UNIFYING CONVENTIONAL MATRIX FACTORIZATION AND HADAMARD DECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Matrix factorization (MF) is inherently constrained by a low-rank bottleneck, whereas the recently proposed Hadamard decomposition (HD) can represent high-rank matrices at the cost of a nearly rigid spectral profile. We first characterize when HD achieves rank amplification: the Hadamard product of two generic rank-r matrices has rank $\min(M, N, r^2)$ almost surely. However, under generic orthonormal factors, its Frobenius energy scales as $\mathcal{O}(r^2/N^2)$, while its average nonzero singular value scales as $\mathcal{O}(1/N)$, revealing an energy-dilution effect that we also verify empirically. Motivated by this analysis, we propose Rada-MF, a rank-adaptive framework that, under a fixed parameter budget, selects among MF, HD, and their additive hybrid according to the entropy-based effective rank of the target spectrum. Rada-MF provides a unified higher-level formulation that subsumes both conventional low-rank MF and HD. On controlled spectra with effective ranks ranging from 8 to 96, Rada-MF-auto, which automatically selects the decomposition regime, closely tracks the best fixed model, achieving a mean oracle gap of 0.012. Truncated SVD performs best at low effective ranks, the SVD-core+HD-residual hybrid is strongest in the intermediate regime, and HD dominates at high effective ranks. Experiments on binary matrices provide further support for the proposed framework.

1 INTRODUCTION

Matrix factorization (MF) is a fundamental linear-algebraic tool in artificial intelligence. It decomposes high-dimensional data into low-rank components to reduce model complexity, suppress noise, and uncover latent patterns within data. As indispensable mathematical building blocks, mainstream matrix-decomposition techniques underpin key AI applications including dimensionality reduction, computer vision, and recommender systems.

The most prevalent form of matrix factorization is given by:

$$\mathbf{M} \approx \mathbf{X}^\top \mathbf{Y}, \tag{1}$$

where $\mathbf{M} \in \mathbb{R}^{M\times N}$ denotes the target matrix, which can represent raw data (e.g., an image) or encode pairwise relations among objects (e.g., user-item interaction matrices). $\mathbf{X} \in \mathbb{R}^{r\times M}$ and $\mathbf{Y} \in \mathbb{R}^{r\times N}$ with $r \ll M,\ N$ are the low-dimensional feature matrices yielded by MF. Matrix factorization compresses the MN raw entries of $\mathbf{M}$ into a total of $(M+N)r$ parameters stored in $\mathbf{X}$ and $\mathbf{Y}$. Eq. 1 appears in many classic algorithms such as Singular Value Decomposition (SVD), Principal Component Analysis (PCA (Wold et al., 1987)), and Non-negative Matrix Factorization (NMF), all of which leverage low-rank priors to mitigate data sparsity and redundant feature dimensions.

The Low-Rank Bottleneck. By construction, $\mathrm{rank}(\mathbf{X}^\top\mathbf{Y}) \le r$. In practical scenarios, the target matrix $\mathbf{M}$ is often high-rank. Consequently, even the best low-rank approximation $\widehat{\mathbf{M}}$ may incur substantial information loss relative to the original matrix.

To address this limitation, **Hadamard Decomposition (HD)** (Ciaperoni et al., 2024) has been proposed. HD factorizes $\mathbf{M}$ as the Hadamard product of B rank-r matrices, enabling the representation to retain high-rank information. Previous work has shown that HD can achieve a rank as large as r^B. This boosted expressive power, however, does not come for free. In this paper, we aim to investigate

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the fundamental pattern and limitations of HD in modeling matrices by answering the following research questions:

- Under what conditions can the constructed matrix actually attain rank r^B?
- What structural patterns characterize matrices generated via Hadamard product?
- What are the respective advantageous regimes of HD and conventional MF?
- Is there a formula that unifies HD and MF?

Main Results. Our contributions are fourfold. (i) We prove that for generic rank-r matrices $\mathbf{A}$ and $\mathbf{B}$ with $r^2 \leq \min(M, N)$, $rank(\mathbf{A} \odot \mathbf{B}) = \min(M, N, r^2)$ with probability 1, and we characterize the spectrum of $\mathbf{A} \odot \mathbf{B}$: its total Frobenius energy scales as $\mathcal{O}(r^2/N^2)$ and its average nonzero singular value as $\mathcal{O}(N^{-1})$, indicating that the amplified rank is paid for by diluted energy (Proposition 4.5, verified by controlled experiments in Fig. 2). (ii) We explain the ill-conditioning of Hadamard products via a Marchenko–Pastur reference scaling, reported as Proposition 4.7. (iii) We propose Rada-MF, a rank-adaptive MF formula that unifies conventional MF and HD. Technically, it conducts selection among conventional MF, HD, and an additive SVD-core+HD-residual hybrid under an identical parameter budget, driven by comparing the budget with the entropy effective rank of the target spectrum. (iv) Across 900 controlled spectra and application benchmarks, Rada-MF-auto approaches the oracle among fixed candidates (mean gap 0.012 RRSE), recovering SVD-level accuracy at low effective rank and HD-level expressiveness at high effective rank.

2 Related Works

Conventional Matrix Factorization. The SVD (Du et al., 2023) stands as the gold standard for low-rank matrix approximation. SVD factors $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ and the truncated SVD, which retains the top r singular values, yields the optimal rank-r approximation in the Frobenius norm. PCA (Wold et al., 1987) builds upon the SVD framework to project high-dimensional data onto a lower-dimensional subspace that captures the directions of maximal variance. For data matrices with rapidly decaying singular spectra, SVD and PCA provide compact and informative representations that have proven highly effective. Non-negative Matrix Factorization (NMF) imposes non-negativity constraints on the factor matrices (Lee & Seung, 1999; Gillis, 2020). This additive, parts-based representation aligns well with human intuition for interpretability and has found widespread success in image analysis, document clustering and bioinformatics. Despite their widespread adoption and empirical success, all the above approaches share a fundamental limitation: they inherently follow the low-rank assumption on the original matrices.

Hadamard Decomposition. In the literature, HD was first explicitly proposed in Ciaperoni et al. (2024). It adopts gradient-based algorithms for the HD problem, with results demonstrating competitive approximation accuracy compared to SVD for several datasets and promising applicability to recommender systems. Xiong et al. (2024) proposes a graph embedding technique whose core idea falls into HD, demonstrating the strong capability of HD in modeling graph data. Ding et al. (2025) applies HD to compressing signals in power systems, achieving better compression efficiency and reconstruction accuracy. Different from MF, HD does not have a closed-form optimal solution. Recently, researchers have been developing efficient algorithms to obtain a good HD solution, e.g. scaled gradient descent (Ciaperoni et al., 2024), manifold-based gradient descent (Gillis et al., 2026), and alternating least-squares-type methods (Wertz et al., 2025).

Summary and Positioning. Instead of trying to improve the quality of the HD solution as recent works Wertz et al. (2025); Gillis et al. (2026), we focus on the analysis of the essential limitations of the representative power of HD and propose a rank-adaptive MF scheme which unifies conventional MF and HD. Table 1 summarizes the key differences between conventional MF, HD, and our Rada-MF.

3 Background

3.1 Notations and Preliminaries

Vectors and Matrices. Normal letters represent scalars (e.g. x). Bold uppercase letters represent matrices (e.g. $\mathbf{W}$). Bold lowercase letters represent column vectors (e.g. $\mathbf{x}$).

Table 1: Comparison of Rada-MF with conventional matrix factorization (MF) and Hadamard Decomposition (HD). Note that both HD and Rada-MF can be equally transformed into an MF-based formulation, as illustrated in Fig. 1.

Method	Form	Rank	Parameters	Pattern Limitations
Conventional MF	$\mathbf{X}^\top\mathbf{Y}$	r	$\mathcal{O}\big((M+N)r\big)$	Low-rank bottleneck
HD (Ciaperoni et al., 2024)	$\bigodot_{b=1}^{B}\mathbf{X}_b^\top\mathbf{Y}_b$	$\min(M,N,r^B)$	$\mathcal{O}\big((M+N)rB\big)$	1. Almost fixed high rank 2. Spectrum distortion
Rada-MF	$\sum_{l=1}^{L}\bigodot_{b=1}^{l}\mathbf{X}_{lb}^\top\mathbf{Y}_{lb}$	Adaptive	$\mathcal{O}\big((M+N)rBL\big)$	–

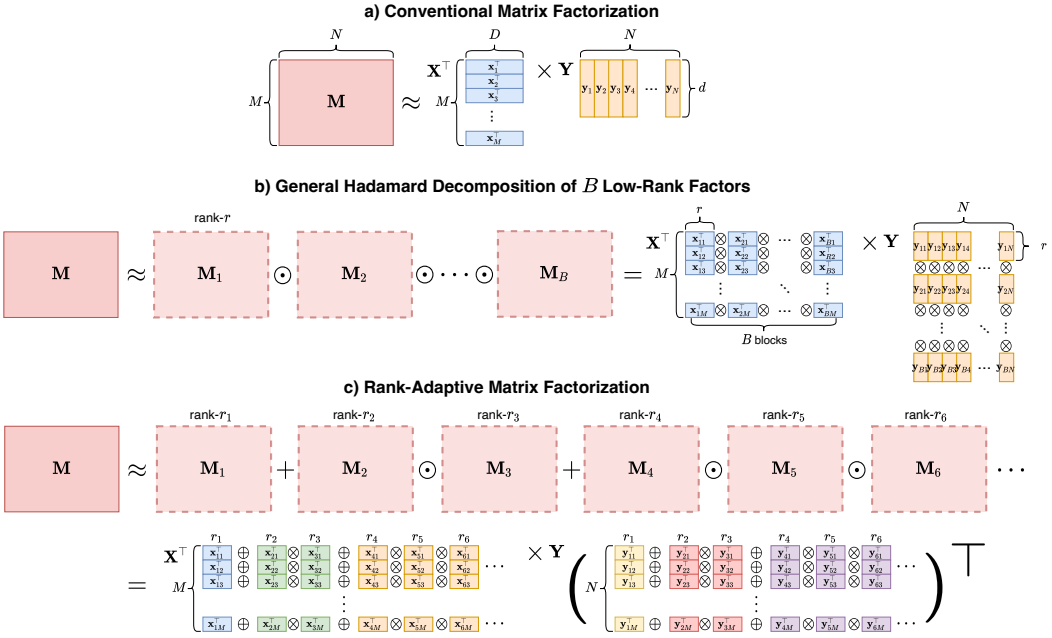


Figure 1: Illustrative differences in decomposition schemes among a) conventional MF, b) HD, and c) our proposed Rada-MF. All of them can be written in the MF form as $\mathbf{X}^\top\mathbf{Y}$.

Tensor Product. Let $\mathbf{a}\in\mathbb{R}^{d_1}$ and $\mathbf{b}\in\mathbb{R}^{d_2}$ be two vectors. The tensor product $\mathbf{a}\otimes\mathbf{b}$ is a vector in $\mathbb{R}^{d_1d_2}$, where the i-th element is $[\mathbf{a}\otimes\mathbf{b}]_{(i-1)*d_2+j}=[\mathbf{a}]_i[\mathbf{b}]_j$.

Hadamard Product. Given two matrices (or vectors) $\mathbf{A}$, $\mathbf{B}$ of the same shape, the Hadamard product is defined by the element-wise multiplication, i.e. $[\mathbf{A}\odot\mathbf{B}]_{ij}=[\mathbf{A}]_{ij}\cdot[\mathbf{B}]_{ij}$.

Concatenation. Given $\mathbf{a}\in\mathbb{R}^{d_1}$ and $\mathbf{b}\in\mathbb{R}^{d_2}$, the concatenation $\mathbf{c}=\mathbf{a}\oplus\mathbf{b}$ is a vector in $\mathbb{R}^{d_1+d_2}$ with the first d_1 dimensions being $\mathbf{a}$ and the last d_2 dimensions being $\mathbf{b}$.

3.2 Error Bound of Conventional Low-Rank MF Methods

The Singular Value Decomposition (SVD) of a rank-r matrix $\mathbf{M}$ is given by $\mathbf{M}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U}\in\mathbb{R}^{M\times M}$, $\mathbf{\Sigma}\in\mathbb{R}^{M\times N}$, and $\mathbf{V}\in\mathbb{R}^{N\times N}$. Here, $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices, and $\mathbf{\Sigma}=diag(\sigma_1,\sigma_2,\ldots,\sigma_r,0,\ldots,0)$ is a diagonal matrix containing the singular values of $\mathbf{M}$.

The truncated SVD of $\mathbf{M}$, denoted $\mathbf{M}^k$, retains only the first k largest singular values and corresponding singular vectors. It is defined as $\mathbf{M}^k=\mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^\top$, where $\mathbf{U}_k$ and $\mathbf{V}_k$ contain the first k left and right singular vectors, respectively, and $\mathbf{\Sigma}_k\in\mathbb{R}^{k\times k}$ is a diagonal matrix containing the first k singular values. The truncated SVD provides the best rank-k approximation of $\mathbf{M}$ with respect to the Frobenius norm, with the error given by $\|\mathbf{M}-\mathbf{M}^k\|_F=\sqrt{\sum_{i=k+1}^{r}\sigma_i^2}$, where $\sigma_{k+1},\ldots,\sigma_r$

are the discarded singular values. The optimal relative error bound for the rank-k approximation is:

$$\min_{\text{rank}(\widehat{\mathbf{M}})\leq k} \frac{\|\mathbf{M}-\widehat{\mathbf{M}}\|_F}{\|\mathbf{M}\|_F} = \frac{\|\mathbf{M}-\mathbf{M}_k\|_F}{\|\mathbf{M}\|_F} = \frac{\sqrt{\sum_{i=k+1}^{r}\sigma_i^2}}{\sqrt{\sum_{i=1}^{r}\sigma_i^2}} \leq \sqrt{\frac{r-k}{r}}, \tag{2}$$

where the upper bound is attained when all r nonzero singular values are equal. This inequality demonstrates that the low-rank methods may result in significant information loss when applied to complex, high-dimensional problems.

3.3 Hadamard Decomposition

Definition 3.1 (Hadamard Decomposition)**.** Given a matrix $\mathbf{M}$, Hadamard Decomposition (HD) seeks to factorize $\mathbf{M}$ as the Hadamard product of low-rank matrices. A rank-r HD is formulated as the following optimization problem:

$$\min_{rank(\mathbf{M}_b)\leq r} \|\mathbf{M}-\mathbf{M}_1\odot\mathbf{M}_2\odot\cdots\odot\mathbf{M}_B\|_F. \tag{3}$$

Most existing studies (Ciaperoni et al., 2024; Wertz et al., 2025) predominantly refer to the $B=2$ case as the standard HD.

HD in the MF form (illustrated in Fig. 1 b). The generalized HD decomposes matrix $\mathbf{M}$ into B rank-r matrices such that $\mathbf{M}\approx\bigodot_{b=1}^{B}\mathbf{M}_b$. This decomposition can also be cast in MF form $\mathbf{M}\approx\mathbf{X}^\top\mathbf{Y}$, where

$$\mathbf{X}=\Big[\bigotimes_{b=1}^{B}\mathbf{x}_{b1},\ \bigotimes_{b=1}^{B}\mathbf{x}_{b2},\ \ldots,\ \bigotimes_{b=1}^{B}\mathbf{x}_{bM}\Big],\quad \mathbf{Y}=\Big[\bigotimes_{b=1}^{B}\mathbf{y}_{b1},\ \bigotimes_{b=1}^{B}\mathbf{y}_{b2},\ \ldots,\ \bigotimes_{b=1}^{B}\mathbf{y}_{bN}\Big], \tag{4}$$

with $\mathbf{x}_{bi},\ \mathbf{y}_{bj}\in\mathbb{R}^r$. Here $\mathbf{X}\in\mathbb{R}^{r^B\times M}$ and $\mathbf{Y}\in\mathbb{R}^{r^B\times N}$; $\{\mathbf{x}_{bi}\}$ and $\{\mathbf{y}_{bj}\}$ denote r-dimensional column vectors. Each rank-r HD factor $\mathbf{M}_b$ is given by the matrix product

$$\mathbf{M}_b=\mathbf{X}_b^\top\mathbf{Y}_b,\quad \mathbf{X}_b=\big[\mathbf{x}_{b1},\mathbf{x}_{b2},\ldots,\mathbf{x}_{bM}\big]\in\mathbb{R}^{r\times M},\ \mathbf{Y}_b=\big[\mathbf{y}_{b1},\mathbf{y}_{b2},\ldots,\mathbf{y}_{bN}\big]\in\mathbb{R}^{r\times N} \tag{5}$$

Parameter Count. The total number of parameters of $\{\mathbf{x}_{ri}\}$ and $\{\mathbf{y}_{rj}\}$ is $(M+N)Br$.

Review of the High-rank Expressiveness. By the property of the Hadamard product $rank(\mathbf{A}\odot\mathbf{B})\leq rank(\mathbf{A})*rank(\mathbf{B})$, the rank of $\widehat{\mathbf{M}}=\bigodot_{b=1}^{B}\mathbf{M}_b$ is bounded by r^B. Prior works have given elaborate discussions on this point (Ciaperoni et al., 2024; Gillis et al., 2026).

4 Theoretical Analysis of the Hadamard Product

4.1 Full-Rank with Probability 1

A key drawback of the Hadamard product is the lack of explicit rank-control over decomposed factors, a capability readily available in classic matrix-decomposition approaches. Theorem 4.1 provides a formal proof for this argument.

Previous works have briefly shown that, given two rank-r matrices $\mathbf{A}$ and $\mathbf{B}$, the rank of the Hadamard product $rank(A\odot B)\leq r^2$. We extend this point to the following theorem.

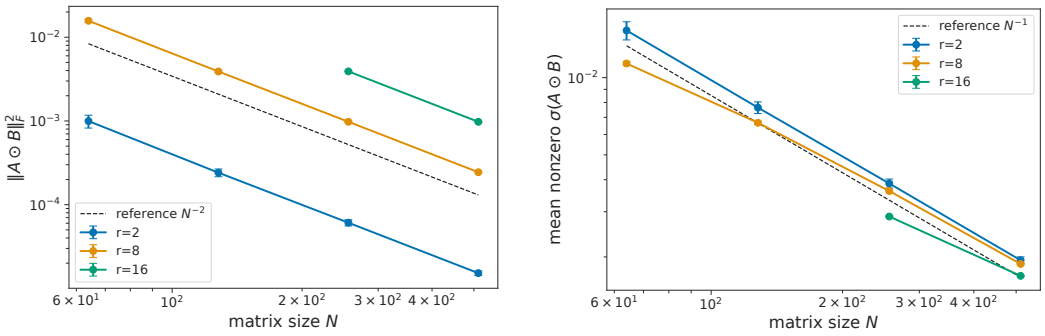
Theorem 4.1. *Let $\mathbf{A},\mathbf{B}\in\mathbb{R}^{M\times N}$ be two matrices of rank r with generic entries, and assume $r^2\leq\min(M,N)$. Then*

$$rank(\mathbf{A}\odot\mathbf{B})=\min(M,\ N,\ r^2) \tag{6}$$

with probability 1. When $r^2>\min(M,N)$, the same quantity is an upper bound: $rank(\mathbf{A}\odot\mathbf{B})\leq\min(M,N,r^2)$.

Proof of Theorem 4.1. The proof is organized into three parts. First, we express $A\odot B$ as a sum of r^2 rank-one matrices by Proposition 4.2. Second, we show that the factor matrices in this decomposition have full column rank by Proposition 4.3. Finally, we use these facts to conclude.

Proposition 4.2 (Decomposition into Rank-One Terms (proven in Appendix A.1))**.** *$\mathbf{A}\odot\mathbf{B}$ can be represented as the sum of r^2 rank-1 matrices.*



(a) Total Frobenius energy $\|\mathbf{A} \odot \mathbf{B}\|_F^2$.

(b) Mean nonzero singular value.

Figure 2: Empirical evidence for Proposition 4.5: generic independent rank-r factors with unit nonzero singular values, $N \in \{64, 128, 256, 512\}$, $r \in \{2, 8, 16\}$, 100 seeds per point. (a) The total energy follows the N^{-2} reference (fitted slopes -2.01 for r=2 and -2.00 for r=8). (b) The mean nonzero singular value follows N^{-1} for r=2 (slope -0.99); at r=8 the fitted slope is -0.87 over the available sizes, a finite-size deviation reported as-is; r=16 admits only $N \in \{256, 512\}$ because $r^2 \leq N$, so no slope is fitted.

Proposition 4.3 (Full Column Rank of the Factor Matrices (proven in Appendix A.2))**.** *Define the matrices* $\mathbf{W} = [\mathbf{w}_{11}, \mathbf{w}_{12}, \ldots, \mathbf{w}_{rr}] \in \mathbb{R}^{M \times r^2}$ *and* $\mathbf{Z} = [\mathbf{z}_{11}, \mathbf{z}_{12}, \ldots, \mathbf{z}_{rr}] \in \mathbb{R}^{N \times r^2}$*. If* $M \geq r^2$ *and* $\mathbf{U}, \tilde{\mathbf{U}}$ *are generic, then* $rank(\mathbf{W}) = r^2$*; likewise, if* $N \geq r^2$ *and* $\mathbf{V}, \tilde{\mathbf{V}}$ *are generic, then* $rank(\mathbf{Z}) = r^2$*.*

From the decomposition given by Proposition 4.2, the column space of $\mathbf{A} \odot \mathbf{B}$ is contained in the span of $\{\mathbf{w}_{kl}\}$ and the row space is contained in the span of $\{\mathbf{z}_{kl}\}$. Equivalently, vectorizing gives $vec(\mathbf{A} \odot \mathbf{B}) = \sum_{k=1}^{r} \sum_{l=1}^{r} \mathbf{z}_{kl} \otimes \mathbf{w}_{kl}$. We now use the following standard fact:

Lemma 4.4 (Rank of column-wise Kronecker product (proven in Appendix A.3))**.** *Let* $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K] \in \mathbb{R}^{d_1 \times K}$ *and* $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_K] \in \mathbb{R}^{d_2 \times K}$*. If* $rank(\mathbf{W}) = K$ *and* $rank(\mathbf{Z}) = K$*, then the matrix* $[\mathbf{z}_1 \otimes \mathbf{w}_1, \ldots, \mathbf{z}_K \otimes \mathbf{w}_K] \in \mathbb{R}^{d_1 d_2 \times K}$ *has rank* K*.*

By Proposition 4.3, both $\mathbf{W}$ and $\mathbf{Z}$ have rank r^2 (when $M \geq r^2$ and $N \geq r^2$); applying Lemma 4.4 with $K = r^2$ then gives $rank(\mathbf{A} \odot \mathbf{B}) = r^2$. If $M < r^2$, the column vectors $\mathbf{w}_{kl} \in \mathbb{R}^M$ can have rank at most M, so $rank(\mathbf{A} \odot \mathbf{B}) \leq M$. Similarly, if $N < r^2$, then $rank(\mathbf{A} \odot \mathbf{B}) \leq N$. So, we have

$$rank(\mathbf{A} \odot \mathbf{B}) = \min(N, M, r^2). \tag{7}$$

□

4.2 Spectrum Distortion

The second limitation of the expressiveness of Hadamard product is that the spectrum will be distorted. It comes from the two aspects:

- **Energy Dilution.** The average singular value decreases very sharply.
- **Spectral Collapse.** The modeled matrices are very likely to be ill-conditioned.

Energy Dilution. While the Hadamard product models high-rank matrices, it inherently suffers from a phenomenon we term Spectral Energy Dilution.

Proposition 4.5 (Spectral Energy Dilution (Fig. 2, proven in Appendix A.4))**.** *Let* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{N \times N}$ *be two generic, independent rank-r matrices (*$r \ll N$*) with perfectly conditioned, non-zero singular values* $\sigma_i(\mathbf{A}) = \sigma_i(\mathbf{B}) = 1$*. As* $N \to \infty$*, the expected total spectral energy (Frobenius norm squared) of their Hadamard product* $\mathbf{A} \odot \mathbf{B}$ *decays asymptotically as* $\mathcal{O}(r^2/N^2)$*, and the average squared singular value* $\bar{\sigma}^2$ *collapses to* $\mathcal{O}(1/N^2)$*.*

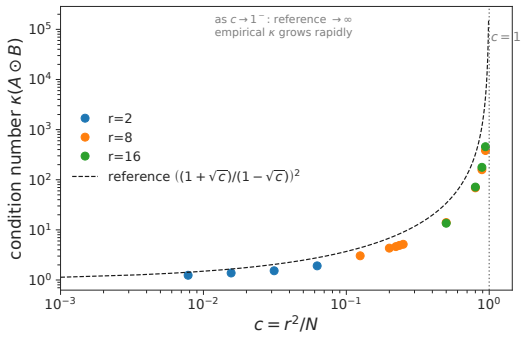


Figure 3: Condition number $\kappa(\mathbf{A} \odot \mathbf{B})$ versus $c = r^2/N$. $\mathbf{A}$ and $\mathbf{B}$ are random generic matrices. The figure shows that when $c < 1$ the condition number mainly depends on $c = r^2/N$.

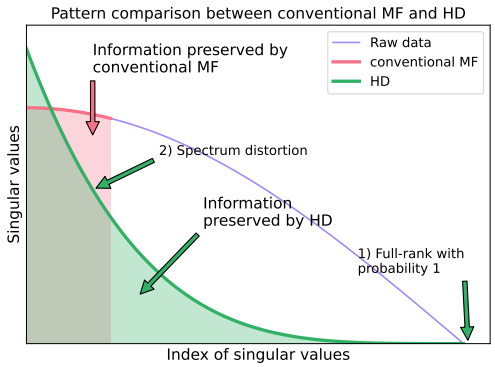


Figure 4: Illustrative pattern comparison of singular spectrum of conventional MF and HD.

Remark 4.6. Proposition 4.5 mathematically demonstrates that Hadamard-based interaction models suffer from a fundamental representation bottleneck. While the original matrices $\mathbf{A}$ and $\mathbf{B}$ contain $\mathcal{O}(1)$ singular values, the Hadamard product scatters this information across an r^2-dimensional space, driving the average singular value down to $\mathcal{O}(1/N)$.

Spectral Collapse. To mathematically explain the sharp decay and to estimate the condition number $\kappa = \frac{\sigma_{\max}}{\sigma_{\min}}$, we can decompose the Hadamard product into latent factors and apply Random Matrix Theory.

Here is the derivation showing why the Hadamard product naturally destroys well-conditioned spectra, along with the precise estimation of the decay.

Proposition 4.7 (Spectral Collapse (Fig. 3, proven in Appendix A.5))**.** *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{N \times N}$ be two generic, independent well-conditioned rank-r matrices ($r^2 \leq N$). The condition number κ of $\mathbf{A} \odot \mathbf{B}$ is approximately $\left(\frac{1+r/\sqrt{N}}{1-r/\sqrt{N}}\right)^2$.*

4.3 Illustrative Pattern Comparison of Conventional MF and Hadamard Decomposition

Based on the above discussion, we conclude that, unlike conventional MF, which preserves low-rank information accurately, HD reconstructs a full-rank matrix with very high probability. Meanwhile, the singular-value spectrum of its reconstructed matrix suffers from distortion relative to that of the original matrix, namely energy dilution and spectral collapse, as elaborated in the preceding subsections. We give an illustration of the patterns of conventional MF and HD in Fig. 4.

5 Rank-Adaptive MF Unifying Conventional MF and HD

In previous sections, we have known that MF suffers from the low-rank bottleneck, whilst HD suffers from the almost fixed high rank and distortion of the spectrum. Neither MF nor HD alone deals well with all the cases from low-rank to high-rank. To make the matrix compression algorithm rank-adaptive, the basic idea is to adopt the factorization formula named as **Rada-MF** that unifies both MF and HD:

$$\mathbf{M} \approx \mathbf{X}_{11}^\top \mathbf{Y}_{11} + (\mathbf{X}_{21}^\top \mathbf{Y}_{21}) \odot (\mathbf{X}_{22}^\top \mathbf{Y}_{22}) + \cdots = \sum_{l=1}^{L} \bigodot_{b=1}^{l} \mathbf{X}_{lb}^\top \mathbf{Y}_{lb}, \tag{8}$$

where $\mathbf{X}_{11}^\top \mathbf{Y}_{11}$ is the MF term and the following terms are the HD terms of different orders. Eq. 8 can also be written in the pure MF form with concatenation and tensor product, as illustrated in Fig. 1. When $L = 1$, Rada-MF degenerates to conventional MF; and when $L = 2$ with no budget for $\mathbf{X}_{11}$ and $\mathbf{Y}_{11}$, Rada-MF degenerates to the standard second-order HD.

The core philosophy of Rada-MF is to dynamically select the factorization strategy based on the sufficiency of the dimension budget d. We partition the problem space into three regimes:

- **MF Regime**: With a large d, the capacity of MF is sufficient to model the principal components directly. We employ standard MF (e.g., Truncated SVD) to minimize reconstruction error without the risk of spectral collapse associated with HD.
- **HD Regime**: When the budget is severely limited, standard MF fails to capture the high-rank nature of the data. We employ pure HD to leverage its exponential rank amplification capability (rank $\sim r^R$), prioritizing structural expressivity over spectral precision.
- **Hybrid Regime**: In the intermediate regime, we adopt the hybrid MF&HD framework. This approach decomposes $\mathbf{M}$ into a sum of a low-rank component $\mathbf{M}_{\text{low}}$ and a high-rank residual component $\mathbf{M}_{\text{high}}$, balancing stability and expressivity.

This strategy ensures that we leverage the stability of MF when resources are abundant, exploit the rank amplification of Hadamard products when resources are scarce, and balance both in intermediate scenarios.

5.1 Model Selection via the Effective Rank

Given the singular values of $\mathbf{M}$, we use the entropy effective rank $r_{\text{eff}} = \exp(-\sum_i p_i \ln p_i)$ with $p_i = \sigma_i^2 / \sum_j \sigma_j^2$ as a continuous measure of spectral complexity. Rada-MF-auto compares the budget D with r_{eff} only: $D > r_{\text{eff}}$ returns the analytic truncated SVD; $r_{\text{eff}} > 2D$ returns pure HD; the intermediate band returns an additive hybrid that splits the budget between an SVD core of rank $d_{\text{mf}} \in \{3D/4,\ D/2,\ D/4\}$ (by deficit severity) and a second-order HD term fitted on the residual, with $B{=}2$ fixed throughout. At $D{=}32$ the three splits give the $(24, 4{+}4)$, $(16, 8{+}8)$ and $(8, 12{+}12)$ configurations of Rada-MF-1/2/3. The rule costs one SVD of $\mathbf{M}$ and involves no per-dataset tuning.

6 Experiments

Setup. When compared with MF baselines we fix the total budget D so that all models share the same number of trainable parameters ($(M{+}N)D$; $D = 32$, i.e., 8192 parameters at $M{=}N{=}128$). The MF baseline is the analytic rank-D truncated SVD computed in float64 on CPU; the HD baseline is the official scaled alternating gradient descent implementation (Ciaperoni et al., 2024) run on CPU (NumPy/Numba) with learning rate 0.01 and $T = 10000$ iterations ($T = 100000$ is used only in the sensitivity check of Appendix B); its occasional "loss is increasing – halving learning rate" messages are part of the official adaptive step-size rule. The hybrid models combine an analytic SVD core with the official HD procedure applied to the residual. Application experiments (Sec. 6.2) train the HD branches with Adam. **Hardware:** an Intel Xeon Silver 4210R CPU and NVIDIA RTX 2080 Ti GPUs; official HD runs on CPU.

6.1 Real-Valued Data Compression

Training Objective. We use the relative root square error (RRSE) loss as the training objective. It is defined as $J_{RRSE}(\widehat{\mathbf{M}}, \mathbf{M}) = \frac{\|\mathbf{M}-\widehat{\mathbf{M}}\|_F}{\|\mathbf{M}\|_F}$.

Data (controlled spectra). We generate $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ with $\mathbf{U}, \mathbf{V}$ independent random orthogonal matrices ($M{=}N{=}128$). Instead of fixing the algebraic rank, we control the entropy effective rank r_{eff} (Sec. 5) by numerically inverting the spectrum parameters of three families: exponential ($\sigma_i \propto e^{-ai}$), power-law ($\sigma_i \propto i^{-p}$), and head-tail ($\sigma_1{=}\ldots{=}\sigma_8$ followed by a power-law tail). Targets are $r_{\text{eff}} \in \{8, 24, 32, 36, 40, 44, 48, 64, 80, 96\}$; each (family, target) pair uses 30 independent seeds (900 matrices in total), each normalized to $\|\mathbf{M}\|_F = 1$, with $|\text{achieved} - \text{target}| < 10^{-13}$. **Models.** All models share budget $D = 32$ (8192 parameters): MF-SVD (analytic truncated SVD), HD2 (official scaled alternating gradient descent), Rada-MF-1/2/3 (SVD cores of rank 24/16/8 plus HD residuals with factors $4{+}4/8{+}8/12{+}12$), and Rada-MF-auto (selection rule in Sec. 5.1).

Result analysis. Table 2 and Fig. 5 report the mean RRSE over 30 seeds on the controlled spectra. (1) Low complexity ($r_{\text{eff}} < 32$): truncated SVD is the best model on the regime mean and in five of the six cells (the exception, power-law at $r_{\text{eff}} = 24$, is won by Rada-MF-3 by 0.012); on the rank-8 head-tail spectra it reaches machine precision, while HD2 incurs a 0.24–0.33 error across the $r_{\text{eff}} = 8$ cells for its fixed high-rank profile. (2) Intermediate ($32 \le r_{\text{eff}} \le 64$): the best fixed hybrid improves over both pure models on the power-law and head-tail spectra (by up to 0.05 RRSE at $r_{\text{eff}} = 40$–48, shrinking to parity at $r_{\text{eff}} = 64$), whereas on fast-decaying exponential spectra SVD

Table 2: Mean RRSE over 30 seeds on controlled spectra ($M{=}N{=}128$, budget $D{=}32$, all models use 8192 parameters). Targets $r_{\text{eff}} \in \{8, 24, 32, 36, 40, 44, 48, 64, 80, 96\}$ are achieved within $|\Delta r_{\text{eff}}| < 10^{-13}$. MF-SVD is the analytic rank-32 truncated SVD (float64); HD2 is the official scaled alternating gradient descent (two rank-16 factors, lr 0.01, $T{=}10000$, CPU); Rada-MF-1/2/3 are analytic SVD cores of rank 24/16/8 plus official HD on the residual with factors $4{+}4/8{+}8/12{+}12$; Rada-MF-auto follows the selection rule of Sec. 5.1. Bold marks the best model per row.

r_{eff}	Spectrum	MF-SVD	HD2	Rada-MF-1	Rada-MF-2	Rada-MF-3	Rada-MF-auto	Selected
8	exponential	**0.0042**	0.2920	0.0133	0.0374	0.0989	**0.0042**	MF
8	head-tail	**0.0000**	0.2395	0.0005	0.0017	0.0030	**0.0000**	MF
8	power-law	**0.1906**	0.3328	0.1968	0.2066	0.2265	**0.1906**	MF
24	exponential	**0.1631**	0.4330	0.1914	0.2476	0.3274	**0.1631**	MF
24	head-tail	**0.2785**	0.4179	0.2810	0.2803	0.3184	**0.2785**	MF
24	power-law	0.3724	0.4374	0.3730	0.3606	**0.3605**	0.3724	MF
32	exponential	**0.2568**	0.4565	0.2782	0.3270	0.3884	0.2782	Rada-MF
32	head-tail	0.3429	0.4450	0.3440	**0.3387**	0.3733	0.3440	Rada-MF
32	power-law	0.4258	0.4598	0.4233	0.4066	**0.3959**	0.4233	Rada-MF
36	exponential	**0.2988**	0.4633	0.3169	0.3594	0.4103	0.3169	Rada-MF
36	head-tail	0.3869	0.4548	0.3856	**0.3701**	0.3893	0.3856	Rada-MF
36	power-law	0.4486	0.4657	0.4456	0.4239	**0.4113**	0.4456	Rada-MF
40	exponential	**0.3374**	0.4720	0.3531	0.3893	0.4308	0.3531	Rada-MF
40	head-tail	0.4087	0.4641	0.4074	**0.3907**	0.4077	0.4074	Rada-MF
40	power-law	0.4697	0.4733	0.4666	0.4412	**0.4229**	0.4666	Rada-MF
44	exponential	**0.3729**	0.4790	0.3853	0.4136	0.4487	0.4136	Rada-MF
44	head-tail	0.4657	0.4707	0.4595	0.4274	**0.4142**	0.4274	Rada-MF
44	power-law	0.4893	0.4760	0.4845	0.4574	**0.4352**	0.4574	Rada-MF
48	exponential	**0.4055**	0.4852	0.4153	0.4357	0.4621	0.4357	Rada-MF
48	head-tail	0.4491	0.4791	0.4461	**0.4264**	0.4407	**0.4264**	Rada-MF
48	power-law	0.5077	0.4788	0.5019	0.4713	**0.4456**	0.4713	Rada-MF
64	exponential	0.5143	**0.4928**	0.5132	0.5101	0.5080	**0.4928**	HD
64	head-tail	0.5299	0.4893	0.5234	0.4933	**0.4891**	**0.4891**	Rada-MF
64	power-law	0.5733	0.4939	0.5645	0.5245	**0.4832**	0.4939	HD
80	exponential	0.6004	**0.5035**	0.5932	0.5682	0.5356	**0.5035**	HD
80	head-tail	0.6138	**0.5012**	0.6019	0.5502	0.5099	**0.5012**	HD
80	power-law	0.6315	**0.5010**	0.6204	0.5679	0.5133	**0.5010**	HD
96	exponential	0.6751	**0.5091**	0.6623	0.6154	0.5616	**0.5091**	HD
96	head-tail	0.6684	**0.5068**	0.6556	0.5987	0.5476	**0.5068**	HD
96	power-law	0.6879	**0.5048**	0.6734	0.6107	0.5422	**0.5048**	HD

remains the best fixed model until $r_{\text{eff}} \approx 60$; across all intermediate cells Rada-MF-auto attains the lowest regime mean (0.418 vs 0.427 for SVD and 0.472 for HD2). (3) High complexity ($r_{\text{eff}} > 64$): HD2 is clearly better (0.504 vs 0.646 for SVD), and Rada-MF-auto selects the HD branch there. (4) Overall, Rada-MF-auto tracks the per-matrix oracle with mean gap 0.0121, median 0.0079 and maximum 0.0572. We emphasize that r_{eff} is a spectral-complexity indicator: these results support spectrum-dependent model-selection trends, not global optimality of any single form.

6.2 Binary-Valued Data Compression

The biggest difference between the binary-valued data and other types of data is that the compression of this type of data does not need to accurately retain the numerical values of "0" or "1", but only needs to distinguish between the data items of "0" and "1".

Training objective and settings. Apart from the reconstruction loss (RRSE), we further adopt a sampled Bayesian Personalized Ranking (BPR) objective: at each step we sample up to 512 positive and 512 negative entries and minimize the softplus surrogate $\mathbb{E}\big[\log \sigma\big(\widehat{\mathbf{M}}_{pos} - \widehat{\mathbf{M}}_{neg}\big)\big]$ over random (pos, neg) pairs. The BPR variant optimizes this ranking objective alone; the RRSE variant optimizes reconstruction alone. **Data.** We consider $\mathbf{M}$ to be an asymmetric binary-valued matrix, with each element drawn independently from a Bernoulli distribution with parameter p, which controls the density of the binary matrices. We vary p from sparse to dense, in the range $\{0.02, 0.04, 0.08, 0.1, 0.2, 0.3\}$. **Metric.** Apart from RRSE, we also report AUC (area under the curve) and ACC

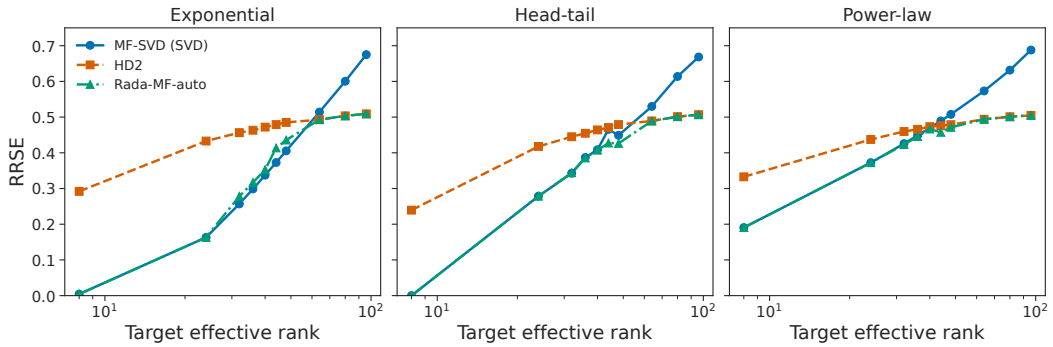


Figure 5: Mean RRSE versus target effective rank on the controlled spectra ($M{=}N{=}128$, budget $D{=}32$; 900 matrices, 30 seeds per point, all models use 8192 parameters). Each panel is one spectrum family; only three budget-matched models are shown for readability. Truncated SVD dominates at low r_{eff}, HD2 at high r_{eff}, and Rada-MF-auto tracks the better of the two throughout while attaining the lowest error in the intermediate band (most visibly on head-tail at $r_{\text{eff}}{=}44$–48).

Table 3: Binary-valued compression under the current protocol (128×128 Bernoulli targets, 4 matrices per density, seed 0; accuracy uses the balanced-count threshold; AUC is the primary metric). Rada-MF-auto computes the entropy effective rank (which varies with density over this range) and follows Sec. 5.1: the HD branch for $D \in \{8, 16\}$ and HD or the hybrid for $D \in \{24, 32\}$. Bold marks the best accuracy and the best AUC per (density, D).

Density	Method	$D=8$		$D=16$		$D=24$		$D=32$	
		Acc ↑	AUC ↑	Acc ↑	AUC ↑	Acc ↑	AUC ↑	Acc ↑	AUC ↑
$p=0.1$	SVD	0.8386	0.5922	0.8493	0.6178	0.8561	0.6246	0.8609	0.6283
	HD2 (RRSE)	**0.9034**	0.8599	**0.9651**	0.9684	0.9948	0.9982	0.9999	0.9999
	HD2 (BPR)	0.9031	**0.9103**	0.9549	**0.9811**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
	Rada-MF-auto (RRSE)	**0.9034**	0.8599	**0.9651**	0.9684	0.9948	0.9982	0.9999	0.9999
	Rada-MF-auto (BPR)	0.9031	**0.9103**	0.9549	**0.9811**	**1.0000**	**1.0000**	**1.0000**	**1.0000**
$p=0.2$	SVD	0.7124	0.5798	0.7280	0.6049	0.7413	0.6182	0.7512	0.6261
	HD2 (RRSE)	0.8048	0.7939	0.8942	0.9178	**0.9550**	**0.9774**	0.9890	0.9973
	HD2 (BPR)	**0.8103**	**0.8246**	**0.8979**	**0.9440**	0.9327	0.9729	**1.0000**	**1.0000**
	Rada-MF-auto (RRSE)	0.8048	0.7939	0.8942	0.9178	**0.9550**	**0.9774**	0.9862	0.9982
	Rada-MF-auto (BPR)	**0.8103**	**0.8246**	**0.8979**	**0.9440**	0.9327	0.9729	1.0000	1.0000
$p=0.3$	SVD	0.6184	0.5687	0.6384	0.5953	0.6528	0.6103	0.6649	0.6191
	HD2 (RRSE)	0.7303	0.7604	0.8274	0.8772	0.9061	0.9515	0.9594	0.9870
	HD2 (BPR)	**0.7350**	**0.7816**	**0.8427**	**0.9102**	0.9078	0.9630	0.9045	0.9541
	Rada-MF-auto (RRSE)	0.7303	0.7604	0.8274	0.8772	0.9130	0.9673	0.9453	0.9853
	Rada-MF-auto (BPR)	**0.7350**	**0.7816**	**0.8427**	**0.9102**	**0.9193**	**0.9728**	**0.9981**	**0.9999**

(classification accuracy) by treating the problem as a binary classification problem to measure how well the positive and negative terms are distinguished in the reconstructed $\widehat{\mathbf{M}}$.

Result analysis. Results are given in Table 3, showing the superiority of Rada-MF in preserving information of binary-valued data matrices. The utilization of BPR loss further boosts Rada-MF's capability of compressing the binary data when D is small.

7 Conclusions and Limitations

We analyzed the rank amplification and spectral distortion of Hadamard products and proposed Rada-MF, a rank-adaptive selection among MF, HD, and their additive hybrid under a fixed parameter budget. Controlled experiments over 900 spectra and applications on binary matrices show that Rada-MF-auto approaches the oracle of its fixed candidates.

Limitations and outlook. The rank-amplification and energy-dilution laws are proven and empirically verified; the condition-number scaling is kept as a Marchenko–Pastur reference. The regime framework invites further exploration of its advantages: adaptive allocation and order selection within the hybrid band, longer HD optimization, solver-stability effects, and broader applications.

AI USE STATEMENT

In this work, we used generative AI tools for implementing, optimizing, and running the experimental code. We have not used generative AI tools for cleaning and reformatting the dataset or proposing hypotheses. Additionally, we used generative AI tools for editing the research paper to improve readability. We have reviewed all AI-assisted work. Code and textual edits generated by AI were reviewed and verified by the authors against the archived experimental outputs. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Martino Ciaperoni, Aristides Gionis, and Heikki Mannila. The hadamard decomposition problem. Data Mining and Knowledge Discovery, 38(4):2306–2347, 2024. doi: 10.1007/s10618-024-01033-y. URL https://doi.org/10.1007/s10618-024-01033-y.

Zhi Ding, Tianyao Ji, and Mengshi Li. Improved hadamard decomposition and its application in data compression in new-type power systems. Mathematics, 13(4):671, 2025.

Ke-Lin Du, MNS Swamy, Zhang-Quan Wang, and Wai Ho Mow. Matrix factorization techniques in machine learning, signal processing, and statistics. Mathematics, 11(12):2674, 2023.

Nicolas Gillis. Nonnegative matrix factorization. SIAM, 2020.

Nicolas Gillis, Subhayan Saha, Stefano Sicilia, and Arnaud Vandaele. Manifold-based algorithms for the hadamard decomposition. arXiv preprint arXiv:2605.28980, 2026.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. nature, 401(6755):788–791, 1999.

Samuel Wertz, Arnaud Vandaele, and Nicolas Gillis. Efficient algorithms for the hadamard decomposition. In 2025 IEEE 35th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1–6. IEEE, 2025.

Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. Chemometrics and Intelligent Laboratory Systems, 2(1):37–52, 1987. ISSN 0169-7439. doi: https://doi.org/10.1016/0169-7439(87)80084-9. URL https://www.sciencedirect.com/science/article/pii/0169743987800849. Proceedings of the Multivariate Statistical Workshop for Geologists and Geochemists.

Hao Xiong, Yehui Tang, Yunlin He, Wei Tan, and Junchi Yan. Node2ket: Efficient high-dimensional network embedding in quantum hilbert space. In The Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=lROh08eK6n.

Appendix

CONTENTS

A Proofs **12**

A.1 Proof of Proposition 4.2 12

A.2 Proof of Proposition 4.3 13

A.3 Proof of Lemma 4.4 13

A.4 Proof of Proposition 4.5 14

A.5 Proof of Proposition 4.7 15

B Controlled-Spectrum Experiments: Additional Results **16**

C Spectral Evidence: Additional Statistics **19**

D Ablations under the Controlled Protocol **20**

E Singular-Value Dynamics of the HD Branch **22**

A PROOFS

A.1 PROOF OF PROPOSITION 4.2

Proposition A.1 (Decomposition into Rank-One Terms)**.** $\mathbf{A} \odot \mathbf{B}$ *can be represented as the sum of* r^2 *rank-1 matrices.*

Proof. Let $\mathbf{A} = \mathbf{U}\mathbf{V}^\top$ and $\mathbf{B} = \tilde{\mathbf{U}}\tilde{\mathbf{V}}^\top$ be rank decompositions, where $\mathbf{U}, \tilde{\mathbf{U}} \in \mathbb{R}^{M\times r}$ and $\mathbf{V}, \tilde{\mathbf{V}} \in \mathbb{R}^{N\times r}$; with the notation of Eq. 4, $\mathbf{U} = \mathbf{X}_b^\top$ and $\mathbf{V} = \mathbf{Y}_b^\top$.

Denote the columns of $\mathbf{U}$ by $\mathbf{u}_1, \ldots, \mathbf{u}_r \in \mathbb{R}^M$, the columns of $\tilde{\mathbf{U}}$ by $\tilde{\mathbf{u}}_1, \ldots, \tilde{\mathbf{u}}_r \in \mathbb{R}^M$, and similarly for $\mathbf{V}, \tilde{\mathbf{V}}$ with columns in $\mathbb{R}^N$.

The (i, j)-entry of $\mathbf{A} \odot \mathbf{B}$ is:

$$[\mathbf{A}\odot\mathbf{B}]_{ij} = [\mathbf{A}]_{ij}[\mathbf{B}]_{ij} = \Big(\sum_{k=1}^{r}[\mathbf{U}]_{ik}[\mathbf{V}]_{jk}\Big)\Big(\sum_{l=1}^{r}[\tilde{\mathbf{U}}]_{il}[\tilde{\mathbf{V}}]_{jl}\Big) = \sum_{k=1}^{r}\sum_{l=1}^{r}([\mathbf{U}]_{ik}\,[\tilde{\mathbf{U}}]_{il})\,([\mathbf{V}]_{jk}\,[\tilde{\mathbf{V}}]_{jl}). \tag{9}$$

For each pair (k, l) with $1 \le k \le r$ and $1 \le l \le r$, we define $\mathbf{w}_{kl} = \mathbf{u}_k \odot \tilde{\mathbf{u}}_l \in \mathbb{R}^M$, and $\mathbf{z}_{kl} = \mathbf{v}_k \odot \tilde{\mathbf{v}}_l \in \mathbb{R}^N$.

Then

$$\mathbf{A} \odot \mathbf{B} = \sum_{k=1}^{r}\sum_{l=1}^{r}\mathbf{w}_{kl}\,\mathbf{z}_{kl}^\top, \tag{10}$$

which proves that $\mathbf{A} \odot \mathbf{B}$ is a sum of r^2 rank-one matrices.

□

A.2 Proof of Proposition 4.3

Proposition A.2 (Full Column Rank of the Factor Matrices)**.** *Define the matrices* $\mathbf{W} = [\mathbf{w}_{11}, \mathbf{w}_{12}, \ldots, \mathbf{w}_{rr}] \in \mathbb{R}^{M \times r^2}$ *and* $\mathbf{Z} = [\mathbf{z}_{11}, \mathbf{z}_{12}, \ldots, \mathbf{z}_{rr}] \in \mathbb{R}^{N \times r^2}$*. If* $M \geq r^2$ *and* $\mathbf{U}, \tilde{\mathbf{U}}$ *are generic, then* $rank(\mathbf{W}) = r^2$*; likewise, if* $N \geq r^2$ *and* $\mathbf{V}, \tilde{\mathbf{V}}$ *are generic, then* $rank(\mathbf{Z}) = r^2$*.*

Proof. We prove the claim for $\mathbf{W}$; the claim for $\mathbf{Z}$ follows by the same argument with $(\mathbf{U}, \tilde{\mathbf{U}})$ replaced by $(\mathbf{V}, \tilde{\mathbf{V}})$ and M by N.

Suppose $\sum_{k,l} c_{kl}\, \mathbf{w}_{kl} = \mathbf{0}$ for some coefficients c_{kl}. Let $\mathbf{C} = (c_{kl}) \in \mathbb{R}^{r \times r}$.

Let $\mathbf{x}_i^\top$ be the i-th row of $\mathbf{U}$ and $\tilde{\mathbf{x}}_i^\top$ the i-th row of $\tilde{\mathbf{U}}$. The i-th component of the linear combination is

$$\Big[\sum_{k,l} c_{kl}\, \mathbf{w}_{kl}\Big]_i = \sum_{k=1}^{r}\sum_{l=1}^{r} c_{kl}[\mathbf{x}_i]_k[\tilde{\mathbf{x}}_i]_l = \mathbf{x}_i^\top \mathbf{C}\, \tilde{\mathbf{x}}_i = 0 \qquad \text{for all } i = 1, \ldots, M. \tag{11}$$

By the cyclic property of the trace,

$$\mathbf{x}_i^\top \mathbf{C}\, \tilde{\mathbf{x}}_i = \mathrm{Tr}(\mathbf{x}_i^\top \mathbf{C}\, \tilde{\mathbf{x}}_i) = \mathrm{Tr}(\mathbf{C}\, \tilde{\mathbf{x}}_i\, \mathbf{x}_i^\top) = \langle \mathbf{C},\, \mathbf{X}_i \rangle_F, \tag{12}$$

where $\mathbf{X}_i = \tilde{\mathbf{x}}_i\, \mathbf{x}_i^\top \in \mathbb{R}^{r \times r}$ is a rank-one matrix and $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product. Thus

$$\langle \mathbf{C},\, \mathbf{X}_i \rangle_F = 0 \qquad \text{for all } i = 1, \ldots, M. \tag{13}$$

The matrices $\mathbf{X}_i$ live in $\mathbb{R}^{r \times r}$, a space of dimension r^2. Since $\mathbf{U}$ and $\tilde{\mathbf{U}}$ are generic, their rows are in general position, so the M matrices $\mathbf{X}_1, \ldots, \mathbf{X}_M$ span the entire r^2-dimensional space with probability 1 (using $M \geq r^2$).

Eq. 13 then forces $\mathbf{C} = \mathbf{0}$, i.e., all $c_{kl} = 0$. Hence the columns of $\mathbf{W}$ are linearly independent (and identically for $\mathbf{Z}$). □

A.3 Proof of Lemma 4.4

Lemma A.3 (Rank of column-wise Kronecker product)**.** *Let* $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K] \in \mathbb{R}^{d_1 \times K}$ *and* $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_K] \in \mathbb{R}^{d_2 \times K}$*. If* $rank(\mathbf{W}) = K$ *and* $rank(\mathbf{Z}) = K$*, then the matrix* $[\mathbf{z}_1 \otimes \mathbf{w}_1, \ldots, \mathbf{z}_K \otimes \mathbf{w}_K] \in \mathbb{R}^{d_1 d_2 \times K}$ *has rank* K*.*

Proof of Lemma 4.4. Suppose $\sum_{k=1}^{K} c_k(z_k \otimes w_k) = 0$. Reshaping each $z_k \otimes w_k$ into the $d_1 \times d_2$ matrix $w_k z_k^\top$, this is equivalent to

$$\sum_{k=1}^{K} c_k\, \mathbf{w}_k \mathbf{z}_k^\top = 0.$$

Writing $\mathbf{D} = diag(c_1, \ldots, c_K)$, this becomes $\mathbf{W}\mathbf{D}\mathbf{Z}^\top = \mathbf{0}$.

Since $\mathbf{W}$ has full column rank, its kernel is trivial, so $\mathbf{W}(\mathbf{D}\mathbf{Z}^\top) = \mathbf{0}$ implies $\mathbf{D}\mathbf{Z}^\top = \mathbf{0}$.

Since $\mathbf{Z}$ has full column rank, $\mathbf{Z}^\top$ has full row rank, so $\mathbf{Z}^\top$ is surjective. Then $\mathbf{D}(\mathbf{Z}^\top \mathbf{x}) = \mathbf{0}$ for all $\mathbf{x}$ implies $\mathbf{D} = \mathbf{0}$, i.e., $c_1 = \cdots = c_K = 0$. □

A.4 Proof of Proposition 4.5

Proposition A.4 (Spectral Energy Dilution). *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{N \times N}$ be two generic, independent rank-r matrices ($r \ll N$) with perfectly conditioned, non-zero singular values $\sigma_i(\mathbf{A}) = \sigma_i(\mathbf{B}) = 1$. As $N \to \infty$, the expected total spectral energy (Frobenius norm squared) of their Hadamard product $\mathbf{A} \odot \mathbf{B}$ decays asymptotically as $\mathcal{O}(r^2/N^2)$, and the average squared singular value $\bar{\sigma}^2$ collapses to $\mathcal{O}(1/N^2)$.*

Proof. Let $\mathbf{A}$ and $\mathbf{B}$ be decomposed via their truncated Singular Value Decompositions (SVD) as $\mathbf{A} = \mathbf{U}_A \mathbf{V}_A^\top$ and $\mathbf{B} = \mathbf{U}_B \mathbf{V}_B^\top$, where the latent basis matrices $\mathbf{U}_A, \mathbf{V}_A, \mathbf{U}_B, \mathbf{V}_B \in \mathbb{R}^{N \times r}$ possess orthonormal columns and are mutually independent. Since the singular values are identically 1, the total spectral energy of both $\mathbf{A}$ and $\mathbf{B}$ is exactly r:

$$\mathbb{E}\big[\|\mathbf{A}\|_F^2\big] = \text{Tr}(\mathbf{U}_A \mathbf{V}_A^\top \mathbf{V}_A \mathbf{U}_A^\top) = \text{Tr}(\mathbf{U}_A \mathbf{U}_A^\top) = r$$

Assuming generic, unstructured bases (e.g., representing complex, dense real-world data), the matrices $\mathbf{U}_A, \mathbf{V}_A$ can be modeled as random matrices whose entries are asymptotically distributed with mean 0 and variance $\frac{1}{N}$. The entries of $\mathbf{A}$ are given by the inner product of the rows: $\mathbf{A}_{ij} = \sum_{k=1}^{r} [\mathbf{U}_A]_{ik} [\mathbf{V}_A]_{jk}$.

Because $\mathbf{U}_A$ and $\mathbf{V}_A$ are independent, the expected value of any entry $\mathbf{A}_{ij}$ is 0, and its variance is strictly governed by the sum of the variances of the products:

$$\mathbb{E}\big[\mathbf{A}_{ij}^2\big] = \text{Var}(\mathbf{A}_{ij}) = \sum_{k=1}^{r} \text{Var}\big([\mathbf{U}_A]_{ik}\big)\text{Var}\big([\mathbf{V}_A]_{jk}\big) \approx r \left(\frac{1}{N}\right)\left(\frac{1}{N}\right) = \frac{r}{N^2}$$

Now, consider the Hadamard product $\mathbf{A} \odot \mathbf{B}$ where $[\mathbf{A} \odot \mathbf{B}]_{ij} = \mathbf{A}_{ij}\mathbf{B}_{ij}$. Because $\mathbf{A}$ and $\mathbf{B}$ are drawn independently, the variance of the entries in $\widehat{\mathbf{M}}$ attenuates multiplicatively:

$$\mathbb{E}\big[[\mathbf{A} \odot \mathbf{B}]_{ij}^2\big] = \mathbb{E}\big[\mathbf{A}_{ij}^2\big]\mathbb{E}\big[\mathbf{B}_{ij}^2\big] \approx \left(\frac{r}{N^2}\right)\left(\frac{r}{N^2}\right) = \frac{r^2}{N^4}. \tag{14}$$

To find the expected total spectral energy of the Hadamard product, we sum the expected squared magnitudes over all N^2 entries:

$$\mathbb{E}\big[\|\mathbf{A} \odot \mathbf{B}\|_F^2\big] = \sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}\big[[\mathbf{A} \odot \mathbf{B}]_{ij}^2\big] \approx N^2 \left(\frac{r^2}{N^4}\right) = \frac{r^2}{N^2}. \tag{15}$$

While the rank of $\mathbf{A} \odot \mathbf{B}$ expands to r^2 (via the combinatorial expansion of the Khatri-Rao product), the total Frobenius energy collapses to $\mathcal{O}(r^2/N^2)$. Since the Frobenius energy is equivalent to the sum of the squared singular values ($\|\mathbf{M}\|_F^2 = \sum_{i=1}^{r^2} \sigma_i^2$), the average squared singular value $\bar{\sigma}^2$ of the resulting matrix of Hadamard Product is strictly bounded:

$$\bar{\sigma}^2 \approx \frac{\mathbb{E}[\|\mathbf{A} \odot \mathbf{B}\|_F^2]}{\text{rank}(\mathbf{A} \odot \mathbf{B})} = \frac{r^2/N^2}{r^2} = \frac{1}{N^2}. \tag{16}$$

This concludes the proof. □

A.5 Proof of Proposition 4.7

Theorem A.5 (Spectral Collapse). *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{N \times N}$ be two generic, independent well-conditioned rank-r matrices ($r^2 \leq N$). The condition number κ of $\mathbf{A} \odot \mathbf{B}$ is approximately $\left(\frac{1+r/\sqrt{N}}{1-r/\sqrt{N}}\right)^2$.*

Proof. **1. The Latent Factor Decomposition (a review of Sec. 3.3)**

Assume $\mathbf{A}$ and $\mathbf{B}$ are well-conditioned rank-r matrices ($r^2 < N$). We can write their truncated SVDs as:

$$\mathbf{A} = \mathbf{U}_A \mathbf{V}_A^\top \quad \text{and} \quad \mathbf{B} = \mathbf{U}_B \mathbf{V}_B^\top, \tag{17}$$

where $\mathbf{U}_A, \mathbf{V}_A, \mathbf{U}_B, \mathbf{V}_B \in \mathbb{R}^{N \times r}$ have orthonormal columns. Let $\star$ denote the row-wise Khatri–Rao product, i.e., the rows of $\mathbf{U}_A \star \mathbf{U}_B$ collect all pairwise products of the corresponding rows of $\mathbf{U}_A$ and $\mathbf{U}_B$. The Hadamard product can then be rewritten as

$$\mathbf{P} = \mathbf{U}_A \star \mathbf{U}_B \in \mathbb{R}^{N \times r^2}, \quad \mathbf{Q} = \mathbf{V}_A \star \mathbf{V}_B \in \mathbb{R}^{N \times r^2}, \tag{18}$$

which allows us to express the Hadamard product as a standard matrix product:

$$\mathbf{A} \odot \mathbf{B} = \mathbf{P}\mathbf{Q}^\top. \tag{19}$$

2. The Marchenko-Pastur "Spread"

Even though $\mathbf{U}_A$ and $\mathbf{U}_B$ are perfectly conditioned (singular values all exactly 1), their entry-wise product $\mathbf{P}$ behaves like a random matrix.

Because the columns of $\mathbf{U}_A$ are orthonormal, their entries roughly follow a distribution with zero mean and variance $\frac{1}{N}$. Consequently, the entries of $\mathbf{P}$ have a variance of roughly $\frac{1}{N^2}$.

According to the Marchenko-Pastur Law, the singular values of a generic $N \times r^2$ random matrix are not flat; they spread out into a specific distribution governed entirely by the aspect ratio $c = \frac{r^2}{N}$. For the matrix $\mathbf{P}$, the singular values are bounded by:

$$\sigma_{\max}(\mathbf{P}) \approx \frac{1}{\sqrt{N}}(1 + \sqrt{c}), \quad \sigma_{\min}(\mathbf{P}) \approx \frac{1}{\sqrt{N}}(1 - \sqrt{c}). \tag{20}$$

3. Estimating the Decay (The Condition Number)

Because $\mathbf{A} \odot \mathbf{B} = \mathbf{P}\mathbf{Q}^\top$, and $\mathbf{P}$ and $\mathbf{Q}$ are drawn independently, the singular values of $\mathbf{A} \odot \mathbf{B}$ are formed by the multiplicative interaction of the singular values of $\mathbf{P}$ and $\mathbf{Q}$. We can estimate the maximum and minimum non-zero singular values of $\mathbf{A} \odot \mathbf{B} = \mathbf{P}\mathbf{Q}^\top$ as:

$$\begin{aligned} \sigma_{\max}(\mathbf{A} \odot \mathbf{B}) &\approx \sigma_{\max}(\mathbf{P})\sigma_{\max}(\mathbf{Q}) \approx \frac{1}{N}(1 + \sqrt{c})^2 \\ \sigma_{\min}(\mathbf{A} \odot \mathbf{B}) &\approx \sigma_{\min}(\mathbf{P})\sigma_{\min}(\mathbf{Q}) \approx \frac{1}{N}(1 - \sqrt{c})^2 \end{aligned} \tag{21}$$

This gives us the final estimation for the condition number (the severity of the decay):

$$\frac{\sigma_{\max}}{\sigma_{\min}} \approx \left(\frac{1 + \sqrt{c}}{1 - \sqrt{c}}\right)^2 \quad \text{where} \quad c = \frac{r^2}{N} \tag{22}$$

□

Table 4: Regime-level summary of the controlled experiments (mean RRSE, mean oracle gap, mean wall-clock time).

Regime	Model	RRSE	Oracle gap	Time (s)
MF ($r_{\text{eff}} < 32$)	MF-SVD	0.1681	0.0026	0.00
MF ($r_{\text{eff}} < 32$)	HD2	0.3588	0.1933	2.55
MF ($r_{\text{eff}} < 32$)	Rada-MF-1	0.1760	0.0105	1.53
MF ($r_{\text{eff}} < 32$)	Rada-MF-2	0.1890	0.0235	1.79
MF ($r_{\text{eff}} < 32$)	Rada-MF-3	0.2225	0.0570	2.28
MF ($r_{\text{eff}} < 32$)	Rada-MF-auto	0.1681	0.0026	0.00
Intermediate ($32 \leq r_{\text{eff}} \leq 64$)	MF-SVD	0.4268	0.0278	0.00
Intermediate ($32 \leq r_{\text{eff}} \leq 64$)	HD2	0.4722	0.0731	3.04
Intermediate ($32 \leq r_{\text{eff}} \leq 64$)	Rada-MF-1	0.4286	0.0295	2.38
Intermediate ($32 \leq r_{\text{eff}} \leq 64$)	Rada-MF-2	0.4226	0.0235	2.60
Intermediate ($32 \leq r_{\text{eff}} \leq 64$)	Rada-MF-3	0.4309	0.0319	3.01
Intermediate ($32 \leq r_{\text{eff}} \leq 64$)	Rada-MF-auto	0.4182	0.0192	2.59
HD ($r_{\text{eff}} > 64$)	MF-SVD	0.6462	0.1421	0.00
HD ($r_{\text{eff}} > 64$)	HD2	0.5044	0.0003	2.22
HD ($r_{\text{eff}} > 64$)	Rada-MF-1	0.6344	0.1303	1.54
HD ($r_{\text{eff}} > 64$)	Rada-MF-2	0.5852	0.0811	1.78
HD ($r_{\text{eff}} > 64$)	Rada-MF-3	0.5350	0.0309	2.28
HD ($r_{\text{eff}} > 64$)	Rada-MF-auto	0.5044	0.0003	2.22

Table 5: Branch counts of Rada-MF-auto per target effective rank (30 seeds per cell; at $r_{\text{eff}} = 64$ the achieved values straddle the $2D = 64$ boundary).

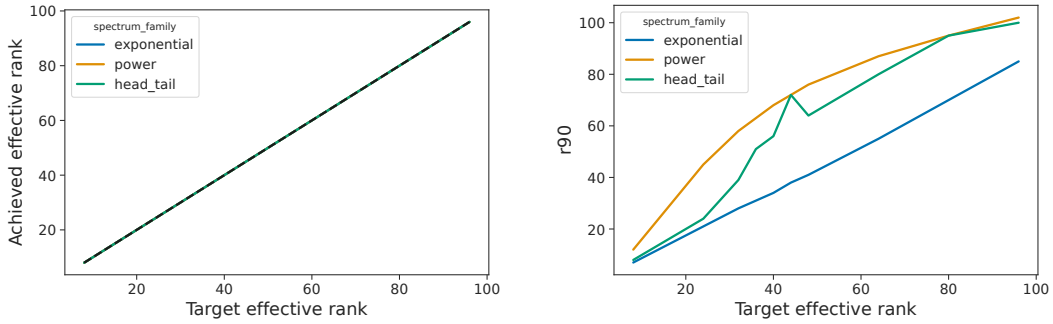
r_{eff}	Selected branch	Matrices
8	MF	90
24	MF	90
32	Rada-MF	90
36	Rada-MF	90
40	Rada-MF	90
44	Rada-MF	90
48	Rada-MF	90
64	HD	60
64	Rada-MF	30
80	HD	90
96	HD	90

B Controlled-Spectrum Experiments: Additional Results

This section collects supplementary results for the controlled experiments of Sec. 6.1. Table 4 summarizes the three regimes: truncated SVD is strongest at low complexity (mean RRSE 0.168 vs. 0.359 for HD2), Rada-MF-auto attains the lowest intermediate-regime mean (0.418 vs. 0.427 for SVD and 0.472 for HD2), and HD2 dominates the high-complexity regime (0.504 vs. 0.646); the analytic SVD costs negligible time while the official HD runs take 2–3 s per matrix. Table 5 reports the branch actually selected by Rada-MF-auto per target: MF for $r_{\text{eff}} \leq 24$, the hybrid for 32–48, and HD for $r_{\text{eff}} \geq 80$, with the $r_{\text{eff}} = 64$ cells straddling the $2D = 64$ boundary (60 HD vs. 30 hybrid). Table 6 checks the sensitivity of the official HD to the iteration budget: with $T = 100000$ instead of the main-protocol $T = 10000$, HD2 improves mildly (e.g., $0.309 \to 0.253$ at r_{eff}=8 exponential and $0.473 \to 0.447$ at r_{eff}=48 power-law), so the regime ordering is not an artifact of under-trained HD. Fig. 6 verifies the r_{eff} control itself: the achieved values track the targets, and r_{90} grows with the target as expected.

Table 6: Official-HD iteration sensitivity: rank-16+16 HD2 with $T = 10000$ (main protocol) vs. $T = 100000$ on representative spectra (seed 0).

Spectrum	r_{eff}	T	RRSE	Time (s)
exponential	8	10000	0.3087	10.2
exponential	8	100000	0.2533	20.4
power	48	10000	0.4734	3.4
power	48	100000	0.4472	20.2
head-tail	96	10000	0.5123	3.4
head-tail	96	100000	0.4758	20.3



(a) Achieved vs. target r_{eff}.

(b) r_{90} vs. target r_{eff}.

Figure 6: r_{eff} control and its relation to energy-truncation ranks: (a) the spectrum-parameter inversion hits every target within 10^{-13}; (b) r_{90} differs substantially from r_{eff}, confirming that r_{eff} is a spectral-complexity proxy rather than a truncation rank.

Table 7: Log-log regression of the Hadamard-product evidence (100 seeds per size; slopes with 95% CIs). r=16 has only two admissible sizes ($r^2 \leq N$), so no slope is fitted.

r	Quantity	Reference	Fitted slope	CI95	Sizes
2	$\|A \odot B\|_F^2$	N^{-2}	-2.009	[-2.027, -1.991]	64,128,256,512
2	mean nonzero σ	N^{-1}	-0.991	[-0.995, -0.987]	64,128,256,512
2	mean nonzero σ^2	N^{-2}	-2.009	[-2.027, -1.991]	64,128,256,512
8	$\|A \odot B\|_F^2$	N^{-2}	-2.002	[-2.007, -1.997]	64,128,256,512
8	mean nonzero σ	N^{-1}	-0.867	[-0.922, -0.812]	64,128,256,512
8	mean nonzero σ^2	N^{-2}	-2.002	[-2.007, -1.997]	64,128,256,512
16	$\|A \odot B\|_F^2$	N^{-2}	–	–	256,512
16	mean nonzero σ	N^{-1}	–	–	256,512
16	mean nonzero σ^2	N^{-2}	–	–	256,512

C Spectral Evidence: Additional Statistics

Table 7 reports the log–log regression slopes with 95% confidence intervals for the reference scalings of Proposition 4.5. For r=2 the fitted slopes are -2.009 (total energy, reference N^{-2}) and -0.991 (mean nonzero singular value, reference N^{-1}), matching the theory to the third digit. At r=8 the energy slope remains on reference (-2.002) while the mean-singular-value slope flattens to -0.867 over the available sizes, a finite-size deviation we report as-is rather than fit away; r=16 admits only $N \in \{256, 512\}$ under the constraint $r^2 \leq N$, so no slope is fitted for that column.

Table 8: Initialization strategies (mean RRSE over 10 seeds).

Initialization	normal	uniform	constant
Avg. RRSE	0.5330	0.5422	1.2407

Table 9: Seed stability (normal initialization, order 2).

Seed	0	1	2	3	4
Avg. RRSE	0.5385	0.5587	0.5154	0.5302	0.4984
Seed	5	6	7	8	9
Avg. RRSE	0.5521	0.5358	0.5529	0.5217	0.5264

Table 10: Loss functions (mean RRSE over 10 seeds). On the controlled targets $\|\mathbf{M}\|_F = 1$, so RSE and RRSE coincide up to normalization, and MSE differs only by scale.

Loss	RRSE	RSE	MSE
Avg. RRSE	0.5330	0.5337	10.1137

Table 11: Hadamard order (mean RRSE over 10 seeds): order 4 spends the same budget on four rank-8 factors and is consistently worse under this protocol.

Order R	2	4
Avg. RRSE	0.5330	0.8688

D Ablations under the Controlled Protocol

All ablations in this section use the controlled power-law target (N=64, r_{eff}=32), budget D=32, Adam with lr 0.01, 1200 iterations, and 10 seeds (0–9).

Table 8 shows that normal initialization is the strongest of the three strategies (0.533 vs. 0.542 for uniform, with constant initialization failing outright at 1.241), and Table 9 that the seed-to-seed spread is modest (per-seed means range from 0.498 to 0.559). Table 10 compares loss functions: RSE and RRSE are statistically indistinguishable (0.533 vs. 0.534; on the controlled targets $\|\mathbf{M}\|_F = 1$, so the two coincide up to normalization), whereas the scale of MSE slows optimization. Table 11 fixes the Hadamard order: spending the same budget on four rank-8 factors (order 4) is consistently worse than two rank-16 factors (0.869 vs. 0.533), which is why B=2 is used throughout. Finally, Table 12 sweeps the HD2 initialization scale: the theory value $s^* = (qN)^{-1/4}$ matches the target magnitude at iteration 0, but smaller scales converge to lower final RRSE, and the unnormalized default $N(0, 1)$ start is orders of magnitude off the target scale yet still converges over the full iteration budget.

Table 12: HD2 initialization scale s (std of factor entries; $s^* = (qN)^{-1/4} \approx 0.149$ matches the target magnitude). "init" is the iteration-0 RRSE, "best" the best RRSE within 3000 iterations (mean of 3 seeds). Bold marks the best final RRSE per spectrum family.

	exponential (r_{eff}=8)		power (r_{eff}=48)		head-tail (r_{eff}=96)	
Init scale	best ↓	init	best ↓	init	best ↓	init
0.001 ($0.01 \times s^*$)	**0.1101**	1	**0.3949**	1	**0.4439**	1
0.004 ($0.03 \times s^*$)	0.1176	1	0.4007	1	0.4440	1
0.015 ($0.1 \times s^*$)	0.1700	1	0.4122	1	0.4487	1
0.045 ($0.3 \times s^*$)	0.2471	1	0.4181	1	0.4519	1
0.149 (s^*, theory)	0.3041	1.4	0.4229	1.4	0.4546	1.4
0.446 ($3 \times s^*$)	0.3398	81	0.4582	81	0.4911	81
1.000 (default $N(0,1)$)	0.6224	2e+03	0.7166	2e+03	0.7434	2e+03
1.487 ($10 \times s^*$)	0.9451	1e+04	0.9626	1e+04	0.9696	1e+04

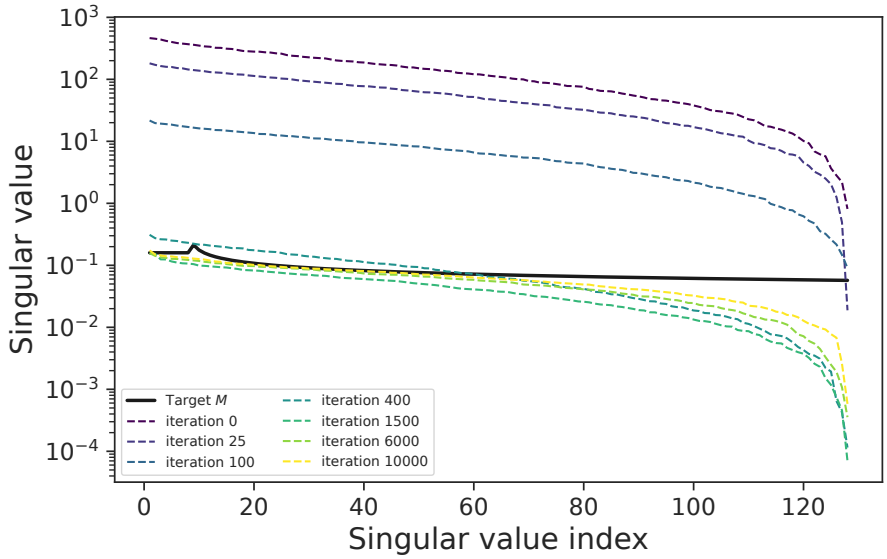


Figure 7: Singular values of the HD2 reconstruction (two rank-16 factors, budget D=32, Adam lr 0.01) while fitting a head-tail spectrum with r_{eff}=96 (M=N=128, seed 0). Snapshots at iterations $0, 25, 100, 400, 1500, 6000, 10000$; RRSE decreases from 2086 to 0.558. The reconstruction spectrum spreads toward the long tail (rank amplification) while the tail singular values stay small (energy dilution).

E Singular-Value Dynamics of the HD Branch

Fig. 7 traces the singular values of the HD2 reconstruction (two rank-16 factors, budget D=32, Adam with $\text{lr} = 0.01$) while fitting a head-tail spectrum with r_{eff}=96. During optimization, the reconstruction spectrum spreads toward the long tail—the rank amplification of Theorem 4.1—while tail singular values stay small, a signature of energy dilution from Proposition 4.5; the RRSE falls from 2086 at initialization to 0.558 at iteration 10,000. This shows that both theoretical effects operate jointly during training, not merely in the asymptotic regime.