi}\mathbf{D} \qquad \text{and} \qquad \mathbf{H}' = \mathbf{D}^{-1}\mathbf{\Pi}^\top\mathbf{H}. \tag{1.8}$$

As the name suggests, there can be two different solutions to the same exact NMF problem that are NOT equivalent. Such a pair can be created if the non-negative constraints are "loose" enough to allow us to rotate the interior matrices.

Example 1.4.2: Non-equivalent solutions

Let $r = 3$, and let $\mathbf{W}$ and $\mathbf{H}$ be the following:

$$\mathbf{W} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{H} = \begin{bmatrix} 0.5 & 0.4 & 0.1 \\ 0.1 & 0.5 & 0.4 \\ 0.4 & 0.1 & 0.5 \end{bmatrix}. \tag{1.9}$$

Since $\mathbf{W}$ is the identity matrix, $\mathbf{A} = \mathbf{H}$.
Now, consider the following matrix $\mathbf{Q}$:

$$\mathbf{Q} = \begin{bmatrix} 1 & 0.1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{Q}^{-1} = \begin{bmatrix} 1 & -0.1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{1.10}$$

The second pair can be created and calculated as follows:

$$\mathbf{W}' = \mathbf{W}\mathbf{Q} = \begin{bmatrix} 1 & 0.1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \mathbf{H}' = \mathbf{Q}^{-1}\mathbf{H} = \begin{bmatrix} 0.49 & 0.35 & 0.06 \\ 0.1 & 0.5 & 0.4 \\ 0.4 & 0.1 & 0.5 \end{bmatrix}. \tag{1.11}$$

It is easy to check that $(\mathbf{W}', \mathbf{H}')$ is a valid pair of solutions. $\mathbf{W}$ and $\mathbf{W}'$ are not essentially unique because $\mathbf{W}'$ has 0.1 off the diagonal, and there is no permutation and scaling matrix such that it maps $\mathbf{W}$ to $\mathbf{W}'$ or vice versa.

When the entire solution set of the Exact NMF of fixed input $\mathbf{A}$ and fixed size r contains only one equivalent class, we say it is unique. We state the formal definition here.

Definition 1.4.3: (Essential) Uniqueness of Exact NMF (Definition 4.1 in Gillis (2020))

The Exact NMF $(\mathbf{W}, \mathbf{H})$ of $\mathbf{A} \in \mathbb{R}_+^{m \times n}$ of size r is said to be **(essentially) unique** if and only if any other Exact NMF $(\mathbf{W}', \mathbf{H}')$ of the same $\mathbf{A}$ of size r is equivalent with $(\mathbf{W}, \mathbf{H})$.

1.4.1 Uniqueness guarantee

Exact NMF is, in general, highly non-unique and non-identifiable. However, essential uniqueness holds under structured assumptions. We highlight a few conditions in which the answer becomes positive.

The case when rank$(\mathbf{A}) \leq 2$

We have the following guarantee:

Theorem 1.4.4: Uniqueness under rank$(\mathbf{A}) \leq 2$ (Theorem 4.2 in Gillis (2020))

Let $\mathbf{A} \in \mathbb{R}_+^{m \times n}$ with $\text{rank}(\mathbf{A}) = 2$. The exact NMF of $\mathbf{A}$ of size 2 is unique if and only if (**iff**) $\mathbf{A}$ contains a 2 by 2 diagonal matrix with positive diagonal elements as a sub-matrix

(which is not required to be a minor). Concretely, there exist indices i, j, k, l such that

$$\mathbf{A}\left(\{i,j\},\{k,l\}\right) = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}, \tag{1.12}$$

for some $\alpha > 0$ and $\beta > 0$.

Roughly speaking, if the rank is 2, normalizing the columns of $\mathbf{A}$ projects the data points onto a one-dimensional line segment. Again, finding an exact NMF means finding a bounding segment (defined by the two columns of $\mathbf{W}$) that completely covers the data points while fitting entirely inside the non-negative orthant.

If $\mathbf{A}$ contains a 2 by 2 diagonal sub-matrix, it means the dataset contains two "pure" extreme points that sit exactly on the boundaries (i.e. the axes) of the non-negative orthant. In this case, the bounding segments cannot be pushed inwards or outwards, hence guaranteeing the uniqueness of the solution.

The case for $r <$ rank$(\mathbf{A})$ and $r >$ rank$(\mathbf{A})$

The former case is the one we usually perform (approximate) low-rank factorization with. However, in such a regime, exact NMF would be impossible, and this is the result due to submultiplicativity of the rank function. To see this:

$$\text{rank}(\mathbf{A}) = \text{rank}(\mathbf{WH}) \leq \min\{\text{rank}(\mathbf{W}), \text{rank}(\mathbf{H})\}. \tag{1.13}$$

Now, obviously the rank of a matrix cannot exceed its smallest dimension (and here recall that we have assumed $r \leq m, n$), and so

$$\text{rank}(\mathbf{W}) \leq r \qquad \text{and} \qquad \text{rank}(\mathbf{H}) \leq r. \tag{1.14}$$

This yields an obvious contradiction since $r < \text{rank}(\mathbf{A}) \leq r$.

In the overparametrized regime $r >$ rank$(\mathbf{A})$, the exact NMF is trivially non-unique. One could create an example like the following: suppose $\mathbf{A} \in \mathbb{R}_+^{m\times n}$ has an exact NMF of size $r_0 = \text{rank}(\mathbf{A})$, that is, there exists $\mathbf{W}_0 \in \mathbb{R}^{m\times r_0}$ and $\mathbf{H}_0 \in \mathbb{R}_+^{r_0\times n}$ such that

$$\mathbf{A} = \mathbf{W}_0\mathbf{H}_0 = \sum_{i=1}^{r_0} \mathbf{W}_0(:,i)\mathbf{H}_0(i,:). \tag{1.15}$$

Note that from here we can construct a continuum of exact NMF solutions of size $r = r_0 + 1$. We do so by artificially splitting any single rank-1 component from our base case into two separate components.

Let $\theta \in (0,1)$, and we can re-write the first component as

$$\mathbf{W}_0(:,1)\mathbf{H}_0(1,:) = \left[\theta\mathbf{W}_0(:,1)\right]\mathbf{H}_0(1,:) + \left[(1-\theta)\mathbf{W}_0(:,1)\right]\mathbf{H}_0(1,:). \tag{1.16}$$

Via this trick, we can augment $\mathbf{W}_0$ and $\mathbf{H}_0$ with another dimension:

$$\mathbb{R}_+^{m\times(r_0+1)} \ni \mathbf{W}(\theta) \triangleq \left[\theta\mathbf{W}_0(:,1) \quad (1-\theta)\mathbf{W}_0(:,1) \quad \mathbf{W}_0(:,2) \quad \cdots \quad \mathbf{W}(:,r_0)\right], \tag{1.17}$$

$$\mathbb{R}_+^{(r_0+1)\times n} \ni \mathbf{H} \triangleq \begin{bmatrix} \mathbf{H}_0(1,:) \\ \mathbf{H}_0(1,:) \\ \mathbf{H}_0(2,:) \\ \vdots \\ \mathbf{H}_0(r_0,:) \end{bmatrix}. \tag{1.18}$$

One can check that $\mathbf{W}(\theta)$ and $\mathbf{H}$ are a family of valid exact NMF solutions of size $r_0 + 1$, which is not unique.

Uniqueness under $r = \mathbf{rank}(\mathbf{A})$

In the rank- 2 case, we require $\mathbf{A}$ to contain a 2 by 2 diagonal matrix with positive diagonal. The same reasoning holds for the general case where we require the matrix to contain the identity as a sub-matrix (up to scaling and permutation). Rigorously, this translates to the following separability condition.

Definition 1.4.5: Separability (Definition 4.10 in Gillis (2020))

The matrix $\mathbf{H} \in \mathbb{R}_+^{r\times n}$ is separable if $\text{cone}(\mathbf{H}) = \mathbb{R}_+^r$.

Theorem 1.4.6: Uniqueness under separability (Theorem 4.12 in Gillis (2020))

If $\mathbf{A}$ admits an exact NMF $(\mathbf{W}, \mathbf{H})$ of size $r = \text{rank}(\mathbf{A})$ where $\mathbf{W}^\top$ and $\mathbf{H}$ are separable, then the exact NMF $(\mathbf{W}, \mathbf{H})$ of $\mathbf{A}$ is unique.[a]

[a] It is also worth noting that we need both factors to be separable to guarantee uniqueness.

Again, separability is a very demanding condition. The Sufficiently Scattered Condition (SSC) is a more relaxed condition that only requires the data to be "spread out" enough.

Definition 1.4.7: Sufficiently Scattered Condition (Definition 4.15 in Gillis (2020))

The matrix $\mathbf{H} \in \mathbb{R}_+^{r\times n}$ is sufficiently scattered if the following two conditions are satisfied:

- **SSC1**: $\mathcal{C} = \{\mathbf{x} \in \mathbb{R}_+^r \mid \mathbf{e}^\top \mathbf{x} \geq \sqrt{r-1}\|\mathbf{x}\|_2\} \subseteq \text{cone}(\mathbf{H})$.
- **SSC2**: There does not exist any orthogonal matrix $\mathbf{Q}$ such that $\text{cone}(\mathbf{H}) \subseteq \text{cone}(\mathbf{Q})$, except for permutation matrices.

SSC1 requires the canonical hull of the data to be wide enough to engulf a specific second order cone $\mathcal{C}$ defined as above. Since this reference cone $\mathcal{C}$ is tangent to every single facet of the non-negative orthant, **SSC1** implicitly forces the data to have some degree of sparsity to push up against these boundaries. It is also worth noting that checking the **SSC1** condition is NP-hard.

SSC2 is a geometric and regularity check. It ensures the cone generated by the data is robust enough that it cannot be trivially enclosed by an arbitrary orthogonal rotation.

When both factors satisfy the SSC, their mutual geometric spread is robust enough to eliminate any rotational and translational slack, hence guaranteeing uniqueness.

Theorem 1.4.8: Uniqueness under SSC (Theorem 4.21 in Gillis (2020))

If $\mathbf{W}^\top$ and $\mathbf{H}$ are sufficiently scattered, then the exact NMF $(\mathbf{W}, \mathbf{H})$ of $\mathbf{A}$ of size $r = \text{rank}(\mathbf{A})$ is unique.

While understanding the theoretical limits of identifiability and uniqueness is crucial, the practical utility of NMF hinges on our ability to actually compute these factorizations efficiently. Exact NMF is generally NP-hard, forcing us to rely on approximate, iterative optimization techniques. The next section provides a comprehensive tour of the most prominent iterative algorithms designed to tackle the non-convex NMF objective.

1.5 Zoo of iterative algorithms

Here, we will do a quick tour of iterative methods to solve NMF. Instead of proving the convergence and iteration complexity rigorously, we will only derive most of the methods heuristically. **In addition, since NMF objective is globally non-convex, most of the existing complexities for these algorithms in the convex regime do not carry over directly to the non-convex regime.**

The most ad-hoc methods for NMF include Multiplicative Update (**MU**) Lee and Seung (2000) and Hierarchical Alternating Least Squares (HALS) Cichocki et al. (2007). The former can be viewed either under the context of Majorization-Minimization (**MM**) or as a form of (entrywise) scaled/adaptive GD. The latter can be viewed as just a special case of block coordinate descent.

1.5.1 MU

Again, the goal is to minimize

$$\begin{aligned} F(\mathbf{W}, \mathbf{H}) &= \frac{1}{2}\|\mathbf{A} - \mathbf{W}\mathbf{H}\|_F^2 \\ &= \frac{1}{2}\sum_{i,j}(\mathbf{A}_{ij} - (\mathbf{W}\mathbf{H})_{ij})^2, \end{aligned} \tag{1.19}$$

under the constraints $\mathbf{W} \geq \mathbf{0}$ and $\mathbf{H} \geq \mathbf{0}$.

Now, to make taking the derivative easier, it helps to rewrite the Frobenius norm using the trace, where $\|\mathbf{A}\|_F^2 = \text{Tr}(\mathbf{A}\mathbf{A}^\top)$:

$$\begin{aligned} F(\mathbf{W}, \mathbf{H}) &= \frac{1}{2}\text{Tr}\left[(\mathbf{A} - \mathbf{W}\mathbf{H})(\mathbf{A} - \mathbf{W}\mathbf{H})^\top\right] \\ &= \frac{1}{2}\text{Tr}\left(\mathbf{A}\mathbf{A}^\top - \mathbf{A}\mathbf{H}^\top\mathbf{W}^\top - \mathbf{W}\mathbf{H}\mathbf{A}^\top + \mathbf{W}\mathbf{H}\mathbf{H}^\top\mathbf{W}^\top\right) \\ &= \frac{1}{2}\text{Tr}(\mathbf{A}\mathbf{A}^\top) - \text{Tr}(\mathbf{W}\mathbf{H}\mathbf{A}^\top) + \frac{1}{2}\text{Tr}(\mathbf{W}\mathbf{H}\mathbf{H}^\top\mathbf{W}^\top). \end{aligned} \tag{1.20}$$

The gradients can be easily calculated as follows:

$$\nabla_{\mathbf{W}} F = -\mathbf{A}\mathbf{H}^\top + \mathbf{W}\mathbf{H}\mathbf{H}^\top \qquad \text{and} \qquad \nabla_{\mathbf{H}} F = -\mathbf{W}^\top \mathbf{A} + \mathbf{W}^\top \mathbf{W}\mathbf{H}. \tag{1.21}$$

Now, suppose for now we are only updating the $\mathbf{W}$ block; then the standard GD takes the form

$$\mathbf{W} \leftarrow \mathbf{W} - \eta_{\mathbf{W}} \nabla F = \mathbf{W} + \eta_{\mathbf{W}}(\mathbf{A}\mathbf{H}^\top - \mathbf{W}\mathbf{H}\mathbf{H}^\top), \tag{1.22}$$

and the most obvious issue with using a fixed learning rate $\eta_{\mathbf{W}}$ is that it can cause the elements of $\mathbf{W}$ to become negative and hence violate the non-negativity constraints. One fix is to use an (entrywise) adaptive learning rate so that the entrywise non-negativity constraints are respected, and this leads to one of the perspectives for the MU method.

Derivation using the MM framework

The analysis here follows Lee and Seung (2000). For the sake of simplicity, we consider optimizing a single column of $\mathbf{H} \in \mathbb{R}^{d \times n}$:

- Let $\mathbf{a} \in \mathbb{R}^m_+$ be a single column of the data matrix $\mathbf{A}$.
- Let $\mathbf{h} \in \mathbb{R}^d_{++}$ be the corresponding column of $\mathbf{H}$. Here we assume strict positivity to prevent division by zero in the update rules.
- Let $\mathbf{W} \in \mathbb{R}^{m \times d}_+$ be the non-negative weight matrix.
- Let $\mathbf{M} = \mathbf{W}^\top \mathbf{W} \in \mathbb{R}^{d \times d}$, because $\mathbf{W}_{ia} \geq \mathbf{0}$ for all i, a; and it follows that $\mathbf{M}_{ab} = \sum_{i=1}^m \mathbf{W}_{ia}\mathbf{W}_{ib} \geq 0$ for all a, b.

Therefore, the exact objective for this column is

$$f(\mathbf{h}) = \frac{1}{2}\|\mathbf{a} - \mathbf{W}\mathbf{h}\|_2^2 = \frac{1}{2}\mathbf{a}^\top \mathbf{a} - \mathbf{a}^\top \mathbf{W}\mathbf{h} + \frac{1}{2}\mathbf{h}^\top \mathbf{M}\mathbf{h}. \tag{1.23}$$

The gradient of $f(\mathbf{h})$ with respect to (**w.r.t.**) $\mathbf{h}$ is

$$\nabla f(\mathbf{h}) = -\mathbf{W}^\top \mathbf{a} + \mathbf{M}\mathbf{h}. \tag{1.24}$$

Definition 1.5.1: Majorization function

A function $G(\mathbf{h}, \mathbf{h}^t)$ is a majorization function for $f(\mathbf{h})$ if it satisfies the following conditions for all $\mathbf{h}, \mathbf{h}^t \in \mathbb{R}^d_+$:

- The global upper bound: $G(\mathbf{h}, \mathbf{h}^t) \geq f(\mathbf{h})$.
- Tightness at the fixed point: $G(\mathbf{h}^t, \mathbf{h}^t) = f(\mathbf{h}^t)$.

Here we treat $\mathbf{h}$ as the free variable and $\mathbf{h}^t$ as the fixed parameter.

Lemma 1.5.2: The descent guarantee

If $G(\mathbf{h}, \mathbf{h}^t)$ is a majorization function for $f(\mathbf{h})$, then $f(\mathbf{h})$ is non-increasing under the update rule $\mathbf{h}^{t+1} = \arg\min_{\mathbf{h}} G(\mathbf{h}, \mathbf{h}^t)$.

Proof:

By the definition of the minimum, $G(\mathbf{h}^{t+1}, \mathbf{h}^t) \leq G(\mathbf{h}^t, \mathbf{h}^t)$.

By the properties of the auxiliary function:

$$f(\mathbf{h}^{t+1}) \leq G(\mathbf{h}^{t+1}, \mathbf{h}^t) \leq G(\mathbf{h}^t, \mathbf{h}^t) = f(\mathbf{h}^t). \tag{1.25}$$

Therefore, $f(\mathbf{h}^{t+1}) \leq f(\mathbf{h}^t)$. This establishes monotonic non-increase. ■

Now the question that remains is how to construct a proper auxiliary majorization function for our $f(\mathbf{h})$.

Lemma 1.5.3: Construction of the majorization function

Define a diagonal matrix $K(\mathbf{h}^t) \in \mathbb{R}^{d \times d}$ such that its diagonal elements are:

$$K_{aa}(\mathbf{h}^t) = \frac{(\mathbf{M}\mathbf{h}^t)_a}{\mathbf{h}_a^t}. \tag{1.26}$$

The function $G(\mathbf{h}, \mathbf{h}^t)$ defined below is a majorization function for $f(\mathbf{h})$:

$$G(\mathbf{h}, \mathbf{h}^t) = f(\mathbf{h}^t) + (\mathbf{h} - \mathbf{h}^t)^\top \nabla f(\mathbf{h}^t) + \frac{1}{2}(\mathbf{h} - \mathbf{h}^t)^\top K(\mathbf{h}^t)(\mathbf{h} - \mathbf{h}^t). \tag{1.27}$$

Proof:

First, we verify the equality condition. If $\mathbf{h} = \mathbf{h}^t$, the second and the third terms become zero, yielding $G(\mathbf{h}, \mathbf{h}) = f(\mathbf{h})$.

Second, we show $G(\mathbf{h}, \mathbf{h}^t) \geq f(\mathbf{h})$: the Taylor expansion of the quadratic function $f(\mathbf{h})$ around $\mathbf{h}^t$ is

$$f(\mathbf{h}) = f(\mathbf{h}^t) + (\mathbf{h} - \mathbf{h}^t)^\top \nabla f(\mathbf{h}^t) + \frac{1}{2}(\mathbf{h} - \mathbf{h}^t)^\top \mathbf{M}(\mathbf{h} - \mathbf{h}^t). \tag{1.28}$$

Comparing $G(\mathbf{h}, \mathbf{h}^t)$ and $f(\mathbf{h})$, the inequality $G(\mathbf{h}, \mathbf{h}^t) \geq f(\mathbf{h})$ holds iff:

$$(\mathbf{h} - \mathbf{h}^t)^\top K(\mathbf{h}^t)(\mathbf{h} - \mathbf{h}^t) \geq (\mathbf{h} - \mathbf{h}^t)^\top \mathbf{M}(\mathbf{h} - \mathbf{h}^t). \tag{1.29}$$

This requires proving the matrix $K(\mathbf{h}^t) - \mathbf{M}$ is positive semi-definite (**p.s.d.**), which amounts to showing

$$\mathbf{x}^\top K(\mathbf{h}^t)\mathbf{x} \geq \mathbf{x}^\top \mathbf{M}\mathbf{x}. \tag{1.30}$$

Expanding $\mathbf{x}^\top K(\mathbf{h}^t)\mathbf{x}$:

$$\mathbf{x}^\top K(\mathbf{h}^t)\mathbf{x} = \sum_{a=1}^{d} \mathbf{x}_a^2 K_{aa}(\mathbf{h}^t) = \sum_{a=1}^{d} \mathbf{x}_a^2 \frac{(\mathbf{M}\mathbf{h}^t)_a}{\mathbf{h}_a^t} = \sum_{a=1}^{d}\sum_{b=1}^{d} \mathbf{M}_{ab}\mathbf{h}_b^t \frac{\mathbf{x}_a^2}{\mathbf{h}_a^t}. \tag{1.31}$$

$\mathbf{M}$ is symmetric, we rewrite the sum by pairing symmetric terms and split the sum into two halves and swap the dummy indices a and b on the second half:

$$\sum_{a=1}^{d}\sum_{b=1}^{d} \mathbf{M}_{ab}\frac{\mathbf{h}_b^t}{\mathbf{h}_a^t}\mathbf{x}_a^2 = \sum_{a=1}^{d}\sum_{b=1}^{d} \mathbf{M}_{ab}\left(\frac{1}{2}\frac{\mathbf{h}_b^t}{\mathbf{h}_a^t}\mathbf{x}_a^2 + \frac{1}{2}\frac{\mathbf{h}_a^t}{\mathbf{h}_b^t}\mathbf{x}_b^2\right). \tag{1.32}$$

By the inequality $A^2 + B^2 \geq 2AB$, we can show that

$$\frac{\mathbf{h}_b^t}{\mathbf{h}_a^t}\mathbf{x}_a^2 + \frac{\mathbf{h}_a^t}{\mathbf{h}_b^t}\mathbf{x}_b^2 \geq 2\mathbf{x}_a\mathbf{x}_b, \tag{1.33}$$

with $A \triangleq \frac{\mathbf{h}_b^t}{\mathbf{h}_a^t}\mathbf{x}_a^2$ and $B \triangleq \frac{\mathbf{h}_a^t}{\mathbf{h}_b^t}\mathbf{x}_b^2$.

Multiplying the non-negative constant $\frac{1}{2}\mathbf{M}_{ab}$ on both sides, we have

$$\mathbf{M}_{ab}\left(\frac{1}{2}\frac{\mathbf{h}_b^t}{\mathbf{h}_a^t}\mathbf{x}_a^2 + \frac{1}{2}\frac{\mathbf{h}_a^t}{\mathbf{h}_b^t}\mathbf{x}_b^2\right) \geq \mathbf{M}_{ab}\mathbf{x}_a\mathbf{x}_b. \tag{1.34}$$

Summing this inequality over all a and b gives:

$$\mathbf{x}^\top K(\mathbf{h}^t)\mathbf{x} \geq \sum_{a=1}^{d}\sum_{b=1}^{d}\mathbf{M}_{ab}\mathbf{x}_a\mathbf{x}_b = \mathbf{x}^\top\mathbf{M}\mathbf{x}. \tag{1.35}$$

This shows that $K(\mathbf{h}^t) - \mathbf{M}$ is p.s.d. Therefore $G(\mathbf{h}, \mathbf{h}^t) \geq f(\mathbf{h})$. ■

Now we are able to derive the MU rule: to find $\mathbf{h}^{t+1} = \arg\min_{\mathbf{h}} G(\mathbf{h}, \mathbf{h}^t)$, we set the gradient of G w.r.t. $\mathbf{h}$ to zero:

$$\nabla_{\mathbf{h}} G(\mathbf{h}, \mathbf{h}^t) = \nabla f(\mathbf{h}^t) + K(\mathbf{h}^t)(\mathbf{h} - \mathbf{h}^t) = \mathbf{0}. \tag{1.36}$$

Solving for $\mathbf{h}$:

$$K(\mathbf{h}^t)\mathbf{h} = K(\mathbf{h}^t)\mathbf{h}^t - \nabla f(\mathbf{h}^t). \tag{1.37}$$

Because $K(\mathbf{h}^t)$ is a diagonal matrix with strictly positive entries, it is invertible. We may isolate a single element $\mathbf{h}_a$:

$$\mathbf{h}_a = \mathbf{h}_a^t - \frac{(\nabla f(\mathbf{h}^t))_a}{K_{aa}(\mathbf{h}^t)}. \tag{1.38}$$

Substitute $\nabla f(\mathbf{h}^t) = -\mathbf{W}^\top\mathbf{a} + \mathbf{M}\mathbf{h}^t$ and $K_{aa}(\mathbf{h}^t) = \frac{(\mathbf{M}\mathbf{h}^t)_a}{\mathbf{h}_a^t}$:

$$\begin{aligned}
\mathbf{h}_a &= \mathbf{h}_a^t - \frac{\mathbf{h}_a^t}{(\mathbf{M}\mathbf{h}^t)_a}\left(-(\mathbf{W}^\top\mathbf{a})_a + (\mathbf{M}\mathbf{h}^t)_a\right) \qquad (1.39)\\
&= \mathbf{h}_a^t + \mathbf{h}_a^t\frac{(\mathbf{W}^\top\mathbf{a})_a}{(\mathbf{M}\mathbf{h}^t)_a} - \mathbf{h}_a^t\\
&= \mathbf{h}_a^t\frac{(\mathbf{W}^\top\mathbf{a})_a}{(\mathbf{M}\mathbf{h}^t)_a}.
\end{aligned}$$

Substitute $\mathbf{M} = \mathbf{W}^\top\mathbf{W}$ back into the original equation to yield the MU for a single column:

$$\mathbf{h}_a = \mathbf{h}_a^t\frac{(\mathbf{W}^\top\mathbf{a})_a}{(\mathbf{W}^\top\mathbf{W}\mathbf{h}^t)_a}. \tag{1.40}$$

Extending this to the full matrix notation, we have:

$$\mathbf{H}^{t+1} = \mathbf{H}^t \odot \frac{\mathbf{W}^\top\mathbf{A}}{\mathbf{W}^\top\mathbf{W}\mathbf{H}^t}. \tag{1.41}$$

A symmetrical derivation yields the update for the $\mathbf{W}$ block:

$$\mathbf{W}^{t+1} = \mathbf{W}^t \odot \frac{\mathbf{A}\mathbf{H}^\top}{\mathbf{W}\mathbf{H}\mathbf{H}^\top}. \tag{1.42}$$

Algorithm 1 MU for NMF

Input: Non-negative target matrix $\mathbf{A} \in \mathbb{R}_+^{m \times n}$, factorization rank r
Output: Converged non-negative factors $\mathbf{W} \in \mathbb{R}_+^{m \times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r \times n}$

1: // Initialization
2: Initialize $\mathbf{W}^0 > \mathbf{0}$ and $\mathbf{H}^0 > \mathbf{0}$

3: // Main Loop
4: **for** $t = 0, 1, 2, \cdots$ **do**
5: $\quad \mathbf{H}^{t+1} \leftarrow \mathbf{H}^t \odot \dfrac{(\mathbf{W}^t)^\top \mathbf{A}}{(\mathbf{W}^t)^\top \mathbf{W}^t \mathbf{H}^t}$
6: $\quad \mathbf{W}^{t+1} \leftarrow \mathbf{W}^t \odot \dfrac{\mathbf{A}(\mathbf{H}^{t+1})^\top}{\mathbf{W}^t \mathbf{H}^{t+1} (\mathbf{H}^{t+1})^\top}$
7: **end for**

1.5.2 HALS: Hierarchical Alternating Least Squares

We can rewrite matrix multiplication $\mathbf{WH}$ as the sum of r rank-one matrices. Let $\mathbf{W}(:, k) \in \mathbb{R}^m$ be the k-th column of $\mathbf{W}$, and $\mathbf{H}(k, :) \in \mathbb{R}^n$ denote the k-th row of $\mathbf{H}$:

$$\mathbf{WH} = \sum_{k=1}^{r} \mathbf{W}(:, k)\mathbf{H}(k, :). \tag{1.43}$$

Define $\mathbf{R}$ as the residual matrix $\mathbf{R} = \mathbf{A} - \mathbf{WH}$, and define its component **EXCLUDING** the k-th component as $\mathbf{R}_k$, i.e.:

$$\mathbf{R}_k = \mathbf{A} - \sum_{j \neq k} \mathbf{W}(:, j)\mathbf{H}(j, :). \tag{1.44}$$

The global objective function, viewed solely as a function of $\mathbf{H}(k, :)$, becomes a local objective $f(\mathbf{H}(k, :))$, i.e.:

$$f(\mathbf{H}(k, :)) = \frac{1}{2}\|\mathbf{R}_k - \mathbf{W}(:, k)\mathbf{H}(k, :)\|_F^2. \tag{1.45}$$

We would like to further decouple the formulation. Again, $\mathbf{H}(k, l)$ is the l-th scalar element of the row vector $\mathbf{H}(k, :)$, and $\mathbf{R}_k(:, l)$ is the l-th column vector of the residual matrix $\mathbf{R}_k$. We may further decompose this local objective by running this down column by column, that is:

$$f(\mathbf{H}(k, :)) = \frac{1}{2}\sum_{l=1}^{n} \|\mathbf{R}_k(:, l) - \mathbf{W}(:, k)\mathbf{H}(k, l)\|_2^{2\,2}. \tag{1.46}$$

Note that since $\mathbf{H}(k, l)$ only appears in the l-th term of the sum, each element of $\mathbf{H}(k, :)$ can be optimized completely independently of the others.

Let's minimize a single term from the sum above w.r.t. the scalar $h \triangleq \mathbf{H}(k, l)$. Denote $\mathbf{r}_l \triangleq \mathbf{R}_k(:, l)$. We would then like to minimize a simple quadratic function $g(h)$:

$$\begin{aligned} g(h) &\triangleq \frac{1}{2}\|\mathbf{r}_l - \mathbf{W}(:, k)h\|_2^2 \\ &= \frac{1}{2}(\mathbf{r}_l^\top \mathbf{r}_l - 2h\mathbf{W}(:, k)^\top \mathbf{r}_l + h^2 \mathbf{W}(:, k)^\top \mathbf{W}(:, k)). \end{aligned} \tag{1.47}$$

[2] Note that the change from Frobenius norm to Euclidean norm is valid here since we are summing column vectors.

Take the first-order derivative and set it to zero:

$$\frac{dg}{dh} = -\mathbf{W}(:,k)^\top \mathbf{r}_l + h\|\mathbf{W}(:,k)\|_2^2 = 0 \implies h = \frac{\mathbf{W}(:,k)^\top \mathbf{r}_l}{\|\mathbf{W}(:,k)\|_2^2}. \tag{1.48}$$

Here we required $h \geq 0$, and since $g(h)$ is a strictly convex parabola (opening upwards, assuming $\|\mathbf{W}(:,k)\| > 0$), the constrained minimum is simply the unconstrained minimum projected onto the non-negative orthant, and therefore the update for $\mathbf{H}(k,l)$ should be

$$\mathbf{H}(k,l)^* = \max\left\{0, \frac{\mathbf{W}(:,k)^\top \mathbf{R}_k(:,l)}{\|\mathbf{W}(:,k)\|_2^2}\right\}. \tag{1.49}$$

Now for the entire row $\mathbf{H}(k,:)$, we simply have:

$$\mathbf{H}(k,:) \leftarrow \max\left\{0, \frac{\mathbf{W}(:,k)^\top \mathbf{R}_k}{\|\mathbf{W}(:,k)\|_2^2}\right\}. \tag{1.50}$$

If we further substitute $\mathbf{R}_k = \mathbf{A} - \mathbf{W}\mathbf{H} + \mathbf{W}(:,k)\mathbf{H}(k,:)$:

$$\mathbf{H}(k,:) \leftarrow \max\left\{\mathbf{0}, \mathbf{H}(k,:) + \frac{\mathbf{W}(:,k)^\top \mathbf{A} - (\mathbf{W}(:,k)^\top \mathbf{W})\mathbf{H}}{\|\mathbf{W}(:,k)\|_2^2}\right\}. \tag{1.51}$$

The update for the k-th column of $\mathbf{W}$ can be derived in the same manner:

$$\mathbf{W}(:,k) \leftarrow \max\left\{\mathbf{0}, \mathbf{W}(:,k) + \frac{\mathbf{A}\mathbf{H}(k,:)^\top - \mathbf{W}(\mathbf{H}\mathbf{H}(k,:)^\top)}{\|\mathbf{H}(k,:)\|_2^2}\right\}. \tag{1.52}$$

Algorithm 2 Hierarchical Alternating Least Squares (HALS)

Input: Non-negative target matrix $\mathbf{A} \in \mathbb{R}_+^{m\times n}$, factorization rank r
Output: Converged non-negative factors $\mathbf{W} \in \mathbb{R}_+^{m\times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r\times n}$

// Initialization
Initialize $\mathbf{W} \geq \mathbf{0}$ and $\mathbf{H} \geq \mathbf{0}$

// Main Loop
repeat
  for $k = 1, 2, \cdots, r$ **do**
    $\mathbf{H}(k,:) \leftarrow \max\left\{\mathbf{0}, \mathbf{H}(k,:) + \frac{\mathbf{W}(:,k)^\top \mathbf{A} - (\mathbf{W}(:,k)^\top \mathbf{W})\mathbf{H}}{\|\mathbf{W}(:,k)\|_2^2}\right\}$
  end for
  for $k = 1, 2, \cdots, r$ **do**
    $\mathbf{W}(:,k) \leftarrow \max\left\{\mathbf{0}, \mathbf{W}(:,k) + \frac{\mathbf{A}\mathbf{H}(k,:)^\top - \mathbf{W}(\mathbf{H}\mathbf{H}(k,:)^\top)}{\|\mathbf{H}(k,:)\|_2^2}\right\}$
  end for
until convergence

1.5.3 PGD, AltPGD, and PALM

Inexact and exact Block Coordinate Descent

HALS as we have seen is a special case of the (Gauss-Seidel/cyclic) Block Coordinate Descent Method. BCD is a decomposition method that simplifies multivariable optimization by updating

a subset of variables at each step while holding the remaining blocks fixed. Specifically, we are interested in minimizing the following composite function:

$$F(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p) = f(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p) + \sum_{i=1}^{p} g_k(\mathbf{x}_k), \tag{1.53}$$

where for all $k \in [p]$, $\mathbf{x}_k \in \mathbb{E}_k$. The mappings f and g are real valued, i.e. $f : \mathbb{E}_1 \times \cdots \times \mathbb{E}_p \to \mathbb{R}$ and $g_k : \mathbb{E}_k \to \mathbb{R}$. Here $\mathbb{E}_k$ denotes the Euclidean space of dimension $d_k \geq 1$ for all $k \in [p]$.

We further define the function $g : \mathbb{E}_1 \times \cdots \times \mathbb{E}_p \to \mathbb{R}$ as

$$g(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p) \equiv \sum_{i=1}^{p} g_i(\mathbf{x}_i). \tag{1.54}$$

Let $\nabla_i f$ denote the gradient w.r.t. the i-th block, whenever f is differentiable, and the total gradient of f is therefore

$$\nabla f(\mathbf{x}) = \left(\nabla_1 f(\mathbf{x}), \nabla_2 f(\mathbf{x}), \cdots, \nabla_p f(\mathbf{x})\right). \tag{1.55}$$

We will need to isolate and analyze the change w.r.t. a single block. To this end, we define the linear transformation that projects a single block to the whole ambient space. That is $\mathcal{U}_i : \mathbb{E}_i \to \mathbb{E}$, where

$$\mathcal{U}_i(\mathbf{d}) = (\mathbf{0}, \cdots, \mathbf{0}, \underbrace{\mathbf{d}}_{i-\text{th block}}, \mathbf{0}, \cdots, \mathbf{0}) \qquad \text{where } \mathbf{d} \in \mathbb{E}_i. \tag{1.56}$$

With the shorthand notation $\mathbf{x} \equiv (\mathbf{x}_1; \cdots; \mathbf{x}_p)$ and $\mathbb{E} \equiv \mathbb{E}_1 \times \cdots \times \mathbb{E}_p$, our problem becomes

$$\min_{\mathbf{x} \in \mathbb{E}} F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x}). \tag{1.57}$$

Denote by $\mathbf{x}^t = (\mathbf{x}_1^t, \mathbf{x}_2^t, \cdots, \mathbf{x}_p^t)$ the iterate at time t, the BCD method then updates the $(t+1)$-th iterate $\mathbf{x}^{t+1}$ by updating the p components sequentially. That is, within the $(t+1)$-th iteration, we have p-sub-iterations: at the k-th sub-iteration, where $k \in [p]$, we update the k-th component of the k-th iterate $\mathbf{x}^t$. For a given $\mathbf{x}^t$, denote by $\mathbf{x}^{t,k}$ the k-th sub-iterate when updating from $\mathbf{x}^t$ to $\mathbf{x}^{t+1}$. That is

$$\mathbf{x}^{t,k} = (\mathbf{x}_1^{t+1}, \mathbf{x}_2^{t+1}, \cdots, \mathbf{x}_{k-1}^{t+1}, \mathbf{x}_k^{t+1}, \mathbf{x}_{k+1}^{t}, \cdots, \mathbf{x}_p^{t}) \qquad \text{for all } k \in [p], \tag{1.58}$$

where $\mathbf{x}_i^{t,k}$ is the i-th sub-block of the iterate $\mathbf{x}^{t,k}$. Therefore, at the k-th sub-iteration, we are only updating the variable $\mathbf{x}_k^{t,k}$. Then $\mathbf{x}^{t,0} \equiv \mathbf{x}^t$ and $\mathbf{x}^{t,p} \equiv \mathbf{x}^{t+1}$.

There are two types of BCD methods: if we are seeking the **absolute minimum** of the objective at every sub-iteration, this will be **exact BCD**. That is:

$$\mathbf{x}_i^{t+1} \in \arg\min_{\mathbf{x}_i \in \mathbb{E}_i} F(\mathbf{x}_1^{t+1}, \cdots, \mathbf{x}_{i-1}^{t+1}, \mathbf{x}_i, \mathbf{x}_{i+1}^{t}, \cdots, \mathbf{x}_p^{t}). \tag{1.59}$$

Inexact BCD seeks to minimize some surrogate model (as in the case of MM) of F as we will see in the following. We will need the following assumptions and definitions:

Assumption 1.5.4: Inexact BCD (Assumption 11.1 in Beck (2017))

(A). $g_i : \mathbb{E}_i \to (-\infty, \infty]$ is proper, closed, and convex for all $i \in [p]$.

(B). $f : \mathbb{E} \to (-\infty, \infty]$ is proper and closed, and $\text{dom}(f)$ is convex; $\text{dom}(g) \subseteq \text{int}\,(\text{dom}(f))$, and f is differentiable over $\text{int}(\text{dom}(f))$.

(C). f is L_f-smooth over $\text{int}\,(\text{dom}(f))$, where $L_f > 0$[a].

(D). There exist $L_1, L_2, \cdots, L_p > 0$ such that for any $i \in \{1, 2, \cdots, p\}$ it holds that

$$\|\nabla_i f(\mathbf{x}) - \nabla_i f(\mathbf{x} + \mathcal{U}_i(\mathbf{d}))\| \leq L_i \|\mathbf{d}\|, \tag{1.60}$$

for all $\mathbf{x} \in \text{int}(\text{dom}(f))$, and $\mathbf{d} \in \mathbb{E}_i$ for which $\mathbf{x} + \mathcal{U}_i(\mathbf{d}) \in \text{int}\,(\text{dom}(f))$.

(E). The optimal set of the problem is nonempty and denoted by X^*. The optimal value is denoted by F_*.

[a] Here, L_f is the global Lipschitz constant. The block-wise Lipschitz constant L_i's can be significantly smaller than L_f in practice.

Definition 1.5.5: Partial prox-gradient step (Definition 11.3 in Beck (2017))

Suppose that f and $g_1, g_2, \cdots, g_p$ satisfy properties (A) and (B) of Assumption 11.1, $L > 0$, and let $i \in \{1, 2, \cdots, p\}$. Then the i-th prox-grad step is the operator $T_L^i : \text{int}(\text{dom}(f)) \to \mathbb{R}^{d_i}$ defined by

$$T_L^i = \mathbf{prox}_{\frac{1}{L} g_i} \left(\mathbf{x}_i - \frac{1}{L} \nabla_i f(\mathbf{x}) \right). \tag{1.61}$$

For fundamentals of proximal operators, readers may refer to Bauschke and Combettes (2017). In light of operator splitting methods, T_L^i may also be viewed as the forward-backward operator.

Algorithm 3 Inexact BCD (The Cyclic Block Proximal Gradient Method, Beck (2017))

Input: Objective function $F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$, block Lipschitz constants $\{L_k\}_{k=1}^p$, number of blocks p

Output: Converged solution $\mathbf{x}^*$

// Initialization
Pick $\mathbf{x}^0 = (\mathbf{x}_1^0, \mathbf{x}_2^0, \cdots, \mathbf{x}_p^0) \in \text{int}(\text{dom}(f))$

// Main Loop
for $t = 0, 1, 2, \cdots$ **do**
    $\mathbf{x}^{t,0} \leftarrow \mathbf{x}^t$
    for $k = 1, 2, \cdots, p$ **do**
        $\mathbf{x}^{t,k} \leftarrow \mathbf{x}^{t,k-1} + \mathcal{U}_k(T_{L_k}^k(\mathbf{x}^{t,k-1}) - \mathbf{x}_k^{t,k-1})$
    end for
    $\mathbf{x}^{t+1} \leftarrow \mathbf{x}^{t,p}$
end for

To put it in context, the partial gradient step evaluated at $\mathbf{x}^{t,k-1}$ is:

$$T_{L_k}^k(\mathbf{x}^{t,k-1}) = \mathbf{prox}_{\frac{1}{L_k} g_k} \left(\mathbf{x}_k^{t,k-1} - \frac{1}{L} \nabla_k f(\mathbf{x}^{t,k-1}) \right). \tag{1.62}$$

One may expand the **prox** operator using standard techniques such as completing the squares, and plug it back into the update expression for $\mathbf{x}^{t,k}$ to get:

$$\mathbf{x}_k^{t,k} \in \arg\min_{\mathbf{y}\in\mathbb{E}_k} \left\{ \underbrace{\langle \nabla_k f(\mathbf{x}^{t,k-1}), \mathbf{y} - \mathbf{x}_k^{t,k-1} \rangle}_{\text{Linearization of } f} + \underbrace{\frac{L_k}{2}\|\mathbf{y} - \mathbf{x}_k^{t,k-1}\|^2}_{\text{Proximity Penalty}} + \underbrace{g_k(y)}_{\text{Regularizer}} \right\}. \tag{1.63}$$

Theorem 1.5.6: Complexity for inexact BCD (Theorem 11.18 in Beck (2017))

Let $\{\mathbf{x}^t\}_{t\geq 0}$ be the sequence generated by the Inexact BCD (Cyclic Block Proximal Gradient) method. Then for any $k \geq 2$

$$F(\mathbf{x}^t) - F_{\text{opt}} \leq \max\left\{ \left(\frac{1}{2}\right)^{(t-1)/2} (F(\mathbf{x}^0) - F_{\text{opt}}), \frac{8p(L_f + L_{\max})^2 R^2}{L_{\min}(t-1)} \right\}, \tag{1.64}$$

where $L_{\min} = \min_{i\in[p]} L_i$ and $L_{\max} = \max_{i\in[p]} L_i$, and $R = R_{F(\mathbf{x}^0)}$.
In addition, if $n \geq 2$ satisfies

$$n \geq N_{\text{Inexact}} \triangleq \max\left\{ \frac{2}{\log(2)} \left[\log(F(\mathbf{x}^0) - F_{\text{opt}}) + \log\left(\frac{1}{\varepsilon}\right) \right], \frac{8p(L_f + L_{\max})^2 R^2}{L_{\min}\varepsilon} \right\} + 1, \tag{1.65}$$

we can achieve $F(\mathbf{x}^n) - F_{\text{opt}} \leq \varepsilon$.

Algorithm 4 Exact BCD (The Alternating Minimization Method, Beck (2017))

Input: Objective function $F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$, number of blocks p
Output: Converged solution $\mathbf{x}^*$

// Initialization
Pick $\mathbf{x}^0 = (\mathbf{x}_1^0, \mathbf{x}_2^0, \cdots, \mathbf{x}_p^0) \in \text{dom}(F)$

// Main Loop
for $k = 0, 1, 2, \cdots$ **do**
  for $i = 1, 2, \cdots, p$ **do**
    $\mathbf{x}_i^{k+1} \in \arg\min_{\mathbf{x}_i\in\mathbb{E}_i} F(\mathbf{x}_1^{k+1}, \cdots, \mathbf{x}_{i-1}^{k+1}, \mathbf{x}_i, \mathbf{x}_{i+1}^k, \cdots, \mathbf{x}_p^k)$
  end for
end for

Assumption 1.5.7: Exact BCD (Assumption 14.10 in Beck (2017))

(A). $g_i \in \mathbb{E}_i \to (-\infty, \infty]$ is proper, closed, and convex for any $i \in [p]$.

(B). $f : \mathbb{E} \to \mathbb{R}$ is convex and L_f-smooth.

(C). For any $\alpha > 0$, there exists $R_\alpha > 0$ such that

$$\max_{\mathbf{x},\mathbf{x}^*\in\mathbb{E}} \{\|\mathbf{x} - \mathbf{x}^*\| : F(\mathbf{x}) \leq \alpha, \mathbf{x}^* \in X^*\} \leq R_\alpha. \tag{1.66}$$

(D). The optimal set of the problem X^* is nonempty. The optimal value is denoted by F_{opt}.

Theorem 1.5.8: Complexity of exact BCD (Theorem 14.11 in Beck (2017))

Suppose that Assumption 1.5.7 holds, and let $\{\mathbf{x}^t\}_{t\geq 0}$ be the sequence generated by the alternating minimization method. Then for all $t \geq 0$:

$$F(\mathbf{x}^t) - F_{\text{opt}} \leq \max\left\{\left(\frac{1}{2}\right)^{(t-1)/2}(F(\mathbf{x}^0) - F_{\text{opt}}), \frac{8L_f p^2 R^2}{t-1}\right\}, \tag{1.67}$$

where $R = R_{F(\mathbf{x}^0)}$.

By standard arguments, one may explicitly calculate [3] the lower bound for n to achieve ε-accuracy:

$$n \geq N_{\text{Exact}} \triangleq \max\left\{\frac{2}{\log(2)}\left[\log(F(\mathbf{x}^0) - F_{\text{opt}}) + \log\left(\frac{1}{\varepsilon}\right)\right], \frac{8pL_f p^2 R^2}{\varepsilon}\right\} + 1. \tag{1.68}$$

Here, we compare inexact and exact BCD **purely from the perspective of iteration complexity**. Notice that there is a transient term $\frac{2}{\log(2)}\left[\log(F(\mathbf{x}^0) - F_{\text{opt}}) + \log(1/\varepsilon)\right]$ and a sub-linear term on the order of $1/\varepsilon$ in both N_{Inexact} and N_{exact}. In the case when we need low accuracy or the initialization error is large, then the transient term will dominate and N_{Inexact} and N_{Exact} will be roughly the same regardless of the sub-linear terms. However, in the case when we need high accuracy (ε is small), the sub-linear term will matter.

Precisely, we may examine the following ratio of the Inexact and Exact sub-linear rate:

$$\rho = \frac{8p(L_f + L_{\max})^2 R^2}{L_{\min}\varepsilon} \Big/ \frac{8pL_f p^2 R^2}{\varepsilon} = \frac{(L_f + L_{\max})^2}{pL_f L_{\min}}, \tag{1.69}$$

then if $\rho < 1$, inexact BCD is better, whereas exact BCD will prevail when $\rho > 1$.

There are concrete scenarios for these two cases:

Case 1: Highly uncoupled and identical blocks

Suppose the variables barely interact with each other across blocks, and hence all blocks are equally smooth. In this case, $L_{\max} = L_{\min} = L_f$, and this would yield:

$$\rho = \frac{4}{p}, \tag{1.70}$$

meaning that in the case when we have $p > 4$ independent, identical blocks, $\rho < 1$ and inexact BCD requires fewer iterations than exact BCD.

Case 2: Highly coupled and poorly conditioned block

In this case, one simplification would be $L_{\max} \gg L_{\min}$, and for simplicity we may take $L_f = pL_{\max}$. This would yield:

$$\rho = \frac{(p+1)^2}{p^2}\left(\frac{L_{\max}}{L_{\min}}\right). \tag{1.71}$$

In this case, since $(p+1) > p$, we would have $\rho > 1$, and exact BCD would yield vastly fewer iterations.

[3] The lower bound for n in this case is not provided in the original text.

However, an explicit and rigorous comparision on the iteration complexity for HALS (exact proximal BCD) and AltPGD (inexact proximal BCD) would be difficult as NMF objective is globally non-convex. Existing literatures on exact and inexact BCD uses different convergence measure in the non-convex regime. The former mostly uses convergence to limit points Tseng (2001); Xu and Yin (2013) and the latter argues convergence to stationarity Razaviyayn et al. (2013) by showing vanshing gradient.

Detour: PALM - Proximal Alternating Linearized Minimization

We consider the following composite minimization problem:

$$\min_{\mathbf{x},\mathbf{y}} \Psi(\mathbf{x},\mathbf{y}) \equiv f(\mathbf{x},\mathbf{y}) + g(\mathbf{x}) + h(\mathbf{y}). \tag{1.72}$$

In standard Proximal Alternating Minimization, we update the blocks by minimizing the entire smooth function f alongside the non-smooth function and a proximal penalty, that is:

- $\mathbf{x}$-update:
$$\mathbf{x}^{t+1} \in \arg\min_{\mathbf{x}} \left\{ f(\mathbf{x},\mathbf{y}^t) + g(\mathbf{x}) + \frac{c_t}{2}\|\mathbf{x}-\mathbf{x}^t\|^2 \right\}. \tag{1.73}$$
- $\mathbf{y}$-update:
$$\mathbf{y}^{t+1} \in \arg\min_{\mathbf{y}} \left\{ f(\mathbf{x}^{t+1},\mathbf{y}) + h(\mathbf{y}) + \frac{d_t}{2}\|\mathbf{y}-\mathbf{y}^t\|^2 \right\}. \tag{1.74}$$

In NMF, this would be the case where $g(\mathbf{W})$ and $h(\mathbf{H})$ are indicator functions of the non-negative orthant , i.e.

$$g(\mathbf{W}) = \delta_+(\mathbf{W}) \qquad \text{and} \qquad h(\mathbf{H}) = \delta_+(\mathbf{H}), \tag{1.75}$$

and we would have, assuming $(\mathbf{W}^t, \mathbf{H}^t)$ is obtained:

- $\mathbf{W}$-update:
$$\mathbf{W}^{k+1} = \arg\min_{\mathbf{W}\geq\mathbf{0}} \left\{ \frac{1}{2}\|\mathbf{A}-\mathbf{W}\mathbf{H}^t\|_F^2 + \frac{c_t}{2}\|\mathbf{W}-\mathbf{W}^t\|_F^2 \right\}. \tag{1.76}$$
- $\mathbf{H}$-update:
$$\mathbf{H}^{t+1} = \arg\min_{\mathbf{H}\geq\mathbf{0}} \left\{ \frac{1}{2}\|\mathbf{A}-\mathbf{W}^{t+1}\mathbf{H}\|_F^2 + \frac{d_t}{2}\|\mathbf{H}-\mathbf{H}^t\|_F^2 \right\}. \tag{1.77}$$

We refer to this method as the plain Proximal Alternating Method (**PAM**). The bottleneck is the presence of the coupling term inside the $\arg\min\{\cdot\}$. Unless $f(\cdot,\cdot)$ is of simple structure, evaluating it explicitly in $\arg\min\{\cdot\}$ at every iteration could be expensive. This would normally require the use of active set methods like non-negative least squares (**NNLS**) by Lawson and Hanson (1974).

PALM Bolte et al. (2014) replaces the smooth terms $f(\mathbf{x},\mathbf{y}^t)$ and $f(\mathbf{x}^{t+1},\mathbf{y})$ by their first order Taylor expansion, i.e. it uses the linearization technique to eliminate the coupling effect due to the presence of f inside $\arg\min$. Concretely, they are:

- $\mathbf{x}$-update:

$$\begin{aligned}\mathbf{x}^{t+1} &\in \arg\min_{\mathbf{x}} \left\{ \langle \nabla_{\mathbf{x}} f(\mathbf{x}^t, \mathbf{y}^t), \mathbf{x} - \mathbf{x}^t \rangle + \frac{c_t}{2} \|\mathbf{x} - \mathbf{x}^t\|_2^2 + g(\mathbf{x}) \right\} \\ &= \mathbf{prox}_{g/c_t} \left(\mathbf{x}^t - \frac{1}{c_t} \nabla_{\mathbf{x}} f(\mathbf{x}^t, \mathbf{y}^t) \right).\end{aligned} \tag{1.78}$$

- $\mathbf{y}$-update:

$$\begin{aligned}\mathbf{y}^{t+1} &\in \arg\min_{\mathbf{y}} \left\{ \langle \nabla_{\mathbf{y}} f(\mathbf{x}^{t+1}, \mathbf{y}^t), \mathbf{y} - \mathbf{y}^t \rangle + \frac{d_t}{2} \|\mathbf{y} - \mathbf{y}^t\|_2^2 + h(\mathbf{y}) \right\} \\ &= \mathbf{prox}_{h/d_t} \left(\mathbf{y}^t - \frac{1}{d_t} \nabla_{\mathbf{y}} f(\mathbf{x}^{t+1}, \mathbf{y}^t) \right).\end{aligned} \tag{1.79}$$

Notice that PALM completely decouples the smooth coupling function f from the non-smooth counterparts g and h. In the case where g and h are indicator functions, PALM recovers AltPGD exactly. Precisely,

- $\mathbf{W}$-update:

$$\mathbf{W}^{t+1} = \max\{\mathbf{0}, \mathbf{W}^t - \alpha_t \nabla_{\mathbf{W}} f(\mathbf{W}^t, \mathbf{H}^t)\}, \tag{1.80}$$

  with $\alpha_t = 1/\|\mathbf{H}^t(\mathbf{H}^t)^\top\|$.

- $\mathbf{H}$-update:

$$\mathbf{H}^{t+1} = \max\{\mathbf{0}, \mathbf{H}^t - \beta_t \nabla_{\mathbf{H}} f(\mathbf{W}^{t+1}, \mathbf{H}^t)\}, \tag{1.81}$$

  with $\beta_t = 1/\|(\mathbf{W}^{t+1})^\top \mathbf{W}^{t+1}\|$.

Algorithm 5 Alternating Projected Gradient Descent (**AltPGD**)

Input: Non-negative target matrix $\mathbf{A} \in \mathbb{R}_+^{m \times n}$, factorization rank r
Output: Converged non-negative factors $\mathbf{W} \in \mathbb{R}_+^{m \times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r \times n}$

// Initialization
Initialize $\mathbf{W}^0 \geq \mathbf{0}$ and $\mathbf{H}^0 \geq \mathbf{0}$

// Main Loop
for $t = 0, 1, 2, \cdots$ **do**
    $\alpha_t \leftarrow 1/\|\mathbf{H}^t(\mathbf{H}^t)^\top\|$
    $\mathbf{W}^{t+1} \leftarrow \max\{\mathbf{0}, \mathbf{W}^t - \alpha_t \nabla_{\mathbf{W}} f(\mathbf{W}^t, \mathbf{H}^t)\}$
    $\beta_t \leftarrow 1/\|(\mathbf{W}^{t+1})^\top \mathbf{W}^{t+1}\|$
    $\mathbf{H}^{t+1} \leftarrow \max\{\mathbf{0}, \mathbf{H}^t - \beta_t \nabla_{\mathbf{H}} f(\mathbf{W}^{t+1}, \mathbf{H}^t)\}$
end for

1.5.4 ADMM

The Alternating Direction Method of Multipliers (**ADMM**) was developed in the mid-1970s by Glowinski and Marroco (1975) and Gabay and Mercier (1976) primarily as an extension of the augmented Lagrangian theory. A few years later, it was formalized that ADMM is equivalent to applying Douglas-Rachford splitting to the dual of the optimization problem Eckstein and Bertsekas (1992). Here we state the general formulation of ADMM and its instantiation in NMF.

General formulation of ADMM

Consider the following standard equality constrained problem:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{z}} \quad & f(\mathbf{x}) + g(\mathbf{z}) \\ \text{s.t.} \quad & \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{c}, \end{aligned} \tag{1.82}$$

where $\mathbf{x}$ lives in $\mathbb{R}^n$ and $\mathbf{z}$ lives in $\mathbb{R}^m$. The constraint matrices $\mathbf{A}$ and $\mathbf{B}$ are of size $p \times n$ and $p \times m$ respectively. The functions $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $g : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ are closed, convex, and proper (CCP). Also, they are possibly non-smooth.

The standard Lagrangian associated with the equality constrained problem will be:

$$L(\mathbf{x}, \mathbf{z}, \mathbf{y}) = f(\mathbf{x}) + g(\mathbf{z}) + \mathbf{y}^\top(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}). \tag{1.83}$$

The augmented Lagrangian adds a quadratic penalty term to the standard Lagrangian, with $\rho > 0$:

$$L_\rho(\mathbf{x}, \mathbf{z}, \mathbf{y}) = f(\mathbf{x}) + g(\mathbf{z}) + \mathbf{y}^\top(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}) + \frac{\rho}{2}\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}\|_2^2. \tag{1.84}$$

There are a couple of advantages to this quadratic term. The standard Lagrangian relies on Dual Ascent. In the case f and g are not **strictly convex** and hence not **strongly convex**, the Fenchel conjugate function functions f^* and g^* are possibly non-smooth. By adding this quadratic term, we are forcing the Lagrangian to be strongly convex w.r.t. the constraint $\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z}$, where ρ is precisely the strong convexity parameter.

Now, the standard Method of Multipliers (**MoM**) minimizes this Augmented Lagrangian jointly w.r.t. $\mathbf{x}$ and $\mathbf{z}$, and then updates the dual variable $\mathbf{y}$:

1. $(\mathbf{x}^{t+1}, \mathbf{z}^{t+1}) = \arg\min_{\mathbf{x},\mathbf{z}} L_\rho(\mathbf{x}, \mathbf{z}, \mathbf{y}^t)$, and
2. $\mathbf{y}^{t+1} = \mathbf{y}^t + \rho(\mathbf{A}\mathbf{x}^{t+1} + \mathbf{B}\mathbf{z}^{t+1} - \mathbf{c})$

In some cases, jointly minimizing $(\mathbf{x}, \mathbf{z})$ might be computationally expensive or even impossible to separate if $\mathbf{A}$ and $\mathbf{B}$ couple the variables. ADMM fixes this by splitting the minimization.

1. $\mathbf{x}$-update: $\mathbf{x}^{t+1} = \arg\min_{\mathbf{x}} L_\rho(\mathbf{x}, \mathbf{z}^t, \mathbf{y}^t)$,
2. $\mathbf{z}$-update: $\mathbf{z}^{t+1} = \arg\min_{\mathbf{z}} L_\rho(\mathbf{x}^{t+1}, \mathbf{z}, \mathbf{y}^t)$, and
3. $\mathbf{y}$-update: $\mathbf{y}^{t+1} = \mathbf{y}^t + \rho(\mathbf{A}\mathbf{x}^{t+1} + \mathbf{B}\mathbf{z}^{t+1} - \mathbf{c})$.

ADMM formulation for the $\mathbf{W}$-subproblem

$$\begin{aligned} \min_{\mathbf{W}} \quad & \frac{1}{2}\|\mathbf{A} - \mathbf{W}\mathbf{H}\|_F^2 \\ \text{s.t.} \quad & \mathbf{W} \geq \mathbf{0}. \end{aligned} \tag{1.85}$$

To apply ADMM, we will need to introduce an auxiliary slack variable $\mathbf{S}$ to separate the least-squares term from the non-negative constraint. Let $\delta_+(\mathbf{S})$ be the indicator, and so the above problem becomes

$$\begin{aligned} \min_{\mathbf{W}} \quad & \frac{1}{2}\|\mathbf{A} - \mathbf{W}\mathbf{H}\|_F^2 + \delta_+(\mathbf{S}) \\ \text{s.t.} \quad & \mathbf{W} - \mathbf{S} = \mathbf{0}. \end{aligned} \tag{1.86}$$

The **scaled** augmented Lagrangian with $\mathbf{U}$ being the dual variable becomes:

$$L_\rho(\mathbf{W}, \mathbf{S}, \mathbf{U}) = \frac{1}{2}\|\mathbf{A} - \mathbf{W}\mathbf{H}\|_F^2 + \delta_+(\mathbf{S}) + \frac{\rho}{2}\|\mathbf{W} - \mathbf{S} + \mathbf{U}\|_F^2. \tag{1.87}$$

Going from the standard augmented Lagrangian to the scaled augmented Lagrangian

While the standard formulation of ADMM relies on the unscaled augmented Lagrangian, it is common practice to use a scaled version. The scaled version combines the linear and quadratic penalty terms into a single squared Frobenius norm. This simplification is highly advantageous because it isolates the primal variables within a standard least-squares framework, making the $\mathbf{W}$ and $\mathbf{S}$ update steps easier to derive and implement.
To see how we arrive at this scaled form for the $\mathbf{W}$-subproblem, we begin with our constrained objective:

$$\begin{aligned} \min_{\mathbf{W},\mathbf{S}} \quad & \frac{1}{2}\|\mathbf{A} - \mathbf{W}\mathbf{H}\|_F^2 + \delta_+(\mathbf{S}) \\ \text{s.t.} \quad & \mathbf{W} - \mathbf{S} = \mathbf{0}. \end{aligned} \tag{1.88}$$

Let $\mathbf{Y} \in \mathbb{R}^{m\times r}$ denote the standard dual variable matrix. The unscaled augmented Lagrangian is defined as:

$$L_\rho(\mathbf{W}, \mathbf{S}, \mathbf{Y}) = \frac{1}{2}\|\mathbf{A} - \mathbf{W}\mathbf{H}\|_F^2 + \delta_+(\mathbf{S}) + \langle \mathbf{Y}, \mathbf{W} - \mathbf{S}\rangle + \frac{\rho}{2}\|\mathbf{W} - \mathbf{S}\|_F^2, \tag{1.89}$$

where $\langle \mathbf{X}, \mathbf{Z}\rangle = \mathrm{Tr}(\mathbf{X}^\top \mathbf{Z})$ represents the standard trace inner product for matrices.
To complete the square for the penalty terms, we introduce a scaled dual variable $\mathbf{U} \triangleq (1/\rho)\mathbf{Y}$, which implies $\mathbf{Y} = \rho\mathbf{U}$ [a]. Substituting this into the linear penalty term gives:

$$\langle \rho\mathbf{U}, \mathbf{W} - \mathbf{S}\rangle + \frac{\rho}{2}\|\mathbf{W} - \mathbf{S}\|_F^2 = \rho\langle \mathbf{U}, \mathbf{W} - \mathbf{S}\rangle + \frac{\rho}{2}\|\mathbf{W} - \mathbf{S}\|_F^2. \tag{1.90}$$

Factoring out the penalty parameter $\frac{\rho}{2}$, we can complete the square by adding and subtracting $\|\mathbf{U}\|_F^2$:

$$\begin{aligned} \rho\langle \mathbf{U}, \mathbf{W} - \mathbf{S}\rangle + \frac{\rho}{2}\|\mathbf{W} - \mathbf{S}\|_F^2 &= \frac{\rho}{2}\left(\|\mathbf{W} - \mathbf{S}\|_F^2 + 2\langle \mathbf{W} - \mathbf{S}, \mathbf{U}\rangle\right) \\ &= \frac{\rho}{2}\left(\|\mathbf{W} - \mathbf{S}\|_F^2 + 2\langle \mathbf{W} - \mathbf{S}, \mathbf{U}\rangle + \|\mathbf{U}\|_F^2 - \|\mathbf{U}\|_F^2\right) \\ &= \frac{\rho}{2}\|\mathbf{W} - \mathbf{S} + \mathbf{U}\|_F^2 - \frac{\rho}{2}\|\mathbf{U}\|_F^2. \end{aligned} \tag{1.91}$$

Substituting this algebraically equivalent expression back into the full Lagrangian yields the scaled form:

$$L_\rho(\mathbf{W}, \mathbf{S}, \mathbf{U}) = \frac{1}{2}\|\mathbf{A} - \mathbf{W}\mathbf{H}\|_F^2 + \delta_+(\mathbf{S}) + \frac{\rho}{2}\|\mathbf{W} - \mathbf{S} + \mathbf{U}\|_F^2 - \frac{\rho}{2}\|\mathbf{U}\|_F^2. \tag{1.92}$$

When optimizing L_ρ w.r.t. the primal variables $\mathbf{W}$ or $\mathbf{S}$, the final term $-\rho/2\|\mathbf{U}\|_F^2$ is independent of the variables of interest and acts solely as a constant. Therefore, it is safely dropped during the $\arg\min$ updates, allowing us to concisely write the working objective for the $\mathbf{W}$-subproblem as:

$$L_\rho(\mathbf{W}, \mathbf{S}, \mathbf{U}) = \frac{1}{2}\|\mathbf{A} - \mathbf{W}\mathbf{H}\|_F^2 + \delta_+(\mathbf{S}) + \frac{\rho}{2}\|\mathbf{W} - \mathbf{S} + \mathbf{U}\|_F^2. \tag{1.93}$$

This scaling absorbs the multiplier into the quadratic penalty, leading to the derivations in equations (1.94) and (1.95), and resulting in a dual update step (1.96) that simply tracks the running sum of the primal residuals.

[a] see Ryu and Yin (2022) and Bauschke and Combettes (2017) for this technique.

The ADMM updates therefore for $\mathbf{W}$ at inner iteration t are:

- $\mathbf{W}$-update: by setting $\nabla_{\mathbf{W}} L_\rho = 0$:
$$\mathbf{W}^{t+1} = (\mathbf{A}\mathbf{H}^\top + \rho\mathbf{S}^t - \rho\mathbf{U}^t)(\mathbf{H}\mathbf{H}^\top + \rho\mathbf{I})^{-1}. \tag{1.94}$$
- $\mathbf{S}$-update: this is the projection onto the non-negative orthant:
$$\begin{aligned}\mathbf{S}^{t+1} &= \arg\min_{\mathbf{S}} \left[\delta_+(\mathbf{S}) + \frac{\rho}{2}\|\mathbf{W}^{t+1} - \mathbf{S} + \mathbf{U}^t\|_F^2\right] \\ &= \max\{\mathbf{0}, \mathbf{W}^{t+1} + \mathbf{U}^t\}.\end{aligned} \tag{1.95}$$
- $\mathbf{U}$-update:
$$\mathbf{U}^{t+1} = \mathbf{U}^t + \mathbf{W}^{t+1} - \mathbf{S}^{t+1}. \tag{1.96}$$

By symmetry, one could derive the update for the $\mathbf{H}$ update. Here we denote the slack variable as $\mathbf{T}$ and the dual variable as $\mathbf{V}$, and the update for the $\mathbf{H}$ subproblem would then be:

- $\mathbf{H}$-update:
$$\mathbf{H}^{t+1} = (\mathbf{W}^\top\mathbf{W} + \rho\mathbf{I})^{-1}(\mathbf{W}^\top\mathbf{A} + \rho\mathbf{T}^t - \rho\mathbf{V}^t). \tag{1.97}$$
- $\mathbf{T}$-update:
$$\mathbf{T}^{t+1} = \max\{\mathbf{0}, \mathbf{H}^{t+1} + \mathbf{V}^t\}. \tag{1.98}$$
- $\mathbf{V}$-update:
$$\mathbf{V}^{t+1} = \mathbf{V}^t + \mathbf{H}^{t+1} - \mathbf{T}^{t+1}. \tag{1.99}$$

1.5.5 Extrapolation and accelerated versions

Extrapolation and acceleration

Recall that standard GD:
$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\nabla f(\mathbf{x}_t), \tag{1.100}$$
as $-\nabla f(\mathbf{x}_t)$ represents the direction in which the descent is the fastest. This, for some less desirable function classes or less well-chosen step-sizes, might lead to oscillations. A more conservative choice is to incorporate information from prior updates, i.e. the momentum $\mathbf{x}_t - \mathbf{x}_{t-1}$, and this leads to the **Heavy Ball Method** in Polyak (1964):
$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\nabla f(\mathbf{x}_t) + \beta(\mathbf{x}_t - \mathbf{x}_{t-1}), \tag{1.101}$$
with β and η properly chosen.

Nesterov (1983) uses the momentum as an intermediate step, that is
$$\mathbf{y}_t \triangleq \mathbf{x}_t + \beta_{t+1}(\mathbf{x}_t - \mathbf{x}_{t-1}), \tag{1.102}$$

Algorithm 6 ADMM for NMF

Input: Non-negative target matrix $\mathbf{A} \in \mathbb{R}_+^{m \times n}$, factorization rank r, penalty parameter $\rho > 0$
Output: Converged non-negative factors $\mathbf{W} \in \mathbb{R}_+^{m \times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r \times n}$

// Initialization
Initialize $\mathbf{W}^0, \mathbf{H}^0, \mathbf{S}^0, \mathbf{T}^0, \mathbf{U}^0, \mathbf{V}^0$

// Main Loop
for $t = 0, 1, 2, \cdots$ **do**
    // $\mathbf{W}$-subproblem updates
    $\mathbf{W}^{t+1} \leftarrow (\mathbf{A}(\mathbf{H}^t)^\top + \rho \mathbf{S}^t - \rho \mathbf{U}^t)(\mathbf{H}^t(\mathbf{H}^t)^\top + \rho \mathbf{I})^{-1}$
    $\mathbf{S}^{t+1} \leftarrow \max\{\mathbf{0}, \mathbf{W}^{t+1} + \mathbf{U}^t\}$
    $\mathbf{U}^{t+1} \leftarrow \mathbf{U}^t + \mathbf{W}^{t+1} - \mathbf{S}^{t+1}$

    // $\mathbf{H}$-subproblem updates
    $\mathbf{H}^{t+1} \leftarrow ((\mathbf{W}^{t+1})^\top \mathbf{W}^{t+1} + \rho \mathbf{I})^{-1}((\mathbf{W}^{t+1})^\top \mathbf{A} + \rho \mathbf{T}^t - \rho \mathbf{V}^t)$
    $\mathbf{T}^{t+1} \leftarrow \max\{\mathbf{0}, \mathbf{H}^{t+1} + \mathbf{V}^t\}$
    $\mathbf{V}^{t+1} \leftarrow \mathbf{V}^t + \mathbf{H}^{t+1} - \mathbf{T}^{t+1}$
end for

where the descent happens at this intermediate point $\mathbf{y}_{t+1}$:

$$\mathbf{x}_t = \mathbf{y}_t - \eta \nabla f(\mathbf{y}_{t+1}). \tag{1.103}$$

Here $\eta = 1/L$ and β_{t+1} is chosen carefully to achieve the optimal $O(1/t^2)$ convergence rate. Many variants and new interpretations are proposed. Such an acceleration scheme could also be applied to the proximal method, leading to the celebrated FISTA method Beck and Teboulle (2009). In general, the choice for the hyper-parameters and the proof technique in accelerated optimization is highly non-trivial.

E-A-HALS: extrapolated A-HALS

In HALS, we need to compute $\mathbf{H}\mathbf{H}^\top$ and $\mathbf{A}\mathbf{H}^\top$ when updating the $\mathbf{W}$ block, and $\mathbf{W}^\top\mathbf{W}$ and $\mathbf{W}^\top\mathbf{A}$ when updating $\mathbf{H}$. Instead of updating one row of $\mathbf{H}$ then immediately switching to updating columns of $\mathbf{W}$, an efficient way Gillis and Glineur (2012) is to update multiple or all rows of $\mathbf{H}$ while reusing $\mathbf{W}^\top\mathbf{W}$ and $\mathbf{W}^\top\mathbf{A}$ to avoid redundant calculations; this is known as A-HALS.

E-A-HALS Ang and Gillis (2019), on top of re-using those matrices, adds a block-wise extrapolation, and this is one of the most efficient methods for NMF. **Although there are complexity guarantees for non-convex extrapolated BCD (and its proximal versions) in Xu and Yin (2013), E-A-HALS does not make the objective function monotonically decrease at each step due to its non-standard and hybrid scheme of extrapolation and projection so the existing complexities do not carry over**.

Algorithm 7 Accelerated HALS for NMF

Input: Data matrix $\mathbf{A} \in \mathbb{R}_+^{m \times n}$, factorization rank r, parameters $\alpha \geq 0, \epsilon \geq 0$
Output: Converged non-negative factors $\mathbf{W} \in \mathbb{R}_+^{m \times r}$ and $\mathbf{H} \in \mathbb{R}_+^{r \times n}$

1: // Initialization
2: Initialize $\mathbf{W}^0 \in \mathbb{R}_+^{m \times r}$ and $\mathbf{H}^0 \in \mathbb{R}_+^{r \times n}$
3: Let K be the number of nonzero entries in $\mathbf{A}$. Calculate $\rho_W = 1 + \frac{K+nr}{mr+m}$ and $\rho_H = 1 + \frac{K+mr}{nr+n}$

4: // Main Loop
5: **for** $t = 0, 1, 2, \cdots$ **do**
6: // $\mathbf{W}$-subproblem updates
7: Compute $\mathbf{P} = \mathbf{A}(\mathbf{H}^t)^\top$ and $\mathbf{Q} = \mathbf{H}^t(\mathbf{H}^t)^\top$; $\mathbf{W}^{t,0} \leftarrow \mathbf{W}^t$
8: **for** $l = 1$ **to** $\lfloor 1 + \alpha\rho_W \rfloor$ **do**
9: Compute $\mathbf{W}^{t,l}$ using HALS
10: **if** $\|\mathbf{W}^{t,l} - \mathbf{W}^{t,l-1}\|_F \leq \epsilon \|\mathbf{W}^{t,1} - \mathbf{W}^{t,0}\|_F$ **then**
11: **break**
12: **end if**
13: **end for**
14: $\mathbf{W}^{t+1} \leftarrow \mathbf{W}^{t,l}$

15: // $\mathbf{H}$-subproblem updates
16: Compute $\mathbf{H}^{t+1}$ from $\mathbf{H}^t$ and $\mathbf{W}^{t+1}$ using a symmetrically adapted version of steps 7–14
17: **end for**

IBPG: Inertial Block Proximal Gradient (Inertial BCD)

As seen before, Nesterov uses a single extrapolation point for evaluating the gradient. **Inertial momentum** uses a two-point extrapolation scheme, namely, one for evaluating the gradient and one for evaluating the proximal term. Again, we are looking at

$$\min_{\mathbf{x} \in \mathbb{R}^d} F(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x}), \tag{1.104}$$

where f is smooth, having a Lipschitz continuous gradient, and g is a proper, lower semi-continuous mapping from $\mathbb{R}^d$ to $\mathbb{R} \cup \{+\infty\}$, and is possibly non-smooth. The two-point extrapolation proposed is as follows, where:

$$\widehat{\mathbf{y}}_t \triangleq \mathbf{x}_t + \beta(\mathbf{x}_t - \mathbf{x}_{t-1}) \qquad \text{and} \qquad \widetilde{\mathbf{y}}_t \triangleq \mathbf{x}_t + \gamma(\mathbf{x}_t - \mathbf{x}_{t-1}). \tag{1.105}$$

Then in single block notation, the proximal update step is:

$$\begin{aligned} &\mathbf{x}_{t+1} = \mathbf{prox}_{\eta g}(\widetilde{\mathbf{y}}_t - \eta \nabla f(\widehat{\mathbf{y}}_t)) \\ \iff\ &\mathbf{x}_{t+1} \in \arg\min_{\mathbf{x}} \left(\langle \nabla f(\widehat{\mathbf{y}}_t), \mathbf{x} - \mathbf{x}_t \rangle + g(\mathbf{x}) + \frac{1}{2\eta} \|\mathbf{x} - \widetilde{\mathbf{y}}_t\|_2^2 \right). \end{aligned} \tag{1.106}$$

In Hien et al. (2020), the authors formalize this inertial block proximal gradient method (IBPG) using this two-point extrapolation scheme. IBPG has been shown to outperform E-A-HALS empirically.

As demonstrated by the wide array of methods from MU to ADMM and accelerated BCD, algorithm design for NMF heavily exploits the block-multi-convex structure of the objective. However,

analyzing the performance of these algorithms fundamentally depends on the inherent difficulty of the NMF problem itself. To contextualize these algorithmic efforts, we conclude this chapter by briefly reviewing the computational complexity of solving NMF.

1.6 Some complexity results

In Vavasis (2009), it is shown that exact NMF is NP-hard. This is achieved by establishing equivalence (in the sense of polynomial time reduction) between the exact NMF problem and an intermediate simplex problem. The latter is equivalent to 3-SAT which is known to be NP-hard (complete) Karp (1972). However, this only rules out the possibility of an algorithm whose running time is polynomial in all parameters (n, m, r). Subsequently, Arora et al. (2012) improved this by providing an exact (and approximate) polynomial-time algorithm for every constant r, under the separability condition.

Chapter 2

Spectral techniques for initializations

Before diving into specific algorithmic techniques, it is instructive to review the foundational problems that motivate the two-stage framework. These examples highlight how structural properties of the target matrix are leveraged during initialization.

2.1 A few canonical problems under the two-stage approach

For MF problems, they can be cast as optimization with a function $f(\mathbf{X}, \mathbf{Y})$ over two variables $\mathbf{X} \in \mathbb{R}^{m\times d}$ and $\mathbf{Y} \in \mathbb{R}^{n\times d}$, where $\mathbf{X}$ and $\mathbf{Y}$ quite often represent low-rank factors that best approximate the ground-truth $\mathbf{A}_*$. In the context of low-rank factorizations, the **ground-truth** typically refers to the true, underlying low-rank matrix that generates the observed data or measurements. We show a few canonical examples summarized from the fantastic expository Chi et al. (2019) in table 2.1.

2.1.1 Why spectral initialization works well (under gradient methods)?

Having established these foundational problem classes, we naturally turn our attention to the reasons behind the empirical success of spectral initializations[1] in solving them. There are two components to such arguments:

1. **Proximity to the ground truth**: An intermediate matrix $\mathbf{M}_0$ constructed using the spectral information of the ground truth $\mathbf{A}_*$ will be very close to it. $\mathbf{M}_0$ is achieved via $\mathbf{X}_0$ and $\mathbf{Y}_0$ that live in the factorized spaces which directly use the spectral information of $\mathbf{A}_*$, usually by SVD. Via matrix perturbation arguments, one can often show that the distance $\|\mathbf{M}_0 - \mathbf{A}_*\|$ is bounded by the signal $\sigma_{\min}(\mathbf{A}_*)$ of the ground truth.
2. **Local curvature/regularity conditions (and benignness of the landscape)**: Within the neighborhood or **the basin of attraction** of the ground-truth $\mathbf{A}_*$, which is dictated by $\sigma_{\min}(\mathbf{A}_*)$, the objective will satisfy some local regularity and curvature conditions. The iterates that land anywhere in this basin will converge under simple first-order methods. This will be the main recipe for establishing the convergence guarantee.

[1] Most tractable and provable results for initialization in matrix optimization will use spectral information of the target matrix. Here spectral initialization narrowly means SVD-based initializations

Table 2.1: Common matrix recovery problems and their objectives.

Problems and Objectives	Key Set-ups and Assumptions
Matrix Sensing $\min f(\mathbf{X},\mathbf{Y}) \triangleq \frac{1}{4p}\sum_{i=1}^{p}\left(\langle \mathbf{A}_i, \mathbf{X}\mathbf{Y}^\top\rangle - y_i\right)^2$ Here $\mathbf{X}$ and $\mathbf{Y}$ relate to $\mathbf{A}_*$ via the measurement matrices $\mathbf{A}_i$ and measurement result y_i.	1. **Goal**: recover a rank- r component $\mathbf{A}_*$ from the $p \geq 1$ measurements. 2. The collection of matrices $\{\mathbf{A}_i \in \mathbb{R}^{m\times n}\}_{i=1}^p$ represents the sensing matrices which are known **a priori**. 3. The collection of scalars $y_i \triangleq \langle \mathbf{A}_i, \mathbf{A}_*\rangle$ are the measurement results. 4. The operator $\mathcal{A}: \mathbb{R}^{m\times n} \to \mathbb{R}^p$ represents the sensing operator whose actions can be cast as $$\mathcal{A}(\mathbf{M}) \triangleq [p^{-1/2}\langle \mathbf{A}_i, \mathbf{M}\rangle]_{1\leq i\leq p}. \tag{2.1}$$ 5. The operator $\mathcal{A}$ satisfies the r-RIP condition, i.e., $$(1-\delta_r)\|\mathbf{M}\|_F^2 \leq \|\mathcal{A}(\mathbf{M})\|_2^2 \leq (1+\delta_r)\|\mathbf{M}\|_F^2, \tag{2.2}$$ holds for all $\mathbf{M} \in \mathbb{R}^{m\times n}$ with $\mathrm{rank}(\mathbf{M}) \leq r$ and $\delta_r \in (0,1)$
Phase Retrieval/Quadratic Sensing $\min f(\mathbf{X},\mathbf{X}) \equiv f(\mathbf{X}) \triangleq \frac{1}{4p}\sum_{j=1}^{p}\left(\|\mathbf{a}_i^\top \mathbf{X}\|_2^2 - y_i\right)^2$	1. **Goal**: recover a rank- r component $\mathbf{A}_* = \mathbf{X}_*\mathbf{X}_*^\top \in \mathbb{R}^{n\times n}$ where $\mathbf{X}_* \in \mathbb{R}^{n\times r}$. 2. The collection of vectors $\{\mathbf{a}_i \in \mathbb{R}^n\}_{i=1}^p$ are design/measurement vectors. 3. The collection of scalars $y_i \triangleq \|\mathbf{a}_i^\top \mathbf{X}_*\|_2^2 = \mathbf{a}_i^\top \mathbf{A}_* \mathbf{a}_i$ are the quadratic measurement results. 4. Design/measurement vectors are drawn i.i.d. Gaussian, that is $\mathbf{a}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$. 5. The case when $r = 1$ is **Phase Retrieval**.
Matrix Completion $\min f(\mathbf{X},\mathbf{Y}) \triangleq \frac{1}{4p}\|\mathcal{P}_\Omega(\mathbf{X}\mathbf{Y}^\top - \mathbf{A}_*)\|_F^2$	1. **Goal**: recover $\mathbf{A}_* \in \mathbb{R}^{m\times n}$ from $\mathcal{P}_\Omega(\mathbf{A}_*) \in \mathbb{R}^{m\times n}$ where $\Omega \subseteq [m]\times[n]$ is the sampling index set. 2. $\mathcal{P}_\Omega: \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ is the projection operator onto the sampling index set Ω. That is, $$[\mathcal{P}_\Omega(\mathbf{M})]_{i,j} = \begin{cases} \mathbf{M}_{i,j} & \text{if } (i,j)\in\Omega \\ 0 & \text{else} \end{cases}. \tag{2.3}$$ 3. Sampling is i.i.d. with probability (**w.p.**) $\theta \in (0,1]$ for each location (i,j). That is, $(i,j) \in \Omega$ independently w.p. θ. 4. Eigenvectors of the ground truth $\mathbf{A}_*$ satisfy the **incoherence condition**, c.f. equation $(34a)$ and $(34b)$ in Chi et al. (2019).
Blind Deconvolution $\min f(\mathbf{h},\mathbf{x}) \triangleq \sum_{j=1}^{p}\|\mathbf{b}_j^{\mathsf{H}}\mathbf{h}\mathbf{x}^{\mathsf{H}}\mathbf{a}_j - y_i\|^2$ Here $\cdot^{\mathsf{H}}$ means the complex conjugate.	1. **Goal**: recover two signals $\mathbf{h}_* \in \mathbb{C}^K$ and $\mathbf{x}_* \in \mathbb{C}^N$ or $\mathbf{M}_* \triangleq \mathbf{h}_*\mathbf{x}_*^{\mathsf{H}}$. 2. The collection of vectors $\{\mathbf{b}_i \in \mathbb{C}^K\}_{i=1}^p$ and $\{\mathbf{a}_i \in \mathbb{C}^N\}_{i=1}^p$ are the measurement vectors. 3. $y_i = \mathbf{b}_i^{\mathsf{H}}\mathbf{h}_*\mathbf{x}_*^{\mathsf{H}}\mathbf{a}_i$ are the **bilinear** measurement results. 4. The measurement vectors are constructed using a semi-random scheme. That is, $\mathbf{a}_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, 1/2\mathbf{I}_N) + i\mathcal{N}(\mathbf{0}, 1/2\mathbf{I}_N)$ and $\mathbf{b}_j$'s are constructed using a DFT matrix. 5. $\mathbf{h}_*$ and $\mathbf{b}_j$ satisfy the incoherence condition, c.f. equation (37) in Chi et al. (2019).
Robust PCA $\min_{\mathbf{S}\in\mathcal{S}_\alpha} f(\mathbf{X},\mathbf{Y},\mathbf{S}) \triangleq \frac{1}{4p}\|\mathcal{P}_\Omega(\mathbf{A}_* - \mathbf{X}\mathbf{Y}^\top - \mathbf{S})\|_F^2$	1. **Goal**: Recover $\mathbf{A}_* = \mathbf{L}_* + \mathbf{S}_* \in \mathbb{R}^{m\times n}$, where $\mathbf{L}_*$ is some low-rank component and $\mathbf{S}_*$ is sparse. We have access to partial entries of $\mathbf{A}_*$ through the random sampling model Ω and $\mathcal{P}_\Omega$ as in the case of Matrix Completion. 2. The sparse component $\mathbf{S}_* \subseteq \mathcal{S}_\alpha$ satisfies the sparsity constraints, that is, $$\mathcal{S}_\alpha \triangleq \{\mathbf{S} : \|\mathbf{S}_*(i,:)\|_0 \leq \alpha n \quad \text{and} \quad \|\mathbf{S}_*(:,j)\|_0 \leq \alpha m \ \forall (i,j) \in [m]\times[n]\}, \tag{2.4}$$ where α is a common sparsity ratio shared by the rows and columns.

Here we gave a list of spectral initializations used, as well as the perturbation guarantees. Unless mentioned otherwise:

- $\mathbf{A}_*$ is of rank r.
- $\mathbf{U} \in \mathbb{R}^{n\times r}, \mathbf{\Sigma} \in \mathbb{R}^{r\times r}$, and $\mathbf{V} \in \mathbb{R}^{r\times m}$ are the compact rank- r SVD of the intermediate matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$, **but not** $\mathbf{A}_*$.
- $\zeta > 0$ is a parameter to control the ratio of perturbation w.r.t. the signal $\sigma_{\min}(\mathbf{A}) = \sigma_r(\mathbf{A})$. It affects the precise high probability estimate for the perturbation guarantees. We will not delve into details.
- c_0 and c_1 are sufficiently large constants, which vary by problem.
- μ is the incoherence parameter, which varies by problem.
- κ is the condition number of $\mathbf{A}_*$.
- $\operatorname{dist}(\mathbf{X}, \mathbf{X}_*)$ is the **Procrustes distance** between two matrices $\mathbf{X}, \mathbf{X}_* \in \mathbb{R}^{m\times r}$ of the same size that mods out rotations, that is,

$$\operatorname{dist}(\mathbf{X}, \mathbf{X}_*) = \min_{\mathbf{Q}\in\mathcal{O}^{r\times r}} \|\mathbf{X}\mathbf{Q} - \mathbf{X}_*\|_F, \tag{2.5}$$

  where $\mathcal{O}^{r\times r}$ is the set of all orthogonal matrices of size $r \times r$.

To formalize the intuition behind spectral methods, we must first quantify the initial proximity to the ground-truth target.

2.1.2 Proximity to ground-truth via matrix perturbation

The derivation of the perturbation falls into three steps:

First, we express our surrogate data matrix $\mathbf{M}$ as $\mathbf{M} = \mathbf{M}_* + \Delta$, where

- $\mathbf{M}_*$ is the "clean" large-sample limit (which is often $\mathbf{E}[\mathbf{M}]$ or the ground-truth $\mathbf{A}_*$) whose eigenvectors perfectly align with $\mathbf{A}_*$.
- Δ is the perturbation caused by finite-sample noise or incomplete data.

The purpose of this decomposition is to make use of the Davis and Kahan (1970) and Wedin (1972) $\sin\Theta$ theorems to bound the distance and/or the principle angles between the intermediate matrix $\mathbf{M}$ and the ground-truth $\mathbf{A}_*$. Such a bound will be dictated by the perturbation Δ and the smallest singular value of $\mathbf{A}_*$. Second, for the $\sin\Theta$ theorems to hold, we need Δ to be small. The techniques to do so vary by context. In the case when RIP condition of the sensing operator can hold, as in the case for Matrix Sensing Tu et al. (2016), we can treat $\mathbf{M}_*$ as $\mathbf{A}_*$ and the perturbation $\|\mathbf{M} - \mathbf{A}_*\|$ can be bounded deterministically using algebraic techniques involving the RIP constant δ_{2r}. If not, we must treat $\mathbf{M}_*$ as its large-sample limit $\mathbf{E}[\mathbf{M}]$, and bound $\|\mathbf{M} - \mathbf{E}[\mathbf{M}]\|$ using Matrix Concentration Inequalities Tropp (2015).

Lastly, once we have successfully obtained a tight bound on $\|\Delta\|$, it can be used to plug-in the $\sin\Theta$ theorems. Beyond mere proximity, we must also ensure that the optimization landscape itself is amenable to GD within this neighborhood.

Table 2.2: Surrogate matrices, spectral initializations, and perturbation guarantees for the problems in Table 2.1.

Problems and Surrogate Matrix	Spectral Initializator	Perturbation Guarantee	Additional Assumptions
Matrix Sensing $\mathbf{M} \triangleq \frac{1}{p}\sum_{i=1}^{p} y_i \mathbf{A}_i = \mathcal{A}^*\mathcal{A}(\mathbf{A}_*)$	$\mathbf{X}_0 = \mathbf{U}\mathbf{\Sigma}^{1/2}$ $\mathbf{Y}_0 = \mathbf{V}^\top \mathbf{\Sigma}^{1/2}$	$\mathrm{dist}^2\left(\begin{bmatrix}\mathbf{X}_0 \\ \mathbf{Y}_0\end{bmatrix}, \begin{bmatrix}\mathbf{X}_* \\ \mathbf{Y}_*\end{bmatrix}\right) \le \zeta\sigma_r(\mathbf{A}_*)$ c.f. **Theorem 22** in Chi et al. (2019).	$\mathcal{A}$ satisfies $2r$-RIP with $\delta_{2r} < c_0\sqrt{\zeta}/(\sqrt{r}\kappa)$
Quadratic Sensing $\mathbf{M} \triangleq \frac{1}{p}\sum_{i=1}^{p} y_i \mathbf{a}_i\mathbf{a}_i^\top$	$\mathbf{X}_0 = \mathbf{U}\mathbf{\Sigma}^{1/2}$	$\mathrm{dist}^2(\mathbf{X}_0, \mathbf{X}_*) \le \zeta\sigma_r(\mathbf{A}_*)$ c.f. **Theorem 24** in Chi et al. (2019).	**Sample Complexity**: $p \ge c_0 n r^4 \log(n)$
Matrix Completion $\mathbf{M} \triangleq \sum_{(i,j)\in\Omega} \langle \mathbf{A}_{i,j}, \mathbf{A}_*\rangle \mathbf{A}_{i,j} = \frac{1}{p}\mathcal{P}_\Omega(\mathbf{A}_*)$	$\mathbf{X}_0 = \mathbf{U}\mathbf{\Sigma}^{1/2}$ $\mathbf{Y}_0 = \mathbf{V}^\top \mathbf{\Sigma}^{1/2}$	$\mathrm{dist}^2\left(\begin{bmatrix}\mathbf{X}_0 \\ \mathbf{Y}_0\end{bmatrix}, \begin{bmatrix}\mathbf{X}_* \\ \mathbf{Y}_*\end{bmatrix}\right) \le \zeta\sigma_r(\mathbf{A}_*)$ c.f. **Theorem 26** in Chi et al. (2019).	**Sample Complexity**: $np > c_0\mu^2 r^2 \log(n)$
Blind Deconvolution $\mathbf{M} \triangleq \frac{1}{p}\sum_{i=1}^{p} y_i \mathbf{b}_i\mathbf{a}_i^{\mathsf{H}}$	$\mathbf{h}_0 = \sqrt{\sigma}\boldsymbol{u}$ $\mathbf{x}_0 = \sqrt{\sigma}\mathbf{v}$	$\min_{\alpha\in\mathbb{C},\lvert\alpha\rvert=1} \{\lVert\alpha\mathbf{h}_0 - \mathbf{h}_*\rVert_2 + \lVert\alpha\mathbf{x}_0 - \mathbf{x}_*\rVert_2\} \le \zeta\lVert\mathbf{h}_*\rVert_2$ c.f. **Theorem 25** in Chi et al. (2019).	**Sample Complexity**: $np > c_0\zeta^{-2}\mu^2 K \log(p)$ Here recall K is the dimension of the $\mathbf{h}$ vector.
Robust PCA $\mathbf{M} \triangleq \frac{1}{p}\mathcal{P}_\Omega(\mathbf{\Gamma}_* - \mathbf{S}_0)$ with $\mathbf{S}_0 = \mathcal{H}_{c\alpha np}(\mathcal{P}_\Omega(\mathbf{\Gamma}_*))$ where the truncation operator $\mathcal{H}_l$ is given in equation (86) of Chi et al. (2019).	$\mathbf{X}_0 = \mathbf{U}\mathbf{\Sigma}^{1/2}$ $\mathbf{Y}_0 = \mathbf{V}^\top \mathbf{\Sigma}^{1/2}$	$\mathrm{dist}^2\left(\begin{bmatrix}\mathbf{X}_0 \\ \mathbf{Y}_0\end{bmatrix}, \begin{bmatrix}\mathbf{X}_* \\ \mathbf{Y}_*\end{bmatrix}\right) \le \zeta\sigma_r(\mathbf{A}_*)$ c.f. **Theorem 29** in Chi et al. (2019).	**Sample Complexity**: $n^2 p \ge c_0\mu r^2 n$ **Sparsity Ratio**: $\alpha \le c_1/(\mu r^{3/2})$

2.1.3 Local curvature condition

Convexity and smoothness are standard recipes for the convergence of first-order methods. Beyond this, we would like to highlight a particular condition that also guarantees convergence. Standard iterative methods take the form:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \mathbf{g}(\mathbf{x}_t), \tag{2.6}$$

where $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^d$ could take several forms. In standard GD, $\mathbf{g}(\mathbf{x}) = \nabla f(\mathbf{x})$.

Definition 2.1.1: Regularity condition (Definition 3 in Chi et al. (2019))

The function $\mathbf{g}$ is said to obey the regularity condition $\mathtt{RC}(\mu, \lambda, \zeta)$ for some $\mu, \lambda, \zeta > 0$ if

$$2\langle \mathbf{g}(\mathbf{x}), \mathbf{x} - \mathbf{x}_*\rangle \ge \mu\lVert\mathbf{g}(\mathbf{x})\rVert_2^2 + \lambda\lVert\mathbf{x} - \mathbf{x}_*\rVert_2^2, \tag{2.7}$$

for all $\mathbf{x} \in B(\mathbf{x}_*, \zeta)$, the open ball of radius ζ centered around $\mathbf{x}_*$.[a]

[a] Here $\mathbf{x}_*$ is a root of $\mathbf{g}$. In Chi et al. (2019), it is assumed that $\mathbf{x}_*$ is identifiable. When f is convex, $\mathbf{x}_*$ is unique.

The intuitive interpretation is as follows:

- For gradient methods to work, we want $-\mathbf{g}(\mathbf{x})$ to point roughly towards the optimum $\mathbf{x}_*$. A positive inner product means the angle between the update steps and the path to optimum is acute.
- The term $\mu\|\mathbf{g}(\mathbf{x})\|_2^2$ prevents the gradient from being too large relative to its alignment with the optimum (to prevent overshooting the target).
- The term $\lambda\|\mathbf{x} - \mathbf{x}_*\|$ ensures that the gradient is sufficiently steep to push the iterates toward the optimum when they are far from it.

The regularity condition can be viewed as a marriage between **strong convexity** and **smoothness**. Recall the first-order characterizations via coercivity and co-coercivity:

Definition 2.1.2: α-strong convexity via coercivity (Theorem 5.24 (iii) in Beck (2017))

Let $f : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ be a closed, convex, and proper function. It is said to be α-strongly convex if

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \alpha \|\mathbf{x} - \mathbf{y}\|^2, \tag{2.8}$$

for all $\mathbf{x}, \mathbf{y} \in \text{dom}(\nabla f)$.

Definition 2.1.3: β-smoothness via co-coercivity (Theorem 5.8 (iv) in Beck (2017))

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. We say f is β-smooth if

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq \frac{1}{\beta} \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2, \tag{2.9}$$

for all $\mathbf{x}, \mathbf{y} \in \text{dom}(f)$. [a]

[a] The co-coercivity is characterized using the dual norm $\|\cdot\|_*$ in Beck (2017). Since the 2-norm is self-dual, we just write $\frac{1}{\beta}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_2^2$ instead of $\frac{1}{\beta}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|_*^2$ for simplicity.

Now, if a function f is both α-strongly convex and β-smooth, then evaluating $\mathbf{y}$ at one local optimum $\mathbf{x}_*$ will yield $\nabla f(\mathbf{x}_*) = 0$. As a result, we have

$$\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}_* \rangle \geq \alpha \|\mathbf{x} - \mathbf{x}_*\|_2^2, \tag{2.10}$$

$$\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}_* \rangle \geq \frac{1}{\beta} \|\nabla f(\mathbf{x})\|_2^2. \tag{2.11}$$

Adding these two together gives:

$$2\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}_* \rangle \geq \frac{1}{\beta} \|\nabla f(\mathbf{x}_*)\|_2^2 + \alpha \|\mathbf{x} - \mathbf{x}_*\|_2^2. \tag{2.12}$$

This is the regularity condition with $\mu = 1/\beta$ and $\lambda = \alpha$. Even though the regularity condition is clearly a weaker condition than strong convexity and smoothness combined, gradient methods still enjoy convergence guarantees under this condition.

Proposition 2.1.4: Convergence under regularity condition (Lemma 2 in Chi et al. (2019))

Under $\mathtt{RC}(\mu, \lambda, \zeta)$, the iterative method (2.6) with a constant step-size $\eta_t \equiv \mu$ and $\mathbf{x}_0 \in B(\mathbf{x}_*, \zeta)$ obeys

$$\|\mathbf{x}_t - \mathbf{x}_*\|_2^2 \le (1 - \mu\lambda)^t \|\mathbf{x}_0 - \mathbf{x}_*\|_2^2 \qquad \text{for all } t \ge 0. \tag{2.13}$$

Proof. We proceed by induction: assume $\mathbf{x}_t \in B(\mathbf{x}_*, \zeta)$, then

$$\begin{aligned}
\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2^2 &= \|\mathbf{x}_t - \eta_t \mathbf{g}(\mathbf{x}_t) - \mathbf{x}_*\|_2^2 \\
&= \|\mathbf{x}_t - \mathbf{x}_*\|_2^2 + \eta_t^2 \|\mathbf{g}(\mathbf{x}_t)\|_2^2 - 2\eta_t \langle \mathbf{x}_t - \mathbf{x}_*, \mathbf{g}(\mathbf{x}_t) \rangle \\
&\le \|\mathbf{x}_t - \mathbf{x}_*\|_2^2 + \eta_t^2 \|\mathbf{g}(\mathbf{x}_t)\|_2^2 - \eta_t \left(\mu \|\mathbf{g}(\mathbf{x}_t)\|_2^2 + \lambda \|\mathbf{x}_t - \mathbf{x}_*\|_2^2 \right) \\
&= (1 - \eta\lambda) \|\mathbf{x}_t - \mathbf{x}_*\|_2^2 + \eta_t(\eta_t - \mu) \|\mathbf{g}(\mathbf{x}_t)\|_2^2 \\
&\le (1 - \eta_t\lambda) \|\mathbf{x}_t - \mathbf{x}_*\|_2^2.
\end{aligned} \tag{2.14}$$

The first inequality is a result of the regularity condition. Note that the last inequality is valid for any $\eta_t \in [0, \mu]$. Setting $\eta_t = \mu$, we arrive at:

$$\|\mathbf{x}_{t+1} - \mathbf{x}_*\|_2^2 \le (1 - \mu\lambda) \|\mathbf{x}_t - \mathbf{x}_*\|_2^2. \tag{2.15}$$

This also guarantees that $\mathbf{x}_{t+1} \in B(\mathbf{x}_*, \zeta)$. The claim follows under induction with the base case $\mathbf{x}_0 \in B(\mathbf{x}_*, \zeta)$ being part of our assumptions. □

With the convergence guarantees rooted in the regularity condition, we can now examine a practical application: the Orthogonal Procrustes Flow algorithm.

2.1.4 Putting it together: Orthogonal Procrustes Flow as a special case

We will use the Orthogonal Procrustes Flow Tu et al. (2016)[2], which is a special case of the quadratic sensing problem, as a canonical example of this two-stage approach. To clarify the optimization target, we begin with a general rank minimization problem of the form

$$\min_{\mathbf{A} \in \mathbb{S}_+^{n \times n}} \operatorname{rank}(\mathbf{A}) \quad \text{s.t.} \quad \mathcal{A}(\mathbf{A}) = \mathbf{b}, \tag{2.16}$$

where the affine transformation $\mathcal{A} : \mathbb{S}_+^{n \times n} \to \mathbb{R}^p$ is known a priori. Its k-th entry is given by $[\mathcal{A}(\mathbf{X})]_k \triangleq \langle \mathbf{A}_k, \mathbf{X} \rangle \triangleq \mathbf{tr}\left(\mathbf{A}_k^\top \mathbf{X}\right)$, where each $\mathbf{A}_k \in \mathbb{R}^{n \times n}$.

Directly solving (2.16) is computationally intractable. Therefore, the standard heuristic is to apply the Burer-Monteiro reparameterization, replacing the matrix $\mathbf{A}$ with a low-rank reparametrization $\mathbf{U}\mathbf{U}^\top$. We thereby aim to solve the unconstrained matrix bilinear equation:

$$\text{find } \mathbf{U} \in \mathbb{R}^{n \times r} \text{ s.t. } \mathcal{A}(\mathbf{U}\mathbf{U}^\top) = \mathbf{b}, \tag{2.17}$$

via a local search heuristic. In the case of quadratic measurements, the actual optimization objective we are minimizing is:

$$\min_{\mathbf{U} \in \mathbb{R}^{n \times r}} f(\mathbf{U}) \triangleq \frac{1}{4} \|\mathcal{A}(\mathbf{U}\mathbf{U}^\top) - \mathbf{b}\|_2^2 = \frac{1}{4} \sum_{k=1}^{p} \left(\langle \mathbf{A}_k, \mathbf{U}\mathbf{U}^\top \rangle - b_k \right)^2. \tag{2.18}$$

[2] the name is due to the use of the proscrustes distance in equation (2.5), and has less to do with the so-called Orthogonal Procrustes problem. Readers may refer to Gower and Dijksterhuis (2004) for this problem and we will not delve into details here.

Assuming a proper initialization $\mathbf{U}_0$ is attained, then the GD update we would like to apply is:

$$\mathbf{U}_{t+1} \triangleq \mathbf{U}_t - \frac{\eta_{t+1}}{\|\mathbf{U}_0\|^2}\nabla f(\mathbf{U}_t) = \mathbf{U}_t - \frac{\eta_{t+1}}{\|\mathbf{U}_0\|^2}\left[\sum_{k=1}^{p}\left(\langle \mathbf{A}_k, \mathbf{U}_t\mathbf{U}_t^\top\rangle - b_k\right)\mathbf{A}_k\mathbf{U}_t\right]. \tag{2.19}$$

To safely reach this unconstrained GD phase, we rely on a single, unified algorithm comprised of two distinct stages. The first stage constructs a surrogate matrix via iterative rank-projection to get within the basin of attraction of the ground-truth. Once sufficiently close, an SVD extracts the initialization $\mathbf{U}_0$. The second stage then proceeds with standard GD, exploiting the benign local geometry. This complete two-stage pipeline is summarized in the following algorithm.

Algorithm 8 Procrustes Flow (PF) (Algorithm 1 of Tu et al. (2016))

Require: $\{\mathbf{A}_k\}_{k=1}^p, \{b_k\}_{k=1}^p, \{\alpha_t\}_{t=1}^\infty, \{\eta_t\}_{t=1}^\infty, T_0 \in \mathbb{N}$

// Initialization phase

$\widetilde{\mathbf{M}}_0 \triangleq \mathbf{0}_{n\times n}$

for $t = 0, 1, \cdots, T_0 - 1$ **do**

  // Projection on to rank r PSD matrices

  $\widetilde{\mathbf{M}}_{t+1} \leftarrow \mathcal{P}_r\left[\widetilde{\mathbf{M}}_t - \alpha_{t+1}\sum_{k=1}^p\left(\langle \mathbf{A}_k, \widetilde{\mathbf{M}}_t\rangle - b_k\right)\mathbf{A}_k\right]$

end for

// SVD of $\widetilde{\mathbf{M}}_{T_0}$ with $\mathbf{Q} \in \mathbb{R}^{n\times r}, \mathbf{\Sigma} \in \mathbb{R}^{r\times r}$

$\mathbf{Q}\mathbf{\Sigma}\mathbf{Q}^\top \triangleq \widetilde{\mathbf{M}}_{T_0}$

$\mathbf{U}_0 \triangleq \mathbf{Q}\mathbf{\Sigma}^{1/2}$

// Gradient descent phase

repeat

  $\mathbf{U}_{t+1} \leftarrow \mathbf{U}_t - \frac{\eta_{t+1}}{\|\mathbf{U}_0\|^2}\left[\sum_{k=1}^{p}\left(\langle \mathbf{A}_k, \mathbf{U}_t\mathbf{U}_t^\top\rangle - b_k\right)\mathbf{A}_k\mathbf{U}_t\right]$

until convergence

The first stage is a variant of singular value hard-thresholding Jain et al. (2010), and this particular iterative rank-projection is due to Oymak et al. (2018), which enjoys geometric convergence due to the RIP property of the sensing operator $\mathcal{A}$.

Proposition 2.1.5: Convergence of iterative rank-projection (Lemma 5.2 in Tu et al. (2016) and Theorem 1.2 in Oymak et al. (2018))

Let $\mathbf{A} \in \mathbb{R}^{n_1\times n_2}$ be an arbitrary matrix of rank r. Let $\mathbf{b} = \mathcal{A}(\mathbf{A}) \in \mathbb{R}^p$ be p linear measurements. Consider the iterative updates:

$$\mathbf{Z}_{t+1} \leftarrow \mathcal{P}_r\left(\mathbf{Z}_t - \frac{1}{p}\sum_{k=1}^{p}(\langle \mathbf{A}_k, \mathbf{Z}_t\rangle - b_k)\mathbf{A}_k\right). \tag{2.20}$$

Then

$$\|\mathbf{Z}_t - \mathbf{A}\|_F \le \rho(\mathbf{A})^t\|\mathbf{Z}_0 - \mathbf{A}\|_F, \tag{2.21}$$

holds. Here, $\rho(\mathcal{A})$ is defined as

$$\rho(\mathcal{A}) \triangleq 2 \sup_{\substack{\|\mathbf{P}\|_F=1, \mathrm{rank}(\mathbf{P})\leq 2r \\ \|\mathbf{Q}\|_F=1, \mathrm{rank}(\mathbf{Q})\leq 2r}} |\langle \mathcal{A}(\mathbf{P}), \mathcal{A}(\mathbf{Q})\rangle - \langle \mathbf{P}, \mathbf{Q}\rangle|. \tag{2.22}$$

The contraction parameter ρ here is due to the RIP-ness of the sensing operator $\mathcal{A}$, and a specific bound is given in Candès (2008).

Discrepancies between two results

It is worth noting that the corresponding result in Oymak et al. (2018) is **Theorem 1.2**, which has a different form than what is stated here. That result establishes PGD for an unknown vector $\mathbf{x}$ from measurements $\mathbf{y} = \mathbf{A}\mathbf{x}$. In the noiseless regime ($w = 0$), the proof in Section 6.2 tracks the error $\mathbf{h}_t = \mathbf{z}_t - \mathbf{x}$ at each iteration. Specifically, equation (6.9) in Oymak et al. (2018) explicitly gives the step-by-step construction: $\|\mathbf{h}_{t+1}\|_2 \leq \kappa\rho(\mu_t)\|\mathbf{h}_t\|_2$. The relation between the two results is summarized as follows:

Oymak et al. (2018)	Tu et al. (2016)
True signal vector $\mathbf{x}$	True rank- r matrix $\mathbf{A}$
Iterative estimate vector $\mathbf{z}_t$	Iterative estimate matrix $\mathbf{Z}_t$
Error term $\mathbf{h}_t = \mathbf{z}_t - \mathbf{x}$	Error term $\mathbf{Z}_t - \mathbf{M}$
Step size μ	Step size $1/p$
Constraint set $\mathcal{K}$ and projection $\mathcal{P}_{\mathcal{K}}$	Rank- r manifold and projection $\mathcal{P}_r$

Once a satisfactory initial matrix is formed, local refinement commences, relying heavily on the previously mentioned curvature properties. The second stage local refinement relies on the same regularity condition. Our objective function $f(\mathbf{U}) \triangleq \frac{1}{4}\|\mathcal{A}(\mathbf{U}\mathbf{U}^\top) - \mathbf{b}\|_{\ell_2}^2$ after sufficient numbers of iteration T_0 of stage one, will satisfy the regularity condition $\mathtt{RC}(\mu, \lambda, \zeta)$ with the gradient coefficient μ, the distance coefficient λ, and the radius coefficient ζ being

$$\mu = \frac{64}{425\|\mathbf{X}_*\|}, \quad \lambda = \frac{1}{2}\sigma_r^2(\mathbf{X}_*), \quad \text{and} \quad \zeta = \frac{1}{4}\sigma_r(\mathbf{X}_*). \tag{2.23}$$

Putting them altogether, we get:

Theorem 2.1.6: Convergence guarantee under two-stage method and quadratic measurements (Theorem 3.2 in Tu et al. (2016))

Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be an arbitrary rank-r symmetric p.s.d. matrix with singular values ordered as $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ and condition number $\kappa = \sigma_1/\sigma_r$. Assume $\mathbf{A} = \mathbf{X}_*\mathbf{X}_*^\top$ for some $\mathbf{X}_* \in \mathbb{R}^{n\times r}$ and let $\mathbf{b} = \mathcal{A}(\mathbf{A}) \in \mathbb{R}^p$ be p linear measurements. Furthermore, assume the mapping $\mathcal{A}$ obeys rank- $6r$ RIP with RIP constant $\delta_{6r} \leq 1/10$. Also let $\alpha_t = 1/p$ for all $t = 1, 2, \cdots$. Then using $T_0 \geq \log(\sqrt{r}\kappa) + 2$ iterations of the initialization phase of Procrustes Flow as stated in Algorithm 8 yields a solution $\mathbf{U}_0$ obeying

$$\mathrm{dist}(\mathbf{U}_0, \mathbf{X}_*) \leq \frac{1}{4}\sigma_r(\mathbf{X}_*). \tag{2.24}$$

Furthermore, take a constant step size $\eta_t = \eta$ for all $t = 1, 2, \cdots$, with $\eta \leq 36/425$. Then, starting from any initial solution obeying (2.24), the t-th iterate of Algorithm 8 satisfies:

$$\operatorname{dist}(\mathbf{U}_t, \mathbf{X}_*) \leq \frac{1}{4}\left(1 - \frac{8}{25}\frac{\eta}{\kappa}\right)^{t/2} \sigma_r(\mathbf{X}_*). \tag{2.25}$$

Connection to NMF

While the Orthogonal Procrustes Flow serves as a pristine, canonical example of the two-stage spectral initialization framework, it is instructive to clarify why this exact sequence of analysis, particularly the global reliance on the regularity condition, does not trivially map onto NMF. First, NMF generally lacks a known, unique ground-truth factor $\mathbf{X}_*$ to anchor the regularity condition globally. Second, and crucially, a recent result by Zhang (2026) demonstrates that NMF inherently contains spurious local minima. At these spurious critical points, the objective cannot uniformly satisfy a Polyak-Łojasiewicz (**PŁ**) or regularity condition. Thus, while Procrustes Flow effectively illustrates the *mechanics* of the two-stage pipeline, the unique geometric hurdles of NMF necessitate the more specialized initializations discussed in subsequent sections.

2.1.5 Implicit bias of (small) Gaussian initialization

There is another line of work Stöger and Soltanolkotabi (2021); Ye and Du (2021); Jiang et al. (2023) showing that gradient methods can achieve success even without using explicit spectral decomposition of the target matrix. A noteworthy piece is the one by Stöger and Soltanolkotabi Stöger and Soltanolkotabi (2021). In this work, the dynamics of gradient methods under small Gaussian initialization, with the assumption of **overparametrization**, resembles the dynamics under spectral initialization, especially during the early phase where GD behaves like a power method. Other techniques under Gaussian initialization include the **leave-one-out** analysis for phase retrieval Chen et al. (2021).

2.1.6 Some default options for initializations of NMF

NNDSVD: Non-negative double singular value decomposition (SVD)

Here, we introduce an initialization scheme that is specifically designed for NMF. Again, in the case of low-rank approximation of NMF, where the size of the factorization k is less than the rank of $\mathbf{A}$, our hope is to design $\mathbf{W}_0 \geq \mathbf{0}$ and $\mathbf{H}_0 \geq \mathbf{0}$ so that the residual

$$\mathbf{R}_0 = \mathbf{A} - \mathbf{W}_0\mathbf{H}_0 \qquad \text{with } \mathbf{W}_0 \in \mathbb{R}_+^{m\times k} \text{ and } \mathbf{H}_0 \in \mathbb{R}_+^{k\times n}, \tag{2.26}$$

is as small as possible.

Optimal low-rank approximation via SVD We first examine the fundamental approximation error constructed by the standard SVD. Suppose the target matrix $\mathbf{A} \in \mathbb{R}_+^{m\times n}$ is of rank r, then an explicit SVD of $\mathbf{A}$ is

$$\mathbf{A} = \sum_{j=1}^{r} \sigma_j \mathbf{u}_j \mathbf{v}_j^\top = \sum_{j=1}^{r} \sigma_j \mathbf{C}^{(j)}, \tag{2.27}$$

where

- σ_j is the j-th singular value, i.e. $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$.
- $\mathbf{u}_j \in \mathbb{R}^{m\times 1}$ is the j-th left singular vector.
- $\mathbf{v}_j \in \mathbb{R}^{n\times 1}$ is the j-th right singular vector.
- $\mathbf{C}^{(j)} \triangleq \mathbf{u}_j \mathbf{v}_j^\top$ is the j-th unit rank factor.

In the case of low-rank approximation, the Eckart-Young-Mirsky theorem Mirsky (1960) guarantees that the optimal unconstrained rank- k approximation $k \leq r$ is formed by taking the top k singular components:

$$\mathbf{A}^{(k)} = \sum_{j=1}^{k} \sigma_j \mathbf{C}^{(j)}. \tag{2.28}$$

The leftover approximation error, which we call $\mathbf{E}^{(k)}$, is the sum of the discarded trailing components:

$$\mathbf{E}^{(k)} \triangleq \mathbf{A} - \mathbf{A}^{(k)} = \sum_{j=k+1}^{r} \sigma_j \mathbf{C}^{(j)}. \tag{2.29}$$

From this, a matrix decomposition is

$$\mathbf{A} = \mathbf{A}^{(k)} + \mathbf{E}^{(k)}. \tag{2.30}$$

SVD with uniform truncation and the sign ambiguity Let $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be the SVD of $\mathbf{A}$. To construct a non-negative initialization from the optimal unconstrained approximation $\mathbf{A}^{(k)}$, a naive approach is to equally distribute the spectral information to the left and right blocks and simply truncate the negative values. Taking the top k singular value-vector pairs (denoted by $\mathbf{U}_k$, $\mathbf{\Sigma}_k$, and $\mathbf{V}_k$), one could define a uniform truncation:

$$\mathbf{W}_{\text{uniform}} = [\mathbf{U}_k \mathbf{\Sigma}_k^{1/2}]_+ \qquad \text{and} \qquad \mathbf{H}_{\text{uniform}} = [\mathbf{\Sigma}_k^{1/2} \mathbf{V}_k^\top]_+. \tag{2.31}$$

Notice that the j-th column of $\mathbf{U}\mathbf{\Sigma}^{1/2}$ is simply $\sigma_j^{1/2}\mathbf{u}_j$ and the j-th row of $\mathbf{\Sigma}^{1/2}\mathbf{V}^\top$ is $\sigma_j^{1/2}\mathbf{v}_j^\top$. After truncation, an equivalent column-wise description of $\mathbf{W}_0$ and row-wise description of $\mathbf{H}_0$ is:

$$\mathbf{W}_{\text{uniform}}(:, j) = \sigma_j^{1/2}[\mathbf{u}_j]_+ \qquad \text{and} \qquad \mathbf{H}_{\text{uniform}}(j, :) = \sigma_j^{1/2}[\mathbf{v}_j]_+, \tag{2.32}$$

where the bracket $[\cdot]_+$ acts as an entrywise operation retaining only positive elements.[3]

However, this formulation is inherently ill-defined because the SVD is not unique. Specifically, if $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ is a valid SVD, then so is $\mathbf{A} = (\mathbf{U}\mathbf{K})\mathbf{\Sigma}(\mathbf{V}\mathbf{K})^\top$, where $\mathbf{K} = \text{diag}(\pm 1, \ldots, \pm 1)$

[3]When x is a scalar, $[x]_+ \triangleq \max\{x, 0\}$ and $[x]_- \triangleq |\min\{x, 0\}|$. When applied to a vector, this acts element-wise. Thus, $\mathbf{u}_{j+} \triangleq [\mathbf{u}_j]_+$ and $\mathbf{u}_{j-} \triangleq [\mathbf{u}_j]_-$, and similarly for $\mathbf{v}_{j+}$ and $\mathbf{v}_{j-}$.

is an arbitrary diagonal sign matrix. At the component level, this means $\mathbf{C}^{(j)} = \mathbf{u}_j\mathbf{v}_j^\top = (-\mathbf{u}_j)(-\mathbf{v}_j)^\top$. Consequently, taking the positive part $[\mathbf{u}_j]_+$ versus the negative part $[\mathbf{u}_j]_-$ (which corresponds to the positive elements of $-\mathbf{u}_j$) is completely arbitrary and relies entirely on the sign convention returned by the underlying numerical solver. It is not theoretically sound to assume that $[\mathbf{u}_j]_+$ is inherently a better initialization than $[\mathbf{u}_j]_-$.

NNDSVD NNDSVD Boutsidis and Gallopoulos (2008) avoids blind truncation. It extracts the positive and negative parts for each rank-1 component formed by $(\mathbf{u}_j, \mathbf{v}_j)$ and objectively checks which combination approximates the original term $\sigma_j\mathbf{u}_j\mathbf{v}_j^\top$ better. It calculates the norms of both the positive and negative lobes:

$$\mu_{j+} \triangleq \|\mathbf{u}_{j+}\|\|\mathbf{v}_{j+}\| \qquad \text{and} \qquad \mu_{j-} \triangleq \|\mathbf{u}_{j-}\|\|\mathbf{v}_{j-}\|. \tag{2.33}$$

By the Perron-Frobenius theorem, the first singular vectors $(\mathbf{u}_1, \mathbf{v}_1)$ of a non-negative matrix are strictly non-negative. Therefore, we initialize the first column of $\mathbf{W}_{\text{NNDSVD}}$ and the first row of $\mathbf{H}_{\text{NNDSVD}}$ directly:

$$\mathbf{W}_{\text{NNDSVD}}(:,1) = \sqrt{\sigma_1}\mathbf{u}_1 \qquad \text{and} \qquad \mathbf{H}_{\text{NNDSVD}}(1,:) = \sqrt{\sigma_1}\mathbf{v}_1^\top. \tag{2.34}$$

For subsequent columns and rows, the choice depends on the magnitude of μ_+ and μ_-. We initialize the rows/columns according to whether the positive or negative component is more dominant. Specifically, for $2 \le j \le r$:

If $\mu_{j+} \ge \mu_{j-}$:

$$\mathbf{W}_{\text{NNDSVD}}(:,j) = \sqrt{\sigma_j\mu_{j+}}\frac{\mathbf{u}_{j+}}{\|\mathbf{u}_{j+}\|} \qquad \text{and} \qquad \mathbf{H}_{\text{NNDSVD}}(j,:) = \sqrt{\sigma_j\mu_{j+}}\frac{\mathbf{v}_{j+}^\top}{\|\mathbf{v}_{j+}\|}. \tag{2.35}$$

Similarly if $\mu_{j+} < \mu_{j-}$:

$$\mathbf{W}_{\text{NNDSVD}}(:,j) = \sqrt{\sigma_j\mu_{j-}}\frac{\mathbf{u}_{j-}}{\|\mathbf{u}_{j-}\|} \qquad \text{and} \qquad \mathbf{H}_{\text{NNDSVD}}(j,:) = \sqrt{\sigma_j\mu_{j-}}\frac{\mathbf{v}_{j-}^\top}{\|\mathbf{v}_{j-}\|}. \tag{2.36}$$

With this construction, one obtains the following bound on the initialization error:

Theorem 2.1.7: Initialization error estimate of NNDSVD (Proposition 6 in Boutsidis and Gallopoulos (2008))

Given $\mathbf{A} \in \mathbb{R}_+^{m\times n}$ and the pair $(\mathbf{W}_{\text{NNDSVD}}, \mathbf{H}_{\text{NNDSVD}})$ initialized by the NNDSVD scheme, then the residual $\mathbf{R}_{\text{NNDSVD}} = \mathbf{A} - \mathbf{W}_{\text{NNDSVD}}\mathbf{H}_{\text{NNDSVD}}$ is bounded as follows:

$$\|\mathbf{E}^{(k)}\|_F \le \|\mathbf{R}_{\text{NNDSVD}}\|_F \le \|\mathbf{E}^{(k)}\|_F + \|\widehat{\mathbf{E}}_{\text{NNDSVD}}\|_F, \tag{2.37}$$

where

$$\|\widehat{\mathbf{E}}_{\text{NNDSVD}}\|_F \le \sum_{j=2}^{k}\left(\sigma_2[\mathbf{C}_+^{(j)}] + 1\right)\sigma_j \le 2\sum_{j=2}^{k}\sigma_j. \tag{2.38}$$

Here $\mathbf{C}_+^{(j)}$ denotes the positive section of the j-th rank 1 singular component of $\mathbf{A}$ and $\sigma_2[\mathbf{C}_+^{(j)}]$ denotes the second largest singular value of the matrix $\mathbf{C}_+^{(j)}$.

Successive Projection Algorithm (SPA) under separability

Beyond the standard SVD-based truncations, another family of fast and robust recursive algorithms exists for NMF under the **separability assumption**. This approach, generalizing the Successive Projection Algorithm (SPA), sequentially identifies the extreme rays of the convex cone generated by the columns of the data matrix.

The theoretical guarantees of this algorithm rely heavily on two core assumptions regarding the target matrix $\mathbf{A}$ and the selection function f:

Assumption 2.1.8: Separability (Assumption 1 in Gillis and Vavasis (2014))

The target matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is separable if it can be written as $\mathbf{A} = \mathbf{WH} = \mathbf{W}[\mathbf{I}_r, \mathbf{H}']$, where $\mathbf{W} \in \mathbb{R}^{m \times r}$ has rank r, $\mathbf{H} \in \mathbb{R}_+^{r \times n}$, and the sum of the entries of each column of $\mathbf{H}'$ is at most one.

In the context of hyperspectral unmixing, this is equivalent to the pure-pixel assumption, which dictates that for each constitutive material (endmember), there exists at least one pixel in the image containing exclusively that material.

Assumption 2.1.9: Strong convexity of the selection function (Assumption 2 in Gillis and Vavasis (2014))

The function $f : \mathbb{R}^m \to \mathbb{R}_+$ is strongly convex with parameter $\mu > 0$, its gradient is Lipschitz continuous with constant L, and its global minimizer is the all-zero vector with $f(\mathbf{0}) = 0$.

The most standard and robust choice for the function f that satisfies this assumption is $f(\mathbf{x}) = \|\mathbf{x}\|_2^2$.

Given these assumptions, the recursive algorithm proceeds by selecting the column of the residual matrix $\mathbf{R}$ that maximizes the function f, and subsequently updating $\mathbf{R}$ by projecting all its columns onto the orthogonal complement of the selected column.

Algorithm 9 Recursive algorithm for separable NMF (SPA)

Require: Separable matrix $\mathbf{A} = \mathbf{WH} \in \mathbb{R}_+^{m \times n}$, the number r of columns to be extracted, and a strongly convex function f obeying Assumption 2

Ensure: Set of indices J such that $\mathbf{A}(:, J) = \mathbf{W}$ up to permutation

1: $\mathbf{R} \leftarrow \mathbf{A}$, $J \leftarrow \{\}$, $j \leftarrow 1$
2: **while** $\mathbf{R} \neq \mathbf{0}$ **and** $j \leq r$ **do**
3: $\quad j^* \leftarrow \arg\max_k f(\mathbf{R}_{:k})$
4: $\quad \mathbf{u}_j \leftarrow \mathbf{R}_{:j^*}$
5: $\quad \mathbf{R} \leftarrow \left(\mathbf{I} - \frac{\mathbf{u}_j \mathbf{u}_j^\top}{\|\mathbf{u}_j\|_2^2}\right) \mathbf{R}$
6: $\quad J \leftarrow J \cup \{j^*\}$
7: $\quad j \leftarrow j + 1$
8: **end while**

Convergence in the noiseless regime Because the function f is strongly convex and the matrix $\mathbf{W}$ is full rank, the algorithm is guaranteed to extract a set of indices corresponding exactly to the columns of $\mathbf{W}$ in the noiseless case. This is formally stated as:

Theorem 2.1.10: Noiseless recovery (Theorem 1 in Gillis and Vavasis (2014))

Let the matrix $\mathbf{A} = \mathbf{WH}$ satisfy Assumption 2.1.8 and the function f satisfy Assumption 2.1.9. Then Algorithm 9 recovers a set of indices J such that $\mathbf{A}(:, J) = \mathbf{W}$ up to permutation.

Relation with QR factorization When instantiated with $f(\mathbf{x}) = \|\mathbf{x}\|_2^2$, Algorithm 9 is intimately related to QR factorization with column pivoting. Specifically, Golub and Businger construct QR factorizations by performing, at each step of the algorithm, Householder reflections w.r.t. the specific column of the data matrix whose projection onto the orthogonal complement of the previously extracted columns has the maximum ℓ_2-norm. This aligns exactly with the greedy selection and orthogonal projection steps of SPA.

Robustness guarantee in the noisy regime In practical scenarios, the input data matrix is often perturbed by noise, such that $\mathbf{A}' = \mathbf{A} + \mathbf{N} = \mathbf{WH} + \mathbf{N}$. SPA demonstrates robustness by approximately identifying the columns of $\mathbf{W}$ even under small arbitrary perturbations.

Theorem 2.1.11: Robustness of SPA (Theorem 3 in Gillis and Vavasis (2014))

Let $\mathbf{A}' = \mathbf{WH} + \mathbf{N}$ be the perturbed matrix, and assume the noise is bounded by $\|\mathbf{n}_i\|_2 \leq \epsilon$ for all i. Under appropriate bounds on ϵ relative to the conditioning of $\mathbf{W}$, the index set J extracted by Algorithm 9 guarantees that there exists a permutation P of $\{1, 2, \ldots, r\}$ such that:

$$\max_{1 \leq j \leq r} \|\mathbf{a}'_{J(j)} - \mathbf{w}_{P(j)}\|_2 \leq \bar{\epsilon} = \epsilon \left(1 + 80 \frac{K(\mathbf{W})^2}{\sigma_r^2(\mathbf{W})} \frac{L}{\mu}\right), \tag{2.39}$$

where $K(\mathbf{W}) = \max_i \|\mathbf{w}_i\|_2$, $\sigma_r(\mathbf{W})$ is the r-th singular value of $\mathbf{W}$, and μ, L are the strong convexity and Lipschitz constants of f.

The bound essentially guarantees that if the perturbation is small enough, the extraction error will be tightly bounded by $\mathcal{O}(\epsilon/\sigma_r^2(\mathbf{W}))$.

Transitioning from spectral initializations, we next explore sketching techniques, which offer powerful computational shortcuts by compressing high-dimensional data while preserving essential structural characteristics.

2.2 Sketching techniques in optimization

Sketching is a family of probabilistic techniques for processing usually large matrices into a much smaller surrogate (i.e. "the sketch") while preserving their key mathematical structure, such as the column spaces, singular values, or Euclidean distances (or distances under other metrics) between the rows or columns. Two important motivations are its uses in low-rank approximations and least squares (subspace embedding):

- In low-rank approximation, given a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and its best rank- k approximation $\mathbf{A}^{(k)}$, where $k \leq \min\{m,n\}$, we hope to construct a probabilistic rank- k estimate $\widetilde{\mathbf{A}}^{(k)}$ possessing low complexity and is cheap to compute, while maintaining the approximation quality in the sense that

$$\|\mathbf{A} - \widetilde{\mathbf{A}}^{(k)}\|_2 \leq (1\pm\varepsilon)\|\mathbf{A} - \mathbf{A}^{(k)}\|_2 \quad \text{with high probability (w.h.p.).} \tag{2.40}$$

- In least-squares, given $\mathbf{A} \in \mathbb{R}^{m\times n}$, we are minimizing an over-determined system $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2$ with $m \gg n$, and we would therefore like to construct a probabilistic estimate $\mathbf{x}'$ s.t.

$$\|\mathbf{A}\mathbf{x}' - \mathbf{b}\|_2 \leq (1\pm\varepsilon)\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \qquad \text{w.h.p..} \tag{2.41}$$

These are all achieved by simplifying $\mathbf{A}$ into a simpler probabilistic surrogate. In what follows, we put more emphasis on its applications on the latter. For applications in low-rank approximations, readers may refer to Halko et al. (2011); Martinsson and Tropp (2020).

2.2.1 Some examples of sketches

To build a sketch, one generates a random test matrix $\mathbf{S}$ (i.e. the sketching matrix) and multiplies it with $\mathbf{A}$. These sketches fall into two categories, depending on whether the distribution of $\mathbf{S}$ depends on the data matrix $\mathbf{A}$.

1. **Data-oblivious sketches**: The distribution of the sketch matrix $\mathbf{S}$ does not depend on $\mathbf{A}$. Examples of this category include Gaussian matrices, Subsampled Randomized Hadamard Transform (**SRHT**), and CountSketch matrices. Due to the data-independent nature, they can be generated independently and applied in a streaming setting with minimal communication.
2. **Data-dependent sketches**: The distribution of $\mathbf{S}$ relies on properties of $\mathbf{A}$. The most common example is Leverage Score Sampling (**LESS**), where columns or rows are sampled with probabilities proportional to their statistical leverage scores. These yield tighter sample bounds for specific matrices but require a preliminary pass to approximate the distributions.

Before elaborating these examples, an important notion in measuring the sketch quality is the notion of subspace embedding. We say the sketch matrix $\mathbf{S} \in \mathbb{R}^{d\times m}$ is an ε-subspace embedding (for the column space) of the matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ if

$$\|\mathbf{S}\mathbf{A}\mathbf{x}\|_2^2 = (1\pm\varepsilon)\|\mathbf{A}\mathbf{x}\|_2^2, \forall \mathbf{x} \in \mathbb{R}^n. \tag{2.42}$$

An equivalent characterization is that

$$\|\mathbf{I}_n - (\mathbf{S}\mathbf{U})^\top \mathbf{S}\mathbf{U}\| = \|\mathbf{I}_n - \mathbf{U}^\top\mathbf{S}^\top\mathbf{S}\mathbf{U}\|_2 \leq \varepsilon, \tag{2.43}$$

for any $\mathbf{U} \in \mathbb{R}^{m\times n}$ whose columns form an orthonormal basis for the column space of $\mathbf{A}$.[4]

Gaussian Sketch

An obvious choice that is easy to implement is to let the entries of $\mathbf{S}$ draw entries from i.i.d. standard normal. With suitable scaling, we have the following guarantee:

[4] Notice that this is a deterministic statement.

Theorem 2.2.1: Embedding guarantee of Gaussian sketches (Theorem 2.3 in Woodruff (2014))

Fix $\varepsilon, \delta \in (0,1)$ and $\mathbf{S}(i,j) \sim \mathcal{N}(0, 1/d)$ with $\mathbf{S} \in \mathbb{R}^{d \times m}$. Then if $d = \Theta((n + \log(1/\delta))\varepsilon^{-2})$, for any $m \times n$ matrix $\mathbf{A}$, w.p. at least $1 - \delta$, $\mathbf{S}$ is a ε-subspace embedding of $\mathbf{A}$.

However, if the sketching matrix happens to be dense, then performing the sketch, i.e. the matrix multiplication, will take $\mathcal{O}(nmd)$ operations. The sub-sampled randomized Hadamard transform coming up next is designed to improve this.[5]

SRHT

For an input dimension n and sketch dimension d, the SRHT matrix $\mathbf{S} \in \mathbb{R}^{d \times m}$ is constructed as the product of three matrices

$$\mathbf{S} = \frac{1}{\sqrt{dm}} \mathbf{P} \mathbf{H}_m \mathbf{D}. \tag{2.44}$$

- $\mathbf{P} \in \mathbb{R}^{d \times m}$ is a **Sampling/Projection Matrix** that selects exactly d coordinates out of m uniformly at random.
- $\mathbf{H}_m \in \mathbb{R}^{m \times m}$ is a **Walsh-Hadamard Matrix**, where n is a power of 2. Its entry is defined as:

  $$\mathbf{H}_m(i,j) = \frac{1}{\sqrt{m}} (-1)^{\langle i,j \rangle}, \tag{2.45}$$

  where $\langle i,j \rangle = \sum_{z=1}^{\log_2 m} i_z j_z$ and $(i_{\log_2 m}, \cdots, i_1)$ and $(j_{\log_2 m}, \cdots, j_1)$ are the binary representations of the integer indices i and j.
- $\mathbf{D} \in \mathbb{R}^{m \times m}$ is a **Random Sign Matrix** where each diagonal entry $\mathbf{D}(i,i)$ is drawn independently from $\{-1, +1\}$ with equal probability.

Theorem 2.2.2: Embedding guarantee of SRHT (Theorem 2.11 in Woodruff (2014))

Let $\mathbf{S}$ be constructed as above, then given a precision parameter $\varepsilon > 0$, if the sample number d is of

$$d = \Omega(\varepsilon^{-2}(\log n)(\sqrt{n} + \sqrt{\log m})^2), \tag{2.46}$$

then, for any fixed $m \times n$ matrix $\mathbf{U}$ with orthonormal columns, we have

$$\|\mathbf{I}_n - \mathbf{U}^\top \mathbf{S}^\top \mathbf{S} \mathbf{U}\|_2 \leq \varepsilon \qquad \text{w.h.p.}. \tag{2.47}$$

Here is how these three matrices collaborate with each other:

- The matrix $\mathbf{H}_m$ acts like a Fourier transform. It takes the randomized input and massively mixes each coordinate with all the other coordinates. If the original data vector had one massive spike, $\mathbf{H}_m$ flattens that information uniformly across all m coordinates. In addition, due to the recursive and block-wise pattern, no dense matrix multiplication

[5] In the ideal case when $\mathbf{A}$ is highly sparse, then it only takes $\mathcal{O}(d\texttt{nnz}(\mathbf{A}))$ operations. However, a common pitfall of all sketching frameworks is that sketching destroys the sparsity/sparsity pattern of the data matrix $\mathbf{A}$.

is needed. By the **Fast Walsh-Hadamard Transform (FWHT)**, it takes only $\mathcal{O}(m \log m)$ operations.

- In the ideal case, $\mathbf{H}_m$ already distributed the information of the input vector across all m coordinates. No single coordinate carries too much/too little information, so the uniform sampling matrix $\mathbf{P}$ could preserve the signal properly.
- Notice that $\mathbf{H}_m$ is a deterministic matrix, so one could still design an input vector that resisted the "mixing effect" of $\mathbf{H}_m$. Adding a randomization matrix $\mathbf{D}$ helps the proper mixing of $\mathbf{H}_m$.

In summary, a precise operation count for applying the $d \times m$ SRHT sketching matrix $\mathbf{S}$ to an $m \times n$ data matrix $\mathbf{A}$ is as follows:

1. **Apply $\mathbf{D}$**: Multiplying $\mathbf{D} \cdot \mathbf{A}$ simply flips the signs of the rows of $\mathbf{A}$, requiring exactly $\mathcal{O}(mn)$ operations.
2. **Apply $\mathbf{H}_m$**: Standard matrix multiplication is avoided. Instead, applying the Fast Walsh-Hadamard Transform (FWHT) independently to each of the n columns of length m takes $\mathcal{O}(m \log m)$ per column. For n columns, this takes exactly $\mathcal{O}(mn \log m)$ operations.
3. **Apply $\mathbf{P}$**: Extracting d rows from the transformed matrix takes $\mathcal{O}(dn)$ operations.

Therefore, the total time complexity for the sketching step is $\mathcal{O}(mn \log m)$. Notice that while this operation count is indeed independent of the target sketch dimension d, one cannot simply choose $d = m$. The parameters represent fundamentally different quantities: m is the original, fixed data dimension, while d is the reduced sketch dimension chosen by the user to satisfy an error tolerance ε. Setting $d = m$ would defeat the core purpose of sketching, as the goal is to make $d \ll m$ to drastically reduce the complexity of the *subsequent* downstream computations (e.g., dropping the cost of solving a sketched least-squares problem from $\mathcal{O}(mn^2)$ to $\mathcal{O}(dn^2)$).

LESS

In previous examples, $\mathbf{S}$ is generated independently of the properties of $\mathbf{A}$. One way to perform data-dependent sketching is to pick a subset of its rows. In the case when all the important information is concentrated in a few rows, uniform sampling will miss them. **Leverage scores** measure the importance of each row: (**Definition 2.6** in Woodruff (2014)):

- Again, let $\mathbf{A} \in \mathbb{R}^{m \times n}$ be a skinny matrix ($m \gg n$) with orthonormal columns.
- The i-th **leverage score** of $\mathbf{A}$ is defined as $l_i^2 \triangleq \|\mathrm{e}_i^\top \mathbf{A}\|_2^2$, which is simply the squared Euclidean norm of the i-th row of $\mathbf{A}$.
- Since $\mathbf{A}$ has orthonormal columns, the sum of all leverage scores is exactly n. Therefore, an ideal sampling probability distribution can be defined as $p_i = l_i^2/n$.
- Since computing the exact leverage score can be slow, we allow an approximate distribution $\mathbf{q} = (q_1, \cdots, q_m)$ as $q_i \geq \beta p_i$ for some constant $\beta > 0$.

In the end, we would like to sample d rows, and hence construct the sampling matrix $\mathbf{\Omega} \in \mathbb{R}^{m \times d}$ and $\mathbf{D} \in \mathbb{R}^{d \times d}$ which is a diagonal re-scaling matrix. For each of the d trials, we independently

pick a row index $i \in [m]$ w.p. q_i. If row i is chosen on trial j, we set $\mathbf{\Omega}(i,j) = 1$ and set the diagonal rescaling weight $\mathbf{D}(j,j) = 1/\sqrt{q_i d}$. We have the following guarantee:

Theorem 2.2.3: Embedding guarantee of LESS (Theorem 2.11 in Woodruff (2014))

Suppose $\mathbf{A} \in \mathbb{R}^{m\times n}$ has orthonormal columns. Suppose the number of sampled rows is

$$d > \frac{144n\ln(2n/\delta)}{\beta\varepsilon^2}, \tag{2.48}$$

and $\mathbf{\Omega}$ and $\mathbf{D}$ are constructed as above. Then w.p. $1-\delta$, simultaneously for all i:

$$1-\varepsilon \le \sigma_i^2(\mathbf{D}^\top\mathbf{\Omega}^\top\mathbf{A}) \le 1+\varepsilon, \tag{2.49}$$

where σ_i denotes the i-th largest singular value.

Discrepancy on the definition of embedding

At first glance, the bound on the singular values in (2.49) might appear distinct from the standard ℓ_2 subspace embedding guarantee considered earlier (i.e., $\|\mathbf{SAx}\|_2^2 = (1\pm\varepsilon)\|\mathbf{Ax}\|_2^2$ for all $\mathbf{x}$). However, they are mathematically identical under the theorem's specific assumption that the data matrix $\mathbf{A}$ has orthonormal columns ($\mathbf{A}^\top\mathbf{A} = \mathbf{I}_n$). For any unit vector $\mathbf{x} \in \mathbb{R}^n$, the geometry of orthonormal columns guarantees that $\|\mathbf{Ax}\|_2^2 = 1$. Substituting this into the standard subspace embedding definition simplifies the condition to $\|\mathbf{SAx}\|_2^2 = 1 \pm \varepsilon$ for all unit vectors $\mathbf{x}$. In linear algebra, the minimum and maximum values of $\|\mathbf{SAx}\|_2^2$ over all unit vectors $\mathbf{x}$ correspond exactly to the squared minimum and maximum singular values of the sketched matrix $\mathbf{SA}$. This leads to (2.49).

Difference between a left sketch $\mathbf{SA}$ and a right sketch $\mathbf{AS}$

So far, our setting has been left sketches. Given a matrix $\mathbf{A} \in \mathbb{R}^{n\times m}$, a left-sketch $\mathbf{S} \in \mathbb{R}^{d\times n}$ with $n \gg d$ reduces the number of rows. It acts on the column vectors of $\mathbf{A}$, meaning that it preserves the geometry of the **column space**. It is universally used for over-determined least squares, where the constraint is on the rows (e.g. $\min\|\mathbf{Ax}-\mathbf{b}\| \to \min\|\mathbf{SAx}-\mathbf{Sb}\|$).

In a right sketch, we draw a matrix $\mathbf{\Omega} \in \mathbb{R}^{m\times d}$ with $m \gg d$ and form $\mathbf{A\Omega} \in \mathbb{R}^{n\times d}$. Here we are reducing the number of columns. It captures the action on the **row space**. It is fundamentally used to extract the dominant spectral structure of $\mathbf{A}$ for low-rank approximations.

These sketching paradigms lay the foundation for scalable optimization by strategically embedding high-dimensional data into more manageable spaces.

2.2.2 Sketching techniques in iterative methods

The use of sketching in optimization and iterative methods falls into three categories:

1. **Sketch-and-Solve**: This is the most direct application of dimensionality reduction. When facing an over-determined least squares problem with $\mathbf{A}$ being "tall and skinny", we compress the rows of $\mathbf{A}$ and the target vector $\mathbf{b}$ using a sketching matrix. We then solve this

much smaller, sketched matrix directly using conventional tools. The resulting solution we get is a good approximation to the original solution w.h.p. This is also desirable if we aim for a low-precision solution, i.e. $\varepsilon = 10^{-1}$.

Concretely, in the ordinary least-squares problem we aim to find $\mathbf{x}_*$ that minimizes:

$$\mathbf{x}_* = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2, \tag{2.50}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ with $m \gg n$. We draw a sketching matrix $\mathbf{S} \in \mathbb{R}^{d\times m}$ with $d \ll m$ (e.g. via SRHT or CountSketch). We compute $\mathbf{SA} \in \mathbb{R}^{d\times n}$ and $\mathbf{Sb} \in \mathbb{R}^d$, and solve

$$\widetilde{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{SA}\mathbf{x} - \mathbf{Sb}\|_2. \tag{2.51}$$

2. **Iterative Sketching**:

   As the name suggests, this is a class of methods where we sketch our problem repeatedly and refine the estimate using conventional optimization methods. This usually leads to a moderate level of precision, e.g. $\varepsilon = 10^{-3}$ to 10^{-5}. A powerful framework in this area is the **Sketch-and-Project** framework Gower and Richtárik (2015), which is primarily aimed to solve linear systems but has many equivalent formulations and applications in stochastic and randomized methods. Its general strategy to solve a consistent linear system $\mathbf{Ax} = \mathbf{b}$ is to take a current guess, sketch the residual to find how far off we are in a random subspace, and project out our guess so that it perfectly satisfies the sketched system: let $\mathbf{B} \in \mathbb{S}^n_+$ and it yields the $\mathbf{B}$-inner product and its induced $\mathbf{B}$-norm:

   $$\langle \mathbf{x}, \mathbf{y}\rangle_{\mathbf{B}} \triangleq \langle \mathbf{Bx}, \mathbf{y}\rangle \qquad \text{and} \qquad \|\mathbf{x}\|_{\mathbf{B}} \triangleq \sqrt{\langle \mathbf{x}, \mathbf{x}\rangle_{\mathbf{B}}}. \tag{2.52}$$

   Suppose $\mathbf{x}_t \in \mathbb{R}^n$ is already computed in solving $\mathbf{Ax} = \mathbf{b}$, then $\mathbf{x}_{t+1}$ is generated via a sketch $\mathbf{S}$:

   $$\mathbf{x}_{t+1} = \arg\min_{\mathbf{x}\in\mathbb{R}^n} \|\mathbf{x} - \mathbf{x}_t\|^2_{\mathbf{B}} \qquad \text{such that } \mathbf{SAx} = \mathbf{Sb}. \tag{2.53}$$

   The convergence rate of these methods usually depends on the sketched geometry of the problem. Moreover, this framework establishes equivalences between other stochastic and randomized methods such as the Karczmarz method and Weighted SGD Needell et al. (2016). Other examples of iterative sketching include sketching for quasi-newton methods Pilanci and Wainwright (2017; 2016) as well as sketching for SDPs Yurtsever et al. (2021).

3. **Sketch-and-Precondition**: In the case we require machine-precision level accuracy, Sketch-and-Solve will be slow for small ε. The **Sketch-and-Precondition** framework Rokhlin and Tygert (2008) uses a sketching matrix to downsample the problem, but only to extract the geometric structure of the matrix (usually via a QR decomposition of the sketch). We then use this extracted structure as a pre-conditioner for a standard iterative method (like LSQR or Conjugate Gradient) on the originally massive problem. To be specific, to solve $\mathbf{Ax} = \mathbf{b}$, we again perform the sketch $\mathbf{SA}$, but then proceed to compute the compact QR decomposition of the sketched matrix

   $$\mathbf{SA} = \mathbf{Q}_{\mathbf{S}}\mathbf{R}. \tag{2.54}$$

   Because $\mathbf{SA}$ has the approximately the same sub-space geometry as $\mathbf{A}$, the small upper-triangular matrix $\mathbf{R} \in \mathbb{R}^{m\times m}$ acts as an excellent pre-conditioner. We then solve the well-conditioned system:

   $$\min_{\mathbf{y}} \|(\mathbf{AR}^{-1})\mathbf{y} - \mathbf{b}\|_2, \tag{2.55}$$

   using an iterative method like LSQR. Once we find $\mathbf{y}$, we recover $\mathbf{x} = \mathbf{R}^{-1}\mathbf{y}$.

Dereziński and Mahoney (2024) gives a modern perspective on the uses of sketching in optimization, including using modern RMT tools to perform complexity analysis.

2.3 A sketching based initialization

In this subsection, $r \triangleq \text{rank}(\mathbf{A})$ and d denotes the factorization dimension. By default, $d \geq r$.

Building upon the generic sketching strategies, we now pivot to specific formulations tailored to MF tasks. While an SVD based initialization enjoys good properties, an explicit SVD may be costly when the dimension of the target matrix is large. Specifically, for $\mathbf{A} \in \mathbb{R}^{m\times n}$, the ratio for the cost of a sketching (with $\mathbf{S} \in \mathbb{R}^{d\times m}$, $d \ll m, n$) and a compact SVD is roughly

$$\frac{\texttt{flop}(\mathbf{A} \mapsto \mathbf{SA})}{\texttt{flop}(\mathbf{A} \mapsto \mathbf{U\Sigma V}^\top)} \approx \frac{d}{\min(m,n)}. \tag{2.56}$$

Here, we will be looking at the unconstrained two-block MF problem

$$\min_{\mathbf{X},\mathbf{Y}} f(\mathbf{X},\mathbf{Y}) = \frac{1}{2}\|\mathbf{A} - \mathbf{XY}^\top\|_F^2, \tag{2.57}$$

with $\mathbf{X} \in \mathbb{R}^{m\times d}$ and $\mathbf{Y} \in \mathbb{R}^{n\times d}$.

2.3.1 PŁ-based approach

Ward and Kolda (2023) proposed an unbalanced sketching-based initialization, using **alternating** GD.

Assumption 2.3.1: Alternating GD (AltGD) (Assumption 1 in Ward and Kolda (2023))

For $\eta > 0$ and initial condition $(\mathbf{X}_0, \mathbf{Y}_0)$, the **asymmetric** update is

$$\mathbf{X}_{t+1} = \mathbf{X}_t - \eta\nabla_{\mathbf{X}} f(\mathbf{X}_t, \mathbf{Y}_t) \tag{2.58}$$
$$\mathbf{Y}_{t+1} = \mathbf{Y}_t - \eta\nabla_{\mathbf{Y}} f(\mathbf{X}_{t+1}, \mathbf{Y}_t), \tag{2.59}$$

with $\nabla_{\mathbf{X}} f(\mathbf{X},\mathbf{Y}) = (\mathbf{XY}^\top - \mathbf{A})\mathbf{Y}$ and $\nabla_{\mathbf{Y}} f(\mathbf{X},\mathbf{Y}) = (\mathbf{XY}^\top - \mathbf{A})^\top\mathbf{X}$.

There are several key components and highlights to this approach. First is an unbalanced initialization:

Assumption 2.3.2: Initialization and related quantities (Assumption 2 in Ward and Kolda (2023))

- Let $\mathbf{\Phi}_1, \mathbf{\Phi}_2 \in \mathbb{R}^{n\times d}$ be random matrices with i.i.d. $\mathcal{N}(0, 1/d)$ and $\mathcal{N}(0, 1/n)$ entries.
- Fix $C \geq 1, \nu < 1/3$ and $D \leq \frac{C}{9}\nu$, and η denotes the stepsize. Let

$$\mathbf{X}_0 = \frac{1}{\eta^{1/2} C\sigma_1(\mathbf{A})}\mathbf{A\Phi}_1 \quad \text{and} \quad \mathbf{Y}_0 = \eta^{1/2} D\sigma_1(\mathbf{A})\mathbf{\Phi}_2. \tag{2.60}$$

- For $\tau > 0$, define
$$\rho = \tau\left(1 - \frac{\sqrt{r-1}}{\sqrt{d}}\right). \tag{2.61}$$
- ϵ-optimality will be shown to be dependent on
$$\beta = \left(\frac{\rho}{C}\frac{\sigma_r(\mathbf{A})}{\sigma_1(\mathbf{A})}\right)^2. \tag{2.62}$$
- The probability of finding the ϵ-optimality will depend on
$$\hat{\delta} = (C_1\tau)^{d-r+1} + e^{-C_2 d} + e^{-r/2} + e^{-d/2} + 2e^{-d}, \tag{2.63}$$
where $C_1, C_2 > 0$ are universal constants.

We have the following guarantee:

Proposition 2.3.3: Initialization (Proposition 4.3 in Ward and Kolda (2023))

Assume $\mathbf{X}_0$ and $\mathbf{Y}_0$ are initialized properly according to Assumption 2.3.2
Further assume that $C \geq 1, \nu < 1$, and $D \leq \frac{C\nu}{9}$ (which is equivalent with $\frac{9D}{C} \leq \nu$).
Then the following holds w.h.p.:

1. $\sigma_r(\mathbf{X}_0) \geq \frac{1}{\sqrt{\eta}}\frac{\rho}{C}\frac{\sigma_r(\mathbf{A})}{\sigma_1(\mathbf{A})}$
2. $\sigma_1(\mathbf{X}_0) \leq \frac{3}{C\sqrt{\eta}}$
3. $\sigma_1(\mathbf{Y}_0) \leq \sqrt{\eta}C\nu\sigma_1(\mathbf{A})/3$
4. $(1-3\nu)^2\|\mathbf{A}\|_F^2 \leq f(\mathbf{X}_0, \mathbf{Y}_0) \leq (1+\nu)^2\|\mathbf{A}\|_F^2$

Here, the bounds on the spectrum of $\mathbf{X}_0$ and $\mathbf{Y}_0$ rely on a fundamental result on non-asymptotic random matrix theory (**RMT**):

Proposition 2.3.4: Deviation of extreme singular values of Gaussian matrices (Corollary 5.35 in Vershynin (2012))

Let $A \in \mathbb{R}^{N\times n}$ be a random matrix with i.i.d. standard normal with $N > n$ (hence a 'tall, skinny' matrix), then for all $t \geq 0$, we have:
$$\sqrt{N} - \sqrt{n} - t \leq \sigma_{\min}(\mathbf{A}) \leq \sigma_{\max}(\mathbf{A}) \leq \sqrt{N} + \sqrt{n} + t, \tag{2.64}$$
w.p. at least $1 - 2\exp(-t^2/2)$.

The initialization is defined as $\mathbf{X}_0 \sim \frac{1}{\sqrt{\eta}}\mathbf{A}\mathbf{\Phi}_1$ and $\mathbf{Y}_0 \sim \sqrt{\eta}\mathbf{\Phi}_2$. Since η is chosen to be small, this means $\mathbf{X}_0$ is scaled up to be large and $\mathbf{Y}_0$ is chosen to be small. Because of this extreme difference in scale, the updates to $\mathbf{X}_t$ remain very small relative to its massive initial scale, while the updates to $\mathbf{Y}_t$ happen rapidly relative to its tiny scale at the initialization. Consequently, the AltGD dynamics effectively mimic standard GD on a simpler linear regression problem, $\min g(\mathbf{Y}) = \|\mathbf{X}_0\mathbf{Y}^\top - \mathbf{A}\|_F^2$, where the large $\mathbf{X}$ is treated as if it were being fixed.

The second highlight is the preservation of column space of the $\mathbf{X}$-block.

Lemma 2.3.5: Boundedness of column span (ColSpan) of X_t (Lemma 4.5 in Ward and Kolda (2023))

With the alternating update scheme, if $\text{ColSpan}(\mathbf{X}_0) \subseteq \text{ColSpan}(\mathbf{X})$, then for all $t \geq 1$:

$$\text{ColSpan}(\mathbf{X}_t) \subseteq \text{ColSpan}(\mathbf{A}). \tag{2.65}$$

A defining feature of this initialization is that $\mathbf{X}_0$ is constructed using $\mathbf{A}\mathbf{\Phi}_1$, forcing its columns to reside entirely within the ColSpan of the target matrix $\mathbf{A}$. Analytically, this is achieved by induction arguments which also take advantage of the simple **additive structure** of the gradient update:

$$\begin{aligned}
\text{ColSpan}(\mathbf{X}_{t+1}) &= \text{ColSpan}(\mathbf{X}_t - \eta \nabla_{\mathbf{X}} f(\mathbf{X}_t, \mathbf{Y}_t)) \\
&= \text{ColSpan}(\mathbf{X}_t - \eta \mathbf{X}_t \mathbf{Y}_t^\top \mathbf{Y}_t + \eta \mathbf{A}\mathbf{Y}_t) \\
&\subseteq \text{ColSpan}(\mathbf{X}_t) \cup \text{ColSpan}(\eta \mathbf{X}_t \mathbf{Y}_t^\top) \cup \text{ColSpan}(\eta \mathbf{A}\mathbf{Y}_t) \\
&\subseteq \text{ColSpan}(\mathbf{A}),
\end{aligned} \tag{2.66}$$

where the implication from the second-last line to the last line is due to the fact that the column span of $\mathbf{X}_t$, $\mathbf{X}_t\mathbf{Y}_t^\top$, and $\mathbf{A}\mathbf{Y}_t$ all belong to the column span of $\mathbf{A}$, by induction or construction.

The column space preservation means that the $\mathbf{X}$-block is not rank degenerate throughout the entire trajectory, and therefore $\sigma_r(\mathbf{X}_t) = \sigma_{\min}(\mathbf{X}_t) > 0$ for all t. Also recall that we have a very refined control on $\sigma_{\min}(\mathbf{X}_0)$ at initialization. Also, with a slightly tighter control on a few constants, we have an even more refined control on $\sigma_{\min}(\mathbf{X}_{\widehat{T}})$ up to some finite horizon $\widehat{T}$. Specifically, for all $0 \leq t \leq \widehat{T} = \left\lfloor \frac{\beta}{8\eta^2 f_0} \right\rfloor$, where $f_0 \triangleq f(\mathbf{X}_0, \mathbf{Y}_0)$, we have

$$\sigma_r(\mathbf{X}_t) \geq \frac{1}{2}\sqrt{\frac{\beta}{\eta}}. \tag{2.67}$$

This is elaborated in **Lemma 4.4** of Ward and Kolda (2023). With these conditions, one could have a PŁ inequality on the $\mathbf{Y}$-block.

Lemma 2.3.6: PŁ inequality (Lemma 4.6 in Ward and Kolda (2023))

Suppose $\mathbf{A}$ has rank r, and $\text{ColSpan}(\mathbf{X}_t) \subseteq \text{ColSpan}(\mathbf{A})$ and $\sigma_r(\mathbf{X}_t) > 0$ for all t, then

$$\|\nabla_Y f(\mathbf{X}_t, \mathbf{Y}_{t-1})\|_F^2 \geq 2\sigma_r^2(\mathbf{X}_t) f(\mathbf{X}_t, \mathbf{Y}_{t-1}). \tag{2.68}$$

In other words, the PŁ condition Polyak (1963); Karimi et al. (2016) essentially guarantees that the gradient grows faster than the function value, ensuring that as long as the gradient is non-zero, the algorithm will make strictly positive progress toward the global min. In the asymmetric MF problem, the loss function $f(\mathbf{X}, \mathbf{Y}) = \frac{1}{2}\|\mathbf{X}\mathbf{Y}^\top - \mathbf{A}\|_F^2$ is highly non-convex and does not **globally** satisfy a PŁ inequality. However, linear convergence is still achievable without requiring a PŁ condition on the $\mathbf{X}$ block, and this is due to the presence of:

- **a local PŁ condition on $\mathbf{Y}$**: because the unbalanced initialization and the update invariance restrict $\mathbf{X}_t$ to the ColSpan of $\mathbf{A}$, and this ensures that its r-th singular value $\sigma_r(\mathbf{X}_t)$ remains strictly positive, a local PŁ emerges.
- **the sufficient decrease on one block**: The proof bounds the gradient step w.r.t. $\mathbf{Y}$ strictly away from zero.

$$\|\nabla_{\mathbf{Y}} f(\mathbf{X}_{t+1}, \mathbf{Y}_t)\|_F^2 \geq \frac{\beta}{2\eta} f(\mathbf{X}_{t+1}, \mathbf{Y}_t). \tag{2.69}$$

Now, the PŁ inequality guarantees sufficient decrease on the $\mathbf{Y}$ block. Putting everything together, one would have the final complexity result:

Theorem 2.3.7: Main convergence result for PŁ-based approach (Theorem 5.1 in Ward and Kolda (2023))

Suppose

1. $\mathbf{X}_0$ and $\mathbf{Y}_0$ are initialized properly as in Assumption 2.3.2
2. $C \geq 4$ (Stronger than $C \geq 1$ defined previously)
3. $\eta \leq \frac{9}{4C\nu\sigma_1(\mathbf{A})}$
4. $\hat{T} = \left\lfloor \frac{\beta}{8\eta^2 f_0} \right\rfloor$

Then w.h.p. (at least $1 - \hat{\delta}$), for all $t = 1, \cdots, \hat{T}$, we have

$$\|\mathbf{A} - \mathbf{X}_t \mathbf{Y}_t^T\|_F^2 \leq 2\exp(-\beta t/4) f_0 \leq \exp(-\beta t/4)(1+\nu)^2 \|\mathbf{A}\|_F^2. \tag{2.70}$$

In Theorem 5.1, the alternating updates act like a driver and a passenger. The $\mathbf{Y}$ block does all the "heavy lifting", i.e. the driver. Meanwhile, we only need the $\mathbf{X}$ update to not cause any damage, i.e. a passenger. Concretely:

- **$\mathbf{X}$ block**: Because the objective function is smooth and the gradients are Lipschitz continuous, **Lemma 4.1** in Ward and Kolda (2023) guarantees that as long as the $\mathbf{Y}$ block is not too large ($\|\mathbf{Y}_t\|^2 \leq 1/\eta$), a gradient step on $\mathbf{X}$ will never increase the loss, that is

$$f(\mathbf{X}_{t+1}, \mathbf{Y}_t) \leq f(\mathbf{X}_t, \mathbf{Y}_t). \tag{2.71}$$

  Again, we don't have strict decrease here since no PŁ condition is invoked. It is a "safe", non-increasing step.
- **$\mathbf{Y}$ block**: The $\mathbf{Y}$ block is where the strict geometric drop occurs. Due to the ColSpan argument, the $\mathbf{X}$ block maintains a **strictly positive r-th singular value** in the next step and for all future steps by induction, that is $\sigma_r(\mathbf{X}_{t+1}) = \sigma_{\min}(\mathbf{X}_{t+1}) > 0$. The objective inherently satisfies the PŁ inequality w.r.t. $\mathbf{Y}$. Taking the gradient step on $\mathbf{Y}$ forces a strict linear decrease:

$$f(\mathbf{X}_{t+1}, \mathbf{Y}_{t+1}) \leq (1 - \beta/4)\, f(\mathbf{X}_{t+1}, \mathbf{Y}_t). \tag{2.72}$$

Combining these two inequalities gives a joint decrease on both blocks:

$$f(\mathbf{X}_{t+1}, \mathbf{Y}_{t+1}) \leq (1 - \beta/4)\, f(\mathbf{X}_{t+1}, \mathbf{Y}_t) \leq (1 - \beta/4)\, f(\mathbf{X}_t, \mathbf{Y}_t). \tag{2.73}$$

Having explored the PŁ-based framework and its asymmetric dynamics, we now consider an alternative analytical lens that reveals convergence via fixed-point mappings.

2.3.2 Contraction-based approach

Xu et al. (2024) proposed to use a contraction/fixed-point method to prove the convergence under **roughly** the same sketching scheme, which works both for standard GD and Nesterov accelerated GD (**NAG**). At the heart of which is a careful decomposition of the GD dynamics. The residual dynamics was decomposed into a linear component and a nonlinear component. The linear component acts as a strict contraction map within a specific subspace containing the residual, and this serves as the main driver for convergence. The higher order nonlinear component remains close to their initialization and with a bootstrapping argument we can bound the nonlinear component with the total residual.

Concretely, the initialization is slightly different in the following sense: $\mathbf{X}_0$ is initialized in the same manner, but $\mathbf{Y}_0$ is chosen to be $\mathbf{0}$ for the simplicity of the proof. In Ward and Kolda (2023), $\mathbf{Y}_0$ is a small random initialization, so the empirical effect of such a modification is negligible.

Denote by $\mathbf{R}_t \triangleq \mathbf{X}_t\mathbf{Y}_t^\top - \mathbf{A}$, $f(\mathbf{X}_t, \mathbf{Y}_t) \triangleq \frac{1}{2}\|\mathbf{R}_t\|_F^2$. The authors uses a (single-block) GD of the following form:

$$\begin{bmatrix}\mathbf{X}_{t+1}\\ \mathbf{Y}_{t+1}\end{bmatrix} = \begin{bmatrix}\mathbf{X}_t - \eta\nabla_{\mathbf{X}} f(\mathbf{X}_t, \mathbf{Y}_t)\mathbf{Y}_t\\ \mathbf{Y}_t - \eta\nabla_{\mathbf{Y}} f(\mathbf{X}_t, \mathbf{Y}_t)\mathbf{X}_t\end{bmatrix} = \begin{bmatrix}\mathbf{X}_t - \eta(\mathbf{X}_t\mathbf{Y}_t^\top - \mathbf{A})\mathbf{Y}_t\\ \mathbf{Y}_t - \eta(\mathbf{X}_t\mathbf{Y}_t^\top - \mathbf{A})^\top\mathbf{X}_t\end{bmatrix}. \tag{2.74}$$

For NAG, recall that it is designed to work on a loss function ℓ that is strongly convex and smooth, and the scheme is as follows:

$$z_{t+1} = \widetilde{z}_t - \eta\nabla\ell(\widetilde{z}_t) \qquad \text{and} \qquad \widetilde{z}_{t+1} = z_{t+1} + \beta(z_{t+1} - z_t), \tag{2.75}$$

where again, η is the step-size and β serves as the momentum/extrapolation parameter. In the context of our problem, this becomes

$$\begin{bmatrix}\mathbf{X}_{t+1}\\ \mathbf{Y}_{t+1}\end{bmatrix} = (1+\beta)\begin{bmatrix}\mathbf{X}_t - \eta\mathbf{R}_t\mathbf{Y}_t\\ \mathbf{Y}_t - \eta\mathbf{R}_t^\top\mathbf{X}_t\end{bmatrix} - \beta\begin{bmatrix}\mathbf{X}_{t-1} - \eta\mathbf{R}_{t-1}\mathbf{Y}_{t-1}\\ \mathbf{Y}_{t-1} - \eta\mathbf{R}_{t-1}^\top\mathbf{X}_{t-1}\end{bmatrix}. \tag{2.76}$$

In both cases, we treat $\mathbf{X}_{-1} = \mathbf{X}_0$ and $\mathbf{Y}_{-1} = \mathbf{Y}_0$.

In the following, we will present the proof structure and key results in parallel for GD and NAG.

Initialization

Formally, draw a Gaussian random matrix $\mathbf{\Phi} \in \mathbb{R}^{n\times d}$ with $\mathbf{\Phi}_{ij} \sim \mathcal{N}(0, 1/d)$, and let

$$\mathbf{X}_0 = c\mathbf{A}\mathbf{\Phi} \qquad \text{and} \qquad \mathbf{Y}_0 = \mathbf{0}, \tag{2.77}$$

where c is a large constant specified later in the final complexity analysis for both GD and NAG. Controlling the spectrum of the random matrix $\mathbf{X}_0$ is in the same spirit with Ward and Kolda (2023) so we will not go into details here.

Residual dynamics decomposition

Here $\mathbf{r}_t$ is the vectorized residual matrix $\mathbf{R}_t = \mathbf{A} - \mathbf{X}_t\mathbf{Y}_t^\top$.

Proposition 2.3.8: GD dynamics (Proposition 2 in Xu et al. (2024))

Let $\mathbf{P}_t \triangleq \mathbf{X}_{t+1} - \mathbf{X}_t$ and $\mathbf{Q}_t \triangleq \mathbf{Y}_{t+1} - \mathbf{Y}_t$ for all $t \geq 0$, then GD admits the following dynamics:

$$\mathbf{r}_{t+1} = (\mathbf{I}_{mn} - \eta \mathbf{H}_0)\mathbf{r}_t + \boldsymbol{\xi}_t, \tag{2.78}$$

where $\mathbf{H}_t = (\mathbf{Y}_t \mathbf{Y}_t^\top) \otimes \mathbf{I}_m + \mathbf{I}_n \otimes (\mathbf{X}_t \mathbf{X}_t^\top)$ and $\boldsymbol{\xi}_t = \eta(\mathbf{H}_0 - \mathbf{H}_t)\mathbf{r}_t + \mathrm{vec}(\mathbf{P}_t \mathbf{Q}_t^\top)$.
Here we define the linear operator $\mathbf{T}_{\mathrm{GD}} \triangleq \mathbf{I}_{mn} - \eta \mathbf{H}_0$.

Proposition 2.3.9: NAG dynamics (Proposition 3 in Xu et al. (2024))

Let $\mathbf{P}_t = \mathbf{X}_{t+1} - \mathbf{X}_t$ and $\mathbf{Q}_t = \mathbf{Y}_{t+1} - \mathbf{Y}_t$ denote the update steps for $t \geq 0$. Then NAG admits the following dynamics:

$$\begin{bmatrix} \mathbf{r}_{t+1} \\ \mathbf{r}_t \end{bmatrix} = \begin{bmatrix} (1+\beta)(\mathbf{I}_{mn} - \eta \mathbf{H}_0) & -\beta(\mathbf{I}_{mn} - \eta \mathbf{H}_0) \\ \mathbf{I}_{mn} & 0 \end{bmatrix} \begin{bmatrix} \mathbf{r}_t \\ \mathbf{r}_{t-1} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\xi}_t \\ 0 \end{bmatrix}, \tag{2.79}$$

where $\mathbf{H}_t = (\mathbf{Y}_t \mathbf{Y}_t)^\top \otimes \mathbf{I}_m + \mathbf{I}_n \otimes (\mathbf{X}_t \mathbf{X}_t)^\top$, and $\boldsymbol{\xi}_t = \boldsymbol{\zeta}_t + \boldsymbol{\iota}_t$, with

$$\boldsymbol{\zeta}_t = \mathrm{vec}(\mathbf{P}_t \mathbf{Q}_t^\top) + \beta \mathrm{vec}(\mathbf{P}_{t-1} \mathbf{Q}_{t-1}^\top) + \beta\eta(\mathbf{R}_{t-1} \mathbf{Y}_{t-1} \mathbf{Q}_{t-1}^\top + \mathbf{P}_{t-1} \mathbf{X}_{t-1}^\top \mathbf{R}_{t-1}) \tag{2.80}$$

$$\boldsymbol{\iota}_t = (1+\beta)\eta(\mathbf{H}_0 - \mathbf{H}_t)\mathbf{r}_t - \beta\eta(\mathbf{H}_0 - \mathbf{H}_{t-1})\mathbf{r}_{t-1}. \tag{2.81}$$

Here again, we define $\mathbf{T}_{\mathrm{NAG}} \triangleq \begin{bmatrix} (1+\beta)(\mathbf{I}_{mn} - \eta \mathbf{H}_0) & -\beta(\mathbf{I}_{mn} - \eta \mathbf{H}_0) \\ \mathbf{I}_{mn} & 0 \end{bmatrix}$.

We breakdown proposition 2.3.9 from the following perspectives: the augmented state space in dynamics formulation, and the roles of higher order terms.

How to understand the (state-augmented) formulation of dynamics here?

A first order difference equation in control theory (for instance, in Linear Quadratic Regulator) takes the form

$$\mathbf{x}_{t+1} = \mathbf{Q}\mathbf{x}_t + \mathbf{B}\mathbf{u}. \tag{2.82}$$

where $\mathbf{x}_t, \mathbf{x}_{t+1} \in \mathbf{R}^{n \times 1}$, $\mathbf{u} \in \mathbf{R}^m$, and $\mathbf{Q} \in \mathbf{R}^{n \times n}$ and $\mathbf{B} \in \mathbf{R}^{n \times m}$. Standard GD naturally fits into this first-order form because the next step only depends on the current step: $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)$, where $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{Q}\mathbf{x} + \mathbf{q}^\top \mathbf{x}$. Momentum-based methods are based on a second-difference scheme, meaning that the update $\mathbf{x}_{t+1}$ depends on $\mathbf{x}_t$ and $\mathbf{x}_{t-1}$. Because standard linear system analysis requires the system to only depend on the immediate prior state, we need to augment the state vector to "remember" the past. We create a new augmented state $\mathbf{z}_t$ that stacks current and previous variables.
To illustrate by a concrete example, we consider the Polyak momentum (Heavy Ball Method) on a strictly convex quadratic $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{Q}\mathbf{x} + \mathbf{q}^\top \mathbf{x}$ with $\mathbf{Q} \succ 0$. The update rule is

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t) + \beta(\mathbf{x}_t - \mathbf{x}_{t-1}). \tag{2.83}$$

Here $\beta > 0$ is the momentum parameter. Substituting our gradient $\nabla f(\mathbf{x}) = \mathbf{Q}\mathbf{x}$:

$$\begin{aligned}\mathbf{x}_{t+1} &= \mathbf{x}_t - \eta\mathbf{Q}\mathbf{x}_t + \beta\mathbf{x}_t - \beta\mathbf{x}_{t-1} \\ &= [(1+\beta)\mathbf{I} - \eta\mathbf{Q}]\,\mathbf{x}_t - \beta\mathbf{x}_{t-1}.\end{aligned} \tag{2.84}$$

To turn this into a first order system, we define the augmented state vector:

$$\mathbf{z}_t = \begin{bmatrix}\mathbf{x}_t \\ \mathbf{x}_{t-1}\end{bmatrix}. \tag{2.85}$$

Now we can write the update rule as a single block-matrix equation:

$$\mathbf{z}_t = \begin{bmatrix}\mathbf{x}_{t+1} \\ \mathbf{x}_t\end{bmatrix} = \begin{bmatrix}(1+\beta)\mathbf{I} - \alpha\mathbf{Q} & -\beta\mathbf{I} \\ \mathbf{I} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{x}_t \\ \mathbf{x}_{t-1}\end{bmatrix}. \tag{2.86}$$

Similarly, Nesterov momentum can also be put into this form. More on the state-space augmentation technique can be found in **Section 4** of Lessard et al. (2016). To emphasize, the derivation above assumes a **quadratic** objective so that $\nabla f(\mathbf{x}_t) = \mathbf{Q}\mathbf{x}_t + \mathbf{q}$ (the constant $\mathbf{q}$ can be absorbed by shifting the origin). For a general, possibly non-convex, function f, the gradient $\nabla f(\mathbf{x}_t)$ is not linear in $\mathbf{x}_t$, and therefore cannot be written as $\mathbf{Q}\mathbf{x}_t$ with a fixed matrix $\mathbf{Q}$. Nevertheless, the linear state-space representation remains highly relevant for two reasons:

- **Local analysis**: Near a local minimum $\mathbf{x}^*$ where $\nabla f(\mathbf{x}^*) = 0$, one can linearize the dynamics using the Hessian $\mathbf{H} = \nabla^2 f(\mathbf{x}^*)$. For small steps, $\nabla f(\mathbf{x}_t) \approx \mathbf{H}(\mathbf{x}_t - \mathbf{x}^*)$, reducing the problem to the quadratic case.
- **Global convergence of momentum methods**: For strongly convex functions with Lipschitz gradient, the exact linear form does not hold, but one can still analyze the method using **integral quadratic constraints (IQCs)** or by "lifting" the system to a higher-dimensional space where the nonlinearity is bounded by quadratic inequalities, see more in Lessard et al. (2016).

Thus, while the matrix equation $\mathbf{z}_{t+1} = \mathbf{A}\mathbf{z}_t$ is exact only for quadratics, it serves as the fundamental building block for understanding and extending momentum methods to more general problems.

Finally, $\boldsymbol{\zeta}_t$ and $\boldsymbol{\iota}_t$ serves the following purposes:

The roles of the higher order terms

- $\boldsymbol{\iota}_t$ represents the "dynamic shift" or the "curvature shift". The core of the authors' convergence relies on showing that the residual contracts linearly according to the initial state matrix $\mathbf{H}_0$. However, the actual update depends on the current state matrices $\mathbf{H}_t$ and $\mathbf{H}_{t-1}$. The $\boldsymbol{\iota}_t$ term captures the error introduced by approximating the current, time-varying curvature with the initial curvature $\mathbf{H}_0$. As long as the iterates $\mathbf{X}_t$ and $\mathbf{Y}_t$ do not move too far from their initializations, $\mathbf{H}_t$ remains close to $\mathbf{H}_0$. Therefore this rate is bounded and shrinks at a rate of $\mathcal{O}(\theta^t)$.

- $\boldsymbol{\zeta}_t$ represents the "higher-order terms". Note that it is made up of terms like $\text{vec}(\mathbf{P}_t\mathbf{Q}_t^\top)$ and momentum-delayed versions of these products. Here it is defined that $\mathbf{P}_t \triangleq \mathbf{X}_{t+1} - \mathbf{X}_t$ and $\mathbf{Q}_t \triangleq \mathbf{Y}_{t+1} - \mathbf{Y}_t$. When the residual matrices are multiplied together during the GD updates, it creates interaction terms between variables. If the residual $\mathbf{R}_t$ is shrinking at a rate of θ, the step increments $\mathbf{P}_t$ and $\mathbf{Q}_t$ also shrink. Consequently, their product shrinks faster, yielding a bound of $\mathcal{O}(\theta^{2t})$.

To ensure these error terms remain manageable, it is critical to confine the dynamics within a well-behaved subspace.

Column space preservation

Lemma 2.3.10: Eigensubspace (Lemma 1 in Xu et al. (2024))

Let $\mathcal{H} \subseteq \mathbb{R}^{mn}$ denote the linear subspace containing all eigenvectors of $\mathbf{H}_0$ with positive eigenvalues. If $\mathbf{X}_0$ is initialized as in (2.77), then we have

$$\mathcal{H} = (\text{col}(\mathbf{A}))^n \qquad \text{and} \qquad \{\mathbf{r}_t, \boldsymbol{\xi}_t\}_t \subset \mathcal{H}, \tag{2.87}$$

where $\mathbf{H}_0, \mathbf{r}_t$, and $\boldsymbol{\xi}_t$ are defined as in Proposition 2.3.8.

Again, $\mathcal{H}$ is the linear subspace spanned by all eigenvectors of the initial "pseudo-Hessian" matrix $\mathbf{H}_0$ that have **strictly positive eigenvalues**. To unpack this, recall our unbalanced initialization $\mathbf{X}_0 = c\mathbf{A}\boldsymbol{\Phi}$ and $\mathbf{Y}_0 = \mathbf{0}$:

- $\mathbf{X}_0$ is formed by multiplying the target matrix $\mathbf{A}$ by a random Gaussian sketch $\boldsymbol{\Phi}$, $\text{col}(\mathbf{X}_0) \subseteq \text{col}(\mathbf{A})$. With the fact that these two matrices are of the same rank, $\mathbf{X}_0$ shares the exact same column space as $\mathbf{A}$.
- The matrix $\mathbf{H}_0$ is given by $\mathbf{I}_n \otimes (\mathbf{X}_0\mathbf{X}_0^\top)$.
- Due to properties of the Kronecker product, the positive eigenspace of $\mathbf{H}_0$ is simply the column space of $\mathbf{X}_0$ repeated n times.

For the orthogonal complement to $\mathcal{H}$ is the null space of $\mathbf{H}_0$. In this subspace, eigenvalues of $\mathbf{H}_0$ are exactly zero. This complement is highly problematic if our algorithm ever enters it. To use GD (Proposition 2.3.8) as an example:

- The linear part of the GD update is governed by the transition operator $\mathbf{T}_{\text{GD}} = \mathbf{I}_{mn} - \eta\mathbf{H}_0$.
- If we apply this operator to a vector in the null space $\mathcal{H}^\perp$, the eigenvalue of $\mathbf{T}_{\text{GD}}$ is 1.
- An eigenvalue of 1 means that any error or residual component in this subspace will **never shrink**.
- $\mathbf{H}_0$ does not have full-rank, $\mathbf{T}_{\text{GD}}$ is NOT a global contraction over the whole space $\mathbb{R}^{mn}$.

Contractivity analysis

With the dynamics decomposition above, we can analyze their convergence separately. The actions of the linear components of the dynamics on the effective subspace $\mathcal{H}$ can be bounded with

the smoothness and strong convexity parameters. With proper tuning, they will become strict contractions.

Lemma 2.3.11: GD contractivity (Lemma 2 in Xu et al. (2024))

Let $L = \sigma_1^2(\mathbf{X}_0)$, $\mu = \sigma_r^2(\mathbf{X}_0)$ and $\mathcal{H}$ as defined before. Let $\eta \in (0, 2/L)$, then for any $\mathbf{v} \in \mathcal{H}$:

$$\|\mathbf{T}_{\text{GD}}\mathbf{v}\| \leq \max\{|1-\eta L|, |1-\eta\mu|\}\|\mathbf{v}\|. \tag{2.88}$$

In particular, if $\eta = \dfrac{2}{L+\mu}$, then $\|\mathbf{T}_{\text{GD}}\mathbf{v}\| \leq \dfrac{L-\mu}{L+\mu}\|\mathbf{v}\|$

Proposition 2.3.12: NAG Contractivity (Proposition 4 in Xu et al. (2024))

Let $\eta = \dfrac{1}{L}$ and $\beta = \dfrac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$, then for all $(\mathbf{u}, \mathbf{v}) \in \mathcal{H} \times \mathcal{H}$, we have

$$\left\| \mathbf{T}_{\text{NAG}} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} \right\| \leq \left(1 - \sqrt{\frac{\mu}{L}}\right) \left\| \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} \right\|. \tag{2.89}$$

It is worth noting that the choice for β is common for Nesterov-type momentum Nesterov (2018).

Controlling the nonlinear error

Lemma 2.3.13: GD nonlinear error (Lemma 3 in Xu et al. (2024))

If there exists $\theta \in (0, 1)$, and some constants C_1 and C_2 such that for any $s \leq t$, the GD dynamics yields

$$\|\mathbf{r}_s\| \leq C_1\theta^s \qquad \|\mathbf{X}_s - \mathbf{X}_0\|_F \leq C_2 \qquad \|\mathbf{Y}_s - \mathbf{Y}_0\|_F \leq C_2, \tag{2.90}$$

then we have

$$\|\text{vec}(\mathbf{P}_s\mathbf{Q}_s^\top)\| \leq C_3\theta^{2s}\|\mathbf{r}_0\|^2 \qquad \text{and} \qquad \|\eta(\mathbf{H}_0 - \mathbf{H}_s)\mathbf{r}_s\| \leq C_4\theta^s\|\mathbf{r}_0\|, \tag{2.91}$$

for some constants C_3 and C_4 depending on C_1 and C_2.
Moreover, if C_1 and C_2 satisfy

$$[\max\{\|\mathbf{X}_0\|, \|\mathbf{Y}_0\|\} + C_2]\,\eta C_1\|\mathbf{r}_0\| \leq (1-\theta)C_2, \tag{2.92}$$

then we have $\|\mathbf{X}_{t+1} - \mathbf{X}_0\|_F \leq C_2$ and $\|\mathbf{Y}_{t+1} - \mathbf{Y}_0\|_F \leq C_2$.

Lemma 2.3.14: NAG nonlinear error (Lemma 5 in Xu et al. (2024))

Suppose $0 < \beta \leq \theta^2 < \theta < 1$. If there exist some constants C_1 and C_2 such that for any $s \leq t$, the NAG dynamics yields

$$\left\| \begin{bmatrix} \mathbf{r}_s \\ \mathbf{r}_{s-1} \end{bmatrix} \right\| \leq C_1\theta^s \left\| \begin{bmatrix} \mathbf{r}_0 \\ \mathbf{r}_{-1} \end{bmatrix} \right\| \quad \|\mathbf{X}_s - \mathbf{X}_0\|_F \leq C_2 \quad \text{and} \quad \|\mathbf{Y}_s - \mathbf{Y}_0\|_F \leq C_2, \tag{2.93}$$

then we have

$$\|\boldsymbol{\zeta}_t\| \leq C_3\theta^{2t}\left\|\begin{bmatrix}\mathbf{r}_0\\ \mathbf{r}_{-1}\end{bmatrix}\right\|^2 \quad \text{and} \quad \|\boldsymbol{\iota}_t\| \leq C_4\theta^t\left\|\begin{bmatrix}\mathbf{r}_0\\ \mathbf{r}_{-1}\end{bmatrix}\right\|, \tag{2.94}$$

for some constants C_3 and C_4 depending on C_1 and C_2. Moreover, if C_1 and C_2 satisfy

$$(\max\{\|\mathbf{X}_0\|, \|\mathbf{Y}_0\|\} + C_2)\eta C_1\left\|\begin{bmatrix}\mathbf{r}_0\\ \mathbf{r}_{-1}\end{bmatrix}\right\| \leq (1-\theta)^2 C_2, \tag{2.95}$$

then we have $\|\mathbf{X}_{t+1} - \mathbf{X}_0\|_F \leq C_2$ and $\|\mathbf{Y}_{t+1} - \mathbf{Y}_0\| \leq C_2$.

With the linear components firmly bounded and the non-linear errors adequately controlled, we are prepared to establish the complete convergence rates for both algorithms.

Final Iteration Complexity Analysis

Putting them together gives [6]:

Theorem 2.3.15: GD convergence rate (Theorem 1 in Xu et al. (2024))

For $0 < \tau < c_1$, denote $\delta = 3e^{\left\{-(d+r+1)\min\left\{\log\frac{1}{c_1\tau}, c_2, \frac{1}{2}\right\}\right\}}$, where c_1 and c_2 are universal constants.
Denote $L = \sigma_1^2(\mathbf{X}_0)$, $\mu = \sigma_r^2(\mathbf{X}_0)$. Let $\eta = \frac{2}{L+\mu}$, and

$$c \geq \underline{c} \triangleq \frac{12\sqrt{d}}{\tau(\sqrt{d}-\sqrt{r-1})\sigma_r(\mathbf{A})}\sqrt{\frac{\mathrm{cond}^4(\mathbf{X}_0)\|\mathbf{A}\|_F}{\mathrm{cond}^2(\mathbf{X}_0)-1}}. \tag{2.96}$$

Then with c plugged in the initialization, i.e. $\mathbf{X}_0 = c\mathbf{A}\boldsymbol{\Phi}$ and $\mathbf{Y}_0 = \mathbf{0}$, GD returns $\mathbf{X}_t$ and $\mathbf{Y}_t$ w.p. at least $1-\delta$ such that

$$\|\mathbf{R}_t\|_F \leq \frac{3c^2\sigma_1^2(\mathbf{A})}{64}\left(1-\frac{\mu}{L}\right)^t. \tag{2.97}$$

In particular, if $c = \underline{c}$, then GD finds $\|\mathbf{R}_t\|_F \leq \varepsilon\|\mathbf{A}\|_F$ in

$$T = \mathcal{O}\left(\frac{d^2\kappa^2}{\tau^2(d-r+1)^2}\log\frac{C}{\varepsilon}\right), \tag{2.98}$$

iterations, where $C \triangleq \frac{27\tau^2(d-\tau+1)^2}{16d^2}\frac{\mathrm{cond}^4(\mathbf{X}_0)\kappa^2}{\mathrm{cond}^2(\mathbf{X}_0)-1}$.

[6] In the original paper, there are two harmless typos. Regarding $\underline{c}$, they've written

$$c \geq \underline{c} \triangleq \frac{\sqrt{d}\sigma_r(A)}{12\tau(\sqrt{d}-\sqrt{r-1})}\sqrt{\frac{\mathrm{cond}^4(X_0)\|A\|_F}{\mathrm{cond}^2(X_0)-1}}$$

This is not correct due to the fact that we need equation (16) in their paper, that is, $\mu \geq \frac{144\mathrm{cond}^4(X_0)\|A\|_F}{\mathrm{cond}^2(X_0)-1}$. The correct version should be the one that is stated above.

Theorem 2.3.16: NAG convergence rate (Theorem 2 in Xu et al. (2024))

For $0 < \tau < c_1$, define δ as in Theorem 1, i.e. $\delta = 3e^{-(d-r+1)\min\left\{\log\frac{1}{c_1\tau}, c_2 d, \frac{d}{2}\right\}}$, where c_1 and c_2 are universal constants.
Define $L = \sigma_1^2(\mathbf{X}_0), \mu = \sigma_r^2(\mathbf{X}_0)$. Further fix the following constants:

$$\eta = \frac{1}{L} \quad \text{and} \quad \beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}} \quad \text{and} \quad c \geq \underline{c} \triangleq 29\sqrt{\frac{d(2\sqrt{d}+\sqrt{r})\|\mathbf{A}\|_F\kappa}{\tau^3(\sqrt{d}-\sqrt{r-1})^3\sigma_r^2(\mathbf{A})}}. \tag{2.99}$$

Then with c plugged in initialization, that is, $\mathbf{X}_0 = c\mathbf{A}\mathbf{\Phi}$ and $\mathbf{Y}_0 = \mathbf{0}$, w.p. at least $1-\delta$ such that

$$\|\mathbf{R}_t\|_F \leq \frac{c^2\sigma_1^2(\mathbf{A})}{64\text{cond}(\mathbf{X}_0)}\left(1-\frac{\sqrt{\mu}}{2\sqrt{L}}\right)^t. \tag{2.100}$$

In particular, if $c = \underline{c}$, then NAG finds $\|\mathbf{R}_T\|_F \leq \epsilon\|\mathbf{A}\|_F$ in

$$T = O\left(\frac{d\kappa}{\tau(d-r+1)}\log\frac{C}{\epsilon}\right), \tag{2.101}$$

iterations, where $C = \frac{841d(2\sqrt{d}+\sqrt{r})}{64\tau^3(\sqrt{d}-\sqrt{r-1})^3}\frac{\kappa^3}{\text{cond}(\mathbf{X}_0)}$.

Recall that we have had the constants C_3 and C_4 to bound the geometric decay rate for the dynamic shift term $\boldsymbol{\iota}_t$ and the higher-order term $\boldsymbol{\zeta}_t$ respectively. To eventually roll/absorb them under C_1, the authors defined an auxiliary term c_0 that wraps the non-linear error, which is

$$c_0 \propto C_1(C_3\|\mathbf{r}_0\| + C_4)\|\mathbf{r}_0\|. \tag{2.102}$$

To absorb the error terms and prove that the residual stays inside the $C_1\theta^t$ envelope, they are required to satisfy a second, entirely different condition:

$$a_0 + \frac{c_0}{\theta-\rho} \leq C_1^2\|\mathbf{A}\|_F \qquad \text{eq. (15) for GD in Xu et al. (2024)} \tag{2.103}$$

$$a_0 + \frac{c_0}{\theta-\rho} \leq \sqrt{2}C_1^2\|\mathbf{A}\|_F \qquad \text{eq. (24) for NAG in Xu et al. (2024),} \tag{2.104}$$

where C_3 and C_4 are absorbed by the term $\frac{c_0}{\theta-\rho}$.

Beyond conventional unconstrained MF methods, other base algorithms also benefit from advanced sketching ideas.

2.3.3 Detour: Nyström initialization works well on a different base method

ScaledGD, originally proposed in Mishra et al. (2014); Mishra and Sepulchre (2016) from the lens of Riemannian optimization, and formalized in Tong et al. (2021), was designed to improve the convergence rate of first order methods on ill-conditioned matrix optimization problems such

as Matrix sensing, Robust PCA, and Matrix Completion. Formally, it is:

$$\begin{aligned}\mathbf{X}_{t+1} &= \mathbf{X}_t - \eta \nabla_{\mathbf{X}} f(\mathbf{X}_t, \mathbf{Y}_t)(\mathbf{Y}_t^\top \mathbf{Y}_t)^{-1} \\ &= \mathbf{X}_t - \eta(\mathbf{X}_t \mathbf{Y}_t^\top - \mathbf{A})\mathbf{Y}_t(\mathbf{Y}_t^\top \mathbf{Y}_t)^{-1}, \qquad (2.105) \\ \mathbf{Y}_{t+1} &= \mathbf{Y}_t - \eta \nabla_{\mathbf{Y}} f(\mathbf{X}_t, \mathbf{Y}_t)(\mathbf{X}_t^\top \mathbf{X}_t)^{-1} \\ &= \mathbf{Y}_t - \eta(\mathbf{X}_t \mathbf{Y}_t^\top - \mathbf{A})^\top \mathbf{X}_t(\mathbf{X}_t^\top \mathbf{X}_t)^{-1}. \qquad (2.106)\end{aligned}$$

Motivation for ScaledGD via quotient-manifold optimization

To motivate the derivation of ScaledGD slightly: recall that for the unconstrained two-block MF $\mathbf{A} = \mathbf{X}\mathbf{Y}^\top$ where $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $\mathbf{Y} \in \mathbb{R}^{m\times d}$, then for any $\mathbf{M} \in \mathrm{GL}(d)$, $\mathbf{X}' = \mathbf{X}\mathbf{M}$ and $\mathbf{Y}' = \mathbf{Y}(\mathbf{M}^{-1})^\top$ is also a valid pair of solutions. Standard GD in Euclidean spaces often wastes computational effort by bouncing around different internal representations such as $(\mathbf{X}, \mathbf{Y})$ and $(\mathbf{X}', \mathbf{Y}')$ that all map to the same matrix $\mathbf{A}$. It also suffers from scaling imbalances: if $\mathbf{Y}$ is large and $\mathbf{X}$ is small, standard GD will struggle to make an effective update using a single learning rate. Instead of performing Euclidean GD, we may first mod out these general linear groups and create a smooth quotient manifold $\mathcal{M} = (\mathbb{R}^{n\times d} \times \mathbb{R}^{m\times d})/\mathrm{GL}(d)$. By directly optimizing on this manifold, each step is making more fundamental changes, bypassing the rotational and scaling ambiguity. If one were to derive the Riemannian metric on this manifold, the resulting gradient scheme is precisely ScaledGD as given in (2.105)–(2.106).

More recently in Li et al. (2025), the authors also showed that applying such an initialization scheme improves the known iteration complexity in Tong et al. (2021). Equally important is the establishment of the column span argument (implying the full rankness of $\mathbf{X}_t$ over the entire trajectory) and a refined analysis on the control of $\sigma_{\min}(\mathbf{X}_t^\top \mathbf{X}_t)$.

Chapter 3

Do their frameworks fit for NMF?

3.1 The attempt for the PŁ framework

3.1.1 The need for non-negativity at initialization

Recall that the initialization in the unconstrained setting is

$$\mathbf{X}_0 \sim \frac{1}{\sqrt{\eta}}\mathbf{A}\mathbf{\Phi}_1 \qquad \text{and} \qquad \mathbf{Y}_0 \sim \sqrt{\eta}\mathbf{\Phi}_2. \tag{3.1}$$

Note that $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$ could draw negative values and we need to preserve non-negativity at initialization for this to work with NMF. One obvious fix is to take the absolute value of the entries, so that the distribution of $\mathbf{\Phi}_1$ and $\mathbf{\Phi}_2$ becomes a half-normal (or folded-normal) distribution.

However, standard non-asymptotic concentration results in RMT for sub-Gaussian matrices strictly require the entries to be isotropic, which fundamentally relies on having a centered (zero) mean [1]. Because the half-normal distribution is strictly non-negative, it inherently possesses a strictly positive, non-centered mean. Consequently, the off-diagonal entries of its second moment matrix do not vanish, breaking the isotropic assumption.

Applying existing RMT concentration bounds to sub-Gaussian variables with a non-centered mean requires a careful reworking of the inequalities, if it is not a complete dead end for this specific framework. This probabilistic roadblock illustrates why one cannot simply rely on purely random non-negative distributions.

Instead, a plausible fix is to draw a sketching matrix using standard, mean-zero Gaussians (perfectly preserving the isotropic sub-Gaussian properties) and then apply a controlled, deterministic, strictly positive rank-1 shift (e.g., $\alpha_0 \mathbf{1}_n \mathbf{1}_d^\top$) to safely push the entire initialization into the non-negative orthant. This approach directly motivates the contraction-based framework explored later in this chapter.

[1] see Vershynin (2012) for details

3.1.2 The need (and failure) for a meaningful proximal PŁ inequality

In this section, we provide a counter-example where the proximal PŁ with such initialization cannot hold. Let $\mathbf{A}$ be a full-rank 2×2 matrix, and consider

$$\mathbf{X}_0 = \frac{1}{\eta^{1/2} C \sigma_1(\mathbf{A})} \mathbf{A} \mathbf{\Phi}_1, \tag{3.2}$$

where $\mathbf{\Phi}_1 \in \mathbb{R}^{2\times 2}$ is drawn from a half-normal distribution $\mathcal{N}_+$. The fundamental property of the half-normal distribution is that its support is strictly positive. Therefore, every entry of $\mathbf{\Phi}_1$ is strictly positive w.p. 1. Let us define the scalar constant $c_A = \frac{1}{\eta^{1/2} C \sigma_1(\mathbf{A})}$.

For the proximal-PŁ inequality to hold over the landscape at $\mathbf{X}_0$, there must exist a universal constant $C_{\text{PL}} > 0$ such that for all $\mathbf{Y} \geq 0$:

$$\|G_\eta^{\mathbf{Y}}(\mathbf{X}_0, \mathbf{Y})\|_F^2 \geq C_{\text{PL}} [\sigma_{\min}(\mathbf{X}_0)]^2 f(\mathbf{X}_0, \mathbf{Y}). \tag{3.3}$$

where the proximal partial gradient $G_\eta^{\mathbf{Y}}$ w.r.t. $\mathbf{Y}$ block is defined as:

$$G_\eta^{\mathbf{Y}}(\mathbf{X}, \mathbf{Y}) = \frac{1}{\eta}(\mathbf{Y} - [\mathbf{Y} - \eta \nabla_{\mathbf{Y}} f(\mathbf{X}, \mathbf{Y})]_+).^{2} \tag{3.4}$$

If this inequality is a universal geometric property of the landscape, it must hold at the constrained global minimum of the function. Let $\mathbf{Y}^*$ be the exact matrix that minimizes $f(\mathbf{X}_0, \mathbf{Y})$ subject to $\mathbf{Y} \geq 0$.

The gradient mapping at the constrained minimum

Because $\mathbf{Y}^*$ minimizes the convex function $f(\mathbf{X}_0, \mathbf{Y})$ over the non-negative orthant, it must satisfy the Karush-Kuhn-Tucker (**KKT**) stationarity conditions. By the algebraic definition of the projection operator, satisfying the KKT conditions is exactly equivalent to the projected gradient mapping being zero:

$$\|G_\eta^{\mathbf{Y}}(\mathbf{X}_0, \mathbf{Y}^*)\|_F^2 = 0. \tag{3.5}$$

To see this, recall that we are minimizing the convex function $f(\mathbf{X}_0, \mathbf{Y})$ w.r.t. $\mathbf{Y}$ under the strict constraint $\mathbf{Y} \geq 0$. Let $\mathbf{Y}^*$ be the exact constrained global minimum. KKT states that for every entry (i, j), exactly three conditions must hold simultaneously at the optimum $\mathbf{Y}^*$:

(1) $\mathbf{Y}^*_{i,j} \geq 0$ (Primal feasibility);

(2) $\nabla_{\mathbf{Y}} f(\mathbf{X}_0, \mathbf{Y}^*)_{i,j} \geq 0$ (Dual feasibility);

(3) $\mathbf{Y}^*_{i,j} \nabla_{\mathbf{Y}} f(\mathbf{X}_0, \mathbf{Y}^*)_{i,j} = 0$ (Complementary slackness).

The projected gradient mapping is defined algebraically for each entry as:

$$G_\eta^{\mathbf{Y}}(\mathbf{X}_0, \mathbf{Y}^*)_{i,j} = \frac{1}{\eta}(\mathbf{Y}^*_{i,j} - [\mathbf{Y}^*_{i,j} - \eta \nabla_{\mathbf{Y}} f(\mathbf{X}_0, \mathbf{Y}^*)_{i,j}]_+).$$

We evaluate this mapping under the strict KKT constraints. By Condition (3), there are exactly two possible states for any entry (i, j):

[2] In light of equation (1.61), the prox-gradient step is $T^{\mathbf{Y}}_{1/\eta} = [\mathbf{Y} - \eta \nabla_{\mathbf{Y}} f(\mathbf{X}, \mathbf{Y})]_+$ with g_i being the indicator of the non-negative orthant, then the proximal partial gradient $G_\eta^{\mathbf{Y}}$ is simply $G_\eta^{\mathbf{Y}} = 1/\eta(\mathbf{Y} - T^{\mathbf{Y}}_{1/\eta}(\mathbf{X}, \mathbf{Y}))$

- **Case A:** If $\mathbf{Y}^*_{i,j} > 0$, Condition 3 strictly requires that the gradient is zero: $\nabla_{\mathbf{Y}} f(\mathbf{X}_0, \mathbf{Y}^*)_{i,j} = 0$. Therefore: $G^{\mathbf{Y}}_\eta(\mathbf{X}_0, \mathbf{Y}^*)_{i,j} = \frac{1}{\eta}(\mathbf{Y}^*_{i,j} - \mathbf{Y}^*_{i,j}) = 0$.

- **Case B:** If $\mathbf{Y}^*_{i,j} = 0$, Condition 2 strictly requires that the gradient is non-negative. The term inside the projection is less than or equal to zero, so $G^{\mathbf{Y}}_\eta(\mathbf{X}_0, \mathbf{Y}^*)_{i,j} = 0$. Because every single entry evaluates exactly to 0, the Frobenius norm is exactly zero: $\|G^{\mathbf{Y}}_\eta(\mathbf{X}_0, \mathbf{Y}^*)\|^2_F = 0$. This proves the equivalence.

The Unconstrained minimum and the specific failure of $\mathbf{\Phi}_1$

To evaluate the right side of the PŁ inequality, we must determine if the residual error $f(\mathbf{X}_0, \mathbf{Y}^*)$ reaches 0. The error $f(\mathbf{X}_0, \mathbf{Y})$ equals 0 iff:

$$\mathbf{X}_0\mathbf{Y}^\top = \mathbf{A}. \tag{3.6}$$

Substitute our specific initialization $\mathbf{X}_0 = c_A\mathbf{A}\mathbf{\Phi}_1$:

$$c_A\mathbf{A}\mathbf{\Phi}_1\mathbf{Y}^\top = \mathbf{A}, \tag{3.7}$$

Because $\mathbf{A}$ is a full-rank 2×2 matrix, it is invertible. Multiply both sides by $\mathbf{A}^{-1}$, and this gets us

$$c_A\mathbf{\Phi}_1\mathbf{Y}^\top = \mathbf{I}. \tag{3.8}$$

Solving for the unconstrained zero-error state $\mathbf{Y}^\top_{\text{unc}}$ yields:

$$\mathbf{Y}^\top_{\text{unc}} = \frac{1}{c_A}\mathbf{\Phi}_1^{-1}. \tag{3.9}$$

This is where our specific initialization mathematically breaks the NMF problem.

Let the strictly positive random matrix $\mathbf{\Phi}_1$ be:

$$\mathbf{\Phi}_1 = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad \text{with} \quad a, b, c, d > 0, \tag{3.10}$$

The following probabilistic statements hold:

- Every entry is strictly positive w.p. 1, i.e. $a, b, c, d > 0$.
- The determinant $\det(\mathbf{\Phi}_1) = ad - bc$ is a continuous random variable. The event $ad - bc = 0$ defines a set with a 4-dimensional Lebesgue measure of exactly 0, and the joint probability distribution is absolutely continuous w.r.t. the Lebesgue measure. This means that $\mathbf{\Phi}_1$ is invertible w.p. 1.

The analytic inverse of $\mathbf{\Phi}_1$ is:

$$\mathbf{\Phi}_1^{-1} = \frac{1}{ad - bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}, \tag{3.11}$$

Because a, b, c, d are strictly positive, the signs of the entries inside the matrix depend exclusively on the determinant $ad - bc$:

1. If $ad - bc > 0$, the scalar multiplier is positive. The off-diagonal entries $-b$ and $-c$ remain strictly negative.

2. If $ad - bc < 0$, the scalar multiplier is negative. The diagonal entries d and a become strictly negative.

In all possible continuous outcomes, exactly two entries of $\mathbf{\Phi}_1^{-1}$ are strictly negative. Consequently, the unconstrained matrix $\mathbf{Y}_{\text{unc}}^{\top} = \frac{1}{c_A}\mathbf{\Phi}_1^{-1}$ contains strictly negative entries w.p. 1.

Because the zero-error state $\mathbf{Y}_{\text{unc}}$ requires strictly negative entries, it lies strictly outside the non-negative orthant. Therefore, the constrained non-negative minimum $\mathbf{Y}^*$ cannot equal $\mathbf{Y}_{\text{unc}}$.

Because $\mathbf{Y}^* \neq \mathbf{Y}_{\text{unc}}$, the factorization cannot be exact. The residual error is strictly positive:

$$f(\mathbf{X}_0, \mathbf{Y}^*) > 0. \tag{3.12}$$

Additionally, because $\mathbf{A}$ and $\mathbf{\Phi}_1$ are invertible matrices, their product $\mathbf{X}_0$ is invertible, meaning that

$$\sigma_{\min}(\mathbf{X}_0) > 0. \tag{3.13}$$

Substitute these exact values back into the proximal-PŁ inequality:

$$\begin{aligned} \|G_\eta^{\mathbf{Y}}(\mathbf{X}_0, \mathbf{Y}^*)\|_F^2 &\geq C_{\text{PL}}[\sigma_{\min}(\mathbf{X}_0)]^2 f(\mathbf{X}_0, \mathbf{Y}^*) \\ 0 &\geq C_{\text{PL}}(\text{strictly positive})(\text{strictly positive}). \end{aligned} \tag{3.14}$$

This strict algebraic outcome forces $C_{\text{PL}} \leq 0$. This means that a proximal-PŁ will fail in the 2 by 2 case w.p. 1.

3.2 The attempt for the contraction-based framework

3.2.1 Motivations and re-parametrization techniques

Again, we are trying to see whether we could adapt such a proof framework for NMF. If we were to directly fit

$$F(\mathbf{W}, \mathbf{H}) = \frac{1}{2}\|\mathbf{A} - \mathbf{W}\mathbf{H}\|_F^2 + \delta_+(\mathbf{W}) + \delta_+(\mathbf{H}), \tag{3.15}$$

this would be impossible due to non-smoothness coming from the projection/indicator function. We've considered a variety of smooth re-parametrization technique, among which are:

1. The Hadamard reparametrization technique for non-negativity constraints

$$F(\mathbf{W}, \mathbf{H}) = \frac{1}{2}\|\mathbf{A} - (\mathbf{W} \odot \mathbf{W})(\mathbf{H} \odot \mathbf{H})\|_F^2 + \lambda_{\mathbf{W}}\mathbf{W} \odot \mathbf{W} + \lambda_{\mathbf{H}}\mathbf{H} \odot \mathbf{H}, \tag{3.16}$$

   which resembles the Burer-Monteiro technique Burer and Monteiro (2003; 2005).The impact of such reparametrization on landscape issues often involves KKT analysis, see for instance Ding and Wright (2025) and Li et al. (2023).

2. An entry-wise log-barrier penalty as a soft-surrogate for the projection:

$$F(\mathbf{W}, \mathbf{H}) = \frac{1}{2}\|\mathbf{A} - \mathbf{W}\mathbf{H}\|_F^2 - \lambda_{\mathbf{W}} \sum_{i,j}[\log(\mathbf{W}_{ij})] - \lambda_{\mathbf{H}} \sum_{i,j}[\log(\mathbf{H}_{ij})]. \tag{3.17}$$

We choose the latter mostly for the sake of curiosity. We would like to see to what extent [3] such a barrier-penalized objective fits into the proof framework of Xu et al. (2024).

[3] The current proof is incomplete and has some unverifiable statements that the author does not have the time to completely fix it. It maybe shown in the future that some of those claims are wrong or contradictory.

3.2.2 Set-up

To this end, we need to make a few modifications on top of what is already presented in the barrierized objective. Namely, the left block needs to always stay within the ColSpan of $\mathbf{A}$. For this, we define the a span-restricted objective as follows

$$\widetilde{F}_\lambda(\mathbf{Z},\mathbf{H}) = \frac{1}{2}\|\mathbf{A}-(\mathbf{AZ})\mathbf{H}\|_F^2 - \lambda\sum_{i,j}\log\left[(\mathbf{AZ})_{ij}\right] - \lambda\sum_{i,j}\log(\mathbf{H}_{ij}), \tag{3.18}$$

with domain $\mathrm{dom}(\widetilde{F}_\lambda) = \{(\mathbf{Z},\mathbf{H}) : \mathbf{AZ} > 0, \mathbf{H} > 0\}$. Its gradient may be calculated explicitly as follows:

$$\nabla_{\mathbf{Z}}\widetilde{F}_\lambda(\mathbf{Z},\mathbf{H}) = \mathbf{A}^\top\big((\mathbf{AZ})\mathbf{H}-\mathbf{A}\big)\mathbf{H}^\top - \lambda\mathbf{A}^\top\big((\mathbf{AZ})^{\odot(-1)}\big). \tag{3.19}$$

$$\nabla_{\mathbf{H}}\widetilde{F}_\lambda(\mathbf{Z},\mathbf{H}) = (\mathbf{AZ})^\top\big((\mathbf{AZ})\mathbf{H}-\mathbf{A}\big) - \lambda\mathbf{H}^{\odot(-1)}. \tag{3.20}$$

Here again, we run single-block GD on this barrierized objective:

$$\mathbf{Z}_{t+1} = \mathbf{Z}_t - \eta\nabla_{\mathbf{Z}}\widetilde{F}_\lambda(\mathbf{Z}_t,\mathbf{H}_t) \qquad \text{and} \qquad \mathbf{H}_{t+1} = \mathbf{H}_t - \eta\nabla_{\mathbf{H}}\widetilde{F}_\lambda(\mathbf{Z}_t,\mathbf{H}_t). \tag{3.21}$$

3.2.3 Initialization and positivity with a Gaussian max bound

We take $\mathbf{\Phi} \in \mathbb{R}^{n\times d}$ with i.i.d. entries $\mathcal{N}(0,1/d)$, and set

$$\mathbf{Z}_0 = c\mathbf{\Phi} + \alpha_0\mathbf{1}_n\mathbf{1}_d^\top \qquad \mathbf{W}_0 = \mathbf{AZ}_0 = c\mathbf{A\Phi} + \alpha_0(\mathbf{A1}_n)\mathbf{1}_d^\top. \tag{3.22}$$

For any fixed row i and column j

$$(\mathbf{A\Phi})_{ij} = \langle \mathbf{A}_{i:}, \mathbf{\Phi}_{:j}\rangle \sim \mathcal{N}\Big(0, \frac{1}{d}\|\mathbf{A}_{i:}\|_2^2\Big). \tag{3.23}$$

Proposition 3.2.1: Gaussian max bound

With the above setting, we have, for ANY $\delta > 0$, w.p. at least $1-\delta$:

$$B_{\mathbf{\Phi}} \triangleq \max_{1\le i\le m,\ 1\le j\le d}\|(\mathbf{A\Phi})_{ij}\| \le \|\mathbf{A}\|_{2\to\infty}\sqrt{\frac{2\log\left(\frac{2md}{\delta}\right)}{d}}, \tag{3.24}$$

where $\|\mathbf{A}\|_{2\to\infty} \triangleq \max_i \|\mathbf{A}_{i:}\|_2$, the magnitude of the row having the max 2 norm.

Proof: We will first need the following lemmas.

Lemma 3.2.2: Gaussian inner product

Let $\mathbf{g} \in \mathbb{R}^n$ have i.i.d. entries $g_k \sim \mathcal{N}(0, 1/d)$. Then for any fixed $\mathbf{a} \in \mathbb{R}^n$,

$$\langle \mathbf{a}, \mathbf{g} \rangle \sim \mathcal{N}\left(0, \frac{\|\mathbf{a}\|^2}{d}\right). \tag{3.25}$$

Proof: Write $\langle \mathbf{a}, \mathbf{g} \rangle = \sum_{k=1}^n a_k g_k$. Since the g_k are independent mean zero Gaussians, $\langle \mathbf{a}, \mathbf{g} \rangle$ is Gaussian with mean 0, and its variance can be computed as:

$$\mathrm{Var}(\langle \mathbf{a}, \mathbf{g} \rangle) = \sum_{k=1}^n a_k^2 \,\mathrm{Var}(g_k) = \sum_{k=1}^n a_k^2 \frac{1}{d} = \frac{\|\mathbf{a}\|^2}{d}. \tag{3.26}$$

This proves the lemma. ■

Lemma 3.2.3: Gaussian tail bound (Chernoff)

If $X \sim \mathcal{N}(0, \sigma^2)$, then for any $t > 0$,

$$\mathbb{P}\left(|X| \geq t\right) \;\leq\; 2\exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{3.27}$$

Proof: For any $\lambda > 0$, by Markov and symmetry,

$$\mathbb{P}(X \geq t) = \mathbb{P}\left(e^{\lambda X} \geq e^{\lambda t}\right) \leq e^{-\lambda t}\,\mathbb{E}\left[e^{\lambda X}\right]. \tag{3.28}$$

For $X \sim \mathcal{N}(0, \sigma^2)$, the moment generating function is $\mathbb{E}[e^{\lambda X}] = \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$. Hence

$$\mathbb{P}(X \geq t) \leq \exp\left(-\lambda t + \frac{\lambda^2\sigma^2}{2}\right). \tag{3.29}$$

Optimizing the right side over $\lambda > 0$ yields the minimizer $\lambda^\star = t/\sigma^2$, giving

$$\mathbb{P}(X \geq t) \leq \exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{3.30}$$

By symmetry of X, the same bound holds for $\mathbb{P}(-X \geq t) = \mathbb{P}(X \leq -t)$. Therefore

$$\mathbb{P}\left(|X| \geq t\right) = \mathbb{P}(X \geq t) + \mathbb{P}(X \leq -t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right). \tag{3.31}$$

This proves the lemma. ■

Lemma 3.2.4: Union bound

If $\{E_\ell\}_{\ell=1}^N$ are events in a probability space, then

$$\mathbb{P}\left(\bigcup_{\ell=1}^N E_\ell\right) \;\leq\; \sum_{\ell=1}^N \mathbb{P}(E_\ell). \tag{3.32}$$

Proof: See any basic probability text. ■

Now, applying Lemma 3.2.2 with $\mathbf{a} \triangleq \mathbf{A}_{i:}$ and $\mathbf{g} \triangleq \mathbf{\Phi}_{:j}$ gives

$$(\mathbf{A}\mathbf{\Phi})_{ij} \sim \mathcal{N}\left(0,\ \sigma_{ij}^2\right) \qquad \text{and} \qquad \sigma_{ij}^2 = \frac{\|\mathbf{A}_{i:}\|^2}{d}. \tag{3.33}$$

Next we apply Lemma 3.2.3 to $X = (\mathbf{A}\mathbf{\Phi})_{ij} \sim \mathcal{N}(0, \sigma_{ij}^2)$ to get, for any $t > 0$,

$$\mathbb{P}\left(|(\mathbf{A}\mathbf{\Phi})_{ij}| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sigma_{ij}^2}\right) = 2\exp\left(-\frac{dt^2}{2\|\mathbf{A}_{i:}\|^2}\right). \tag{3.34}$$

Apply Lemma 3.2.4 with events $E_{ij} = \{|(\mathbf{A}\mathbf{\Phi})_{ij}| \geq t\}$, and use the bound from Lemma 3.2.3:

$$\mathbb{P}\left(\max_{i,j} |(\mathbf{A}\mathbf{\Phi})_{ij}| \geq t\right) \leq \sum_{i=1}^{m}\sum_{j=1}^{d} \mathbb{P}\left(|(\mathbf{A}\mathbf{\Phi})_{ij}| \geq t\right) \leq \sum_{i=1}^{m}\sum_{j=1}^{d} 2\exp\left(-\frac{dt^2}{2\|\mathbf{A}_{i:}\|^2}\right). \tag{3.35}$$

By definition, since $\|\mathbf{A}_{i:}\| \leq \|\mathbf{A}\|_{2\to\infty}$ for every i, we get the uniform bound

$$\mathbb{P}\left(\max_{i,j} |(\mathbf{A}\mathbf{\Phi})_{ij}| \geq t\right) \leq 2md\ \exp\left(-\frac{dt^2}{2\|\mathbf{A}\|_{2\to\infty}^2}\right). \tag{3.36}$$

Set the RHS to be at most δ and solve for t:

$$2md\ \exp\left(-\frac{dt^2}{2\|\mathbf{A}\|_{2\to\infty}^2}\right) \leq \delta \iff t^2 \geq \frac{2\|\mathbf{A}\|_{2\to\infty}^2}{d}\ \log\left(\frac{2md}{\delta}\right). \tag{3.37}$$

Therefore, simply choosing $t \triangleq \|\mathbf{A}\|_{2\to\infty}\sqrt{\frac{2}{d}\log\left(\frac{2md}{\delta}\right)}$ yields:

$$\mathbb{P}\left(\max_{i,j} |(\mathbf{A}\mathbf{\Phi})_{ij}| \geq \|\mathbf{A}\|_{2\to\infty}\sqrt{\frac{2\log\left(\frac{2md}{\delta}\right)}{d}}\right) \leq \delta. \tag{3.38}$$

Equivalently, w.p. at least $1-\delta$:

$$\max_{1\leq i\leq m,\ 1\leq j\leq d} |(\mathbf{A}\mathbf{\Phi})_{ij}| \leq \|\mathbf{A}\|_{2\to\infty}\sqrt{\frac{2\log\left(\frac{2md}{\delta}\right)}{d}}. \tag{3.39}$$

We are done. ■

Lemma 3.2.5: Positivity of initialization

Assume the following conditions hold:

- $\mathbf{A} \geq 0$ entrywise (for free, since we are doing NMF)
- **all row–sums are strictly positive** (so $a_{\min} \triangleq \min_i(\mathbf{A}\mathbf{1}_n)_i > 0$), which is less desirable but required.

Then, choosing

$$\alpha_0 \geq \frac{c\|\mathbf{A}\|_{2\to\infty}}{a_{\min}}\sqrt{\frac{2\log\left(\frac{2md}{\delta}\right)}{d}} + \frac{2C_2}{a_{\min}}, \tag{3.40}$$

ensures, w.p. $\geq 1-\delta$, that

$$\min_{i,j}(\mathbf{W}_0)_{ij} \geq 2C_2 \implies \mathbf{W}_0 > 2C_2 \text{ entrywise.} \tag{3.41}$$

At the same time, $\mathbf{H}_0 = \varepsilon \mathbf{1}_d \mathbf{1}_n^\top$ with $\varepsilon > 2C_2$ gives $\mathbf{H}_0 > 2C_2$ entrywise.

Proof: Again, recall that $\mathbf{W}_0 = \mathbf{A}\mathbf{Z}_0 = c\mathbf{A}\mathbf{\Phi} + \alpha_0(\mathbf{A}\mathbf{1}_n)\mathbf{1}_d^\top$; Hence, with that same probability (from the Gaussian max bound):

$$\begin{aligned} c(\mathbf{A}\mathbf{\Phi})_{k,l} &\geq -c \max_{k,l} |(\mathbf{A}\mathbf{\Phi})_{k,l}| \\ &\geq -c\|\mathbf{A}\|_{2\to\infty} \sqrt{\frac{2\log\left(\frac{2md}{\delta}\right)}{d}}. \end{aligned} \tag{3.42}$$

This bound is the same for all (k, l), so when we take $\min_{i,j}$ it stays unchanged and no j terms remain; Further

$$[\alpha_0(\mathbf{A}\mathbf{1}_n)\mathbf{1}_d^\top]_{i,j} = \alpha_0(\mathbf{A}\mathbf{1}_n)_i(\mathbf{1}_d)_j = \alpha_0(\mathbf{A}\mathbf{1}_n)_i, \tag{3.43}$$

In summary, w.p. $1 - \delta$:

$$\begin{aligned} \min_{i,j}(\mathbf{W}_0)_{ij} &= \min_{i,j}\left\{\alpha_0(\mathbf{A}\mathbf{1}_n)_i + c(\mathbf{A}\mathbf{\Phi})_{ij}\right\} \\ &\geq \alpha_0 \min_i(\mathbf{A}\mathbf{1}_n)_i - c\|\mathbf{A}\|_{2\to\infty}\sqrt{\frac{2\log\left(\frac{2md}{\delta}\right)}{d}}. \end{aligned} \tag{3.44}$$

Therefore, applying the assumption $a_{\min} \triangleq \min_i(\mathbf{A}\mathbf{1}_n)_i > 0$ and substituting the choice of α_0 from (3.40) into this lower bound guarantees that $\min_{i,j}(\mathbf{W}_0)_{ij} \geq 2C_2$. The strict entrywise positivity of $\mathbf{H}_0$ follows trivially from its explicit construction. ■

Safety buffer: factor of 2

The factor of 2 acts as a **safety buffer** against the drifts of the iterates. Specifically, the induction hypothesis requires maintaining a positive lower bound $\underline{w} \geq C_2$ for all t. Recall that in our case GD allows the iterates to move away from initialization by a distance up to C_2 for all, i.e. $\|\mathbf{W}_t - \mathbf{W}_0\|_F \leq C_2$. That is,

$$\mathbf{W}_t \geq \mathbf{W}_0 - \|\mathbf{W}_t - \mathbf{W}_0\| \geq 2C_2 - C_2 = C_2, \tag{3.45}$$

and this will ensure $\mathbf{W}_t$ stays strictly positive.

3.2.4 Dynamics decomposition

Proposition 3.2.6: Result 1 on dynamics decomposition

Let $\mathbf{R}_t = (\mathbf{A}\mathbf{Z}_t)\mathbf{H}_t - \mathbf{A}$ and $\mathbf{r}_t = \mathrm{vec}(\mathbf{R}_t)$. Then,

1. For all $t \geq 0$

$$\mathbf{r}_{t+1} = \mathbf{T}_{\text{bar}}\mathbf{r}_t + \lambda \mathbf{b}_t + \boldsymbol{\xi}_t, \tag{3.46}$$

where the linear part is

$$\mathbf{T}_{\text{bar}} \triangleq \mathbf{I} - \eta\mathbf{G}_0^{\text{mat}} \quad \text{and} \quad \mathbf{G}_0^{\text{mat}} \triangleq (\mathbf{I}_n \otimes (\mathbf{W}_0\mathbf{W}_0^\top)) + ((\mathbf{H}_0^\top\mathbf{H}_0) \otimes (\mathbf{A}\mathbf{A})^\top). \tag{3.47}$$

The barrier forcing is

$$\mathbf{b}_t = \eta \mathrm{vec}\big((\mathbf{A}\mathbf{A}^\top \mathbf{W}_t^{\odot(-1)})\mathbf{H}_t\big) + \eta \mathrm{vec}\big(\mathbf{W}_t \mathbf{H}_t^{\odot(-1)}\big), \tag{3.48}$$

with $\mathbf{W}_t = \mathbf{A}\mathbf{Z}_t$. Furthermore, $\boldsymbol{\xi}_t \triangleq \boldsymbol{\psi}_t + \boldsymbol{\iota}_t$ is the vectorized non-linear component that consists of a quadratic remainder $\boldsymbol{\psi}_t$ and a linearization error term $\boldsymbol{\iota}_t$. Concretely, $\boldsymbol{\psi}_t \triangleq \mathrm{vec}(\boldsymbol{\Psi}_t)$, and

$$\boldsymbol{\Psi}_t \triangleq \Delta\mathbf{W}\Delta\mathbf{H} = (\mathbf{W}_{t+1} - \mathbf{W}_t)(\mathbf{H}_{t+1} - \mathbf{H}_t). \tag{3.49}$$

where $\Delta\mathbf{W}$ and $\Delta\mathbf{H}$ are the one-step increments in the $\mathbf{W}$ and $\mathbf{H}$ block respectively. The vectorized linearization error is given by $\boldsymbol{\iota_t} \triangleq \mathrm{vec}(\mathbf{E}_t^{(Z)} + \mathbf{E}_t^{(H)})$, consisting of the sum of linearization errors w.r.t. their respective base points $\mathbf{W}_0$ and $\mathbf{H}_0$. Concretely, they are:

$$\begin{aligned} \mathbf{E}_t^{(Z)} &\triangleq -\eta\mathbf{A}\mathbf{A}^\top\big(\mathbf{R}_t(\mathbf{H}_t^\top - \mathbf{H}_0^\top)\big)\mathbf{H}_t - \eta\mathbf{A}\mathbf{A}^\top(\mathbf{R}_t\mathbf{H}_0^\top)(\mathbf{H}_t - \mathbf{H}_0), \\ \mathbf{E}_t^{(Z)} &\triangleq -\eta(\mathbf{W}_t - \mathbf{W}_0)\mathbf{W}_t^\top\mathbf{R}_t - \eta\mathbf{W}_0(\mathbf{W}_t^\top - \mathbf{W}_0^\top)\mathbf{R}_t. \end{aligned} \tag{3.50}$$

2. Moreover, $\mathbf{r}_t$, $\mathbf{b}_t$, $\boldsymbol{\xi}_t$ all lie in $\mathcal{S} \triangleq (\mathrm{col}(\mathbf{A}))^n$.

Proof:

Once again, throughout the proof, we denote $\mathbf{R}_t \triangleq \mathbf{W}_t\mathbf{H}_t - \mathbf{A}$ as the residual and $\mathbf{r}_t \triangleq \mathrm{vec}(\mathbf{R}_t)$ as the vectorized residual. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{Z}_t \in \mathbb{R}^{n\times d}$, $\mathbf{H}_t \in \mathbb{R}^{d\times n}$,and $\mathbf{W}_t := \mathbf{A}\mathbf{Z}_t \in \mathbb{R}^{m\times d}$.

By the single-block GD update, the one-step increments are:

$$\Delta\mathbf{Z} \triangleq \mathbf{Z}_{t+1} - \mathbf{Z}_t, \qquad \Delta\mathbf{H} \triangleq \mathbf{H}_{t+1} - \mathbf{H}_t. \tag{3.51}$$

Derivation of $\Delta\mathbf{Z}$ and $\Delta\mathbf{W}$

The $\mathbf{Z}$-update induces

$$\begin{aligned} \Delta\mathbf{Z} &= -\eta\nabla_{\mathbf{Z}}\widetilde{F}_\lambda(\mathbf{Z}_t, \mathbf{H}_t) \\ &= -\eta\Big[\mathbf{A}^\top\big((\mathbf{A}\mathbf{Z}_t)\mathbf{H}_t - \mathbf{A}\big)\mathbf{H}_t^\top - \lambda\mathbf{A}^\top\big((\mathbf{A}\mathbf{Z}_t)^{\odot(-1)}\big)\Big] \\ &= -\eta\Big[\mathbf{A}^\top\big(\mathbf{R}_t\mathbf{H}_t^\top\big) - \lambda\mathbf{A}^\top\big(\mathbf{W}_t^{\odot(-1)}\big)\Big], \end{aligned} \tag{3.52}$$

where $\mathbf{W}_t \triangleq \mathbf{A}\mathbf{Z}_t$ and $\mathbf{R}_t = \mathbf{W}_t\mathbf{H}_t - \mathbf{A}$. Now we can map $\Delta\mathbf{Z}$ to $\Delta\mathbf{W}$ through $\mathbf{W} = \mathbf{A}\mathbf{Z}$:

$$\Delta\mathbf{W} \triangleq \mathbf{W}_{t+1} - \mathbf{W}_t = \mathbf{A}\Delta\mathbf{Z} = -\eta\Big[\mathbf{A}\mathbf{A}^\top\big(\mathbf{R}_t\mathbf{H}_t^\top\big) - \lambda\mathbf{A}\mathbf{A}^\top\big(\mathbf{W}_t^{\odot(-1)}\big)\Big]. \tag{3.53}$$

Derivation of $\Delta\mathbf{H}$

The $\mathbf{H}$-update induces

$$\begin{aligned} \Delta\mathbf{H} &= -\eta\nabla_{\mathbf{H}}\widetilde{F}_\lambda(\mathbf{Z}_t, \mathbf{H}_t) \\ &= -\eta\Big[\mathbf{Z}_t^\top\mathbf{A}^\top\big((\mathbf{A}\mathbf{Z}_t)\mathbf{H}_t - \mathbf{A}\big) - \lambda\mathbf{H}_t^{\odot(-1)}\Big] \\ &= -\eta\big[\mathbf{W}_t^\top\mathbf{R}_t - \lambda\mathbf{H}_t^{\odot(-1)}\big]. \end{aligned} \tag{3.54}$$

Residual update $\mathbf{R}_{t+1}$

Start from the residual definition, we have:

$$\begin{aligned}\mathbf{R}_{t+1} &= \mathbf{W}_{t+1}\mathbf{H}_{t+1} - \mathbf{A} \\ &= (\mathbf{W}_t + \Delta\mathbf{W})(\mathbf{H}_t + \Delta\mathbf{H}) - \mathbf{A} \\ &= \underbrace{\mathbf{W}_t\mathbf{H}_t - \mathbf{A}}_{=\mathbf{R}_t} + \underbrace{\Delta\mathbf{W}\mathbf{H}_t}_{\text{I}} + \underbrace{\mathbf{W}_t\Delta\mathbf{H}}_{\text{II}} + \underbrace{\Delta\mathbf{W}\Delta\mathbf{H}}_{\text{III}}. \end{aligned} \tag{3.55}$$

We then compute the three contributions from I, II, III using the derived increments:

- Term I:

$$\begin{aligned}\Delta\mathbf{W}\mathbf{H}_t &= \Big(-\eta[\mathbf{A}\mathbf{A}^\top(\mathbf{R}_t\mathbf{H}_t^\top) - \lambda\mathbf{A}\mathbf{A}^\top(\mathbf{W}_t^{\odot(-1)})]\Big)\mathbf{H}_t \\ &= -\eta\mathbf{A}\mathbf{A}^\top(\mathbf{R}_t\mathbf{H}_t^\top)\mathbf{H}_t + \eta\lambda\mathbf{A}\mathbf{A}^\top(\mathbf{W}_t^{\odot(-1)})\mathbf{H}_t. \end{aligned} \tag{3.56}$$

- Term II:

$$\begin{aligned}\mathbf{W}_t\Delta\mathbf{H} &= \mathbf{W}_t\Big(-\eta[\mathbf{W}_t^\top\mathbf{R}_t - \lambda\mathbf{H}_t^{\odot(-1)}]\Big) \\ &= -\eta\mathbf{W}_t\mathbf{W}_t^\top\mathbf{R}_t + \eta\lambda\mathbf{W}_t\mathbf{H}_t^{\odot(-1)}. \end{aligned} \tag{3.57}$$

- Term III: $\Delta\mathbf{W}\Delta\mathbf{H}$ is quadratic in $\mathbf{R}_t$ and will be kept as a remainder.

Putting pieces together,

$$\begin{aligned}\mathbf{R}_{t+1} = \mathbf{R}_t &- \eta\mathbf{A}\mathbf{A}^\top(\mathbf{R}_t\mathbf{H}_t^\top)\mathbf{H}_t - \eta\mathbf{W}_t\mathbf{W}_t^\top\mathbf{R}_t \\ &+ \lambda\Big[\eta\mathbf{A}\mathbf{A}^\top(\mathbf{W}_t^{\odot(-1)})\mathbf{H}_t + \eta\mathbf{W}_t\mathbf{H}_t^{\odot(-1)}\Big] + \underbrace{\Delta\mathbf{W}\Delta\mathbf{H}}_{\triangleq\,\mathbf{\Psi}_t}. \end{aligned} \tag{3.58}$$

Linearization at the base point via the Lyapunov operator

Introduce the base point $(\mathbf{W}_0, \mathbf{H}_0)$ and define the linear operator:

$$\boldsymbol{\mathcal{G}}_0[\Delta] \triangleq \mathbf{W}_0\mathbf{W}_0^\top\Delta + (\mathbf{A}\mathbf{A}^\top)\Delta\mathbf{H}_0^\top\mathbf{H}_0. \tag{3.59}$$

Add and subtract the base linear terms to separate a smooth linear part from a remainder. Write:

$$\begin{aligned}-\eta\mathbf{A}\mathbf{A}^\top(\mathbf{R}_t\mathbf{H}_t^\top)\mathbf{H}_t &= -\eta(\mathbf{A}\mathbf{A}^\top)\mathbf{R}_t\mathbf{H}_0^\top\mathbf{H}_0 \\ &\quad + \underbrace{-\eta\mathbf{A}\mathbf{A}^\top\big(\mathbf{R}_t(\mathbf{H}_t^\top - \mathbf{H}_0^\top)\big)\mathbf{H}_t - \eta\mathbf{A}\mathbf{A}^\top(\mathbf{R}_t\mathbf{H}_0^\top)(\mathbf{H}_t - \mathbf{H}_0)}_{=:\,\mathbf{E}_t^{(Z)}} \\ -\eta\mathbf{W}_t\mathbf{W}_t^\top\mathbf{R}_t &= -\eta\mathbf{W}_0\mathbf{W}_0^\top\mathbf{R}_t \\ &\quad + \underbrace{-\eta(\mathbf{W}_t - \mathbf{W}_0)\mathbf{W}_t^\top\mathbf{R}_t - \eta\mathbf{W}_0(\mathbf{W}_t^\top - \mathbf{W}_0^\top)\mathbf{R}_t}_{=:\,\mathbf{E}_t^{(H)}}. \end{aligned} \tag{3.60}$$

Hence the residual update becomes:

$$\mathbf{R}_{t+1} = \mathbf{R}_t - \eta\boldsymbol{\mathcal{G}}_0[\mathbf{R}_t] + \lambda\mathbf{Bar}_t + \mathbf{\Xi}_t, \tag{3.61}$$

where

$$\mathbf{Bar}_t \triangleq \eta\mathbf{A}\mathbf{A}^\top(\mathbf{W}_t^{\odot(-1)})\mathbf{H}_t + \eta\mathbf{W}_t\mathbf{H}_t^{\odot(-1)}, \tag{3.62}$$

and the full remainder is

$$\mathbf{\Xi}_t \triangleq \mathbf{\Psi}_t + \mathbf{E}_t^{(Z)} + \mathbf{E}_t^{(H)}. \tag{3.63}$$

Vectorization and the block operator $\mathbf{T}_{\text{bar}}$

Use $\text{vec}(\mathbf{U}\Delta\mathbf{V}) = (\mathbf{V}^\top \otimes \mathbf{U})\text{vec}(\Delta)$ to get the matrix representation of $\boldsymbol{\mathcal{G}}_0$ on vectors:

$$\text{vec}(\boldsymbol{\mathcal{G}}_0[\Delta]) = \Big[\mathbf{I}_n \otimes (\mathbf{W}_0\mathbf{W}_0^\top) + (\mathbf{H}_0^\top\mathbf{H}_0) \otimes (\mathbf{A}\mathbf{A}^\top)\Big]\text{vec}(\Delta) \triangleq \mathbf{G}_0^{\text{mat}}\,\text{vec}(\Delta). \tag{3.64}$$

Define

$$\mathbf{b}_t \triangleq \eta\text{vec}\big(\mathbf{A}\mathbf{A}^\top(\mathbf{W}_t^{\odot(-1)})\mathbf{H}_t\big) + \eta\text{vec}\big(\mathbf{W}_t\mathbf{H}_t^{\odot(-1)}\big), \tag{3.65}$$

and

$$\boldsymbol{\xi}_t \triangleq \text{vec}(\boldsymbol{\Xi}_t). \tag{3.66}$$

Then the vectorized residual recursion is

$$\mathbf{r}_{t+1} = (\mathbf{I} - \eta\mathbf{G}_0^{\text{mat}})\mathbf{r}_t + \lambda\mathbf{b}_t + \boldsymbol{\xi}_t, \tag{3.67}$$

and so

$$\mathbf{r}_{t+1} = \mathbf{T}_{\text{bar}}\mathbf{r}_t + \lambda\mathbf{b}_t + \boldsymbol{\xi}_t. \tag{3.68}$$

Column-span invariance

Each column of $\mathbf{R}_t$ lies in $\text{col}(\mathbf{A})$:

- $\mathbf{R}_t = \mathbf{W}_t\mathbf{H}_t - \mathbf{A} = \mathbf{A}\mathbf{Z}_t\mathbf{H}_t - \mathbf{A}$ is a linear combination of columns of $\mathbf{A}$ minus $\mathbf{A}$, hence in $\text{col}(\mathbf{A})$.
- The updates apply $\mathbf{A}\mathbf{A}^\top$ to $\mathbf{R}_t$ on the left and multiply on the right by $\mathbf{H}$ or its Gram factors; $\mathbf{A}\mathbf{A}^\top$ maps $\text{col}(\mathbf{A})$ to itself. The barrier terms $\mathbf{A}\mathbf{A}^\top(\mathbf{W}_t^{\odot(-1)})$ and $\mathbf{W}_t$ are both in $\text{col}(\mathbf{A})$. Therefore every term in the recursion for $\mathbf{R}_{t+1}$ is a linear combination of columns of $\mathbf{A}$.
- Thus the invariant subspace

  $$\mathcal{H} := \{\Delta \in \mathbb{R}^{m\times n} : \text{each column of } \Delta \text{ lies in } \text{col}(\mathbf{A})\}, \tag{3.69}$$

  contains $\mathbf{R}_t$ for all t, and the stacked state evolves in $\mathcal{H} \times \mathcal{H}$

Each column of $\mathbf{E}_t^{(Z)}$ lives in $\text{col}(\mathbf{A})$:

- By definition, the linearization error from the $\mathbf{Z}$-update is:

  $$\mathbf{E}_t^{(Z)} = -\eta\mathbf{A}\mathbf{A}^\top\big(\mathbf{R}_t(\mathbf{H}_t^\top - \mathbf{H}_0^\top)\big)\mathbf{H}_t - \eta\mathbf{A}\mathbf{A}^\top(\mathbf{R}_t\mathbf{H}_0^\top)(\mathbf{H}_t - \mathbf{H}_0). \tag{3.70}$$

- As both terms are explicitly pre-multiplied by $\mathbf{A}\mathbf{A}^\top$, we can factor out the target matrix $\mathbf{A}$ on the left:

  $$\mathbf{E}_t^{(Z)} = \mathbf{A}\Big[-\eta\mathbf{A}^\top\big(\mathbf{R}_t(\mathbf{H}_t^\top - \mathbf{H}_0^\top)\big)\mathbf{H}_t - \eta\mathbf{A}^\top(\mathbf{R}_t\mathbf{H}_0^\top)(\mathbf{H}_t - \mathbf{H}_0)\Big]. \tag{3.71}$$

  Note that $\mathbf{A}$ factors out perfectly on the left.

Each column of $\mathbf{E}_t^{(H)}$ lives in $\text{col}(\mathbf{A})$:

- The linearization error from the $\mathbf{H}$-update is given by:

$$\mathbf{E}_t^{(H)} = -\eta(\mathbf{W}_t - \mathbf{W}_0)\mathbf{W}_t^\top \mathbf{R}_t - \eta \mathbf{W}_0(\mathbf{W}_t^\top - \mathbf{W}_0^\top)\mathbf{R}_t. \tag{3.72}$$

- By substituting the structural parameterization of the state variables, $\mathbf{W}_t = \mathbf{A}\mathbf{Z}_t$ and $\mathbf{W}_0 = \mathbf{A}\mathbf{Z}_0$, we obtain:

$$\begin{aligned}\mathbf{E}_t^{(H)} &= -\eta(\mathbf{A}\mathbf{Z}_t - \mathbf{A}\mathbf{Z}_0)\mathbf{W}_t^\top \mathbf{R}_t - \eta(\mathbf{A}\mathbf{Z}_0)(\mathbf{W}_t^\top - \mathbf{W}_0^\top)\mathbf{R}_t \\ &= \mathbf{A}\Big[-\eta(\mathbf{Z}_t - \mathbf{Z}_0)\mathbf{W}_t^\top \mathbf{R}_t - \eta \mathbf{Z}_0(\mathbf{W}_t^\top - \mathbf{W}_0^\top)\mathbf{R}_t\Big].\end{aligned} \tag{3.73}$$

  Note that $\mathbf{A}$ factors out perfectly on the left.

Each column of $\mathbf{\Psi}_t$ lives in $\mathrm{col}(\mathbf{A})$:

- The linearization error from the $\mathbf{H}$-update is given by:

$$\mathbf{E}_t^{(H)} = -\eta(\mathbf{W}_t - \mathbf{W}_0)\mathbf{W}_t^\top \mathbf{R}_t - \eta \mathbf{W}_0(\mathbf{W}_t^\top - \mathbf{W}_0^\top)\mathbf{R}_t. \tag{3.74}$$

- By substituting the structural parameterization of the state variables, $\mathbf{W}_t = \mathbf{A}\mathbf{Z}_t$ and $\mathbf{W}_0 = \mathbf{A}\mathbf{Z}_0$, we obtain:

$$\begin{aligned}\mathbf{E}_t^{(H)} &= -\eta(\mathbf{A}\mathbf{Z}_t - \mathbf{A}\mathbf{Z}_0)\mathbf{W}_t^\top \mathbf{R}_t - \eta(\mathbf{A}\mathbf{Z}_0)(\mathbf{W}_t^\top - \mathbf{W}_0^\top)\mathbf{R}_t \\ &= \mathbf{A}\Big[-\eta(\mathbf{Z}_t - \mathbf{Z}_0)\mathbf{W}_t^\top \mathbf{R}_t - \eta \mathbf{Z}_0(\mathbf{W}_t^\top - \mathbf{W}_0^\top)\mathbf{R}_t\Big].\end{aligned} \tag{3.75}$$

  Note that $\mathbf{A}$ factors out perfectly on the left.

Each column of $\mathbf{Bar}_t$ lives in $\mathrm{col}(\mathbf{A})$:

- The exact barrier term introduced by the update dynamics is defined as:

$$\mathbf{Bar}_t = \eta \mathbf{A}\mathbf{A}^\top(\mathbf{W}_t^{\odot(-1)})\mathbf{H}_t + \eta \mathbf{W}_t \mathbf{H}_t^{\odot(-1)}. \tag{3.76}$$

- To prove that each column of $\mathbf{Bar}_t$ lives in $\mathrm{col}(\mathbf{A})$, we examine both terms independently. For the first term, the matrix $\mathbf{A}$ can be trivially factored out from the left due to the explicit projection matrix $\mathbf{A}\mathbf{A}^\top$:

$$\eta \mathbf{A}\mathbf{A}^\top(\mathbf{W}_t^{\odot(-1)})\mathbf{H}_t = \mathbf{A}\Big[\eta \mathbf{A}^\top(\mathbf{W}_t^{\odot(-1)})\mathbf{H}_t\Big]. \tag{3.77}$$

- For the second term, we substitute the algorithmic state parameterization $\mathbf{W}_t = \mathbf{A}\mathbf{Z}_t$:

$$\eta \mathbf{W}_t \mathbf{H}_t^{\odot(-1)} = \eta(\mathbf{A}\mathbf{Z}_t)\mathbf{H}_t^{\odot(-1)} = \mathbf{A}\Big[\eta \mathbf{Z}_t \mathbf{H}_t^{\odot(-1)}\Big]. \tag{3.78}$$

- Because both terms can be written strictly as the matrix $\mathbf{A}$ multiplying another matrix, their sum can also be factored identically:

$$\mathbf{Bar}_t = \mathbf{A}\Big[\eta \mathbf{A}^\top(\mathbf{W}_t^{\odot(-1)})\mathbf{H}_t + \eta \mathbf{Z}_t \mathbf{H}_t^{\odot(-1)}\Big]. \tag{3.79}$$

We are done. ■

3.2.5 Contractivity via spectral analysis of $\mathbf{T}_{\text{bar}}$

Recall that

$$\mathcal{G}_0[\Delta] \triangleq \mathbf{W}_0\mathbf{W}_0^\top\Delta + (\mathbf{A}\mathbf{A}^\top)\Delta\mathbf{H}_0^\top\mathbf{H}_0 \quad \text{and} \quad \mathbf{G}_0^{\text{mat}} \triangleq \mathbf{I}_n \otimes (\mathbf{W}_0\mathbf{W}_0^\top) + (\mathbf{H}_0^\top\mathbf{H}_0) \otimes (\mathbf{A}\mathbf{A}^\top), \tag{3.80}$$

Two structural facts we will use and justify in-line are:

- $\mathbf{G}_0^{\text{mat}}$ is p.s.d., as:
  - $\mathbf{W}_0\mathbf{W}_0^\top$, $\mathbf{H}_0^\top\mathbf{H}_0$, and $\mathbf{A}\mathbf{A}^\top$ are all p.s.d.,
  - kronecker of symmetric matrices is symmetric.
  - the sum of symmetric matrices is symmetric, and hence $\mathbf{G}_0^{\text{mat}}$ is orthogonally diagonalizable.
- The space splits orthogonally as $\mathbb{R}^{mn} = \mathcal{K} \oplus \mathcal{H}$, where

$$\mathcal{K} \triangleq \ker(\mathbf{G}_0^{\text{mat}}) \quad \text{and} \quad \mathcal{H} \triangleq \mathcal{K}^\perp = \text{range}(\mathbf{G}_0^{\text{mat}}). \tag{3.81}$$

One can check that $\mathcal{H} = \{\text{vec}(\Delta) : \Delta_{:j} \in \text{col}(\mathbf{A}) \text{ for all } j\}$ has dimension rn, and $\mathcal{K}$ is its orthogonal complement of dimension $(m-r)n$. In addition, on the subspace $\mathcal{H}$, all eigenvalues s of $\mathbf{G}_0^{\text{mat}}$ satisfy $s \geq s_{\min} > 0$; on $\mathcal{K}$ they are $s = 0$.

Dimension of $\mathcal{H}$

Note here that $\dim(\mathcal{H}) = nr$. The subspace $\mathcal{H}$ is the space of vectorized matrices in $\mathbb{R}^{m\times n}$ whose columns lie in $\text{col}(\mathbf{A})$. Since $\text{col}(\mathbf{A})$ has dimension r and there are n columns, the total dimension of the subspace is $r \times n$.

Eigenvalues of $\mathbf{G}_0^{\text{mat}}$ on $\mathcal{H}$ and $\mathcal{K}$

Note here that NOT all eigenvalues of $\mathbf{G}_0^{\text{mat}}$ are positive. They are positive only on the active subspace $\mathcal{H}$. $\mathbf{G}_0^{\text{mat}}$ has a non-trivial null-space/kernel $\mathcal{K}$ because $\mathbf{W}_0\mathbf{W}_0^\top$ is rank- r (where $r < d$) and $\mathbf{H}_0^\top\mathbf{H}_0$ is rank-1 (where $1 < n$). The eigenvalues are zero on $\mathcal{K}$ and strictly positive only on the subspace $\mathcal{H}$.

Throughout, we analyze the linear part of the first-order stacked residual update. Just to recall, $\mathbf{T}_{bar}$ is defined as:

$$\mathbf{r}_{t+1} = \underbrace{(\mathbf{I} - \eta\mathbf{G}_0^{\text{mat}})}_{\triangleq \mathbf{T}_{\text{bar}}}\mathbf{r}_t \quad + \quad (\text{forcing and higher-order terms}). \tag{3.82}$$

Proposition 3.2.7: Result 2 on spectral analysis

Let $\mathbf{G}_0^{\text{mat}}$ be as above. Define

$$s_{\min} \triangleq \min\{s_i > 0 : s_i \in \text{spec}(\mathbf{G}_0^{\text{mat}})\} \quad \text{and} \quad s_{\max} \triangleq \max \text{spec}(\mathbf{G}_0^{\text{mat}}), \tag{3.83}$$

where $\text{spec}(\cdot)$ denotes the spectrum of a matrix. Denote by ρ_* the spectral radius of $\mathbf{T}_{\text{bar}}|_{\mathcal{H}}$.

(1). On the subspace $\mathcal{H} \triangleq \mathrm{range}(\mathbf{G}_0^{\mathrm{mat}})$ we have the Rayleigh bounds

$$s_{\min} = \mu_X \triangleq \sigma_r(\mathbf{W}_0)^2 \qquad \text{and} \qquad s_{\max} \leq L + \sigma_1(\mathbf{A})^2 \|\mathbf{H}_0\|_2^2, \tag{3.84}$$

(2). Furthermore, if the stepsize η satisfies $0 < \eta < 2/s_{\max}$, then $\rho_* \in (0, 1)$.

Proof:

Let $\Delta \in \mathbb{R}^{m \times n}$, the quadratic form associated with the (pseudo) symmetric operator $\boldsymbol{\mathcal{G}}_0$ is:

$$q(\Delta) \triangleq \langle \Delta, \boldsymbol{\mathcal{G}}_0[\Delta] \rangle = \mathrm{tr}(\Delta^\top \boldsymbol{\mathcal{G}}_0[\Delta]). \tag{3.85}$$

Recall that in (3.64), we already have $\mathrm{vec}(\boldsymbol{\mathcal{G}}_0[\Delta]) = \mathbf{G}_0^{\mathrm{mat}}\, \mathrm{vec}(\Delta)$. In light of this, we have

$$q(\Delta) = \mathrm{vec}(\Delta)^\top \mathrm{vec}(\boldsymbol{\mathcal{G}}_0[\Delta]) = \mathrm{vec}(\Delta)^\top \mathbf{G}_0^{\mathrm{mat}}\, \mathrm{vec}(\Delta). \tag{3.86}$$

Let $\mathbf{v} \triangleq \mathrm{vec}(\Delta)$, then $s_{\min}$ and $s_{\max}$ can be viewed as estimating the quadratic:

$$s_{\min} = \min_{\mathbf{v} \in \mathcal{H}, \|\mathbf{v}\|_2 = 1} \mathbf{v}^\top \mathbf{G}_0^{\mathrm{mat}} \mathbf{v} \qquad \text{and} \qquad s_{\max} = \max_{\mathbf{v} \in \mathcal{H}, \|\mathbf{v}\|_2 = 1} \mathbf{v}^\top \mathbf{G}_0^{\mathrm{mat}} \mathbf{v}. \tag{3.87}$$

In a more compact form, they become

$$s_{\min} = \min_{\|\Delta\|_F = 1} q(\Delta) \qquad \text{and} \qquad s_{\max} = \max_{\|\Delta\|_F = 1} q(\Delta). \tag{3.88}$$

We estimate the extreme singular values according to these two characterizations.

The lower bound $s_{\min} \geq \mu_X$ via curvature information of $\mathbf{W}_0\mathbf{W}_0^\top$

Substituting the definition $\boldsymbol{\mathcal{G}}_0[\Delta] = \mathbf{W}_0\mathbf{W}_0^\top\Delta + (\mathbf{A}\mathbf{A}^\top)\Delta\mathbf{H}_0^\top\mathbf{H}_0$ yields:

$$q(\Delta) = \underbrace{\langle \Delta, \mathbf{W}_0\mathbf{W}_0^\top\Delta \rangle}_{T_1} + \underbrace{\langle \Delta, \mathbf{A}\mathbf{A}^\top\Delta\mathbf{H}_0^\top\mathbf{H}_0 \rangle}_{T_2}. \tag{3.89}$$

We expand the T_1 term:

$$T_1 = \mathrm{tr}\left(\Delta^\top\mathbf{W}_0\mathbf{W}_0^\top\Delta\right) = \mathrm{tr}\left[(\mathbf{W}_0^\top\Delta)^\top(\mathbf{W}_0^\top\Delta)\right] = \|\mathbf{W}_0^\top\Delta\|_F^2. \tag{3.90}$$

This term measures the energy of the residual Δ when projected onto the column space of $\mathbf{W}_0$. Recall that

$$s_{\min} = \min_{\Delta \in \mathcal{H}, \|\Delta\|_F = 1} q(\Delta). \tag{3.91}$$

From here we have:

1. **Restricted subspace**: we have established that the residuals live in
$$\mathcal{H} \triangleq \{\Delta \mid \mathrm{col}(\Delta) \subseteq \mathrm{col}(\mathbf{A})\}.$$
2. **Rank preservation**: by initialization, $\mathbf{W}_0 = \mathbf{A}\mathbf{Z}_0$. If $\mathbf{Z}_0$ is well-conditioned (full-rank in our case), then $\mathrm{col}(\mathbf{W}_0) = \mathrm{col}(\mathbf{A})$.
3. **Strict positivity**: Because Δ lives entirely in $\mathrm{col}(\mathbf{A})$, and $\mathrm{col}(\mathbf{A}) = \mathrm{col}(\mathbf{W}_0)$, the matrix product $\mathbf{W}_0^\top\Delta$ cannot be zero unless Δ is zero. Concretely, for any vector $v \in \mathrm{col}(\mathbf{W}_0)$, we have $\|\mathbf{W}_0^\top v\| \geq \sigma_r(\mathbf{W}_0)\|v\|$. Applying this column-wise to Δ:
$$\|\mathbf{W}_0^\top\Delta\|_F^2 \geq \sigma_r(\mathbf{W}_0)^2\|\Delta\|_F^2. \tag{3.92}$$

Therefore, $\mathbf{W}_0\mathbf{W}_0^\top$ provides the necessary curvature (p.s.d.) to ensure $s_{\min} \geq \sigma_r(\mathbf{W}_0)^2$. Without this term (or if $\mathbf{W}_0$ were rank-deficient), $s_{\min}$ could be zero, destroying the contraction guarantee $\rho_* < 1$.

Attainment of equality $s_{\min} = \mu_X$

Again

$$\mathcal{G}_0[\Delta] = \underbrace{\mathbf{W}_0\mathbf{W}_0^\top\Delta}_{\mathcal{T}_W[\Delta]} + \underbrace{\mathbf{A}\mathbf{A}^\top\Delta\mathbf{H}_0^\top\mathbf{H}_0}_{\mathcal{T}_H[\Delta]}. \tag{3.93}$$

The min eigenvalue $s_{\min}$ on the subspace $\mathcal{H}$ is given by:

$$s_{\min} = \min_{\Delta\in\mathcal{H}\ ,\ \|\Delta\|_F=1} \left(\langle\Delta, \mathcal{T}_W[\Delta]\rangle + \langle\Delta, \mathcal{T}_H[\Delta]\rangle\right). \tag{3.94}$$

We established that $\langle\Delta, \mathcal{T}_W[\Delta]\rangle \geq \mu_X\|\Delta\|_F^2$ and $\langle\Delta, \mathcal{T}_H[\Delta]\rangle \geq 0$, and therefore $s_{\min} \geq \mu_X$. To prove $s_{\min} = \mu_X$, we must find a non-zero matrix $\Delta^* \in \mathcal{H}$ such that

1. $\mathcal{T}_W[\Delta^*] = \mu_X\Delta^*$, i.e. it is an eigenvector of the $\mathbf{W}$-part within min eigenvalue.
2. $\mathcal{T}_H[\Delta^*] = 0$, i.e. it lies in the null space of the $\mathbf{H}$-part.

The subspace $\mathcal{H}$ consists of all matrices $\Delta \in \mathbb{R}^{m\times n}$ whose columns lie in $\text{col}(\mathbf{A})$. Due to ColSpan preservation $\text{col}(\mathbf{W}_0) = \text{col}(\mathbf{A})$ and rank-r initialization assumption, we can construct Δ^* using the singular vectors of $\mathbf{W}_0$. We do so with the following steps:

Step 1: Minimize the $\mathbf{W}$-part

Let $\mathbf{u}_r \in \mathbb{R}^m$ be the **left** singular vector of $\mathbf{W}_0$ w.r.t. its smallest singular value $\sigma_r(\mathbf{W}_0)$, then:

- $\mathbf{u}_r \in \text{col}(\mathbf{W}_0)$, and so it is a valid column for matrices in $\mathcal{H}$, and
- $\mathbf{W}_0\mathbf{W}_0^\top\mathbf{u}_r = \sigma_r(\mathbf{W}_0)^2\mathbf{u}_r = \mu_X\mathbf{u}_r$.

Step 2: Nullify the $\mathbf{H}$-part: The operator $\mathcal{T}_H$ acts on Δ from the right via $\mathbf{H}_0^\top\mathbf{H}_0$:

$$\mathcal{T}_H[\Delta] = (\mathbf{A}\mathbf{A}^\top\Delta)(\mathbf{H}_0^\top\mathbf{H}_0). \tag{3.95}$$

To make this zero, we need Δ to be orthogonal to the row space $\mathbf{H}_0^\top$.Let $\mathbf{v} \in \mathbb{R}^n$ be a vector in the null space of $\mathbf{H}_0$, i.e. $\mathbf{H}_0\mathbf{v} = 0$. At initialization, $\mathbf{H}_0 = \varepsilon_H\mathbf{1}_d\mathbf{1}_n^\top$. This matrix has rank 1. As long as the dimension n (the number of columns of $\mathbf{A}$) is strictly greater than 1, the null space of $\mathbf{H}_0$ is non-empty (which is of dimension $n-1$)

An explicit choice for $\mathbf{v}$ such that $\mathbf{v} \perp \mathbf{1}_n$ is $\mathbf{v} = [1, -1, 0, \cdots, 0]^\top$.

Step 3: Define Δ^*, construct the rank-1 matrix

$$\Delta^* \triangleq \mathbf{u}_r\mathbf{v}^\top. \tag{3.96}$$

We proceed as follows:

- **Check subspace constraint**: The columns of Δ^* are multiples of $\mathbf{u}_r$. In addition, since $\mathbf{u}_r \in \text{col}(\mathbf{W}_0) = \text{col}(\mathbf{A})$, we have $\Delta^* \in \mathcal{H}$.
- **Evaluate $\mathcal{T}_W[\Delta^*]$**:

$$\mathbf{W}_0\mathbf{W}_0^\top(\mathbf{u}_r\mathbf{v}^\top) = (\mathbf{W}_0\mathbf{W}_0^\top\mathbf{u}_r)\mathbf{v}^\top = (\mu_X\mathbf{u}_r)\mathbf{v}^\top = \mu_X\Delta^*. \tag{3.97}$$

- **Evaluate $\boldsymbol{\mathcal{T}}_H[\Delta^*]$:**

$$\mathbf{A}\mathbf{A}^\top(\mathbf{u}_r\mathbf{v}^\top)\mathbf{H}_0^\top\mathbf{H}_0 = (\mathbf{A}\mathbf{A}^\top\mathbf{u}_r)(\mathbf{v}^\top\mathbf{H}_0^\top)\mathbf{H}_0, \tag{3.98}$$

and since we've chosen $\mathbf{v}$ such that$\mathbf{v}^\top\mathbf{H}_0^\top = 0$, the entire term vanishes:

$$\boldsymbol{\mathcal{T}}_H[\Delta^*] = 0. \tag{3.99}$$

Finally, applying the full operator to Δ^* yields:

$$\boldsymbol{\mathcal{G}}_0[\Delta^*] = \boldsymbol{\mathcal{T}}_W[\Delta^*] + \boldsymbol{\mathcal{T}}_H[\Delta^*] = \mu_X\Delta^* + 0 = \mu_X\Delta^*. \tag{3.100}$$

Since $s_{\min}$ is the min eigenvalue and we have found a direction achieving μ_X, and we previously proved $s_{\min} \geq \mu_X$, it follows that

$$s_{\min} = \mu_X. \tag{3.101}$$

We are done with this part. ■

Upper bound $s_{\max} \leq L + \sigma_1(\mathbf{A})^2\|\mathbf{H}_0\|_2^2$

$$s_{\max} = \max_{\|\Delta\|_F=1} \|\boldsymbol{\mathcal{G}}_0[\Delta]\|_F \leq L + \sigma_1(\mathbf{A})^2\|\mathbf{H}_0\|_2^2. \tag{3.102}$$

where $L \triangleq \sigma_1(\mathbf{W}_0)^2$. Here, we seek to bound the operator norm $\boldsymbol{\mathcal{G}}_0$ induced by Frobenius norm:

$$\|\boldsymbol{\mathcal{G}}_0[\Delta]\|_F^2 = \|\mathbf{W}_0\mathbf{W}_0^\top\Delta + \mathbf{A}\mathbf{A}^\top\Delta\mathbf{H}_0^\top\mathbf{H}_0\|_F^2. \tag{3.103}$$

Via triangle inequality, we get

$$\|\boldsymbol{\mathcal{G}}_0[\Delta]\|_F \leq \|\mathbf{W}_0\mathbf{W}_0^\top\Delta\|_F + \|\mathbf{A}\mathbf{A}^\top\Delta\mathbf{H}_0^\top\mathbf{H}_0\|_F. \tag{3.104}$$

For the first term:

$$\begin{aligned}\|\mathbf{W}_0\mathbf{W}_0^\top\Delta\|_F &\leq \|\mathbf{W}_0\mathbf{W}_0^\top\|_2\|\Delta\|_F \\ &= \sigma_1(\mathbf{W}_0\mathbf{W}_0^\top)\|\Delta\|_F \\ &= \sigma_1(\mathbf{W}_0)^2\|\Delta\|_F \\ &= \sigma_1(\mathbf{W}_0)^2\|\Delta\|_F.\end{aligned} \tag{3.105}$$

For the second term:

$$\begin{aligned}\|\mathbf{A}\mathbf{A}^\top\Delta\mathbf{H}_0^\top\mathbf{H}_0\|_F &\leq \|\mathbf{A}\mathbf{A}^\top\|_2\|\Delta\|_F\|\mathbf{H}_0^\top\mathbf{H}_0\|_2 \\ &\leq \sigma_1(\mathbf{A})^2\|\Delta\|_F\|\mathbf{H}_0\|_2^2.\end{aligned} \tag{3.106}$$

Finally, combining two bound yields

$$\|\boldsymbol{\mathcal{G}}_0[\Delta]\|_F \leq \left(L + \sigma_1(\mathbf{A})^2\|\mathbf{H}_0\|_2^2\right)\|\Delta\|_F. \tag{3.107}$$

where again $L = \sigma_1(\mathbf{W}_0)^2$.

Compute $\det(\mathbf{T}_{\text{bar}} - z\mathbf{I}_{mn})$

Fix a scalar $z \in \mathbb{C}$. Then:

$$\mathbf{T}_{\text{bar}} - z\mathbf{I}_{mn} = \mathbf{I} - \eta\mathbf{G}_0^{\text{mat}} - z\mathbf{I}. \tag{3.108}$$

Since $\mathbf{G}_0^{\text{mat}}$ is symmetric, choose an orthonormal basis of $\mathbb{R}^{mn}$ that diagonalizes it:

$$\mathbf{G}_0^{\text{mat}} = \mathbf{Q}\text{diag}(s_1, \ldots, s_{mn})\mathbf{Q}^\top \qquad \text{and} \qquad \mathbf{Q}^\top\mathbf{Q} = \mathbf{I}. \tag{3.109}$$

$$\det(\mathbf{T}_{\text{bar}} - z\mathbf{I}_{mn}) = \prod_{i=1}^{mn}(1 - \eta s_i - z). \tag{3.110}$$

Therefore,

$$\text{spec}(\mathbf{T}_{\text{bar}}) = \{1 - \eta s_i : s_i \in \text{spec}(\mathbf{G}_0^{\text{mat}})\}. \tag{3.111}$$

and the restriction onto $\mathcal{H}$ carries over to both sides:

$$\text{spec}(\mathbf{T}_{\text{bar}}|_{\mathcal{H}}) = \{1 - \eta s_i : s_i \in \text{spec}(\mathbf{G}_0^{\text{mat}}|_{\mathcal{H}})\}. \tag{3.112}$$

Admissible η for contraction on $\mathcal{H}$

Recall that since $\mathbf{G}_0^{\text{mat}}$ is p.s.d., all its eigenvalues are non-negative. The $\mathcal{H}$-subspace restriction of its spectrum $\text{spec}(\mathbf{G}_0^{\text{mat}}|_{\mathcal{H}})$ simply contains only strictly positive values. Then for $\rho_* < 1$, we need

$$-1 < 1 - \eta s < 1 \quad \text{for all } s \in [s_{\min}, s_{\max}], \tag{3.113}$$

which holds iff

$$0 < \eta < \frac{2}{s_{\max}}. \tag{3.114}$$

Thus, with $0 < \eta < 2/s_{\max}$, the restriction of $\mathbf{T}_{\text{bar}}$ to $\mathcal{H}$ is a contraction. We are done. ■.

3.2.6 Control on the nonlinear terms

Before stating results on controlling the nonlinear error and final convergence, we review the choice of our hyper-parameter here: $\|\mathbf{W}_0\|_2 = \sqrt{L}$. The step sizes are $\eta = \frac{2}{s_{\max}+s_{\min}}$. The geometric decay rate is $\theta = \frac{1+\rho_*}{2}$. We define the barrier sequence strictly as $\lambda_t = \lambda_0\theta^t$. The error coefficient functions are strictly defined as:

$$\begin{aligned}
K_1 &\triangleq \eta^2\sigma_1(\mathbf{A})^2(\|\mathbf{H}_0\|_2 + C_2)(\sqrt{L} + C_2),\\
K_2 &\triangleq \eta^2\sigma_1(\mathbf{A})^2\left[(\|\mathbf{H}_0\|_2 + C_2)\frac{\sqrt{nd}}{C_2} + (\sqrt{L} + C_2)\frac{\sqrt{md}}{C_2}\right],\\
K_3 &\triangleq \eta^2\sigma_1(\mathbf{A})^2\frac{\sqrt{md\,nd}}{C_2^2},\\
K_4 &\triangleq C_2\left[\eta\sigma_1(\mathbf{A})^2(2\|\mathbf{H}_0\|_2 + C_2) + \eta(2\sqrt{L} + C_2)\right],\\
C_{\text{bar}} &\triangleq \eta\sigma_1(\mathbf{A})^2(\|\mathbf{H}_0\|_2 + C_2)\frac{\sqrt{md}}{C_2} + \eta(\sqrt{L} + C_2)\frac{\sqrt{nd}}{C_2}, \quad \text{and}\\
\Delta_{\text{gap}} &\triangleq (\theta - \rho_*) - K_4.
\end{aligned} \tag{3.115}$$

Proposition 3.2.8: Result 3 on conditional error bound

We assume the following:

1. **Induction Region**: For a fixed t, assume that for all $0 \leq s \leq t$, the iterates strictly satisfy $\|\mathbf{W}_s - \mathbf{W}_0\|_F \leq C_2$ and $\|\mathbf{H}_s - \mathbf{H}_0\|_F \leq C_2$.
2. **Envelope bound**: Assume for $0 \leq s \leq t$, the stacked residual satisfies $\|\mathbf{r}_s\| \leq C_1\theta^s$.

Then the nonlinearity $\boldsymbol{\xi}_t = \boldsymbol{\psi}_t + \boldsymbol{\iota}_t$ satisfies the exact upper bound

$$\|\boldsymbol{\xi}_t\| \leq K_1(C_1\theta^t)^2 + (K_2\lambda_t + K_4)(C_1\theta^t) + K_3\lambda_t^2. \tag{3.116}$$

where K_1, K_2, K_3, and K_4 are defined as following using C_2 in (3.115).

Proof:

Strict bounds on current state norms

By the Envelope Bound hypothesis, at the current step t, the norm of the stacked residual $\mathbf{r}_t$ is bounded by $C_1\theta^t$. Because $\mathbf{r}_t = \mathrm{vec}(\mathbf{R}_t)$, we possess the strict inequality:

$$\|\mathbf{R}_t\|_F = \|\mathbf{r}_t\| \leq C_1\theta^t, \tag{3.117}$$

From the Induction region hypothesis, we apply the triangle inequality to bound the spectral norms of the current iterates:

$$\|\mathbf{H}_t\|_2 \leq \|\mathbf{H}_0\|_2 + \|\mathbf{H}_t - \mathbf{H}_0\|_2 \leq \|\mathbf{H}_0\|_2 + \|\mathbf{H}_t - \mathbf{H}_0\|_F \leq \|\mathbf{H}_0\|_2 + C_2, \tag{3.118}$$

$$\|\mathbf{W}_t\|_2 \leq \|\mathbf{W}_0\|_2 + \|\mathbf{W}_t - \mathbf{W}_0\|_2 \leq \sqrt{L} + \|\mathbf{W}_t - \mathbf{W}_0\|_F \leq \sqrt{L} + C_2. \tag{3.119}$$

Furthermore, because the absolute value of the difference of any single matrix entry is strictly bounded by the Frobenius norm of the matrix difference, we evaluate the entrywise floors:

$$\min_{i,j}(\mathbf{W}_t)_{ij} \geq \min_{i,j}(\mathbf{W}_0)_{ij} - \|\mathbf{W}_t - \mathbf{W}_0\|_F \geq 2C_2 - C_2 = C_2, \tag{3.120}$$

$$\min_{i,j}(\mathbf{H}_t)_{ij} \geq \min_{i,j}(\mathbf{H}_0)_{ij} - \|\mathbf{H}_t - \mathbf{H}_0\|_F \geq 2C_2 - C_2 = C_2. \tag{3.121}$$

These strict entrywise lower bounds yield exact upper bounds for the Frobenius norms of the element-wise inverses:

$$\|\mathbf{W}_t^{\odot(-1)}\|_F \leq \frac{\sqrt{md}}{C_2} \qquad \text{and} \qquad \|\mathbf{H}_t^{\odot(-1)}\|_F \leq \frac{\sqrt{nd}}{C_2}. \tag{3.122}$$

Strict bounds on increments $\mathbf{P}_t$ and $\mathbf{Q}_t$

The nonlinear error $\boldsymbol{\xi}_t$ is composed of a quadratic term $\boldsymbol{\zeta}_t$ and a linearization error $\boldsymbol{\iota}_t$. We first bound the matrix increments:

$$\mathbf{P}_t \triangleq -\eta\mathbf{A}\mathbf{A}^\top\mathbf{R}_t\mathbf{H}_t^\top + \eta\lambda_t\mathbf{A}\mathbf{A}^\top\mathbf{W}_t^{\odot(-1)}. \tag{3.123}$$

Recall the mixed-norm inequality $\|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_2\|\mathbf{B}\|_F$. Applying it with the triangle inequality yields:

$$\|\mathbf{P}_t\|_F \leq \eta\|\mathbf{A}\mathbf{A}^\top\|_2\|\mathbf{R}_t\|_F\|\mathbf{H}_t^\top\|_2 + \eta\lambda_t\|\mathbf{A}\mathbf{A}^\top\|_2\|\mathbf{W}_t^{\odot(-1)}\|_F. \tag{3.124}$$

Substituting exactly $\|\mathbf{A}\mathbf{A}^\top\|_2 = \sigma_1(\mathbf{A})^2$ and the bounds derived in the previous step yields:

$$\|\mathbf{P}_t\|_F \leq \eta\sigma_1(\mathbf{A})^2(\|\mathbf{H}_0\|_2 + C_2)(C_1\theta^t) + \eta\sigma_1(\mathbf{A})^2\frac{\sqrt{md}}{C_2}\lambda_t. \tag{3.125}$$

Similarly, for the second increment:

$$\mathbf{Q}_t \triangleq -\eta\mathbf{W}_t^\top\mathbf{R}_t + \eta\lambda_t\mathbf{H}_t^{\odot(-1)}. \tag{3.126}$$

Taking the Frobenius norm:

$$\|\mathbf{Q}_t\|_F \leq \eta\|\mathbf{W}_t^\top\|_2\|\mathbf{R}_t\|_F + \eta\lambda_t\|\mathbf{H}_t^{\odot(-1)}\|_F. \tag{3.127}$$

Substituting the strict bounds from Step 1 yields:

$$\|\mathbf{Q}_t\|_F \leq \eta(\sqrt{L} + C_2)(C_1\theta^t) + \eta\frac{\sqrt{nd}}{C_2}\lambda_t. \tag{3.128}$$

Strict bound on the quadratic error $\boldsymbol{\zeta}_t$

The quadratic error component is identically defined as $\boldsymbol{\zeta}_t \triangleq \mathbf{P}_t\mathbf{Q}_t^\top$. By sub-multiplicativity:

$$\|\boldsymbol{\Psi}_t\|_F \leq \|\mathbf{P}_t\|_F\|\mathbf{Q}_t\|_F. \tag{3.129}$$

We explicitly multiply the two upper bounds derived in Step 2:

$$\begin{aligned}\|\boldsymbol{\Psi}_t\|_F \leq &\left[\eta\sigma_1(\mathbf{A})^2(\|\mathbf{H}_0\|_2 + C_2)(C_1\theta^t) + \eta\sigma_1(\mathbf{A})^2\frac{\sqrt{md}}{C_2}\lambda_t\right] \\ &\times \left[\eta(\sqrt{L} + C_2)(C_1\theta^t) + \eta\frac{\sqrt{nd}}{C_2}\lambda_t\right].\end{aligned} \tag{3.130}$$

Expanding this product precisely separates the terms into powers of $(C_1\theta^t)$ and λ_t:

1. The term multiplying $(C_1\theta^t)^2$ exactly matches the definition of K_1.
2. The term multiplying $(C_1\theta^t)\lambda_t$ exactly matches the definition of K_2.
3. The term multiplying λ_t^2 exactly matches the definition of K_3.

Thus, we obtain the exact inequality:

$$\|\boldsymbol{\psi}_t\|_2 \leq K_1(C_1\theta^t)^2 + K_2\lambda_t(C_1\theta^t) + K_3\lambda_t^2. \tag{3.131}$$

Strict bound on the linearization error $\boldsymbol{\iota}_t$

The linearization error is identically $\boldsymbol{\iota}_t \triangleq \mathrm{vec}(\mathbf{E}_t^{(\mathbf{Z})} + \mathbf{E}_t^{(\mathbf{H})})$. We bound each component independently:

$$\mathbf{E}_t^{(\mathbf{Z})} \triangleq -\eta \mathbf{A}\mathbf{A}^\top [\mathbf{R}_t(\mathbf{H}_t^\top - \mathbf{H}_0^\top)]\mathbf{H}_t - \eta \mathbf{A}\mathbf{A}^\top [\mathbf{R}_t \mathbf{H}_0^\top](\mathbf{H}_t - \mathbf{H}_0). \tag{3.132}$$

Taking the Frobenius norm and applying the triangle inequality yields:

$$\begin{aligned}\|\mathbf{E}_t^{(\mathbf{Z})}\|_F \leq & \eta \|\mathbf{A}\mathbf{A}^\top\|_2 \|\mathbf{R}_t\|_F \|\mathbf{H}_t - \mathbf{H}_0\|_2 \|\mathbf{H}_t\|_2 \\ & + \eta \|\mathbf{A}\mathbf{A}^\top\|_2 \|\mathbf{R}_t\|_F \|\mathbf{H}_0\|_2 \|\mathbf{H}_t - \mathbf{H}_0\|_2.\end{aligned} \tag{3.133}$$

Substituting the induction bound $\|\mathbf{H}_t - \mathbf{H}_0\|_2 \leq \|\mathbf{H}_t - \mathbf{H}_0\|_F \leq C_2$:

$$\|\mathbf{E}_t^{(\mathbf{Z})}\|_F \leq \eta \sigma_1(\mathbf{A})^2 (C_1\theta^t) C_2 (\|\mathbf{H}_0\|_2 + C_2) + \eta \sigma_1(\mathbf{A})^2 (C_1\theta^t) \|\mathbf{H}_0\|_2 C_2. \tag{3.134}$$

By factoring the common terms algebraically:

$$\|\mathbf{E}_t^{(\mathbf{Z})}\|_F \leq \eta \sigma_1(\mathbf{A})^2 C_2 (2\|\mathbf{H}_0\|_2 + C_2)(C_1\theta^t). \tag{3.135}$$

Similarly, we evaluate the second component:

$$\mathbf{E}_t^{(\mathbf{H})} \triangleq -\eta(\mathbf{W}_t - \mathbf{W}_0)\mathbf{W}_t^\top \mathbf{R}_t - \eta \mathbf{W}_0 (\mathbf{W}_t - \mathbf{W}_0)^\top \mathbf{R}_t. \tag{3.136}$$

Taking the Frobenius norm and applying identical algebraic substitutions yields:

$$\|\mathbf{E}_t^{(\mathbf{H})}\|_F \leq \eta \|\mathbf{W}_t - \mathbf{W}_0\|_2 \|\mathbf{W}_t\|_2 \|\mathbf{R}_t\|_F + \eta \|\mathbf{W}_0\|_2 \|\mathbf{W}_t - \mathbf{W}_0\|_2 \|\mathbf{R}_t\|_F, \tag{3.137}$$

$$\|\mathbf{E}_t^{(\mathbf{H})}\|_F \leq \eta C_2(\sqrt{L} + C_2)(C_1\theta^t) + \eta\sqrt{L} C_2 (C_1\theta^t), \text{ and} \tag{3.138}$$

$$\|\mathbf{E}_t^{(\mathbf{H})}\|_F \leq \eta C_2 (2\sqrt{L} + C_2)(C_1\theta^t). \tag{3.139}$$

Summing these bounds via the triangle inequality yields exactly K_4:

$$\|\boldsymbol{\iota}_t\|_2 \leq \|\mathbf{E}_t^{(\mathbf{Z})}\|_F + \|\mathbf{E}_t^{(\mathbf{H})}\|_F \leq K_4 (C_1\theta^t). \tag{3.140}$$

Exact total nonlinearity error

The full non-linear error vector identically equals $\boldsymbol{\xi}_t \triangleq \boldsymbol{\psi}_t + \boldsymbol{\iota}_t$. Applying the triangle inequality to the total sum:

$$\|\boldsymbol{\xi}_t\|_2 \leq \|\boldsymbol{\psi}_t\|_2 + \|\boldsymbol{\iota}_t\|_2. \tag{3.141}$$

Substituting the strict upper bounds from Step 3 and Step 4:

$$\|\boldsymbol{\xi}_t\|_2 \leq K_1 (C_1\theta^t)^2 + K_2 \lambda_t (C_1\theta^t) + K_3 \lambda_t^2 + K_4 (C_1\theta^t). \tag{3.142}$$

Factoring the linear common multiplier $(C_1\theta^t)$ from the relevant terms yields exactly:

$$\|\boldsymbol{\xi}_t\|_2 \leq K_1 (C_1\theta^t)^2 + (K_2\lambda_t + K_4)(C_1\theta^t) + K_3 \lambda_t^2. \tag{3.143}$$

We are done. ■

3.2.7 Convergence

Here, we state the **conditional** result on the final iteration complexity analysis.

Proposition 3.2.9: Result 4 on conditional global convergence in infinite horizon

Assume the constants C_1, C_2 and λ satisfy all their respective constraint inequalities explicitly below. Then the following three propositions hold jointly for all $t \geq 0$:

$$\mathcal{P}_1(t) : \|\mathbf{r}_t\| \leq C_1\theta^t, \quad \mathcal{P}_2(t) : \|\mathbf{W}_t - \mathbf{W}_0\|_F \leq C_2, \quad \text{and} \quad \mathcal{P}_3(t) : \|\mathbf{H}_t - \mathbf{H}_0\|_F \leq C_2, \tag{3.144}$$

where $\rho_* \triangleq \rho(\mathbf{T}_{\text{bar}}) \in (0,1)$ with $\eta < 2/s_{\max}$ and $\theta \triangleq \frac{1+\rho_*}{2}$.
In addition, the following are the specific conditions that C_1, C_2 and λ needs to satisfy:

1. Conditions on C_1: Again, C_1 represents the magnitude of the initial stacked residual. It is explicitly defined by the initialization matrices $\mathbf{W}_0$ and $\mathbf{H}_0$:

$$C_1 \triangleq \|\mathbf{r}_0\| = \|\mathbf{W}_0\mathbf{H}_0 - \mathbf{A}\|_F. \tag{3.145}$$

   By substituting the explicit initialization definitions $\mathbf{W}_0 = c\mathbf{A}\mathbf{\Phi} + \alpha_0(\mathbf{A}\mathbf{1}_n\mathbf{1}_d^\top)$ and $\mathbf{H}_0 = \varepsilon_{\mathbf{H}}\mathbf{1}_d\mathbf{1}_n^\top$. Therefore, an explicit choice for C_1 is:

$$C_1 = \left\| \left(c\mathbf{A}\mathbf{\Phi} + \alpha_0(\mathbf{A}\mathbf{1}_n\mathbf{1}_d^\top)\right)\left(\varepsilon_{\mathbf{H}}\mathbf{1}_d\mathbf{1}_n^\top\right) - \mathbf{A} \right\|_F. \tag{3.146}$$

   Conditions to satisfy: To ensure that the quadratic initial noise is strictly bounded by the linear stability gap, C_1 must satisfy the single strict inequality constraint:

$$K_1 C_1 < \Delta_{\text{gap}}. \tag{3.147}$$

2. Conditions on λ_0: The initial barrier weight λ_0 must reside in the interval $\lambda_0 \in (0, \lambda_{\max}]$, where $\lambda_{\max}$ is the minimum of the following three bounds:

   - **Basin bound** λ_{basin}: λ_0 must not exceed the strictly positive real root of the quadratic equation

$$K_3\lambda_0^2 + (C_{\text{bar}} + K_2C_1)\lambda_0 + (K_1C_1^2 - \Delta_{\text{gap}}C_1) = 0, \tag{3.148}$$

     which has an explicit expression:

$$\lambda_{\text{basin}} \triangleq \frac{-(C_{\text{bar}} + K_2C_1) + \sqrt{(C_{\text{bar}} + K_2C_1)^2 - 4K_3(K_1C_1^2 - \Delta_{\text{gap}}C_1)}}{2K_3}. \tag{3.149}$$

   - **Increment bound for the $\mathbf{W}$ block** $\lambda_{\text{inc}}^{\mathbf{W}}$:

$$\lambda_{\text{inc}}^{\mathbf{W}} \triangleq \frac{C_2(\theta - \rho_*) - \eta\sigma_1(\mathbf{A})^2(\|\mathbf{H}_0\|_2 + C_2)C_1}{\eta\sigma_1(\mathbf{A})^2\frac{\sqrt{md}}{C_2}}. \tag{3.150}$$

   - **Increment bound for the $\mathbf{H}$ block** $\lambda_{\text{inc}}^{\mathbf{H}}$:

$$\lambda_{\text{inc}}^{\mathbf{H}} \triangleq \frac{C_2(\theta - \rho_*) - \eta(\sqrt{L} + C_2)C_1}{\eta\frac{\sqrt{nd}}{C_2}}. \tag{3.151}$$

That is, $\lambda_0 \leq \min\{\lambda_{\text{basin}}, \lambda_{\text{inc}}^{\mathbf{W}}, \lambda_{\text{inc}}^{\mathbf{H}}\}$.

3. Conditions on C_2:

- **Lower bounds**:
  - Strict positivity: $C_2 > 0$
  - Capacity for the $\mathbf{W}$-block infinite drift:

$$C_2 > \frac{\eta\sigma_1(\mathbf{A})^2(\|\mathbf{H}_0\|_2 + C_2)C_1}{\theta - \rho_*}, \tag{3.152}$$

    which implicitly assumes

$$C_1 < (\theta - \rho_*)/\eta\sigma_1(\mathbf{A})^2. \tag{3.153}$$

  - Capacity for the $\mathbf{H}$-block infinite drift:

$$C_2 > \frac{\eta(\sqrt{L} + C_2)C_1}{\theta - \rho_*}, \tag{3.154}$$

    which implicitly assumes

$$C_1 < (\theta - \rho_*)/\eta. \tag{3.155}$$

- **Upper bounds**:
  - Linear Gap Positivity $C_2^{\max}$: To ensure that $K_4 < \frac{\theta - \rho_*}{2}$, which guarantees that $\Delta_{\text{gap}} > 0$, C_2 must be bounded by the positive root of a relevant quadratic. Specifically:

$$\begin{aligned} C_2 \leq C_2^{\max} \triangleq & \frac{-\left(2\eta\sigma_1(\mathbf{A})^2\|\mathbf{H}_0\|_2 + 2\eta\sqrt{L}\right)}{2\left(\eta\sigma_1(\mathbf{A})^2 + \eta\right)} \\ & + \frac{\sqrt{S_1 + S_2}}{2\left(\eta\sigma_1(\mathbf{A})^2 + \eta\right)}, \\ \text{where} \quad & S_1 \triangleq \left(2\eta\sigma_1(\mathbf{A})^2\|\mathbf{H}_0\|_2 + 2\eta\sqrt{L}\right)^2, \\ & S_2 \triangleq 2\left(\eta\sigma_1(\mathbf{A})^2 + \eta\right)(\theta - \rho_*). \end{aligned} \tag{3.156}$$

  - **Initialization Floor Bounds**: We need $\varepsilon_{\mathbf{H}} \geq 2C_2$ to ensure $\mathbf{H}_0 \geq 2C_2$ and for $\mathbf{W}_0 \geq 2C_2$ we need:

$$\alpha_0 a_{\min} - cB_{\mathbf{\Phi}} \geq 2C_2, \tag{3.157}$$

    where $B_{\mathbf{\Phi}}$ is given in equation (3.24).

Principal goal of these conditions

Here, the principal goal is NOT to verify the absence of contradictions between these conditions. We only derive and state the necessity of them such that, if no contradiction is present between the upper and lower bounds, then $\mathcal{P}_1$, $\mathcal{P}_2$, and $\mathcal{P}_3$ hold.

Proof:

We prove that the joint hypothesis $\mathcal{H}(t) = \mathcal{P}_1(t) \wedge \mathcal{P}_2(t) \wedge \mathcal{P}_3(t)$ holds for all $t \geq 0$ by induction.

Base case ($t = 0$)**:**

1. $\mathcal{P}_1(0)$ states $\|\mathbf{r}_0\| \leq C_1\theta^0$. By exact definition, $C_1 \triangleq \|\mathbf{r}_0\|$. Note that since $\theta^0 = 1$, we have $\|\mathbf{r}_0\| \leq C_1$, which holds identically.
2. $\mathcal{P}_2(0)$ states $\|\mathbf{W}_0 - \mathbf{W}_0\|_F \leq C_2$. Since $\|\mathbf{0}\|_F = 0$, and by the lower bound $C_2 > 0$, this holds strictly.
3. $\mathcal{P}_3(0)$ states $\|\mathbf{H}_0 - \mathbf{H}_0\|_F \leq C_2$. By identical logic, this holds strictly.

Thus, the base case $\mathcal{H}(0)$ holds.

Inductive step:

Assume the joint hypothesis $\mathcal{H}(s)$ holds for all integers $0 \leq s \leq t$. We must rigorously prove $\mathcal{H}(t+1)$, which requires proving $\mathcal{P}_1(t+1)$, $\mathcal{P}_2(t+1)$, and $\mathcal{P}_3(t+1)$.

Step 1: Prove $\mathcal{P}_1(t+1)$

Because $\mathcal{P}_2(s)$ and $\mathcal{P}_3(s)$ hold for all $0 \leq s \leq t$, the induction region constraints are satisfied. Because $\mathcal{P}_1(s)$ holds, the envelope bound is satisfied. Therefore, we invoke the strict bound on the nonlinearity $\boldsymbol{\xi}_t$ from **Result 3**, evaluated with the exact barrier sequence $\lambda_t = \lambda_0\theta^t$:

$$\|\boldsymbol{\xi}_t\| \leq K_1(C_1\theta^t)^2 + (K_2\lambda_0\theta^t + K_4)(C_1\theta^t) + K_3(\lambda_0\theta^t)^2. \tag{3.158}$$

From the exact one-step state dynamics norm bound, we have:

$$\|\mathbf{r}_{t+1}\| \leq \rho_*\|\mathbf{r}_t\| + \lambda_t C_{\text{bar}} + \|\boldsymbol{\xi}_t\|. \tag{3.159}$$

Substitute $\|\mathbf{r}_t\| \leq C_1\theta^t$, $\lambda_t = \lambda_0\theta^t$ and the strict bound for $\|\boldsymbol{\xi}_t\|$:

$$\|\mathbf{r}_{t+1}\| \leq \rho_* C_1\theta^t + \lambda_0\theta^t C_{\text{bar}} + K_1C_1^2\theta^{2t} + K_2\lambda_0 C_1\theta^{2t} + K_4C_1\theta^t + K_3\lambda_0^2\theta^{2t}. \tag{3.160}$$

Because the geometric rate satisfies $0 < \theta < 1$, we hold the strict inequality $\theta^{2t} \leq \theta^t$. By replacing θ^{2t} with θ^t on the strictly positive quadratic terms, we maintain a strict upper bound. We factor the common term θ^t from the entire right-hand side:

$$\|\mathbf{r}_{t+1}\| \leq \theta^t \left[\rho_* C_1 + C_{\text{bar}}\lambda_0 + K_1C_1^2 + K_2C_1\lambda_0 + K_4C_1 + K_3\lambda_0^2\right]. \tag{3.161}$$

To prove $\mathcal{P}_1(t+1)$, which requires $\|\mathbf{r}_{t+1}\| \leq C_1\theta^{t+1}$, it is sufficient to prove that the bracketed expression is less than or equal to $C_1\theta$:

$$\rho_* C_1 + C_{\text{bar}}\lambda_0 + K_1 C_1^2 + K_2 C_1 \lambda_0 + K_4 C_1 + K_3 \lambda_0^2 \leq C_1\theta. \tag{3.162}$$

We subtract $(\rho_* C_1 + K_4 C_1)$ from both sides:

$$K_3\lambda_0^2 + (C_{\text{bar}} + K_2 C_1)\lambda_0 + K_1 C_1^2 \leq C_1(\theta - \rho_* - K_4). \tag{3.163}$$

By exact definition, $\Delta_{\text{gap}} \triangleq (\theta - \rho_*) - K_4$. We subtract $C_1\Delta_{\text{gap}}$ from both sides to form a quadratic inequality w.r.t. λ_0:

$$K_3\lambda_0^2 + (C_{\text{bar}} + K_2 C_1)\lambda_0 + (K_1 C_1^2 - \Delta_{\text{gap}} C_1) \leq 0. \tag{3.164}$$

By the assumed condition on C_1, i.e. $K_1 C_1 < \Delta_{\text{gap}}$, the constant term of this quadratic is strictly negative. Therefore, a strictly positive root λ_{basin} exists. By the assumed condition on the barrier, $\lambda_0 \leq \lambda_{\text{basin}}$. Because λ_0 falls within the closed interval bounded by the positive root, the quadratic inequality holds strictly.

Consequently, $\|\mathbf{r}_{t+1}\| \leq \theta^t(C_1\theta) = C_1\theta^{t+1}$. Thus, $\mathcal{P}_1(t+1)$ holds.

Step 2: Prove $\mathcal{P}_2(t+1)$

We bound the total distance from initialization by summing the one-step increments of $\mathbf{W}$ via the triangle inequality:

$$\|\mathbf{W}_{t+1} - \mathbf{W}_0\|_F \leq \sum_{s=0}^{t} \|\mathbf{W}_{s+1} - \mathbf{W}_s\|_F. \tag{3.165}$$

The explicit one-step increment for the $\mathbf{W}$ block is:

$$\mathbf{W}_{s+1} - \mathbf{W}_s = -\eta\mathbf{A}\mathbf{A}^\top\mathbf{R}_s\mathbf{H}_s^\top + \eta\lambda_s\mathbf{A}\mathbf{A}^\top\mathbf{W}_s^{\odot(-1)}. \tag{3.166}$$

By taking the Frobenius norm, applying the triangle inequality, and using submultiplicativity:

$$\|\mathbf{W}_{s+1} - \mathbf{W}_s\|_F \leq \eta\|\mathbf{A}\mathbf{A}^\top\|_2\|\mathbf{H}_s^\top\|_2\|\mathbf{R}_s\|_F + \eta\lambda_s\|\mathbf{A}\mathbf{A}^\top\|_2\|\mathbf{W}_s^{\odot(-1)}\|_F. \tag{3.167}$$

By the induction hypotheses for all $s \leq t$, we substitute $\|\mathbf{A}\mathbf{A}^\top\|_2 = \sigma_1(\mathbf{A})^2$, $\|\mathbf{H}_s\|_2 \leq \|\mathbf{H}_0\|_2 + C_2$, and the strict inverse bound $\|\mathbf{W}_s^{\odot(-1)}\|_F \leq \frac{\sqrt{md}}{C_2}$. Furthermore, $\|\mathbf{R}_s\|_F \leq \|\mathbf{r}_s\| \leq C_1\theta^s$.

Therefore, we get:

$$\|\mathbf{W}_{s+1} - \mathbf{W}_s\|_F \leq \eta\sigma_1(\mathbf{A})^2(\|\mathbf{H}_0\|_2 + C_2)C_1\theta^s + \eta\sigma_1(\mathbf{A})^2\frac{\sqrt{md}}{C_2}\lambda_0\theta^s. \tag{3.168}$$

Summing this exact expression from $s = 0$ to t, we get:

$$\|\mathbf{W}_{t+1} - \mathbf{W}_0\|_F \leq \left[\eta\sigma_1(\mathbf{A})^2(\|\mathbf{H}_0\|_2 + C_2)C_1 + \eta\sigma_1(\mathbf{A})^2\frac{\sqrt{md}}{C_2}\lambda_0\right]\sum_{s=0}^{t}\theta^s. \tag{3.169}$$

Because $\theta > 0$, the finite sum is strictly bounded by the infinite geometric series limit:

$$\sum_{s=0}^{t} \theta^s < \sum_{s=0}^{\infty} \theta^s = \frac{1}{1-\theta} = \frac{1}{\theta - \rho_*}. \tag{3.170}$$

Substitute this strictly into the bound, we get:

$$\|\mathbf{W}_{t+1} - \mathbf{W}_0\|_F < \frac{\eta\sigma_1(\mathbf{A})^2(\|\mathbf{H}_0\|_2 + C_2)C_1 + \eta\sigma_1(\mathbf{A})^2\frac{\sqrt{md}}{C_2}\lambda_0}{\theta - \rho_*}. \tag{3.171}$$

To prove $\mathcal{P}_2(t+1)$, we must guarantee this quantity is less than or equal to C_2. We set up the required inequality:

$$\frac{\eta\sigma_1(\mathbf{A})^2(\|\mathbf{H}_0\|_2 + C_2)C_1 + \eta\sigma_1(\mathbf{A})^2\frac{\sqrt{md}}{C_2}\lambda_0}{\theta - \rho_*} \leq C_2. \tag{3.172}$$

Multiply both sides by $(\theta - \rho_*)$ and isolate the term containing λ_0:

$$\eta\sigma_1(\mathbf{A})^2\frac{\sqrt{md}}{C_2}\lambda_0 \leq C_2(\theta - \rho_*) - \eta\sigma_1(\mathbf{A})^2(\|\mathbf{H}_0\|_2 + C_2)C_1. \tag{3.173}$$

Dividing by the exact coefficient of λ_0 produces:

$$\lambda_0 \leq \frac{C_2(\theta - \rho_*) - \eta\sigma_1(\mathbf{A})^2(\|\mathbf{H}_0\|_2 + C_2)C_1}{\eta\sigma_1(\mathbf{A})^2\frac{\sqrt{md}}{C_2}}. \tag{3.174}$$

This right-hand side is the exact definition of $\lambda_{\text{inc}}^{\mathbf{W}}$. The non-negativity of $\lambda_{\text{inc}}^{\mathbf{W}}$ also assumes equations (3.152) and (3.153).Because the assumption explicitly mandates $\lambda_0 \leq \lambda_{\text{inc}}^{\mathbf{W}}$, the inequality holds. Thus, $\|\mathbf{W}_{t+1} - \mathbf{W}_0\|_F \leq C_2$, proving $\mathcal{P}_2(t+1)$.

Step 3: Prove $\mathcal{P}_3(t+1)$

We apply identical algebraic derivation to the $\mathbf{H}$ block. The explicit one-step increment is:

$$\mathbf{H}_{s+1} - \mathbf{H}_s = -\eta\mathbf{W}_s^\top\mathbf{R}_s + \eta\lambda_s\mathbf{H}_s^{\odot(-1)}. \tag{3.175}$$

Taking the Frobenius norm and applying the induction hypotheses exactly as before, with $\|\mathbf{W}_s\|_2 \leq \|\mathbf{W}_0\|_2 + C_2 = \sqrt{L} + C_2$ and $\|\mathbf{H}_s^{\odot(-1)}\|_F \leq \frac{\sqrt{nd}}{C_2}$:

$$\|\mathbf{H}_{s+1} - \mathbf{H}_s\|_F \leq \eta(\sqrt{L} + C_2)C_1\theta^s + \eta\frac{\sqrt{nd}}{C_2}\lambda_0\theta^s. \tag{3.176}$$

Summing from $s = 0$ to t and substituting the strict upper limit for the geometric series:

$$\|\mathbf{H}_{t+1} - \mathbf{H}_0\|_F < \frac{\eta(\sqrt{L} + C_2)C_1 + \eta\frac{\sqrt{nd}}{C_2}\lambda_0}{\theta - \rho_*}. \tag{3.177}$$

We enforce that this strict upper bound is less than or equal to C_2:

$$\frac{\eta(\sqrt{L} + C_2)C_1 + \eta\frac{\sqrt{nd}}{C_2}\lambda_0}{\theta - \rho_*} \leq C_2. \tag{3.178}$$

Isolating the term containing λ_0:

$$\eta\frac{\sqrt{nd}}{C_2}\lambda_0 \leq C_2(\theta-\rho_*) - \eta(\sqrt{L}+C_2)C_1. \tag{3.179}$$

Dividing by the coefficient of λ_0 yields:

$$\lambda_0 \leq \frac{C_2(\theta-\rho_*) - \eta(\sqrt{L}+C_2)C_1}{\eta\frac{\sqrt{nd}}{C_2}}. \tag{3.180}$$

This right-hand side is exactly the definition of $\lambda_{\text{inc}}^{\mathbf{H}}$. The non-negativity of $\lambda_{\text{inc}}^{\mathbf{H}}$ also assumes equations (3.154) and (3.155).Because the assumption explicitly mandates $\lambda_0 \leq \lambda_{\text{inc}}^{\mathbf{H}}$, the inequality holds. Thus, $\|\mathbf{H}_{t+1} - \mathbf{H}_0\|_F \leq C_2$, proving $\mathcal{P}_3(t+1)$.

In summary, the joint hypothesis $\mathcal{H}(t+1)$ holds. We are done. ■

3.2.8 The neccessities (and pitfalls) of those conditional statements

Here, I would like to address the neccessities of a few conditional statements that hasn't been explicitly mentioned or used in the "proof" of proposition 3.2.7, as well as some related pitfalls and methods to relax those conditions.

On the neccessities of equations (3.152) and (3.154)

Recall the $\mathbf{W}$ block increment bound $\lambda_{\text{inc}}^{\mathbf{W}}$ for λ:

$$\lambda_{\text{inc}}^{\mathbf{W}} \triangleq \frac{C_2(\theta-\rho_*) - \eta\sigma_1(\mathbf{A})^2(\|\mathbf{H}_0\|_2 + C_2)C_1}{\eta\sigma_1(\mathbf{A})^2\frac{\sqrt{md}}{C_2}}. \tag{3.181}$$

We need $\lambda \leq \lambda_{\text{inc}}^{\mathbf{W}}$ for $\mathcal{P}_2$ to hold, i.e. the total drift on $\mathbf{W}$ block can be controlled. Equation (3.152) ensures that $\lambda_{\text{inc}}^{\mathbf{W}} \geq 0$. By symmetry, equation (3.154) ensures that $\lambda_{\text{inc}}^{\mathbf{H}} \geq 0$.

On the neccessity and pitfalls of equation (3.147)

Recall the condition:

$$K_1C_1 < \Delta_{\text{gap}}, \tag{3.182}$$

where $\Delta_{\text{gap}} = (\theta-\rho_*) - K_4$. This condition strictly arises to prevent a contradiction between the **lower bound** and the **upper bound** of the initial barrier weight λ_0. Again, the barrier weight λ_0 at initialization is mandated to satisfy:

1. **Lower Bound:** $\lambda_0 > 0$ (The barrier weight must be strictly positive).
2. **Upper Bound:** $\lambda_0 \leq \lambda_{\text{basin}}$ (To ensure the initial residual remains strictly inside the geometric decay envelope during the inductive step).

To avoid a contradiction between these two bounds, the valid interval $(0, \lambda_{\text{basin}}]$ must be non-empty. This requires us to strictly guarantee:

$$\lambda_{\text{basin}} > 0. \tag{3.183}$$

By definition in Result 4, λ_{basin} is the positive real root of the following quadratic equation w.r.t. λ_0:

$$K_3\lambda_0^2 + (C_{\text{bar}} + K_2C_1)\lambda_0 + (K_1C_1^2 - \Delta_{\text{gap}}C_1) = 0. \tag{3.184}$$

Let $f(\lambda_0)$ denote this quadratic function. Note that:

- The leading coefficient K_3 is strictly positive.
- The linear coefficient $(C_{\text{bar}} + K_2C_1)$ is strictly positive.

Because $f(\lambda_0)$ is a parabola opening upwards with a strictly positive slope at $\lambda_0 = 0$, the only way for it to possess a strictly positive real root ($\lambda_{\text{basin}} > 0$) is if the constant term evaluates to a strictly negative value:

$$K_1C_1^2 - \Delta_{\text{gap}}C_1 < 0. \tag{3.185}$$

Because the initial stacked residual C_1 is strictly positive, we divide the entire inequality by C_1 to obtain:

$$K_1C_1 < \Delta_{\text{gap}} \iff K_1C_1 < (\theta - \rho_*) - K_4. \tag{3.186}$$

By substituting the explicit algebraic definitions of K_1, C_1, and K_4 into this strict requirement, we obtain the following inequality that involves the high-probability RMT statement at initialization, which is:

$$\begin{aligned}
&\eta^2\sigma_1(\mathbf{A})^2(\|\mathbf{H}_0\|_2 + C_2)(\sqrt{L} + C_2)\Big[2C_2c\|\mathbf{A}\mathbf{\Phi}\mathbf{1}_d\mathbf{1}_n^\top\|_F \\
&\quad + 2C_2\left(\frac{2C_2 + cB_{\mathbf{\Phi}}}{a_{\min}}\right)d\|\mathbf{A}\mathbf{1}_n\mathbf{1}_n^\top\|_F + \|\mathbf{A}\|_F\Big] \\
&< (\theta - \rho_*) - C_2\left[\eta\sigma_1(\mathbf{A})^2(2\|\mathbf{H}_0\|_2 + C_2) + \eta(2\sqrt{L} + C_2)\right].
\end{aligned} \tag{3.187}$$

In addition to the difficulty of verifying such statement, one of the undesirable things is that we can only show such a C_2 exist, but it would be difficult to pin down a specific and handy value for C_2.

On the neccessity of equation (3.156)

Recall the linear stability gap is defined as $\Delta_{\text{gap}} \triangleq (\theta - \rho_*) - K_4$, and we have required that $K_1C_1 < \Delta_{\text{gap}}$, and this implicitly assumes Δ_{gap} is positive. For this we will need:

$$K_4 < \theta - \rho_*. \tag{3.188}$$

One may check that $K_4(C_2)$ is a strictly increasing function of C_2 with $K_4(0) = 0$. By definition, $\theta \triangleq \frac{1+\rho_*}{2}$ and so $\theta - \rho_* > 0$. Therefore, we define $C_2^{\max}$ to be the root of $K_4(C_2^{\max}) = (\theta - \rho_*)/2$,

so that $\Delta_{\text{gap}} = \frac{\theta-\rho_*}{2}$ at $C_2^{\max}$. [4] Expanding the condition $K_4 = \frac{\theta-\rho_*}{2}$gives us a quadratic in C_2, which is:

$$(\eta\sigma_1(\mathbf{A})^2 + \eta)C_2^2 + (2\eta\sigma_1(\mathbf{A})^2\|\mathbf{H}_0\|_2 + 2\eta\sqrt{L})C_2 - \frac{\theta-\rho_*}{2} = 0. \tag{3.189}$$

Because the leading coefficient is strictly positive and the constant term is strictly negative, we have exactly one positive root and one negative root. Then the positive root $C_2^{\max}$ has a closed form solution in terms of the coefficients of this qradratic, which is precisely equation (3.156).

Untangling the geometric rate θ from the linear contraction ρ_*

In Xu et al. (2024), they authors have nice plug-ins for ρ_* and θ, so that it happens that θ is the mid-point between 1 and ρ_*. Here, due to the fact that we do not know the precise value of $s_{\max}$, but only an upper bound in proposition 3.2.7, we only obtain an upper bound on ρ_*. This also limits our capacity to pin down a specific value of θ. In their proof, due to θ being exactly at the midpoint, the distance to the linear rate $(\theta - \rho_*)$ and the distance to the stability ceiling $(1 - \theta)$ are forced to be algebraically identical. While this simplifies the resulting bounds by allowing a single denominator to describe both the one-step induction gap and the infinite series limit, it artificially restricts the flexibility of the proof.

To derive a tighter and more generalized feasibility region for our hyperparameters, we untangle these two rates. We define θ as a free parameter bounded strictly within the active interval:

$$\rho_* < \theta < 1. \tag{3.190}$$

By decoupling θ, we explicitly separate the two fundamental bottlenecks of the trajectory analysis:

1. **The Induction Gap $(\theta - \rho_*)$:** This quantity represents the one-step buffer available to absorb the non-linear residual $\boldsymbol{\xi}_t$. Maximizing this gap (by pushing $\theta \to 1$) allows the algorithm to tolerate a larger initial error C_1, thereby widening the theoretical basin of attraction.
2. **The Infinite Drift Limit $(1 - \theta)$:** This quantity controls the absolute distance the state matrices $\mathbf{W}_t$ and $\mathbf{H}_t$ can drift from their initialization. Because the total accumulated trajectory length is bounded by the geometric series $\sum_{s=0}^{\infty}\theta^s = \frac{1}{1-\theta}$, maximizing this gap (by pushing $\theta \to \rho_*$) minimizes the total landscape drift C_2. This is mathematically critical to ensure the trajectory does not leave the region where the linear approximation $\boldsymbol{\mathcal{G}}_0$ remains valid.

To see exactly how this infinite drift limit dictates the basin of attraction, we trace the accumulation of the matrix state errors. Consider the one-step increment for the $\mathbf{H}$-block. Isolating the primary drift driven by the gradient of the residual, we take the Frobenius norm and apply the sub-multiplicative property:

$$\|\mathbf{H}_{s+1} - \mathbf{H}_s\|_F \leq \eta\|\mathbf{W}_s\|_2\|\mathbf{R}_s\|_F. \tag{3.191}$$

[4] De facto, we have required $K_4 \leq \frac{\theta-\rho_*}{2}$. The choice for this ratio $1/2$ is to balance the restrictions on C_1 and C_2. If this ratio is 0.99, then we do have a looser bound on C_2. We then have $\Delta_{\text{gap}} \geq 0.01(\theta - \rho_*)$. Since we also need $K_1C_1 < \Delta_{\text{gap}}$, this would require $K_1C_1 < 0.01(\theta - \rho_*)$, meaning that the initializations $\mathbf{W}_0$ and $\mathbf{H}_0$ must be exceptionally close to the exact solutions to make C_1 tiny enough. Similar story holds if this ratio is extremely small.

By substituting the induction hypothesis for the state bound $\|\mathbf{W}_s\|_2 \leq \sqrt{L}+C_2$ and the geometric decay envelope of the residual $\|\mathbf{R}_s\|_F \leq C_1\theta^s$, the one-step distance is strictly bounded by:

$$\|\mathbf{H}_{s+1} - \mathbf{H}_s\|_F \leq \eta(\sqrt{L} + C_2)C_1\theta^s. \tag{3.192}$$

To find the maximum possible distance the trajectory can drift from its initialization $\mathbf{H}_0$ over all time steps, we sum these increments using the triangle inequality:

$$\|\mathbf{H}_{t+1} - \mathbf{H}_0\|_F \leq \sum_{s=0}^{t} \|\mathbf{H}_{s+1} - \mathbf{H}_s\|_F \leq \eta(\sqrt{L} + C_2)C_1 \sum_{s=0}^{t} \theta^s. \tag{3.193}$$

Because the geometric rate satisfies $\theta \in (0, 1)$, the finite sum is strictly bounded by the infinite geometric series limit $\sum_{s=0}^{\infty} \theta^s = \frac{1}{1-\theta}$. Substituting this limit provides the absolute maximum trajectory length:

$$\|\mathbf{H}_{t+1} - \mathbf{H}_0\|_F < \frac{\eta(\sqrt{L} + C_2)C_1}{1 - \theta}. \tag{3.194}$$

For the induction hypothesis $\mathcal{P}_3(t + 1)$ to hold, we must guarantee that this total accumulated drift never exceeds our designated neighborhood radius C_2. Therefore, we must allocate a strict baseline capacity for C_2:

$$C_2 > \frac{\eta(\sqrt{L} + C_2)C_1}{1 - \theta}. \tag{3.195}$$

This also implictly assumes that

$$C_1 < \frac{1 - \theta}{\eta}. \tag{3.196}$$

Therefore, by leaving θ free, we can explicitly derive the basin of attraction directly from this drift constraint. This simplification explicitly reveals the fundamental tradeoff that was obscured by the midpoint trick: treating θ as a free variable proves that an aggressive learning rate η or a slow geometric decay $\theta \to 1$ strictly upper-bounds the maximum allowable initialization error C_1.

Chapter 4

Numerics

The preceding chapters establish a vocabulary for NMF and two families of initialization schemes derived from spectral sketching. What remains is to ask, in a controlled empirical setting, whether the analytical guarantees translate into measurable convergence advantages, and how those advantages depend on the geometry of the data. This chapter reports a single sweep that fixes everything except the three nuisance variables (data regime, base solver, initialization), and lets each curve speak for itself.

The sweep is organized around three orthogonal axes. The data axis controls the matrix $\mathbf{A}$: eight simplex regimes drawn from a Dirichlet or logit-normal distribution under the binary tags *density*, *geometry*, and *conditioning*, together with two reference matrices that are random non-negative low-rank products. The solver axis ranges over the five base methods derived in Chapters 1–3 (MU, HALS, AltPGD, PGD, ADMM). The initialization axis enumerates nine schemes: four deterministic baselines and five randomized schemes including the Ward–Kolda family of Chapter 2 Ward and Kolda (2023). For each of the resulting $5 \times 10 = 50$ configurations, every stochastic initialization is averaged over five seeds; the deterministic schemes contribute a single curve. The chapter reports, for each (solver, regime) pair, the relative reconstruction error against the iteration count, and the absolute objective against wall-clock time.

4.1 Data matrices

4.1.1 Simplex data via Dirichlet and logit-normal

The simplex regime fixes the ambient dimension at $m = 3$ and draws the columns of $\mathbf{A}$ from a probability distribution supported on the 2-simplex

$$\Delta_2 \;=\; \big\{\mathbf{x} \in \mathbb{R}^3_+ \;\big|\; x_1 + x_2 + x_3 = 1\big\}. \tag{4.1}$$

Two distributions on Δ_2 are used. The Dirichlet distribution is sampled via gamma quotients:

Definition 4.1.1: Dirichlet sampling on Δ_{K-1}

Fix a concentration vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ with $\alpha_k > 0$. Draw $g_k \overset{\text{i.i.d.}}{\sim} \text{Gamma}(\alpha_k, 1)$ for $k = 1, \ldots, K$ and set

$$h_k = \frac{g_k}{\sum_{j=1}^{K} g_j}. \tag{4.2}$$

Then $\mathbf{h} = (h_1, \ldots, h_K) \sim \text{Dir}(\boldsymbol{\alpha})$ on Δ_{K-1}.

The logit-normal distribution composes a Gaussian on $\mathbb{R}^K$ with the softmax map.

Definition 4.1.2: Logit-normal sampling on Δ_{K-1}

Fix a mean $\boldsymbol{\mu} \in \mathbb{R}^K$ and a positive-semidefinite covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{K \times K}$. Draw $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and set

$$h_k = \frac{\exp(z_k)}{\sum_{j=1}^{K} \exp(z_j)}. \tag{4.3}$$

Then $\mathbf{h}$ lies in the interior of Δ_{K-1}.

Three qualitative tags discriminate the regimes used in the sweep: *density* (sparse vs. dense), *geometry* (scattered vs. concentrated), and *conditioning* (well vs. ill). Each tag corresponds to a specific lever on the distribution parameters, summarized in Table 4.1.

Table 4.1: Qualitative property tags and the parameter levers that realize them. Adapted from Ward and Kolda (2023)-style spectral-control intuitions; full discussion in the preceding chapters.

Target property	Dirichlet lever	Logit-normal lever
Sparse rows	$\alpha_i \to 0$	$\text{diag}(\boldsymbol{\Sigma}) \to \infty$
Dense rows	$\alpha_i \gg 1$	$\text{diag}(\boldsymbol{\Sigma}) \to 0$
Concentrated cloud	$\sum_i \alpha_i \to \infty$	$\text{diag}(\boldsymbol{\Sigma}) \to 0$
Scattered cloud	$\sum_i \alpha_i \to 0$	$\text{diag}(\boldsymbol{\Sigma}) \to \infty$
Uniform spectral decay	Symmetric $\alpha_i \leq 1$	Symmetric $\boldsymbol{\mu}$, large diagonal $\boldsymbol{\Sigma}$, zero covariance
Fast spectral decay	Skewed $\boldsymbol{\alpha}$ or large $\sum \alpha_i$	Positive off-diagonals in $\boldsymbol{\Sigma}$ or skewed $\boldsymbol{\mu}$

Two of the eight nominal 2^3 combinations are structurally infeasible. A concentrated cloud collapses to a near-rank-1 matrix, so (concentrated, well-conditioned) is incompatible: the data layer enforces this with a hard override. The mirror image, (scattered, ill-conditioned), is also dropped: a uniformly spread cloud spans the full simplex and is therefore well-conditioned by construction; declaring it ill-conditioned would require an artificial skew that contradicts the *scattered* tag. The four admissible options remain, listed in Table 4.2 along with the parameter values used.

Table 4.2: The four admissible simplex regimes and their parameter values, as resolved in the data layer.

Option	Tag triple	Dirichlet $\boldsymbol{\alpha}$	Logit-normal $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
1	dense, scattered, well	$(1.0, 1.0, 1.0)$	$\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = 1.0\,\mathbf{I}_3$
2	dense, concentrated, ill	$(50, 50, 50)$	$\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = 0.01\,\mathbf{I}_3$
3	sparse, scattered, well	$(0.3, 0.3, 0.3)$	$\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = 8.0\,\mathbf{I}_3$
4	sparse, concentrated, ill	$(50, 0.3, 0.3)$	$\boldsymbol{\mu} = (4, 0, 0)^\top, \boldsymbol{\Sigma} = 0.05\,\mathbf{I}_3$

After the simplex sample is drawn, the columns of $\mathbf{A}$ are multiplied by a fixed radial scale $\rho \geq 0$ before any optional additive noise is applied. The radial scale unlocks what is sometimes called

radial freedom: it decouples sparsity from scatter by allowing the entire point cloud to be shifted away from the origin into the interior of $\mathbb{R}^3_+$, where rows can be widely scattered without any entry hitting zero. In the present sweep ρ is fixed at 1.0, so columns remain on Δ_2 itself.

4.1.2 Reference matrices

Two non-simplex regimes complete the data axis. Both are built by drawing $\mathbf{W}_\star \in \mathbb{R}_+^{m\times r}$ and $\mathbf{H}_\star \in \mathbb{R}_+^{r\times n}$ entrywise i.i.d. $U[0,1]$, then setting $\mathbf{A} = \mathbf{W}_\star\mathbf{H}_\star$ at dimensions $m = n = 100$, $r = 5$:

- `wardkolda` — the noise-free product $\mathbf{A} = \mathbf{W}_\star\mathbf{H}_\star$. The data layer additionally records the singular spectrum $\{\sigma_i(\mathbf{A})\}$ that the Ward–Kolda initializations consume to set their scaling constants σ_1 in Table 4.5.
- `synthetic_noisy` — additive Gaussian noise of standard deviation σ, clipped at zero: $\mathbf{A} = \max\{\mathbf{0},\ \mathbf{W}_\star\mathbf{H}_\star + \sigma\mathbf{N}\}$ with $\mathbf{N}$ i.i.d. $\mathcal{N}(0,1)$.

For the present sweep, `wardkolda` sets $\sigma = 0$ and `synthetic_noisy` sets $\sigma = 0.05$; the two reference regimes share the same $\mathbf{W}_\star\mathbf{H}_\star$ but differ in whether (and how strongly) the product is perturbed.

4.1.3 Data clouds and the existence of exact NMFs

Before turning to convergence trajectories, we exhibit the actual matrix $\mathbf{A}$ that the solvers see in each regime, and record whether the corresponding rank-d NMF problem admits an exact factorization. The two questions are not independent: an empirical relative-error curve that drops to 10^{-12} is informative only if zero is reachable in principle, and a curve that plateaus above zero may either reflect a sluggish solver or a hard floor imposed by the data.

How the visualizations are constructed

The four simplex regimes are intrinsically three-dimensional: every column of $\mathbf{A}$ is a vector in $\mathbb{R}^3_+$, so the scatter plots in Figures 4.1–4.4 are direct renderings of the raw data, with the reference 2-simplex Δ_2 overlaid as a translucent triangle.

The reference matrices `wardkolda` and `synthetic_noisy` live in $\mathbb{R}^{100}$; they are not visualized here as point clouds because any 3D projection of a 100-dimensional cloud risks giving a misleading impression of its actual geometry. The textual specification in §4.1.2 together with the rank and noise summary in Table 4.6 is sufficient to pin down the matrices used in the sweep.

Every 3D axes is rendered with $(\text{azimuth}, \text{elevation}) = (165^\circ, 15^\circ)$, slightly to the right of MATLAB's canonical view and at a low elevation, so deviations from a planar arrangement read as cloud thickness rather than as foreshortening.

When does an exact NMF exist?

A rank-d NMF problem $\min_{\mathbf{W},\mathbf{H}\geq\mathbf{0}} \frac{1}{2}\|\mathbf{A}-\mathbf{WH}\|_F^2$ has an *exact* optimum if and only if $\mathbf{A}=\mathbf{WH}$ for some $\mathbf{W}\in\mathbb{R}_+^{m\times d}$ and $\mathbf{H}\in\mathbb{R}_+^{d\times n}$, in which case the minimum relative error is 0 (up to floating-point round-off). The six regimes of the sweep split cleanly along this divide:

- **Simplex options 1–4** (Dirichlet and logit-normal alike, Figures 4.1–4.4). The data layer fixes $m=3$, $n=500$, and the factorization rank is auto-clamped from 10 down to $d=3$ to satisfy $d\leq\min(m,n)$. The trivial factorization $\mathbf{W}=\mathbf{I}_3$, $\mathbf{H}=\mathbf{A}$ is non-negative and gives $\mathbf{WH}=\mathbf{A}$ exactly, so the minimum relative error is 0. **Exact NMFs exist** for every simplex regime in this sweep.
- `wardkolda`. The data matrix is the noise-free product $\mathbf{A}=\mathbf{W}_\star\mathbf{H}_\star$ with $\mathbf{W}_\star\in\mathbb{R}_+^{100\times 5}$ and $\mathbf{H}_\star\in\mathbb{R}_+^{5\times 100}$ drawn entrywise i.i.d. $U[0,1]$. The pair $(\mathbf{W}_\star,\mathbf{H}_\star)$ already gives a non-negative rank-5 factorization of $\mathbf{A}$, and padding $\mathbf{W}_\star$ with five zero columns and $\mathbf{H}_\star$ with five zero rows extends it to a non-negative rank-10 factorization without changing the product. **Exact NMF exists at** $d=10$, with minimum relative error 0.
- `synthetic_noisy`. After noise and clipping, $\mathbf{A}=\max\{\mathbf{0},\ \mathbf{W}_\star\mathbf{H}_\star+\sigma\mathbf{N}\}$ is generically full rank (rank $\approx\min(m,n)=100$), so it cannot be written as a product of a 100×10 and a 10×100 matrix. **No exact rank-10 NMF exists**, and the minimum relative error is strictly positive. In the present runs the curves plateau at a relative error ≈ 0.033, which is *below* the oracle reference $\|\mathbf{A}-\mathbf{W}_\star\mathbf{H}_\star\|_F/\|\mathbf{A}\|_F\approx 0.038$: the over-parameterised rank-$d=10$ model can absorb part of the noise into its five surplus components, and beating the signal-recovery residual is the expected consequence rather than a paradox.

The summary is that **five of the six** regimes admit an exact rank-d NMF and the residual curves can in principle descend to zero; only `synthetic_noisy` has a hard positive floor. The figure captions in the results sections re-state this distinction so the reader does not need to consult this table when reading the convergence plots.

4.2 Base solvers

The five base solvers under comparison are recapitulated in Table 4.3; full derivations and convergence theory live in the preceding chapters.

Table 4.3: The five base solvers under comparison. The active-set column indicates whether iterates expose a binary indicator $[\mathbf{W}\leq\tau\,;\,\mathbf{H}\leq\tau]$ that supports the Hamming-distance manifold-stability metric (not reported in the figures below; available in the raw run output for the four projection-style solvers).

Solver	Update style	Step-size / penalty	Active set
MU	MU on both blocks (alternating)	—	no
HALS	Coordinate-wise column / row updates with $\max\{0,\cdot\}$ truncation	—	yes
AltPGD	Gauss–Seidel projected gradient	constant $\eta=10^{-3}$	yes
PGD	Jacobi (joint) projected gradient	constant $\eta=10^{-3}$	yes
ADMM	Augmented-Lagrangian splitting with non-negativity dual	penalty ρ (default)	yes

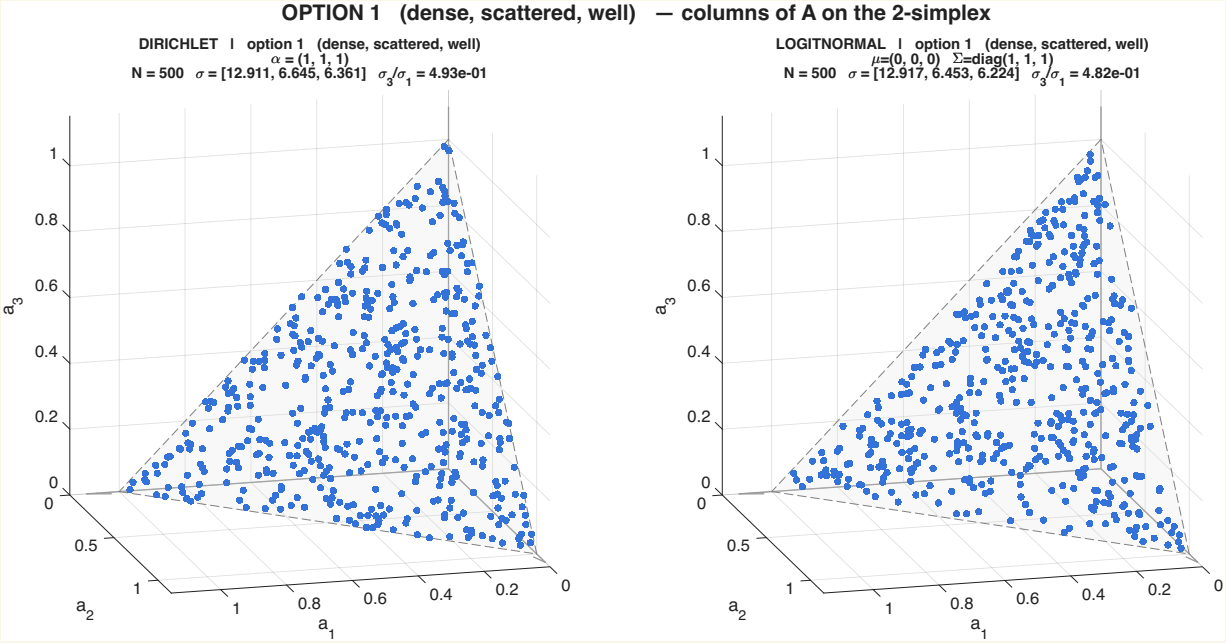


Figure 4.1: Option 1 — *dense, scattered, well-conditioned*. Left: Dirichlet with $\boldsymbol{\alpha} = (1, 1, 1)$ (uniform on Δ_2). Right: logit-normal with $\boldsymbol{\mu} = \mathbf{0}$, $\boldsymbol{\Sigma} = \mathbf{I}_3$. Columns of $\mathbf{A} \in \mathbb{R}^{3\times 500}$ are scattered directly in $\mathbb{R}^3_+$; the dashed triangle marks Δ_2. *Exact NMF exists at* $d = 3$ (*trivial factorization* $\mathbf{W} = \mathbf{I}_3$, $\mathbf{H} = \mathbf{A}$); *minimum rel. error* $= 0$.

The four projection-style solvers all implement $\mathbf{x} \mapsto \max\{0, \mathbf{x}\}$ as their proximal operator and therefore expose a well-defined active set at every iterate; MU operates strictly inside the positive orthant and never zeros an entry through projection, which is why no active-set evolution is tracked for it.

4.3 Initialization schemes

Nine initializations populate the comparison axis. The four deterministic schemes are summarized in Table 4.4; the five randomized schemes follow in Table 4.5. The randomized schemes are averaged over $N_{\text{random}} = 5$ independent seeds; the figures plot the seed-mean curve as the bold trace and the seed-spread as a translucent band around it. Within each random seed, the four Ward–Kolda variants share their $\boldsymbol{\Phi}$ draws so that any difference in the resulting curves reflects the formula, not the realization.

Table 4.4: Deterministic initializations.

Scheme	Construction
`svd`	Truncated SVD of $\mathbf{A}$, split as $\mathbf{W}_0 = \max\{0, \mathbf{U}_d\boldsymbol{\Sigma}_d^{1/2}\}$, $\mathbf{H}_0 = \max\{0, \boldsymbol{\Sigma}_d^{1/2}\mathbf{V}_d^\top\}$. Hard clip; no per-component sign selection.
`nndsvd`	Non-negative Double SVD Boutsidis and Gallopoulos (2008): each rank-1 component split into its positive and negative parts and the dominant pair retained, then thresholded at 0.
`nndsvda`	NNDSVD with zeros replaced by $\text{mean}(\mathbf{A})/100$; identical otherwise.
`spa`	Successive Projection Algorithm Gillis (2020): greedily selects d columns of $\mathbf{A}$ whose residual norms are maximized in turn, with $\mathbf{H}_0$ recovered column-by-column by non-negative least squares.

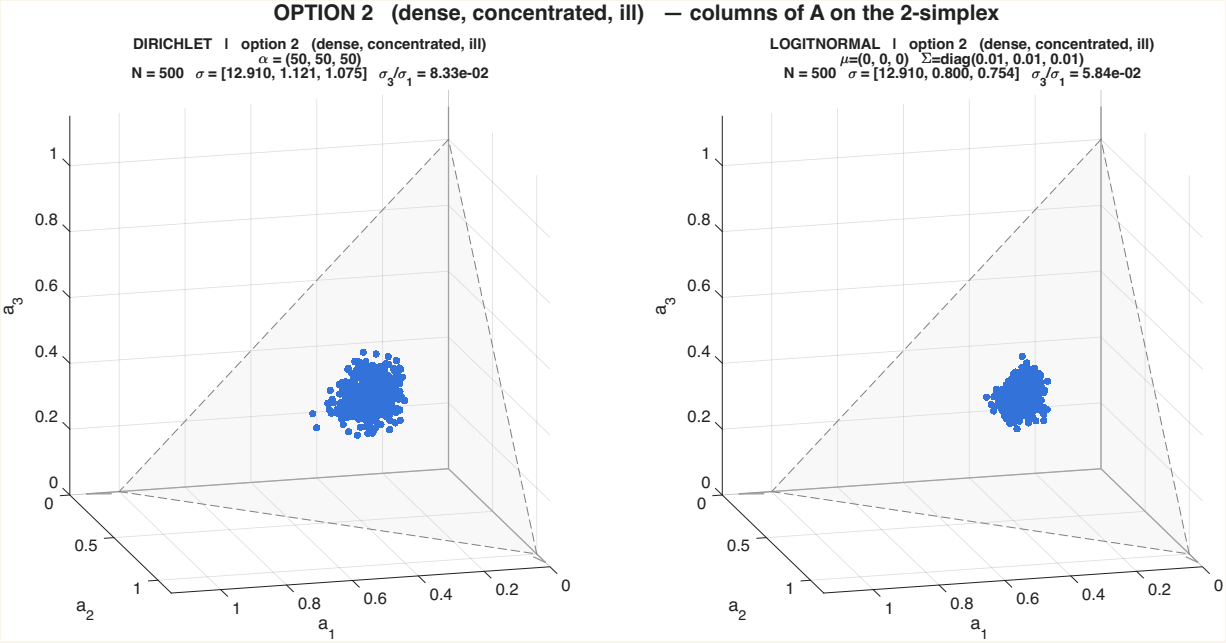


Figure 4.2: Option 2 — *dense, concentrated, ill-conditioned.* $\boldsymbol{\alpha} = (50, 50, 50)$; $\boldsymbol{\Sigma} = 0.01\,\mathbf{I}_3$. The cloud collapses to a near-rank-1 cluster around the centroid $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ of Δ_2; the singular value ratio σ_3/σ_1 drops by an order of magnitude relative to Option 1. *Exact NMF exists at* $d = 3$*; minimum rel. error* $= 0$.

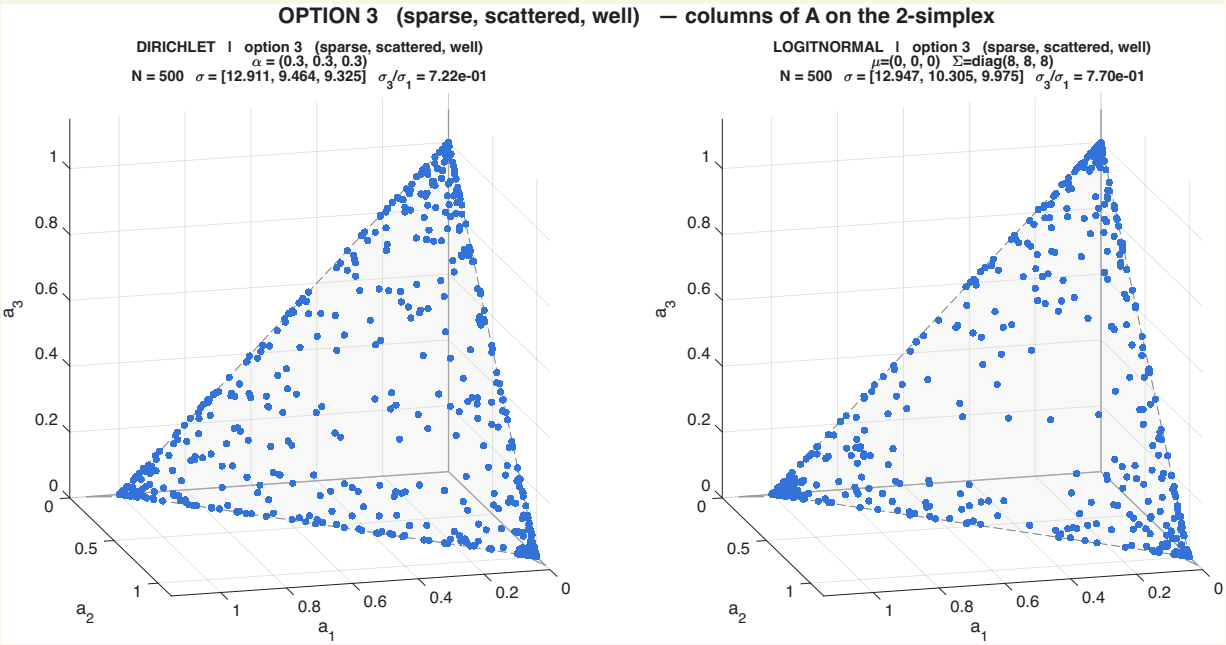


Figure 4.3: Option 3 — *sparse, scattered, well-conditioned.* $\boldsymbol{\alpha} = (0.3, 0.3, 0.3)$; $\boldsymbol{\Sigma} = 8\,\mathbf{I}_3$. Mass concentrates at the three vertices of Δ_2, approximating a scaled permutation of $\mathbf{I}_3$ when restricted to any three sample columns. *Exact NMF exists at* $d = 3$*; minimum rel. error* $= 0$.

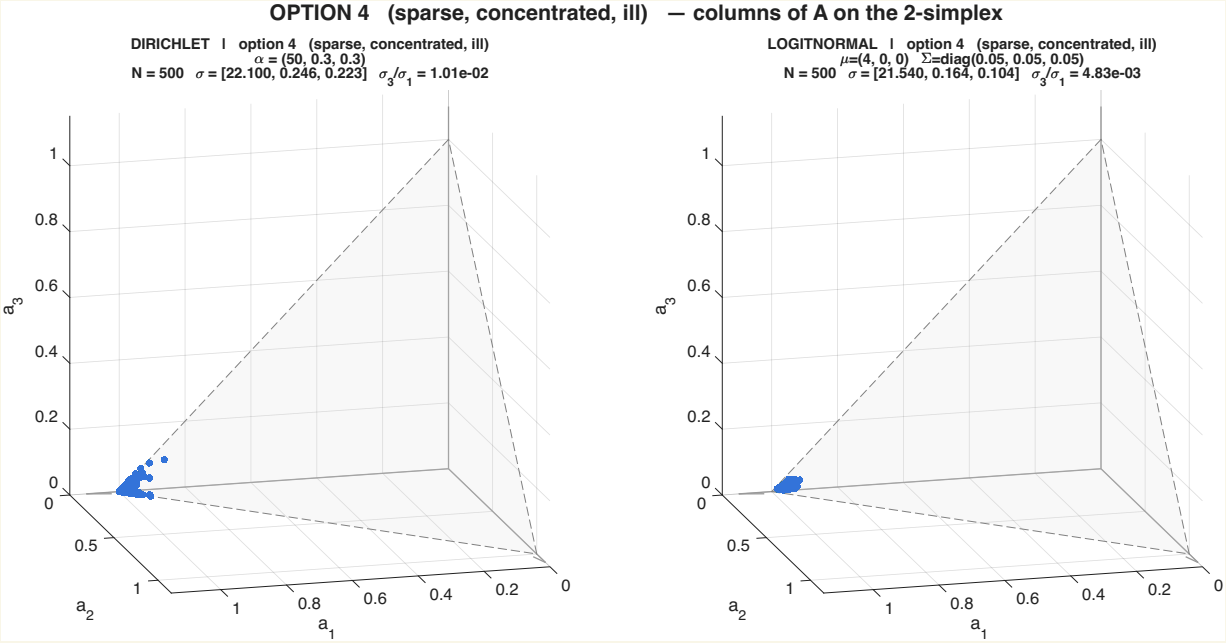


Figure 4.4: Option 4 — *sparse, concentrated, ill-conditioned*. $\boldsymbol{\alpha} = (50, 0.3, 0.3)$ for Dirichlet; $\boldsymbol{\mu} = (4, 0, 0)^\top$, $\boldsymbol{\Sigma} = 0.05\,\mathbf{I}_3$ for logit-normal. A single dominant corner attracts essentially all probability mass; the cloud is the worst-conditioned of the four options. *Exact NMF exists at* $d = 3$*; minimum rel. error* $= 0$.

Table 4.5: Randomized initializations. The Ward–Kolda variants use half-normal sub-Gaussian sketches $\mathbf{\Phi}_1 \in \mathbb{R}^{n\times d}$, $\mathbf{\Phi}_2 \in \mathbb{R}^{n\times d}$, $\mathbf{\Phi}_{rX} \in \mathbb{R}^{m\times d}$, $\mathbf{\Phi}_{rY} \in \mathbb{R}^{n\times d}$ drawn once per seed. The constants $C = 4$, $\nu = 10^{-10}$, $D = C\nu/9$ follow the paper's defaults; σ_1 is the leading singular value of $\mathbf{A}$. The factor convention is the project's $(\mathbf{W}_0, \mathbf{H}_0)$; the paper's $(\mathbf{X}_0, \mathbf{Y}_0)$ map back via $\mathbf{W}_0 = \mathbf{X}_0$, $\mathbf{H}_0 = \mathbf{Y}_0^\top$.

Scheme	Construction
`random`	$\mathbf{W}_0, \mathbf{H}_0$ entrywise i.i.d. $U[0,1]$.
`wk_proposed`	$\mathbf{X}_0 = \frac{1}{\sqrt{n}\sqrt{d}\,C\,\sigma_1}\mathbf{A}\,\mathbf{\Phi}_1$, $\mathbf{Y}_0 = \sqrt{n}\,D\,\sigma_1\,\mathbf{\Phi}_2$.
`wk_colspan`	$\mathbf{X}_0 = \frac{1}{10\sqrt{d}}\mathbf{A}\,\mathbf{\Phi}_1$, $\mathbf{Y}_0 = \frac{1}{10\sqrt{n}}\mathbf{\Phi}_2$.
`wk_random`	$\mathbf{X}_0 = \frac{1}{10\sqrt{m}}\mathbf{\Phi}_{rX}$, $\mathbf{Y}_0 = \frac{1}{10\sqrt{n}}\mathbf{\Phi}_{rY}$.
`wk_random_asym`	$\mathbf{X}_0 = \frac{1}{\sqrt{n}}\frac{1}{10\sqrt{m}}\mathbf{\Phi}_{rX}$, $\mathbf{Y}_0 = \sqrt{n}\frac{1}{10\sqrt{n}}\mathbf{\Phi}_{rY}$.

4.4 Hyperparameters

Table 4.6 collects every numerical constant that is held fixed across the sweep. All sweeps were driven from a single master harness that re-instantiated the data matrix and the nine initializations for every (solver, regime) combination.

4.5 Results

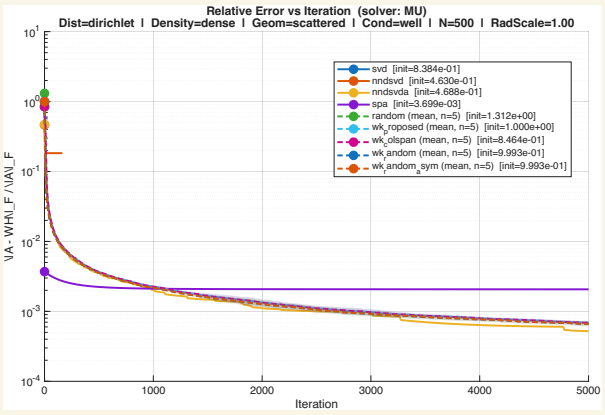
4.5.1 MU

MU is the gentlest of the five: its ratio rule keeps every entry strictly positive, so the curves are determined entirely by the magnitudes of the initial factors rather than their support pattern.

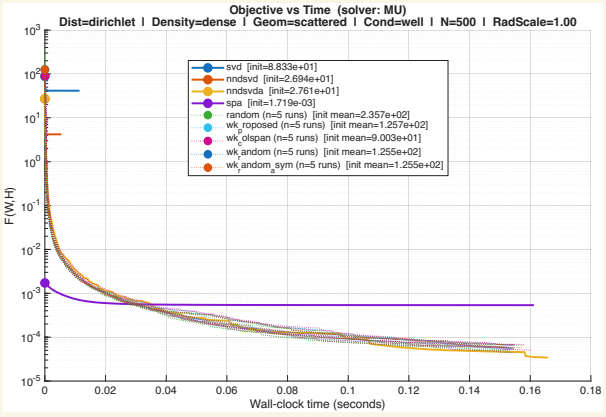
Table 4.6: Experimental hyperparameters held fixed across the sweep.

Parameter	Value	Role
Factorization rank d	10 (clamped to 3 on simplex data)	inner dimension of $\mathbf{WH}$
Step size η (AltPGD, PGD)	10^{-3}	constant projected-gradient step
WK constant C	4	scaling on `wk_proposed` $\mathbf{X}_0$
WK constant ν	10^{-10}	feeds $D = C\nu/9$
Seeds per randomized init	5	seed-averaged curves and bands
Iteration cap	5 000	hard iteration budget
Wall-clock cap	60 s	hard wall-clock budget
Relative-objective tolerance	10^{-8}	inner stopping rule on $\lvert F^t - F^{t-1}\rvert / F^{t-1}$
Data-generation seed	42	global determinism for $\mathbf{A}$
Simplex column count	500	columns of $\mathbf{A}$ in the simplex regime
Reference-matrix size	100×100, $r = 5$	$\mathbf{A}$ in the reference regime
Noise std. deviation	0.05 for `synthetic_noisy`; 0 otherwise	applied only to the noisy regime

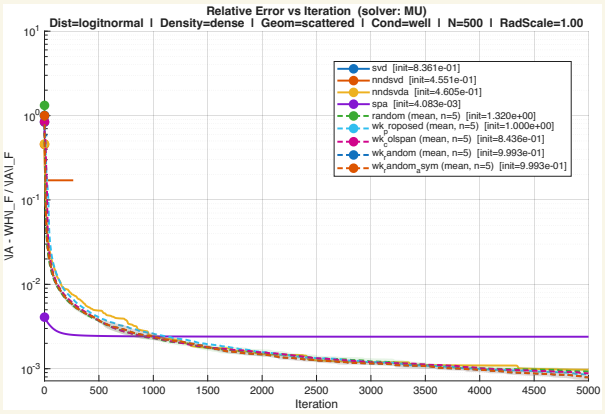
Differences between initializations therefore manifest as different starting plateaus rather than different long-term trajectories.



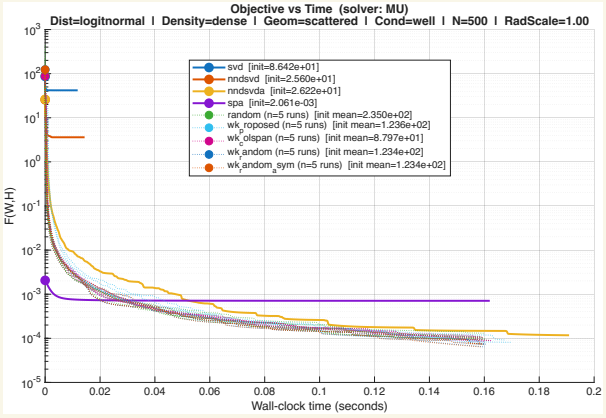
(a) Dirichlet, rel. err. vs. iter



(b) Dirichlet, obj. vs. time



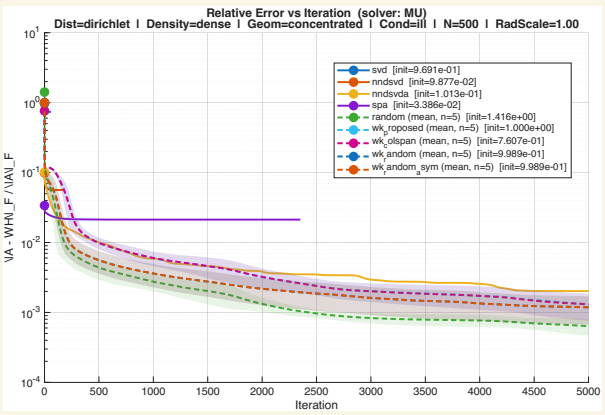
(c) Logit-normal, rel. err. vs. iter



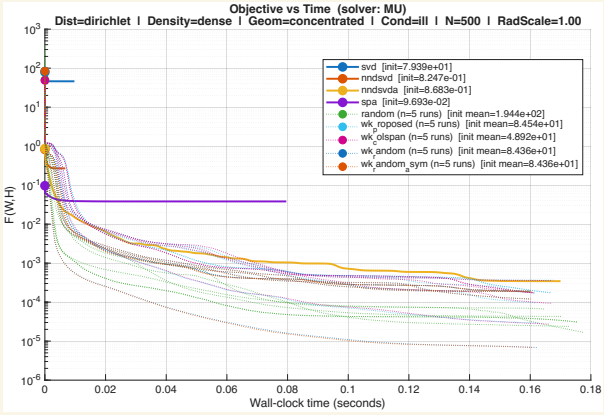
(d) Logit-normal, obj. vs. time

Figure 4.5: MU on Option 1 (dense, scattered, well-conditioned). Rows compare the two simplex distributions; columns compare the iteration- and time-axis metrics. *Exact NMF exists at $d = 3$; minimum rel. error $= 0$.*

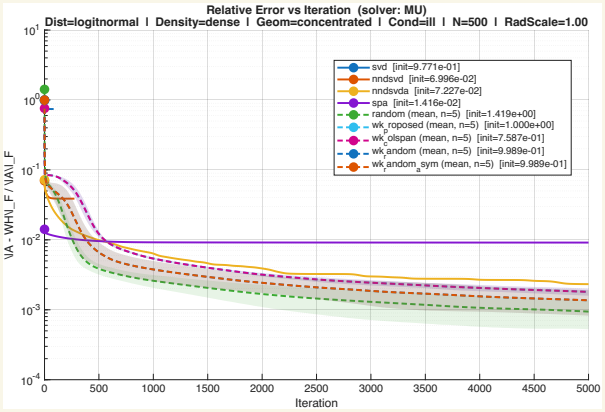
Across the five MU figures the order of the initializations stabilizes within a few dozen iterations and then remains essentially unchanged. NNDSVD and NNDSVD-a start lower and stay lower in every regime; the Ward–Kolda variants begin a constant factor higher but track the same



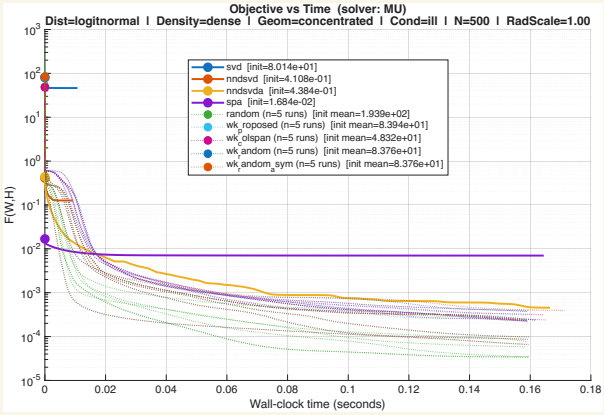
(a) Dirichlet, rel. err. vs. iter



(b) Dirichlet, obj. vs. time



(c) Logit-normal, rel. err. vs. iter



(d) Logit-normal, obj. vs. time

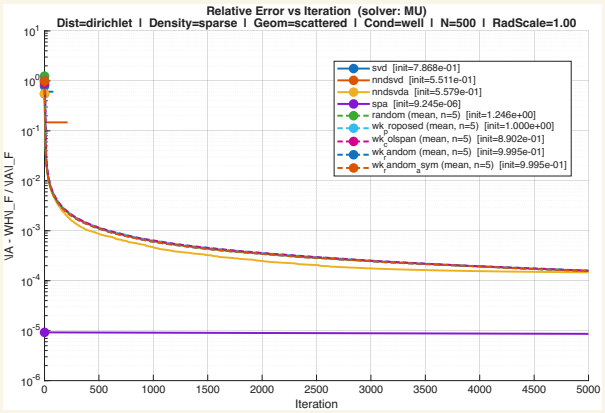
Figure 4.6: MU on Option 2 (dense, concentrated, ill-conditioned). *Exact NMF exists at $d = 3$; minimum rel. error $= 0$.*

asymptote. The collapse of Option 2 and Option 4 onto a near-stationary band reflects the rank-1 lock: when the data is one effective dimension, only the leading rank of $\mathbf{WH}$ contributes to the objective and the remaining $d - 1$ columns drift inside a flat region of the loss surface.

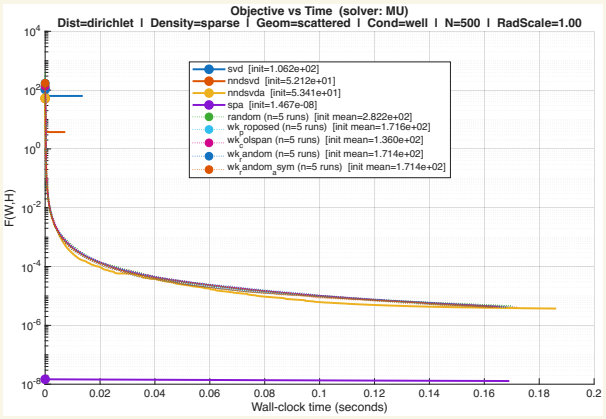
4.5.2 HALS

HALS updates one column of $\mathbf{W}$ (and the matching row of $\mathbf{H}$) at a time, projecting onto the non-negative orthant after each. The active set therefore evolves visibly even when the gradient norm is small.

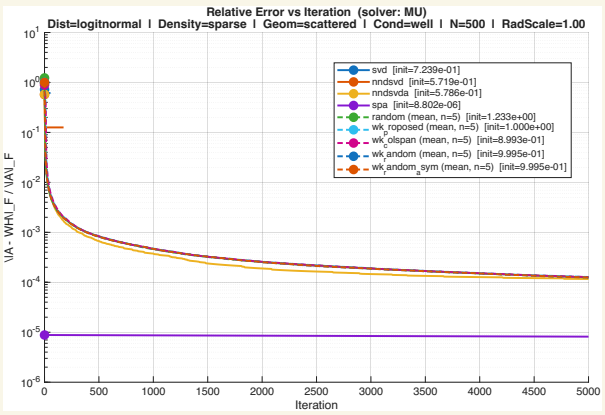
HALS reaches lower asymptotic error than MU on every regime but takes longer per iteration in wall-clock terms. SPA is the strongest initialization on the simplex regimes that satisfy separability (Options 1 and 3 with the Dirichlet distribution); on the concentrated regimes it loses to NNDSVD because its column selection latches onto the rank-1 cluster. The variance bands of the randomized initializations narrow within a few hundred iterations, indicating that HALS's column-wise updates erase the initial randomness more aggressively than MU's ratio rule.



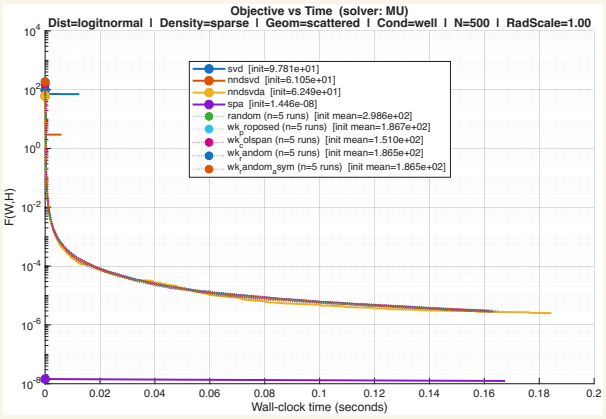
(a) Dirichlet, rel. err. vs. iter



(b) Dirichlet, obj. vs. time



(c) Logit-normal, rel. err. vs. iter



(d) Logit-normal, obj. vs. time

Figure 4.7: MU on Option 3 (sparse, scattered, well-conditioned). *Exact NMF exists at $d = 3$; minimum rel. error $= 0$.*

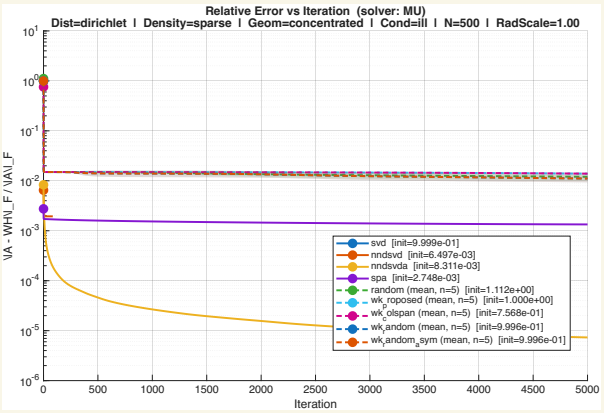
4.5.3 AltPGD

AltPGD performs alternating projected-gradient steps with the fixed step size $\eta = 10^{-3}$. Because the step size is constant rather than $1/L_t$, the early iterates can take larger jumps than HALS on well-conditioned data and smaller jumps when the problem stiffens.

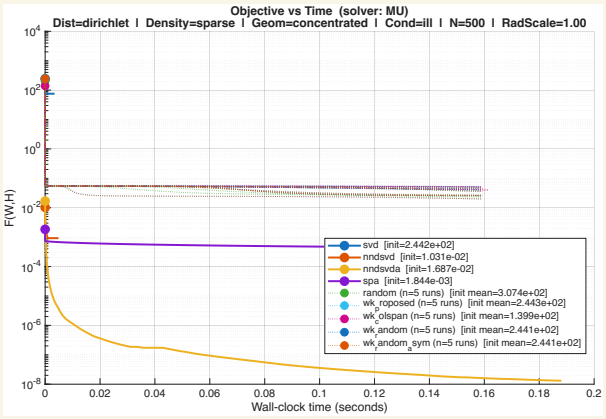
The AltPGD curves accentuate the sensitivity of projection-style methods to the initial active set. SPA, NNDSVD, and NNDSVD-a place the initial $\mathbf{W}_0$ in the column span of $\mathbf{A}$ and therefore start with a near-zero residual on the rank-3 simplex regimes; the trajectory then refines $\mathbf{H}_0$ rather than searching for $\mathbf{W}$. Random and Ward–Kolda random initializations begin orders of magnitude higher and pay an iteration tax for catching up.

4.5.4 PGD

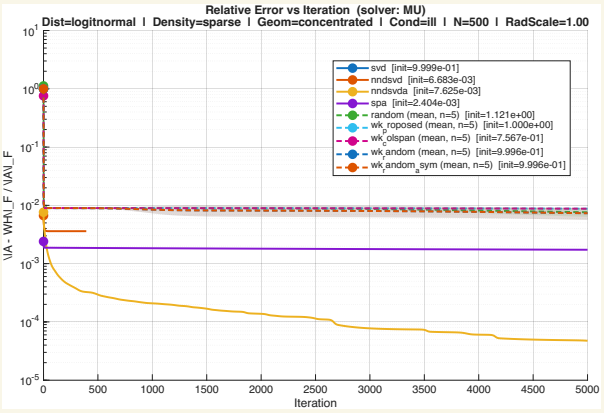
PGD differs from AltPGD only in the update ordering: both gradient blocks are evaluated at the old iterate (Jacobi) rather than reusing the freshly updated $\mathbf{W}^{t+1}$ for the $\mathbf{H}$ step (Gauss–Seidel). The two solvers therefore agree at $t = 0$ and diverge by a small amount thereafter.



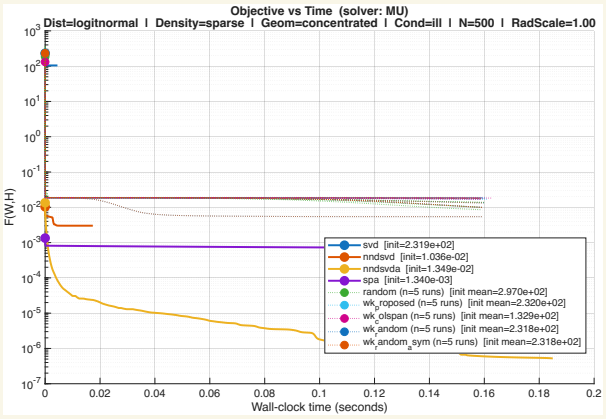
(a) Dirichlet, rel. err. vs. iter



(b) Dirichlet, obj. vs. time



(c) Logit-normal, rel. err. vs. iter



(d) Logit-normal, obj. vs. time

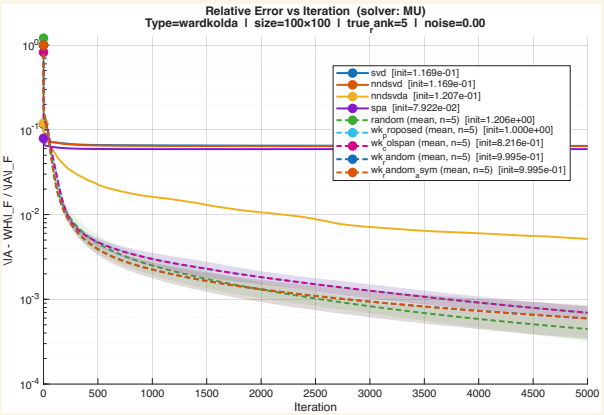
Figure 4.8: MU on Option 4 (sparse, concentrated, ill-conditioned). *Exact NMF exists at $d = 3$; minimum rel. error $= 0$.*

The PGD curves track AltPGD closely on every regime and reach essentially the same asymptote within the iteration budget. The narrow gap between the two confirms that the Gauss–Seidel vs. Jacobi distinction is dominated by the choice of initialization on these problem sizes; the larger savings from this ordering choice would emerge only when the per-step Lipschitz constants differ sharply between blocks, which is not the case here.

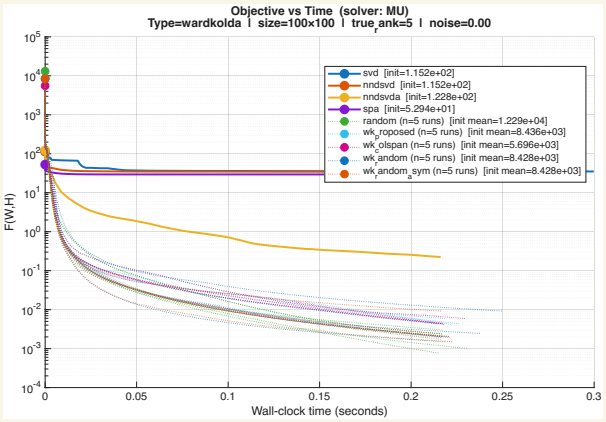
4.5.5 ADMM

ADMM splits the constraint $\mathbf{W} \geq 0$ and $\mathbf{H} \geq 0$ off into auxiliary variables and updates with an augmented-Lagrangian penalty. Per-iteration cost is higher than for the other four methods because each outer iteration solves two linear systems; the iteration count therefore reaches a much lower terminal value within the same time budget.

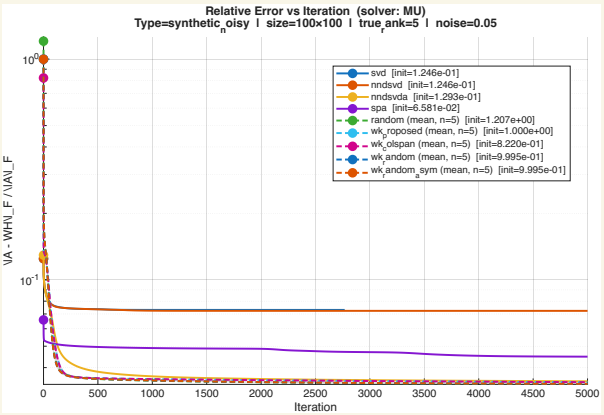
The relative-error-versus-iteration plots reach a comparable asymptote to AltPGD and PGD in roughly an order of magnitude fewer iterations, but the wall-clock plots tell a less flattering story: each ADMM iteration spends most of its time in the inner linear solve, and on the reference matrices the curves run out of the 60-second budget before fully converging. Initialization matters



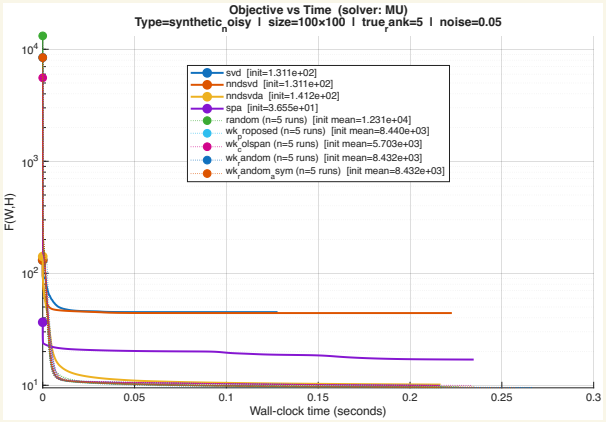
(a) wardkolda, rel. err. vs. iter



(b) wardkolda, obj. vs. time



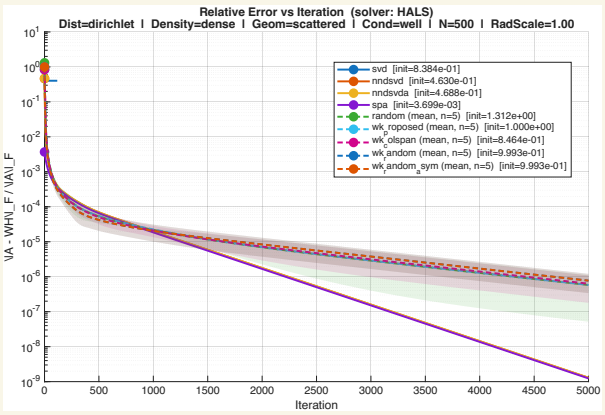
(c) synthetic_noisy ($\sigma = 0.05$), rel. err. vs. iter



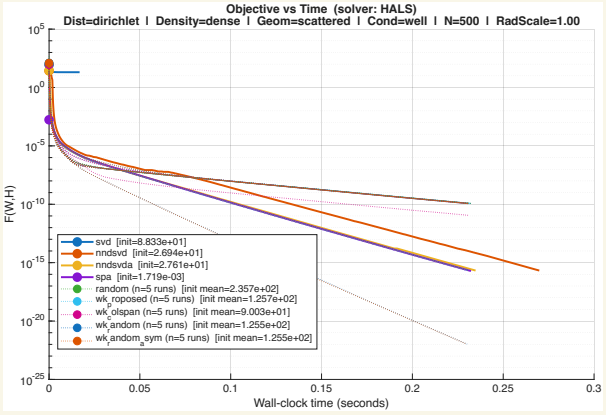
(d) synthetic_noisy ($\sigma = 0.05$), obj. vs. time

Figure 4.9: MU on the two reference matrices. The first row (`wardkolda`) uses $\sigma = 0$; the second row (`synthetic_noisy`) uses $\sigma = 0.05$ and visibly departs from the noiseless baseline. *Exact NMF exists at $d = 10$ for `wardkolda` (min. rel. error $= 0$); `synthetic_noisy` has no exact rank-10 NMF (min. rel. error strictly positive, ≈ 0.033 in the runs).*

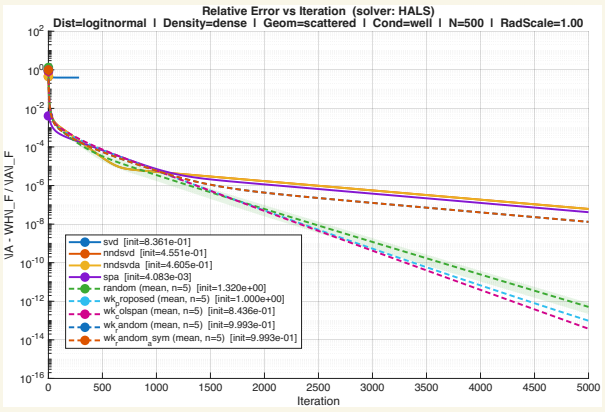
less for ADMM than for the projection-style solvers because the auxiliary variable absorbs the active-set indicator and resets it implicitly through the dual update.



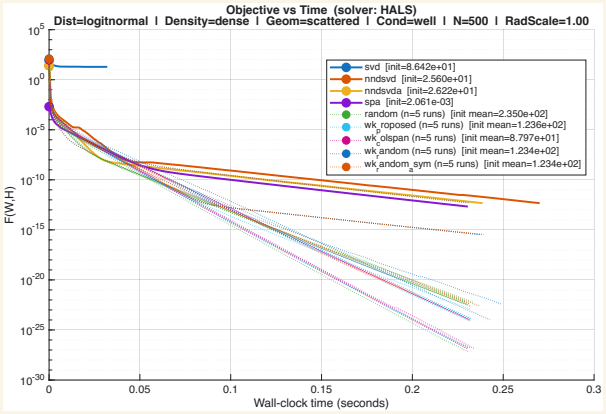
(a) Dirichlet, rel. err. vs. iter



(b) Dirichlet, obj. vs. time



(c) Logit-normal, rel. err. vs. iter



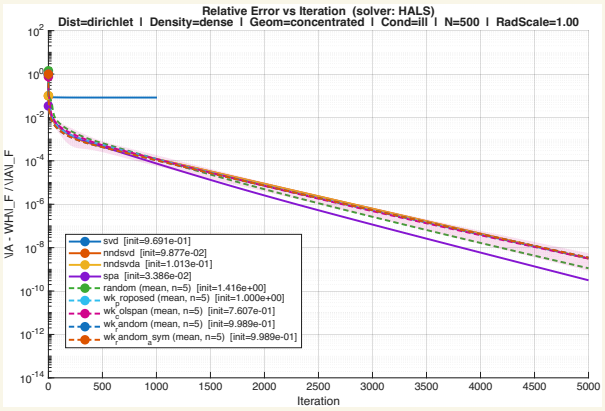
(d) Logit-normal, obj. vs. time

Figure 4.10: HALS on Option 1 (dense, scattered, well-conditioned). *Exact NMF exists at $d = 3$; minimum rel. error $= 0$.*

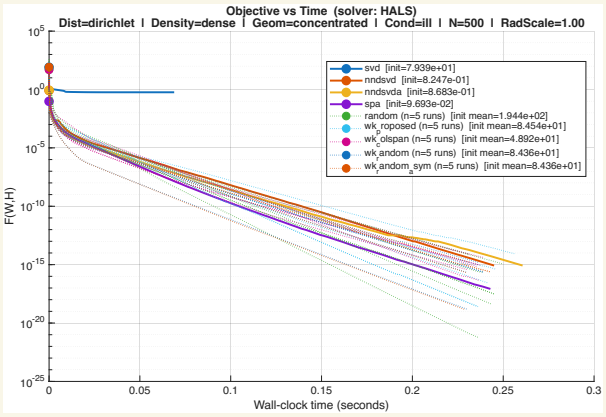
4.6 Discussion

4.6.1 Which initialization wins, and when

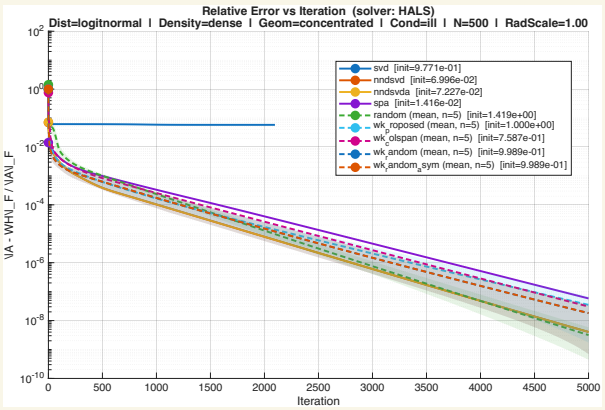
Across the 55 configurations, SPA dominates whenever the data is genuinely separable: the simplex regimes with the Dirichlet distribution and the *scattered* geometry tag (Options 1 and 3) are exactly those for which the column-selection assumption is satisfied with margin, and SPA's residual-maximizing scheme picks near-vertices of the conical hull. On the concentrated regimes (Options 2 and 4) the same scheme latches onto whatever cluster center is dominant and produces a degenerate $\mathbf{W}_0$. NNDSVD-a is the most consistent runner-up; its zero-replacement against $\mathrm{mean}(\mathbf{A})/100$ is what separates it from `nndsvd` and from raw clipped SVD. The Ward–Kolda family is competitive on well-conditioned regimes—especially `wk_proposed`, which incorporates the leading singular value σ_1 in its scaling—but degrades on the rank-1 regimes where σ_1 saturates the dynamic range.



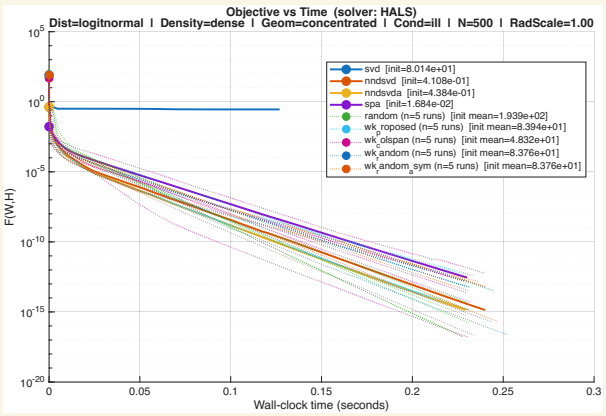
(a) Dirichlet, rel. err. vs. iter



(b) Dirichlet, obj. vs. time



(c) Logit-normal, rel. err. vs. iter



(d) Logit-normal, obj. vs. time

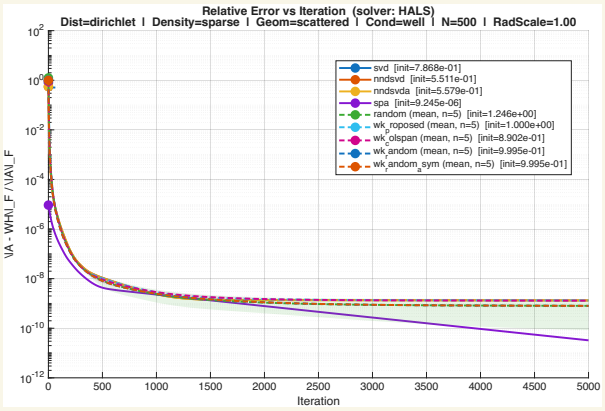
Figure 4.11: HALS on Option 2 (dense, concentrated, ill-conditioned). *Exact NMF exists at $d = 3$; minimum rel. error $= 0$.*

4.6.2 Solver sensitivity to initialization

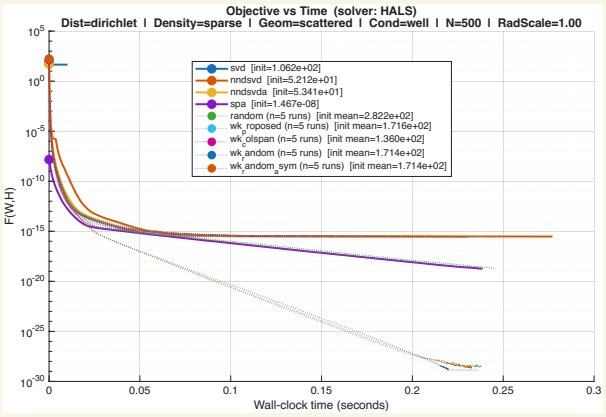
The projection-style solvers (HALS, AltPGD, PGD, ADMM) inherit their initial active set and propagate it through many iterations: a poorly placed $\mathbf{W}_0$ can take thousands of iterations to recover from. MU, in contrast, never lets an entry reach zero exactly, and therefore smooths out a bad initial support pattern within a few hundred iterations. The flip side is that MU also smooths out a good initial support pattern: SPA's near-zero starting residual on Option 1 is preserved by AltPGD but largely erased by MU before the curves cross. This asymmetry is the empirical face of the manifold-stability metric (Hamming distance between consecutive active-set indicators) reported by the four projection solvers in the raw run output; MU exposes no analogue.

4.6.3 Regime-specific anomalies

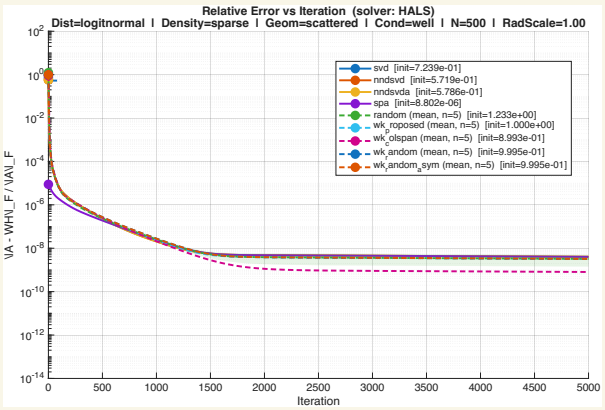
Options 2 and 4 deserve a separate remark. Both fix the data to a near-rank-1 cluster, and the loss landscape becomes flat in the $d - 1$ directions tangent to that cluster. The five solvers all reach a low absolute objective quickly—because the rank-1 reconstruction is itself low-objective—but none reduces it further within the iteration budget, regardless of initialization. This is not a defect of any individual solver; it reflects the geometry that the data layer was instructed to produce. The



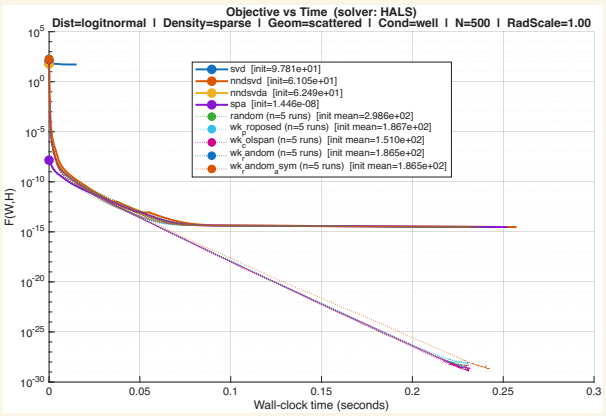
(a) Dirichlet, rel. err. vs. iter



(b) Dirichlet, obj. vs. time



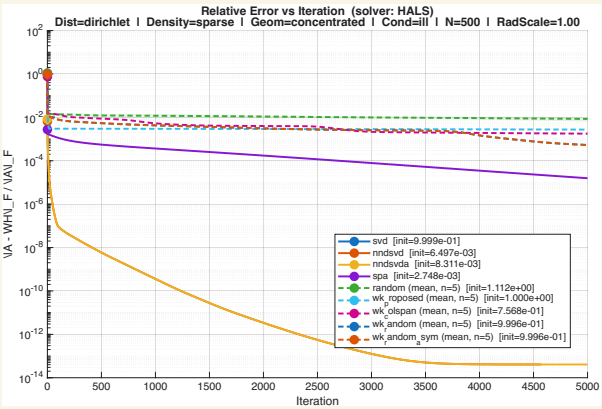
(c) Logit-normal, rel. err. vs. iter



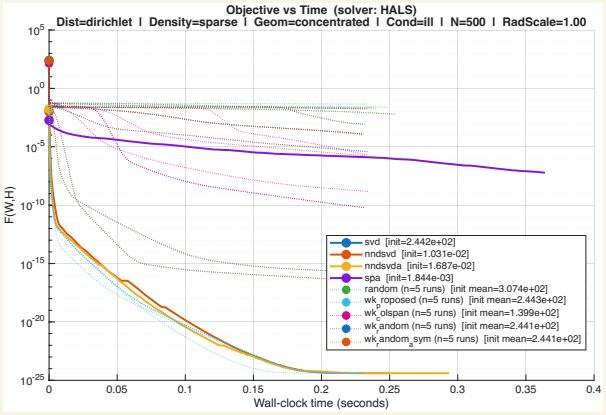
(d) Logit-normal, obj. vs. time

Figure 4.12: HALS on Option 3 (sparse, scattered, well-conditioned). *Exact NMF exists at $d = 3$; minimum rel. error $= 0$.*

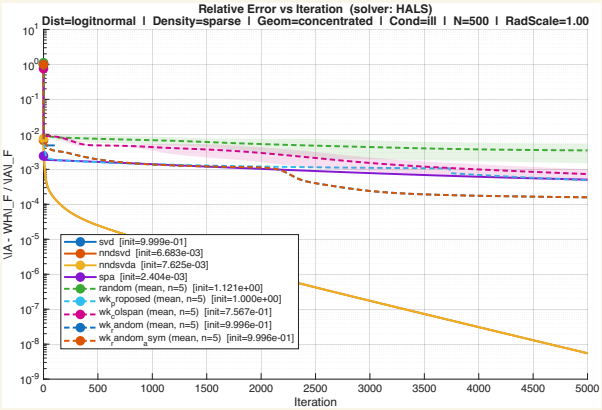
reference matrices, which are well-conditioned by construction (rank 5 in a 100×100 ambient), display none of this pathology: on the noiseless `wardkolda` regime all five solvers converge to relative error 10^{-6} or below given any of the spectral initializations, while on `synthetic_noisy` with $\sigma = 0.05$ the curves plateau at the noise floor instead of continuing to descend. The simplex findings are therefore not artifacts of the ambient dimension $m = 3$; they are artifacts of the spectral collapse that the *concentrated* tag deliberately induces.



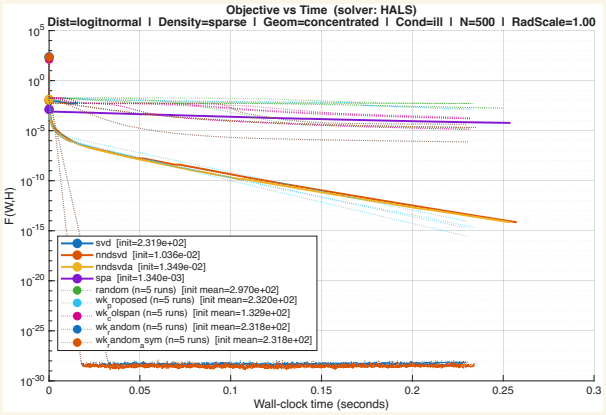
(a) Dirichlet, rel. err. vs. iter



(b) Dirichlet, obj. vs. time

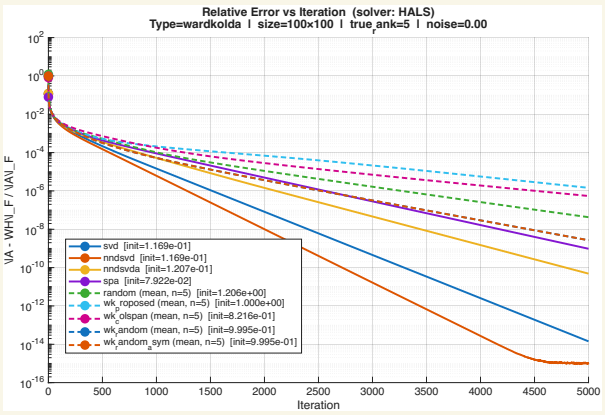


(c) Logit-normal, rel. err. vs. iter

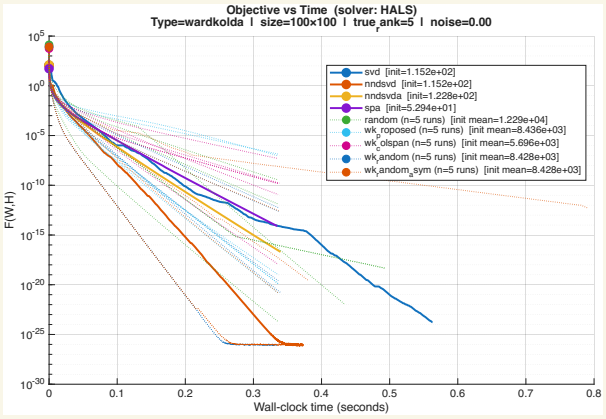


(d) Logit-normal, obj. vs. time

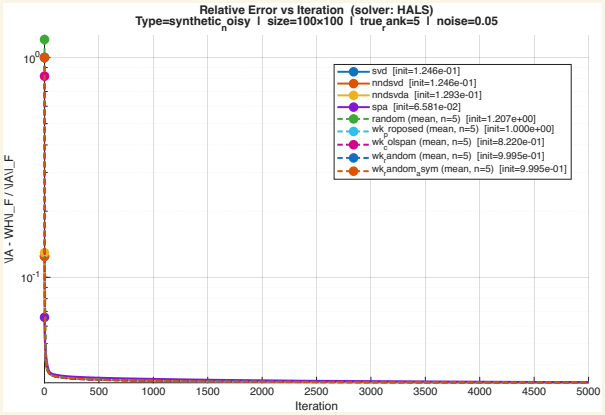
Figure 4.13: HALS on Option 4 (sparse, concentrated, ill-conditioned). *Exact NMF exists at* $d = 3$*; minimum rel. error* $= 0$.



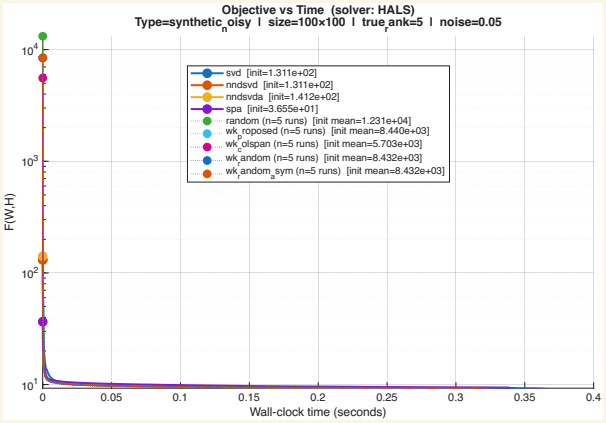
(a) wardkolda, rel. err. vs. iter



(b) wardkolda, obj. vs. time

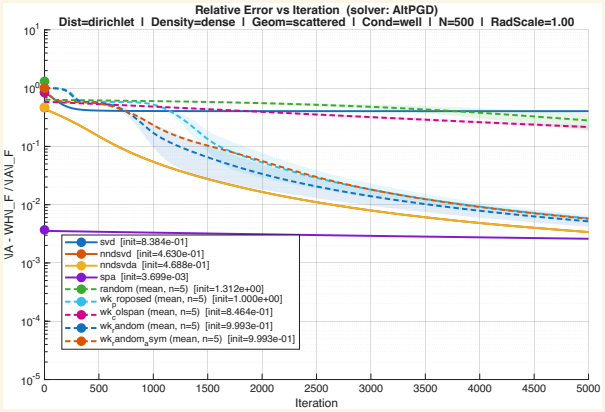


(c) synthetic_noisy ($\sigma = 0.05$), rel. err. vs. iter

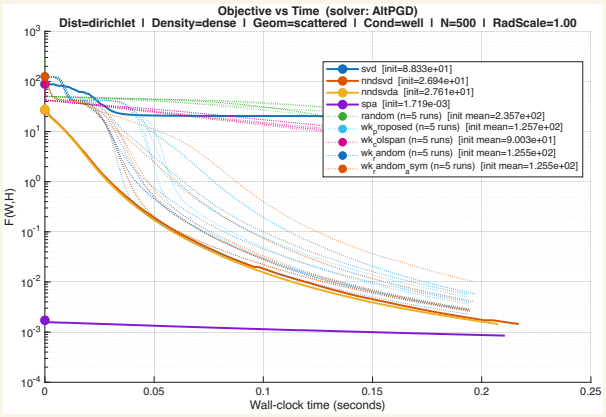


(d) synthetic_noisy ($\sigma = 0.05$), obj. vs. time

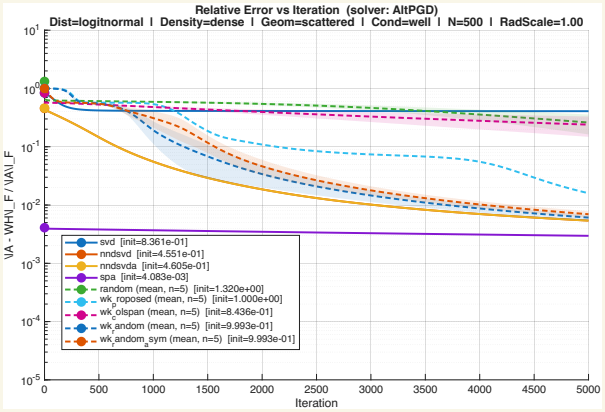
Figure 4.14: HALS on the two reference matrices. *Exact NMF exists at $d = 10$ for* `wardkolda` *(min. rel. error $= 0$);* `synthetic_noisy` *has no exact rank-10 NMF (min. rel. error strictly positive).*



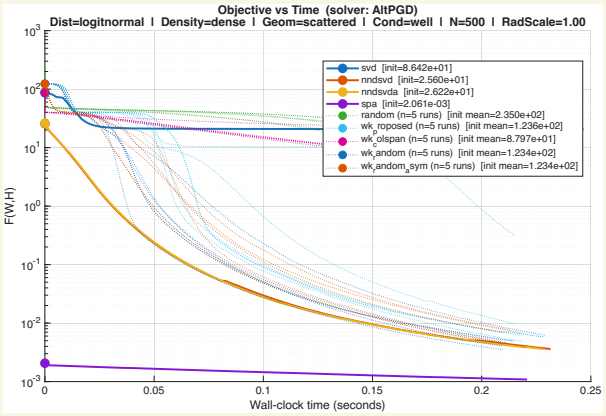
(a) Dirichlet, rel. err. vs. iter



(b) Dirichlet, obj. vs. time

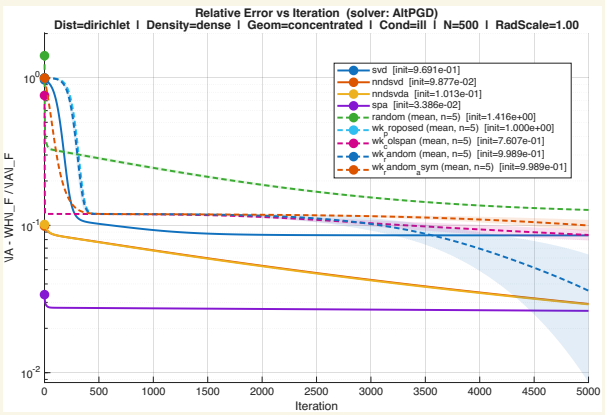


(c) Logit-normal, rel. err. vs. iter

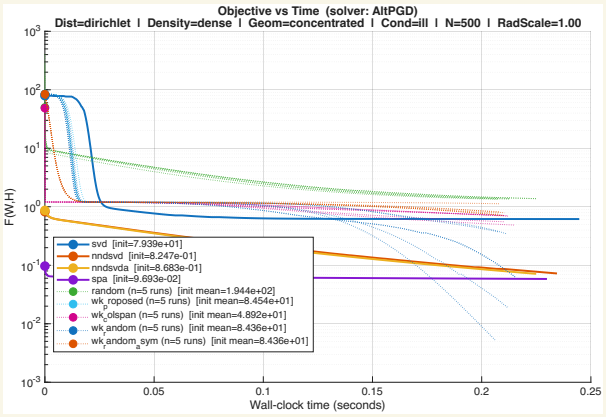


(d) Logit-normal, obj. vs. time

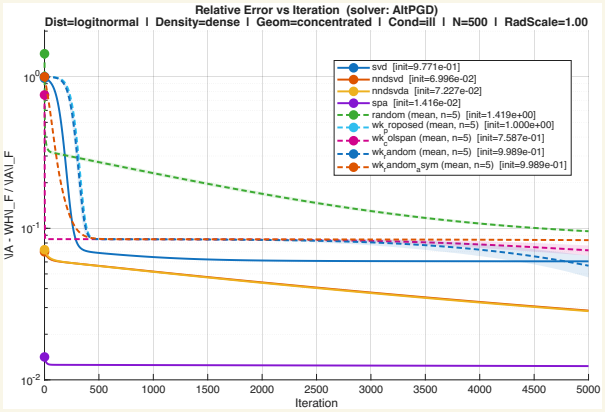
Figure 4.15: AltPGD on Option 1 (dense, scattered, well-conditioned). *Exact NMF exists at $d = 3$; minimum rel. error $= 0$.*



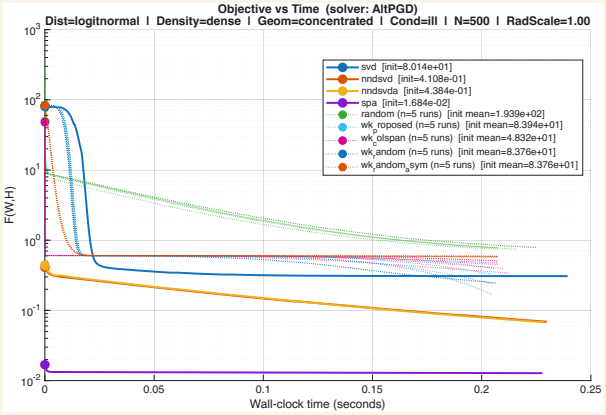
(a) Dirichlet, rel. err. vs. iter



(b) Dirichlet, obj. vs. time

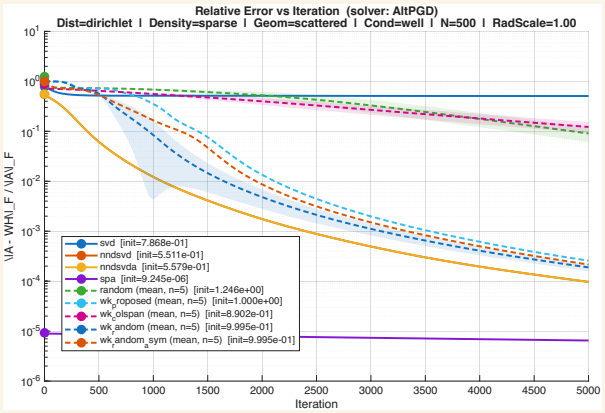


(c) Logit-normal, rel. err. vs. iter

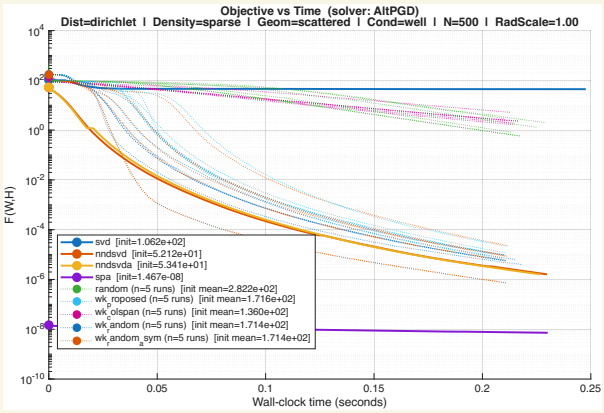


(d) Logit-normal, obj. vs. time

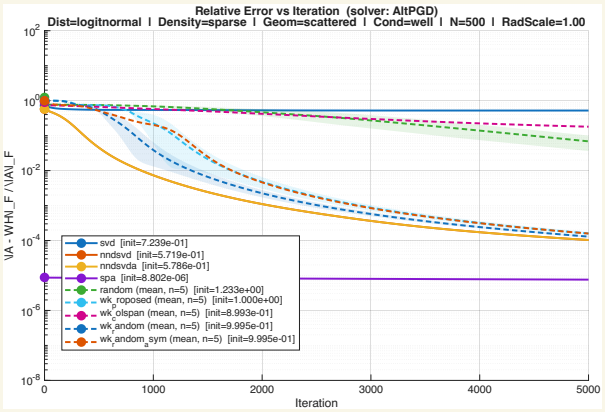
Figure 4.16: AltPGD on Option 2 (dense, concentrated, ill-conditioned). *Exact NMF exists at* $d = 3$*; minimum rel. error* $= 0$.



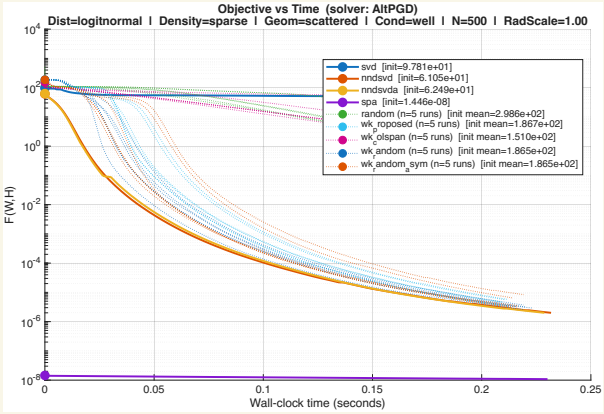
(a) Dirichlet, rel. err. vs. iter



(b) Dirichlet, obj. vs. time

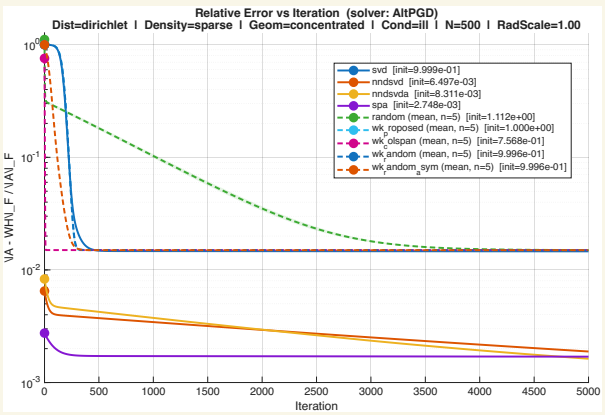


(c) Logit-normal, rel. err. vs. iter

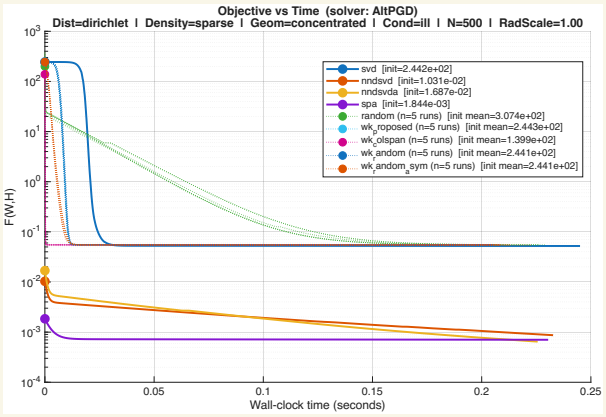


(d) Logit-normal, obj. vs. time

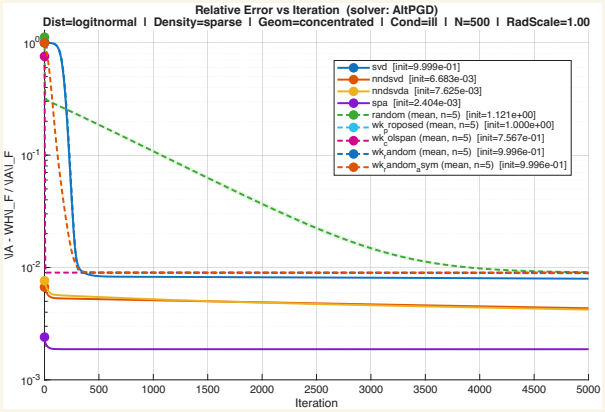
Figure 4.17: AltPGD on Option 3 (sparse, scattered, well-conditioned). *Exact NMF exists at $d = 3$; minimum rel. error $= 0$.*



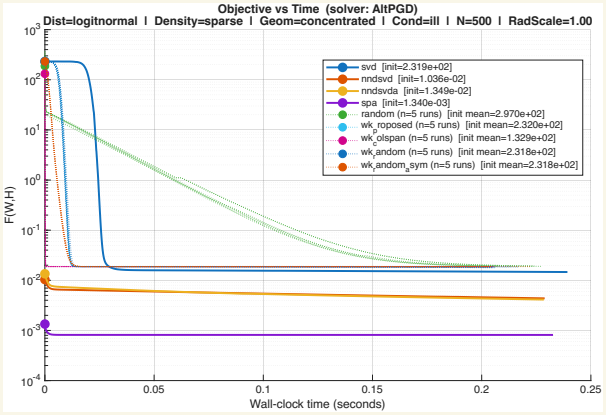
(a) Dirichlet, rel. err. vs. iter



(b) Dirichlet, obj. vs. time

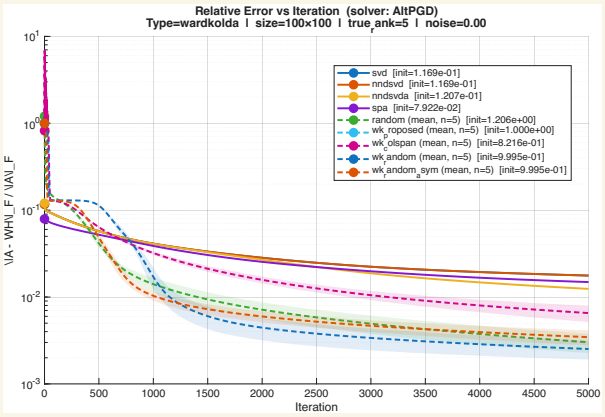


(c) Logit-normal, rel. err. vs. iter

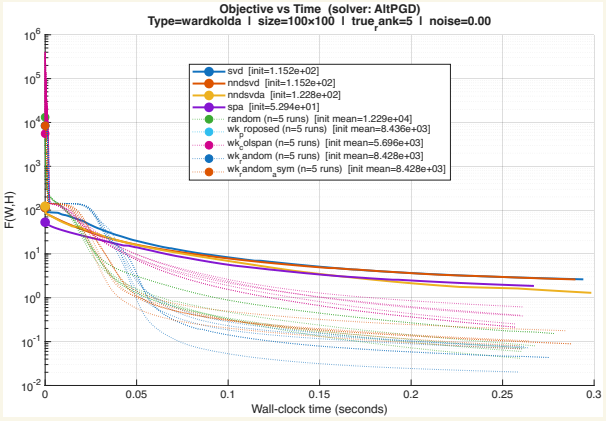


(d) Logit-normal, obj. vs. time

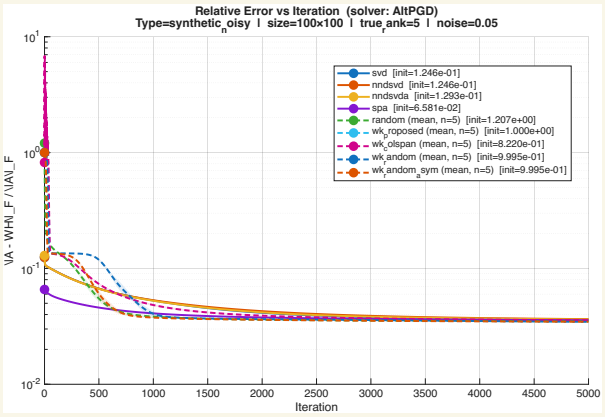
Figure 4.18: AltPGD on Option 4 (sparse, concentrated, ill-conditioned). *Exact NMF exists at $d = 3$; minimum rel. error $= 0$.*



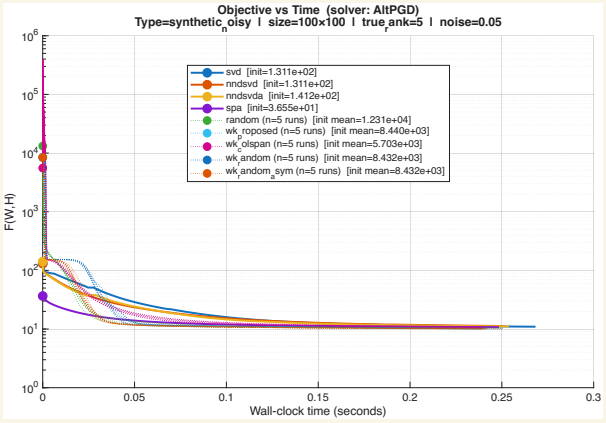
(a) wardkolda, rel. err. vs. iter



(b) wardkolda, obj. vs. time

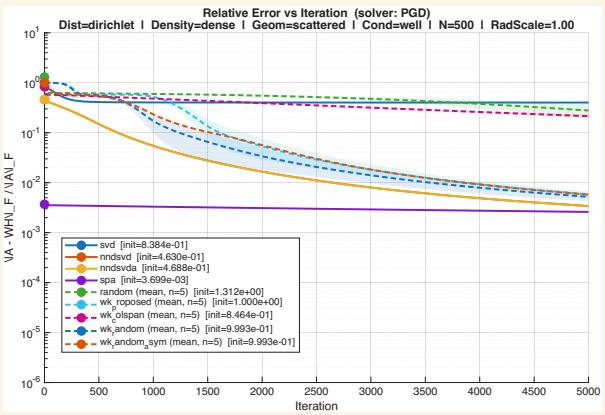


(c) synthetic_noisy ($\sigma = 0.05$), rel. err. vs. iter

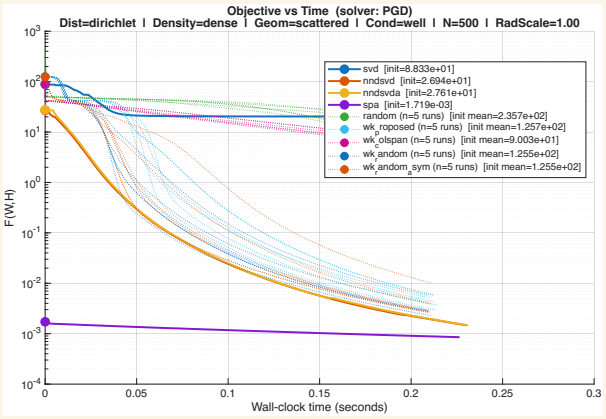


(d) synthetic_noisy ($\sigma = 0.05$), obj. vs. time

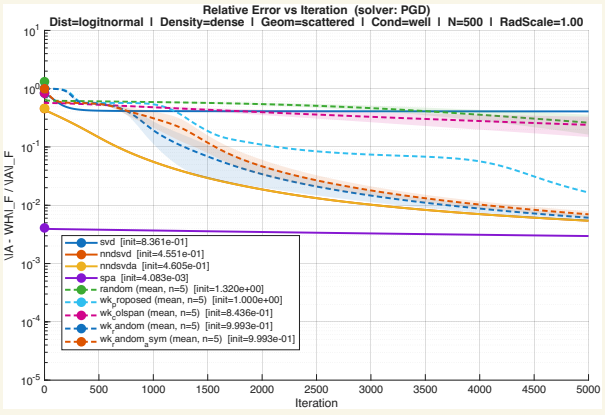
Figure 4.19: AltPGD on the two reference matrices. *Exact NMF exists at* $d = 10$ *for* `wardkolda` *(min. rel. error* $= 0$*);* `synthetic_noisy` *has no exact rank-*10 *NMF (min. rel. error strictly positive).*



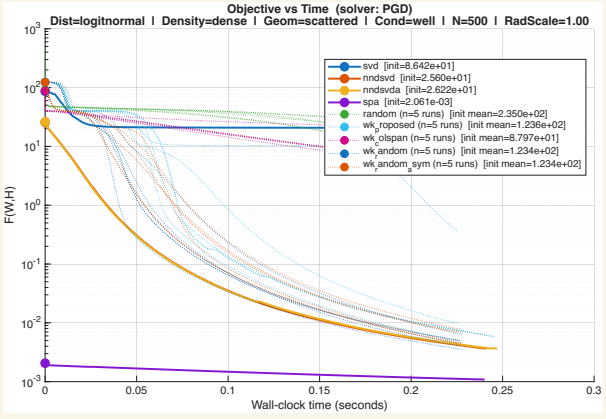
(a) Dirichlet, rel. err. vs. iter



(b) Dirichlet, obj. vs. time

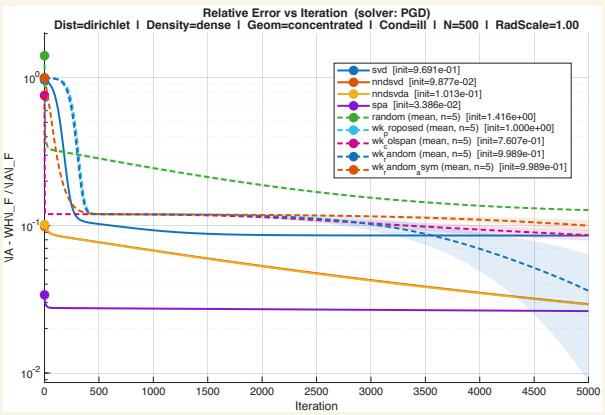


(c) Logit-normal, rel. err. vs. iter

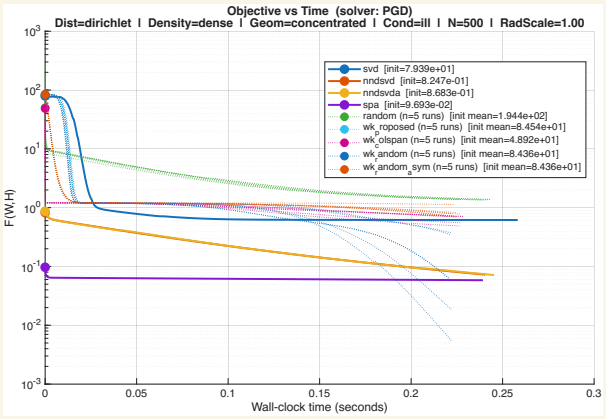


(d) Logit-normal, obj. vs. time

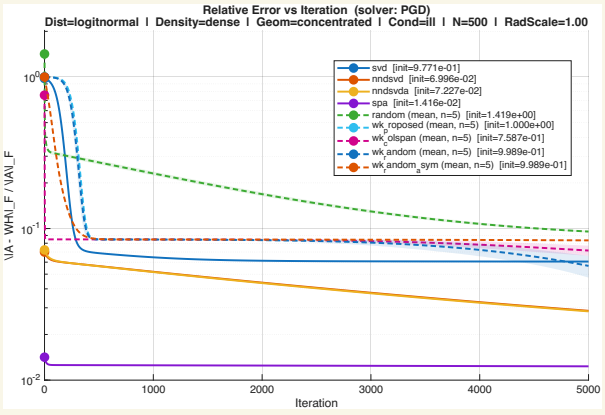
Figure 4.20: PGD on Option 1 (dense, scattered, well-conditioned). *Exact NMF exists at $d = 3$; minimum rel. error $= 0$.*



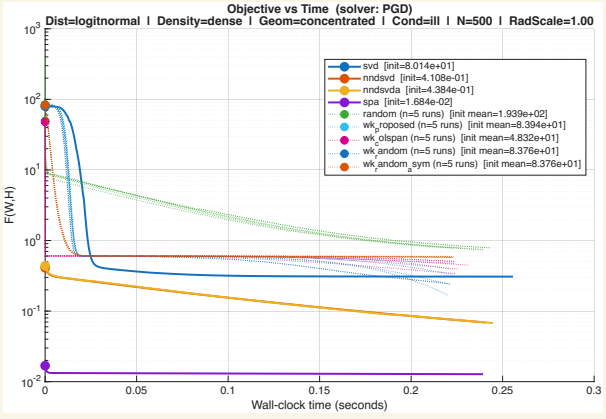
(a) Dirichlet, rel. err. vs. iter



(b) Dirichlet, obj. vs. time

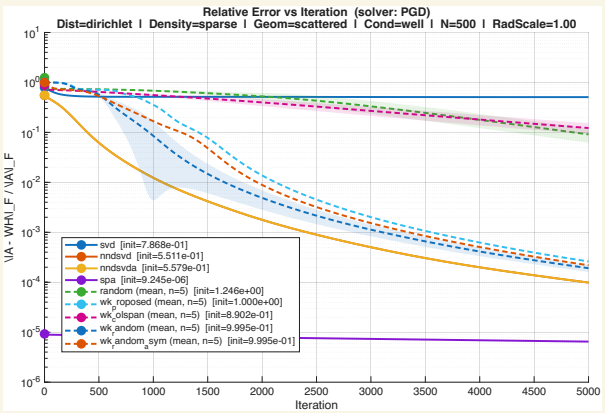


(c) Logit-normal, rel. err. vs. iter

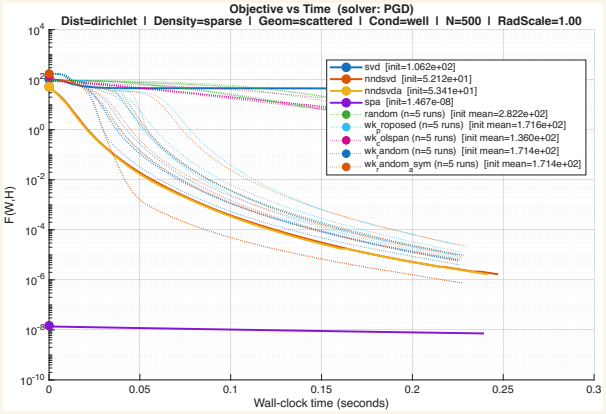


(d) Logit-normal, obj. vs. time

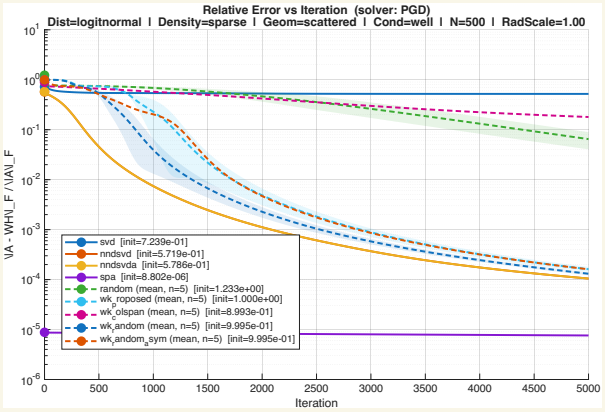
Figure 4.21: PGD on Option 2 (dense, concentrated, ill-conditioned). *Exact NMF exists at $d = 3$; minimum rel. error $= 0$.*



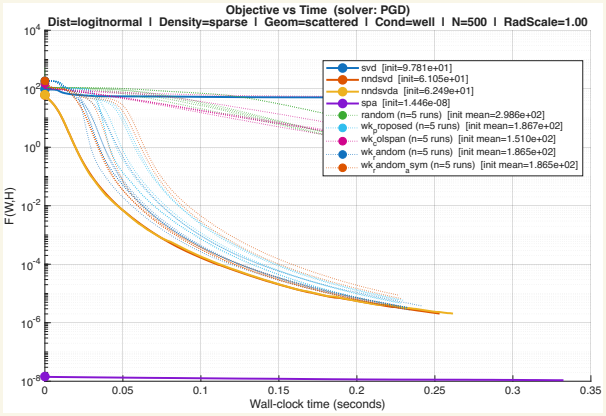
(a) Dirichlet, rel. err. vs. iter



(b) Dirichlet, obj. vs. time

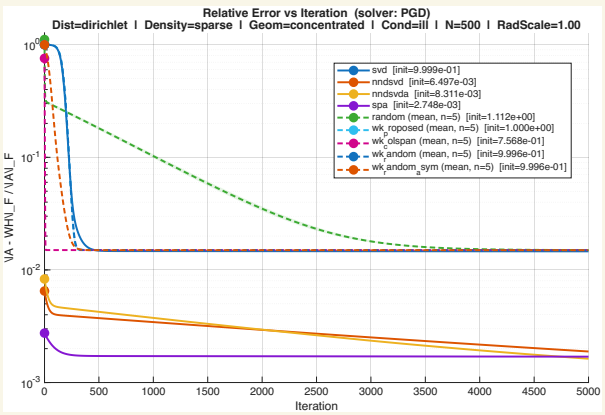


(c) Logit-normal, rel. err. vs. iter

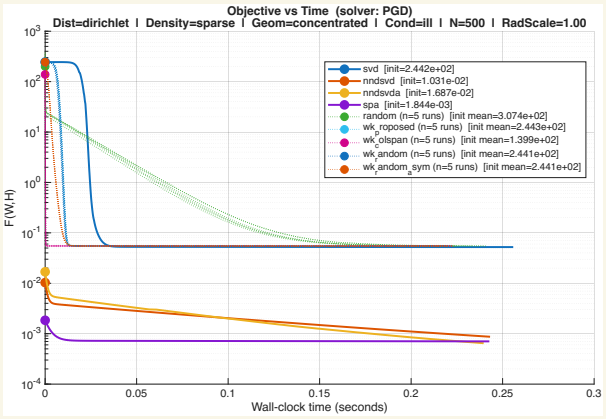


(d) Logit-normal, obj. vs. time

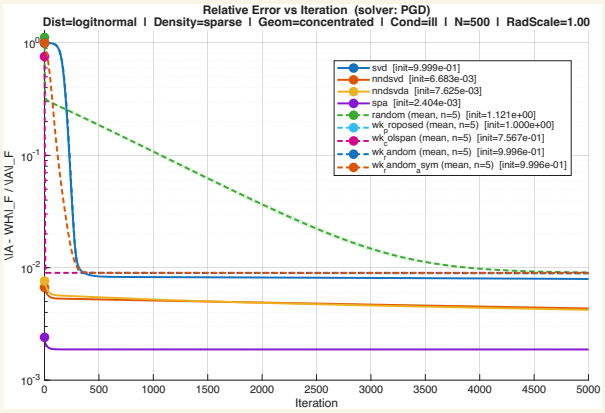
Figure 4.22: PGD on Option 3 (sparse, scattered, well-conditioned). *Exact NMF exists at $d = 3$; minimum rel. error $= 0$.*



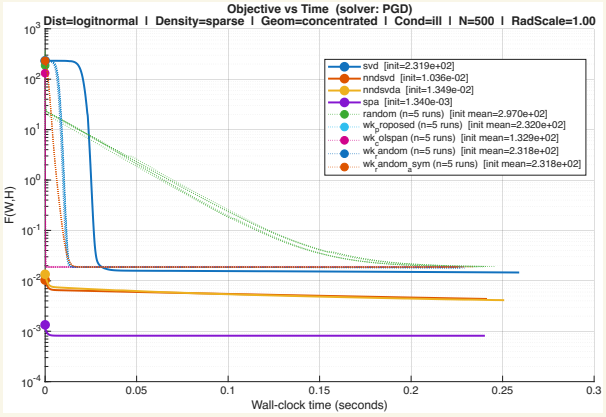
(a) Dirichlet, rel. err. vs. iter



(b) Dirichlet, obj. vs. time

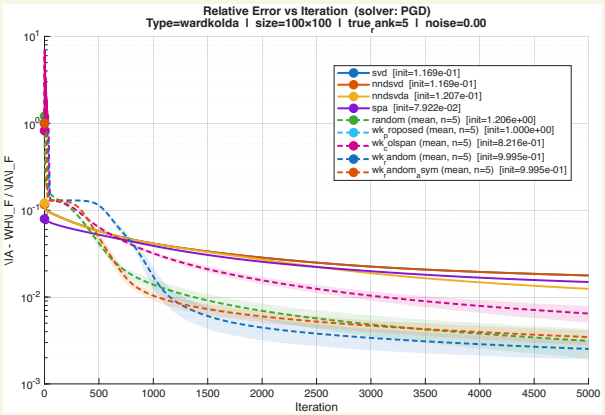


(c) Logit-normal, rel. err. vs. iter

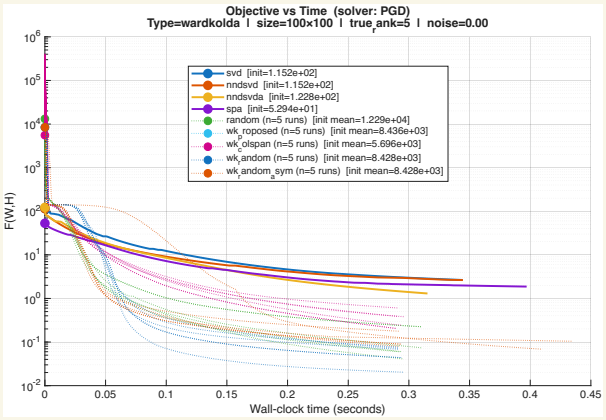


(d) Logit-normal, obj. vs. time

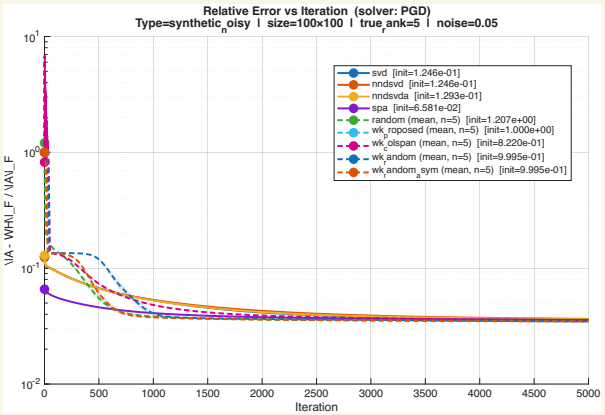
Figure 4.23: PGD on Option 4 (sparse, concentrated, ill-conditioned). *Exact NMF exists at* $d = 3$*; minimum rel. error* $= 0$.



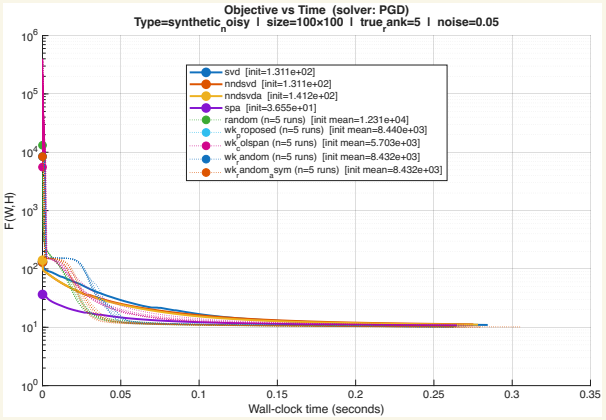
(a) wardkolda, rel. err. vs. iter



(b) wardkolda, obj. vs. time

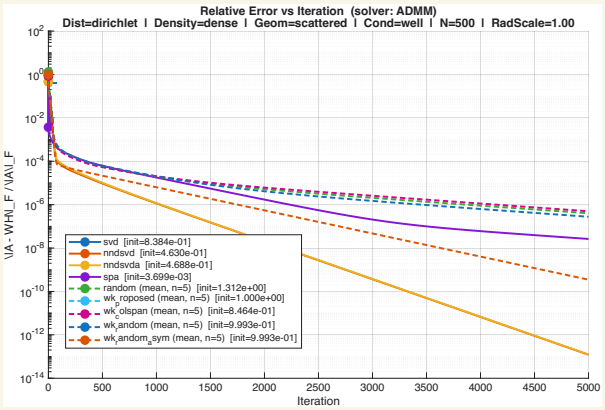


(c) synthetic_noisy ($\sigma = 0.05$), rel. err. vs. iter

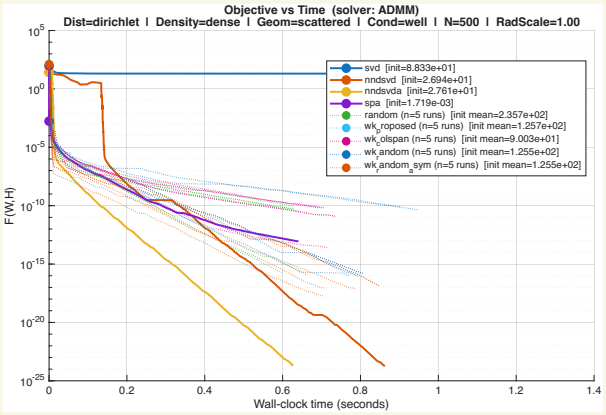


(d) synthetic_noisy ($\sigma = 0.05$), obj. vs. time

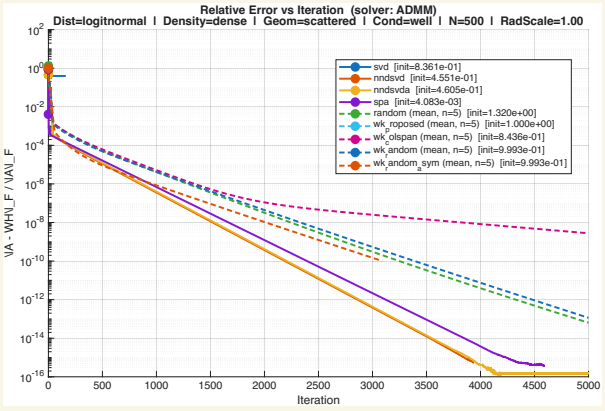
Figure 4.24: PGD on the two reference matrices. *Exact NMF exists at $d = 10$ for* `wardkolda` *(min. rel. error $= 0$);* `synthetic_noisy` *has no exact rank-10 NMF (min. rel. error strictly positive).*



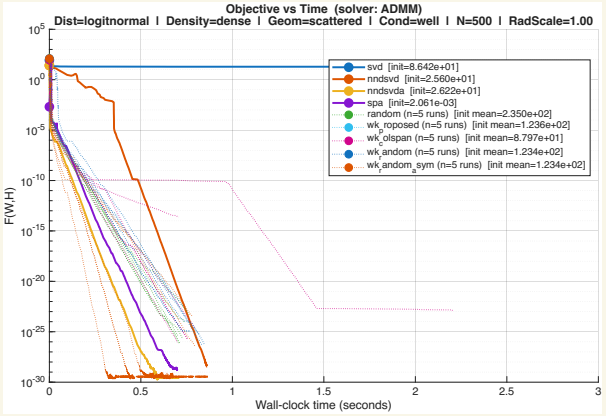
(a) Dirichlet, rel. err. vs. iter



(b) Dirichlet, obj. vs. time

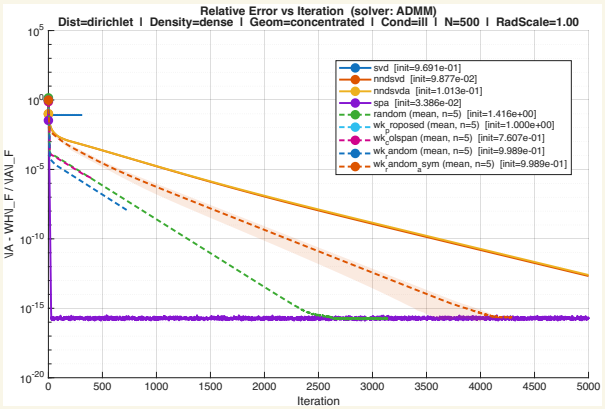


(c) Logit-normal, rel. err. vs. iter

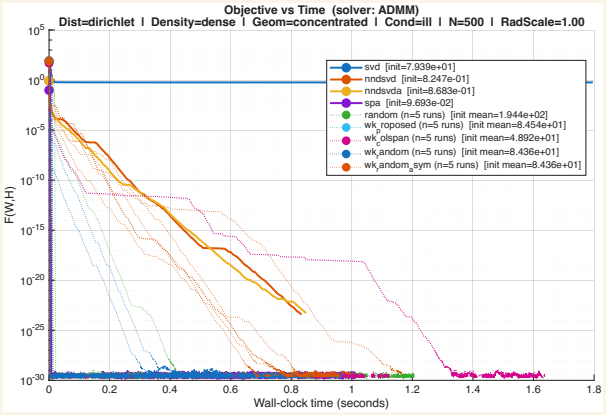


(d) Logit-normal, obj. vs. time

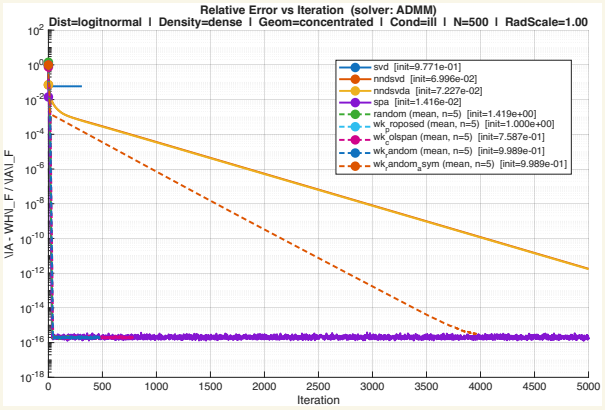
Figure 4.25: ADMM on Option 1 (dense, scattered, well-conditioned). *Exact NMF exists at $d = 3$; minimum rel. error $= 0$.*



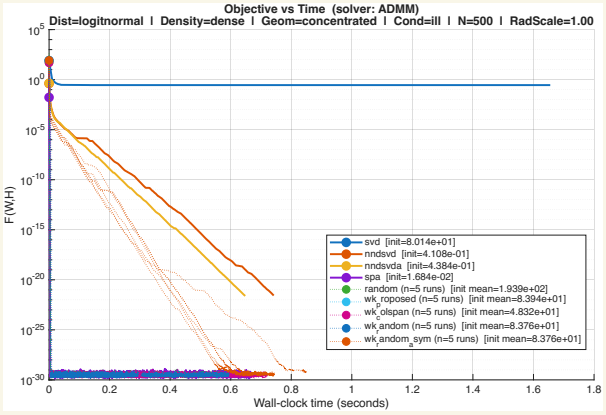
(a) Dirichlet, rel. err. vs. iter



(b) Dirichlet, obj. vs. time

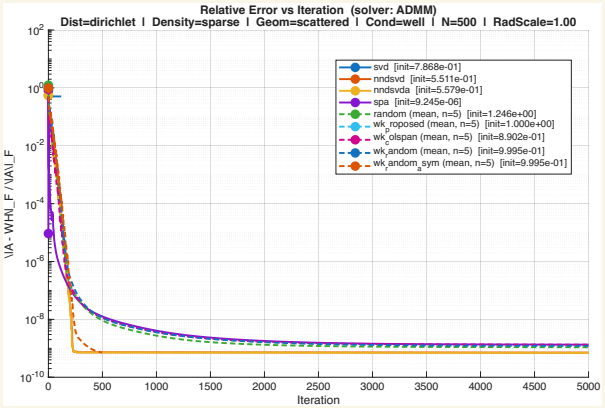


(c) Logit-normal, rel. err. vs. iter

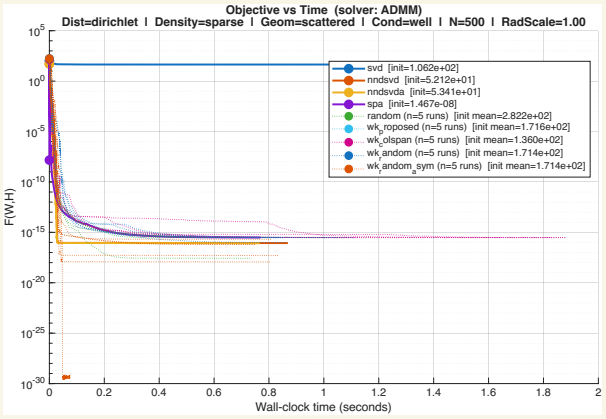


(d) Logit-normal, obj. vs. time

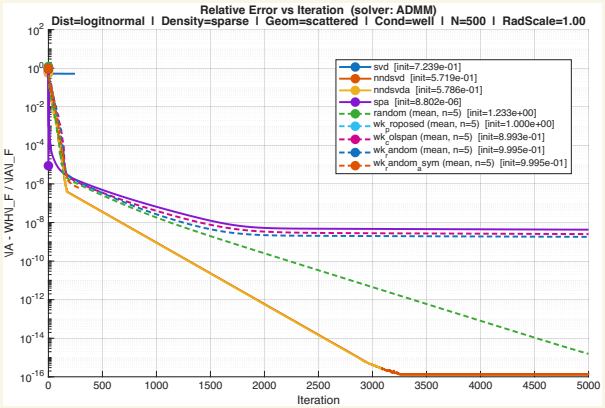
Figure 4.26: ADMM on Option 2 (dense, concentrated, ill-conditioned). *Exact NMF exists at* $d = 3$*; minimum rel. error* $= 0$.



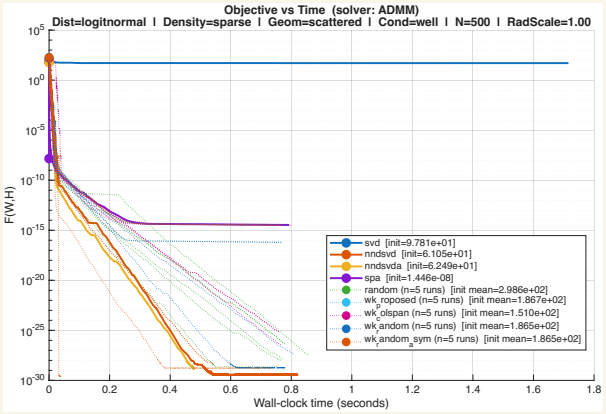
(a) Dirichlet, rel. err. vs. iter



(b) Dirichlet, obj. vs. time

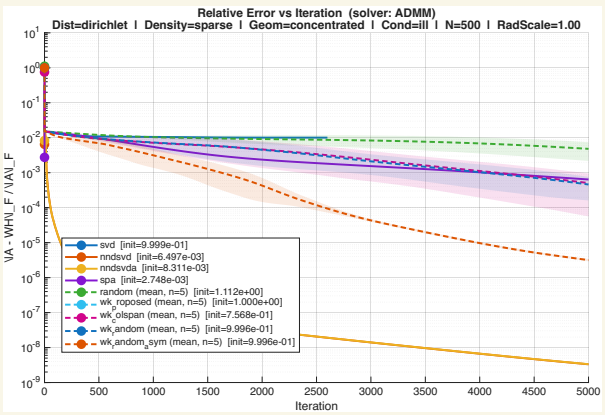


(c) Logit-normal, rel. err. vs. iter

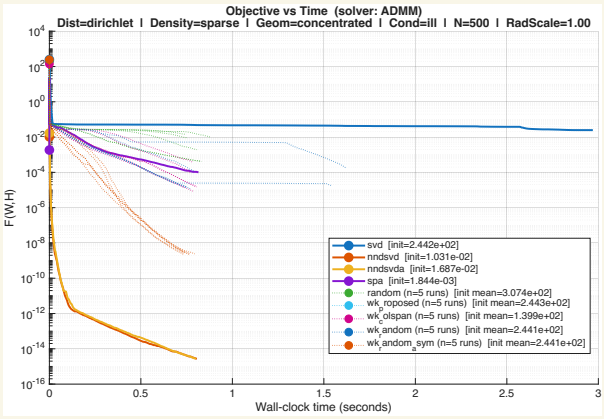


(d) Logit-normal, obj. vs. time

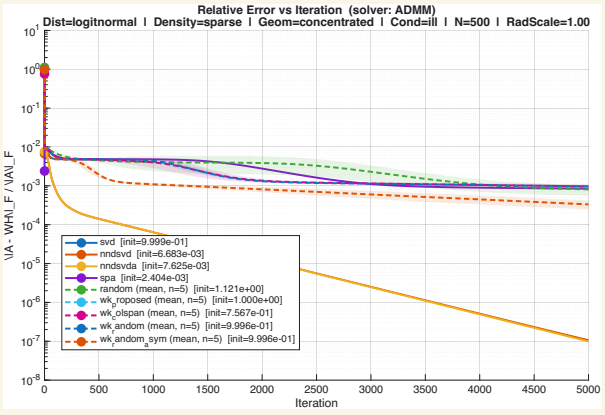
Figure 4.27: ADMM on Option 3 (sparse, scattered, well-conditioned). *Exact NMF exists at $d = 3$; minimum rel. error $= 0$.*



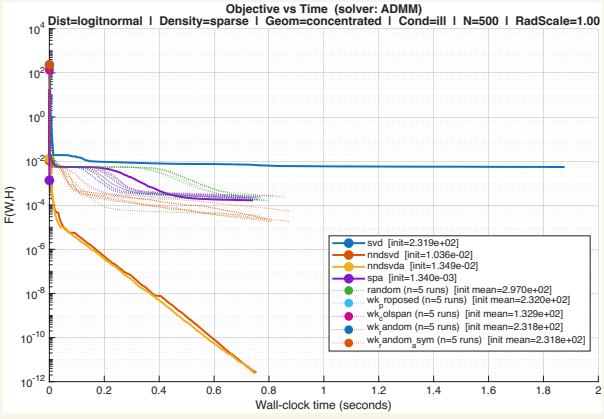
(a) Dirichlet, rel. err. vs. iter



(b) Dirichlet, obj. vs. time

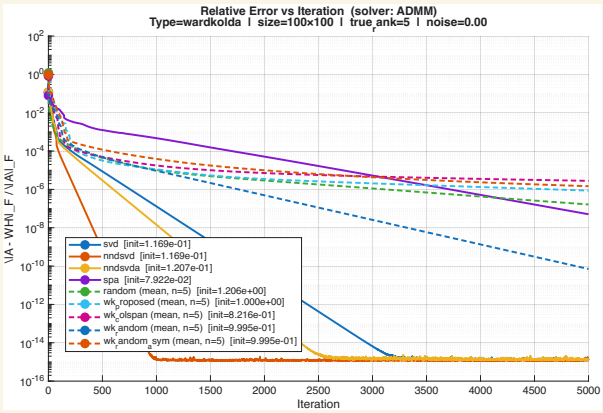


(c) Logit-normal, rel. err. vs. iter

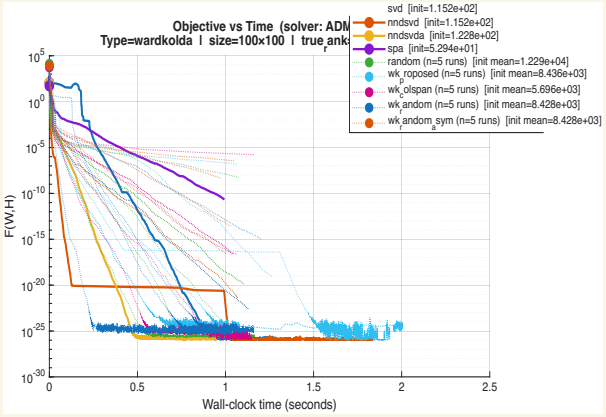


(d) Logit-normal, obj. vs. time

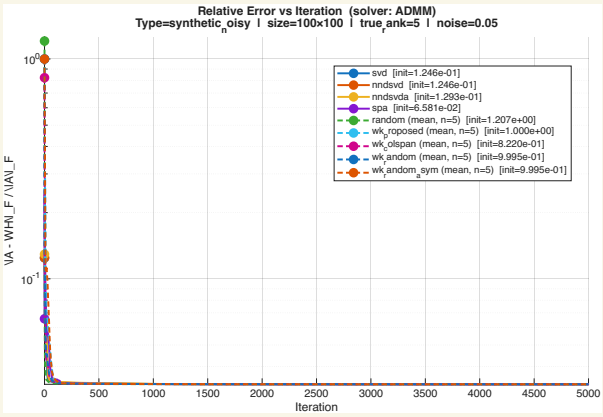
Figure 4.28: ADMM on Option 4 (sparse, concentrated, ill-conditioned). *Exact NMF exists at $d = 3$; minimum rel. error $= 0$.*



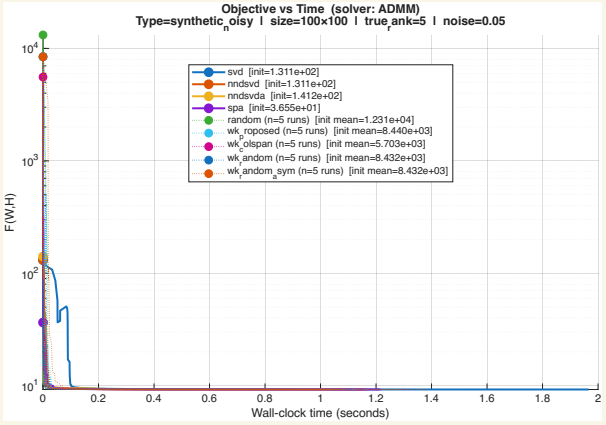
(a) wardkolda, rel. err. vs. iter



(b) wardkolda, obj. vs. time



(c) synthetic_noisy ($\sigma = 0.05$), rel. err. vs. iter



(d) synthetic_noisy ($\sigma = 0.05$), obj. vs. time

Figure 4.29: ADMM on the two reference matrices. *Exact NMF exists at $d = 10$ for* `wardkolda` *(min. rel. error $= 0$);* `synthetic_noisy` *has no exact rank-10 NMF (min. rel. error strictly positive).*

References

Shun'ichi Amari. *Information Geometry and Its Applications*. Springer, 2016. doi: 10.1007/978-4-431-55978-8.

Andersen Man Shun Ang and Nicolas Gillis. Accelerating nonnegative matrix factorization algorithms using extrapolation. *Neural Computation*, 31(2):417–439, 2019. doi: 10.1162/neco_a_01157.

Sanjeev Arora, Rong Ge, Ravindran Kannan, and Ankur Moitra. Computing a nonnegative matrix factorization – provably. In *Proceedings of the 44th Annual ACM Symposium on Theory of Computing*, pages 145–162, 2012. doi: 10.1145/2213977.2213994.

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics. Springer, Cham, 2 edition, 2017. doi: 10.1007/978-3-319-48311-5.

Amir Beck. *First-Order Methods in Optimization*. SIAM, 2017. doi: 10.1137/1.9781611974997.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. doi: 10.1137/080716542.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1):459–494, 2014. doi: 10.1007/s10107-013-0701-9.

Christos Boutsidis and Efstratios Gallopoulos. SVD based initialization: a head start for nonnegative matrix factorization. *Pattern Recognition*, 41(4):1350–1362, 2008. doi: 10.1016/j.patcog.2007.09.010.

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.

Samuel Burer and Renato D. C. Monteiro. Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444, 2005. doi: 10.1007/s10444-004-0564-1.

James V. Burke and Paul Tseng. A unified analysis of Hoffman's bound via Fenchel duality. *SIAM Journal on Optimization*, 6(2):265–282, 1996. doi: 10.1137/0806015.

Emmanuel J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008. doi: 10.1016/j.crma.2008.03.014.

Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Spectral methods for data science: a statistical perspective. *Foundations and Trends in Machine Learning*, 12(5–6):566–806, 2021. doi: 10.1561/2200000079.

Yuejie Chi, Yue M. Lu, and Yuxin Chen. Nonconvex optimization meets low-rank matrix factorization: an overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019. doi: 10.1109/TSP.2019.2937282.

Andrzej Cichocki, Rafal Zdunek, and Shun-ichi Amari. Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization. In *Independent Component Analysis and Signal Separation*, pages 169–176, 2007. doi: 10.1007/978-3-540-74494-8_22.

Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. doi: 10.1137/0707001.

Michał Dereziński and Michael W. Mahoney. Recent and upcoming developments in randomized numerical linear algebra for machine learning. *arXiv preprint arXiv:2406.11151*, 2024. URL https://arxiv.org/abs/2406.11151.

Lijun Ding and Stephen J. Wright. On squared-variable formulations. *SIAM Journal on Optimization*, 35(4):2265–2293, 2025. doi: 10.1137/23M1608343.

Jonathan Eckstein and Dimitri P. Bertsekas. On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1):293–318, 1992. doi: 10.1007/BF01581204.

Daniel Gabay and Bertrand Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Computers & Mathematics with Applications*, 2(1):17–40, 1976. doi: 10.1016/0898-1221(76)90003-1.

Nicolas Gillis. *Nonnegative Matrix Factorization*. SIAM, 2020. doi: 10.1137/1.9781611976410.

Nicolas Gillis and François Glineur. Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization. *Neural Computation*, 24(4):1085–1105, 2012. doi: 10.1162/NECO_a_00256.

Nicolas Gillis and Stephen A. Vavasis. Fast and robust recursive algorithms for separable nonnegative matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(4):698–714, 2014. doi: 10.1109/TPAMI.2013.226.

Roland Glowinski and A. Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, 9(R2):41–76, 1975. doi: 10.1051/m2an/197509r200411.

John C. Gower and Garmt B. Dijksterhuis. *Procrustes Problems*. Oxford Statistical Science Series. Oxford University Press, Oxford, 2004. doi: 10.1093/acprof:oso/9780198510581.001.0001.

Robert M. Gower and Peter Richtárik. Randomized iterative methods for linear systems. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1660–1690, 2015. doi: 10.1137/15M1025487.

Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011. doi: 10.1137/090771806.

Le Thi Khanh Hien, Nicolas Gillis, and Panagiotis Patrinos. Inertial block proximal methods for non-convex non-smooth optimization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 5671–5681, 2020. URL https://proceedings.mlr.press/v119/le20a.html.

Prateek Jain, Raghu Meka, and Inderjit S. Dhillon. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, pages 937–945, 2010. URL https://proceedings.neurips.cc/paper_files/paper/2010/file/08d98638c6fcd194a4b1e6992063e944-Paper.pdf.

Liwei Jiang, Yudong Chen, and Lijun Ding. Algorithmic regularization in model-free overparametrized asymmetric matrix factorization. *SIAM Journal on Mathematics of Data Science*, 5(3):723–744, 2023. doi: 10.1137/22M1519833.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, pages 795–811, 2016. doi: 10.1007/978-3-319-46128-1_50.

Richard M. Karp. Reducibility among combinatorial problems. In Raymond E. Miller, James W. Thatcher, and Jean D. Bohlinger, editors, *Complexity of Computer Computations*, pages 85–103. Springer, Boston, MA, 1972. doi: 10.1007/978-1-4684-2001-2_9.

Charles L. Lawson and Richard J. Hanson. *Solving Least Squares Problems*. SIAM, 1974. doi: 10.1137/1.9781611971217.

Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, pages 535–541, 2000. URL https://proceedings.neurips.cc/paper_files/paper/2000/file/f9d1152547c0bde01830b7e8bd60024c-Paper.pdf.

Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016. doi: 10.1137/15M1009597.

Bingcong Li, Liang Zhang, Aryan Mokhtari, and Niao He. On the crucial role of initialization for matrix factorization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=YTEwJaBdh0.

Qiuwei Li, Daniel McKenzie, and Wotao Yin. From the simplex to the sphere: Faster constrained optimization using the Hadamard parametrization. *Information and Inference: A Journal of the IMA*, 12(3):1898–1937, 2023. doi: 10.1093/imaiai/iaad017.

Per-Gunnar Martinsson and Joel A. Tropp. Randomized numerical Linear Algebra: Foundations and algorithms. *Acta Numerica*, 29:403–572, 2020. doi: 10.1017/S0962492920000021.

Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, 1960. doi: 10.1093/qmath/11.1.50.

Bamdev Mishra and Rodolphe Sepulchre. Riemannian preconditioning. *SIAM Journal on Optimization*, 26(1):635–660, 2016. doi: 10.1137/140970860.

Bamdev Mishra, Gilles Meyer, Silvère Bonnabel, and Rodolphe Sepulchre. Fixed-rank matrix factorizations and Riemannian low-rank optimization. *Computational Statistics*, 29(3):591–621, 2014. doi: 10.1007/s00180-013-0464-z.

Deanna Needell, Rachel Ward, and Nathan Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Mathematical Programming*, 155(1–2):549–573, 2016. doi: 10.1007/s10107-015-0864-7.

Yurii Nesterov. *Lectures on Convex Optimization*. Springer, 2nd edition, 2018. doi: 10.1007/978-3-319-91578-4.

Yurii E. Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. *Doklady Akademii Nauk SSSR*, 269(3):543–547, 1983. URL https://www.mathnet.ru/php/archive.phtml?wshow=paper&jrnid=dan&paperid=46009&option_lang=eng. Original in Russian.

Samet Oymak, Benjamin Recht, and Mahdi Soltanolkotabi. Sharp time-data tradeoffs for linear inverse problems. *IEEE Transactions on Information Theory*, 64(6):4129–4158, 2018. doi: 10.1109/TIT.2017.2773497.

Mert Pilanci and Martin J. Wainwright. Iterative Hessian sketch: Fast and accurate solution approximation for constrained least-squares. *Journal of Machine Learning Research*, 17(53):1–38, 2016. URL http://jmlr.org/papers/v17/14-460.html.

Mert Pilanci and Martin J. Wainwright. Newton sketch: A near linear-time optimization algorithm with linear-quadratic convergence. *SIAM Journal on Optimization*, 27(1):205–245, 2017. doi: 10.1137/15M1021106.

Boris T. Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963. doi: 10.1016/0041-5553(63)90382-3.

Boris T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964. doi: 10.1016/0041-5553(64)90137-5.

Meisam Razaviyayn, Mingyi Hong, and Zhi-Quan Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013. doi: 10.1137/120891009.

Vladimir Rokhlin and Mark Tygert. A fast randomized algorithm for overdetermined linear least-squares regression. *Proceedings of the National Academy of Sciences*, 105(36):13212–13217, 2008. doi: 10.1073/pnas.0804869105.

Ernest K. Ryu and Wotao Yin. *Large-Scale Convex Optimization: Algorithms & Analyses via Monotone Operators*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2022. doi: 10.1017/9781009181167.

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*, volume 34, pages 23831–23843, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/c82836ed448c41094025b4a872c5341e-Abstract.html.

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. URL http://jmlr.org/papers/v22/20-1067.html.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1–2):1–230, 2015. doi: 10.1561/2200000048.

Paul Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109(3):475–494, 2001. doi: 10.1023/A:1017501703105.

Stephen Tu, Ross Boczar, Max Simchowitz, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via Procrustes flow. In *Proceedings of the 33rd International Conference on Machine Learning*, pages 964–973, 2016. URL https://proceedings.mlr.press/v48/tu16.pdf.

Stephen A. Vavasis. On the complexity of nonnegative matrix factorization. *SIAM Journal on Optimization*, 2009. doi: 10.1137/070709967.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and Gitta Kutyniok, editors, *Compressed Sensing: Theory and Applications*, pages 210–268. Cambridge University Press, Cambridge, 2012. doi: 10.1017/CBO9780511794308.006.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Rachel Ward and Tamara G. Kolda. Convergence of alternating gradient descent for matrix factorization. In *Advances in Neural Information Processing Systems*, pages 22369–22382, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/46c10f6c8ea5aa6f267bcdabcb123f97-Paper-Conference.pdf.

Per-Åke Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1):99–111, 1972. doi: 10.1007/BF01932678.

David P. Woodruff. Sketching as a tool for numerical Linear Algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157, 2014. doi: 10.1561/0400000060.

Yangyang Xu and Wotao Yin. A block coordinate descent method for regularized multiconvex optimization with applications to nonnegative tensor factorization and completion. *SIAM Journal on Imaging Sciences*, 6(3):1758–1789, 2013. doi: 10.1137/120887795.

Zhenghao Xu, Yuqing Wang, Tuo Zhao, Rachel Ward, and Molei Tao. Provable acceleration of Nesterov's accelerated gradient for asymmetric matrix factorization and linear neural networks. In *Advances in Neural Information Processing Systems*, pages 33726–33755, 2024. doi: 10.52202/079017-1062.

Tian Ye and Simon Shaolei Du. Global convergence of gradient descent for asymmetric low-rank matrix factorization. In *Advances in Neural Information Processing Systems*, pages 1429–1439, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/0af854284f4ab0cfea8fcfd889cbb41a-Paper.pdf.

Alp Yurtsever, Joel A. Tropp, Olivier Fercoq, Madeleine Udell, and Volkan Cevher. Scalable semidefinite programming. *SIAM Journal on Mathematics of Data Science*, 3(1):171–200, 2021. doi: 10.1137/19M1305045.

Richard Y. Zhang. Nonnegative low-rank matrix recovery can have spurious local minima. *Optimization Letters*, 20(4):683–704, 2026. doi: 10.1007/s11590-026-02290-0.