Geodesics and Low Rank Behavior in the Deep Linear Network

Alan Chen
Advisor: Govind Menon
Reader: Eli Upfal

Spring 2025

Abstract

Deep learning models often generalize surprisingly well despite possessing far more parameters than traditional theories would suggest is optimal. The Deep Linear Network (DLN) is a phenomenological model for over-parameterization that also reveals an intricate geometric structure underlying training dynamics. We address two open problems in the DLN: (1) the explicit computation of geodesics in the full-rank manifold, and (2) a deeper exploration of the "low-rank attraction" observed for quadratic energies. We derive a general system of ordinary differential equations (ODEs) describing geodesics in the DLN and provide closed-form solutions in special cases where endpoint singular vectors are related by rotation. Then, through numerical experiments, we demonstrate the sensitivity of low-rank attraction to factors such as initialization schemes and energy sparsity. Finally, we conclude with an investigation into using an entropic log-volume form related to the geometry on the full-rank manifold as an explicit regularizer for a simple class of energies. Our findings demonstrate how theoretical geometry can interplay with practical insights in over-parameterized networks and conclude by highlighting several open directions for research in expanding the DLN theory and bridging its insights with true deep learning.

Acknowledgments

Firstly, I would like to thank Govind Menon for his invaluable mentorship throughout the thesis and his endless passion and energy for mathematics that, ever since I took his Pattern Theory course in Spring 2022, has continuously inspired me to seek out the beauty in applied mathematics. Austin Stromme (ENSAE/CREST) and Kathryn Lindsey (Boston College) also advised me through various discussions relevant to the Bures-Wasserstein geometry and DLN. I am also deeply indebted to my peers and faculty in the CS and APMA departments at Brown University. Lastly, I am forever grateful towards my family and cat for their unlimited support and encouragement - thank you for always cheering for me.

Contents

Chapter 1

Introduction

1.1 Background

Modern deep learning advances have demonstrated that simple ideas performed at scale can prove to be powerful for approximating functions and solving tasks such as protein folding or natural language generation [Jumper et al., 2021, Achiam et al., 2023, Kaplan et al., 2020]. However, many of these advances are motivated purely by empirics but lack significant grounding in theoretical principles. Rigorously justifying why deep learning models generalize well despite having an enormous number of parameters, defying classical machine learning tradeoffs, remains an open problem [Allen-Zhu et al., 2019, Arora et al., 2020].

A cornerstone of modern deep learning is the multi-layer perceptron (MLP) [Rumelhart et al., 1985]. The MLP consists of $N \in \mathbb{Z}^+$ layers, each which composes affine transformations with elementwise application of a nonlinear activation function $\sigma : \mathbb{R} \to \mathbb{R}$:

$$x_p = \sigma(W_p x_{p-1} + b_p) \tag{1.1}$$

where $x_0 \in \mathbb{R}^{d_0}$ is the data sample and $W_p \in \mathbb{M}_{d_p \times d_{p-1}}$ and $b_p \in \mathbb{R}^{d_p}$. This model is *overparameterized* because the same function can effectively be approximated by multiple choices of values for the W_ps and b_ps (Figure 1.1) - some tasks involve training neural networks with millions of parameters to fit datasets with thousands of points. Another practical example of an overparameterized model is the transformer's attention mechanism [Vaswani et al., 2017]. The self-attention module is overparameterized in the same way the Deep Linear Network will be. For a sequence of n d-dimensional inputs $\mathbf{x} \in \mathbb{R}^{n \times d}$, the self attention operator contains terms like

$$\exp \langle W_Q x_i, W_K x_j \rangle = \exp(x_i^T W_Q^T W_K x_j), \tag{1.2}$$

where W_Q and W_K are the query and key matrices of the module. Clearly the matrix $W_Q^T W_K$ is overparameterized - in fact, many interpretability papers argue it does not make sense in practice to consider these matrices separately [Elhage et al., 2021] and instead should be thought of as a single kernel that modifies the "metric" within this attention layer. However, W_Q and W_K are trained separately because it is what works in practice.

Neural networks are trained via gradient descent on an energy/loss function

$$E : \mathbb{R}^{d_N} \to \mathbb{R}. \tag{1.3}$$

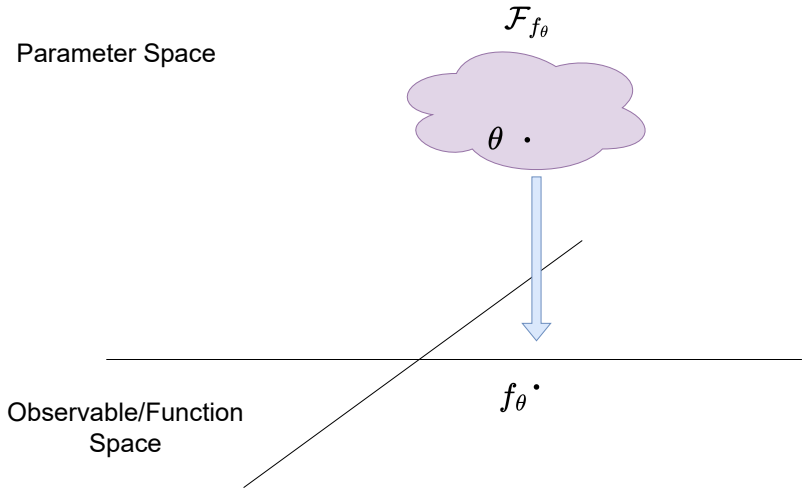


Figure 1.1: Visualization of overparameterization. Multiple sets of parameters appear to induce the same function according to a known distribution, generating a fiber of parameters $\mathcal{F}_{f_\theta}$ above each function f_θ. However, these different parameters may behave differently on unknown data points.

Examples of E include quadratic losses to a specified label or next token cross entropy in an autoregressive language modeling task. In order to train the models, the gradient of E is *backpropagated* via the chain rule onto all parameters W_p and b_p, generating small parameter updates that iteratively take the parameters towards the optimal solution [Rumelhart et al., 1985]. The discrete time updates are applied via an optimizer and repeated over batches of data that approximate the true distribution of data.

1.1.1 The Deep Linear Network

We will use Symm_d to denote $d \times d$ symmetric matrices, O_d to denote the group of $d \times d$ orthogonal matrices, and $\mathbb{M}_d$ to denote the set of $d \times d$ matrices.

The Deep Linear Network (DLN) [Arora et al., 2018a,b, Menon, 2024] is a phenomenological model specifically targeted at understanding what mathematical structure emerges when models are overparameterized. The DLN of *width* $d \in \mathbb{Z}^+$ and *depth* $N \in \mathbb{Z}^+$ consists of two spaces: the "upstairs" parameter space $\mathbb{M}_d^N$ and "downstairs" observable space $\mathbb{M}_d$, linked by the mapping $\phi : \mathbb{M}_d^N \to \mathbb{M}_d$ that sends

$$(W_N, W_{N-1}, \ldots, W_1) \mapsto X = W_N W_{N-1} \cdots W_1. \tag{1.4}$$

The parallel to MLPs is clear - we strip away the nonlinearities and biases to form the absolute simplest model of overparameterization: composition of linear mappings / matrix factorization. Indeed, upstairs, we have Nd^2 dimensions yet the downstairs observable space is just d^2 dimensions, capturing the innate dimensional difference that characterizes an overparameterized system.

Discrete time gradient descent of an energy is simply approximating the continuous time minimizing gradient flow

$$\dot{\mathbf{W}} = -\nabla_{\mathbf{W}} E(X) = -\nabla_{\mathbf{W}} E(\phi(\mathbf{W})). \tag{1.5}$$

Because the loss function only depends on the observable state X, the gradient itself can only push the parameters in d^2 dimensions within the larger Nd^2 dimensional ambient space. A simple application of the chain rule expands the flow into coordinates that makes these "thin" gradients explicit [Arora et al., 2018a]. For all $1 \leq p \leq N$,

$$\dot{W}_p = -(W_N \dots W_{p+1})^T dE(X)(W_{p-1} \dots W_1)^T, \tag{1.6}$$

where

$$dE(X)_{ij} = \frac{\partial E}{\partial X_{ij}}.$$

Notice that the pre- and post-factors are fixed by over-parameterization. The only quantity that changes with E is $dE(X)$, a d^2 dimensional matrix.

This gradient flow is the fundamental object of study in the DLN. Fully understanding the surprisingly intricate structure is likely of utmost importance to generalizing results to true deep learning. In this thesis, we will offer initial work towards deeper theoretical and experimental understanding through computation of geodesics and numerical simulations. A summary of this thesis is as follows:

1. We will review the fundamental results describing the DLN, including discussion of invariant manifolds and Riemannian geometry (Section 2.1). The important results are the flow along a "stable" invariant manifold (Theorem 2.1.2) that we will revisit repeatedly and recognizing the Riemannian geometry underlying said flow (Theorem 2.2.2).

2. We present new intuition for the geometries at play through results on the geodesics in the DLN (Section 3.1). We motivate our computations via first deriving the geodesics on the classic Bures-Wasserstein geometry (Theorem 3.2.3) and then extending the derivations to the DLN in Theorems 3.3.2 and 3.4.2.

3. Finally, we conduct principled numerical experiments to replicate and extend existing understanding on low rank attraction and provide new simulations for a thermodynamic perspective on the DLN (Section 4.1). We investigate sensitivity of low rank attraction to initialization, dimensionality, and sparsity and present new results using an entropic volume-based regularizer.

Proofs will be presented when novel or appropriate (or charming). Otherwise, readers will be referred to the original sources.

Chapter 2

The Deep Linear Network

2.1 Preliminaries

We begin with a continued overview of the DLN model formulation and fundamental results in the DLN. A core foundation of any good dynamical system study is the identification of conserved quantities through symmetries.

To this end, we partition the ambient $\mathbb{M}_d^N$ into algebraic *varieties* $\mathcal{M}_{\mathbf{G}}$, where

$$\mathbf{G} = (G_{N-1}, G_{N-2}, \ldots, G_1) \in \text{Symm}_d^{N-1}$$

characterizes a solution set in $\mathbb{M}_d^N$ to the quadratic equations

$$\mathcal{M}_{\mathbf{G}} = \{\mathbf{W} \in \mathbb{M}_d^N | W_{p+1}^T W_{p+1} = W_p W_p^T - G_p, 1 \leq p \leq N-1\}. \tag{2.1}$$

The first result is that these varieties are invariant under the gradient flow from Equation 1.5 (Figure 2.1).

Theorem 2.1.1 (DLN Invariant Manifolds [Arora et al., 2018b])**.** Assume $\mathbf{W}(t)$ obeys the gradient flow in Equation 1.5 with initial condition $\mathbf{W}_0$ that induces an initial algebraic variety $\mathbf{G}_0$. Then, the following two properties hold:

1. $\mathbf{W}(t) \in \mathcal{M}_{\mathbf{G}_0}$ for all $t \in \mathbb{R}$ (i.e. $\frac{d}{dt} G_p = 0$).
2. The downstairs matrix $X(t) = \phi(\mathbf{W}(t))$ satisfies

$$\dot{X}(t) = -\sum_{p=1}^{N} (A_{p+1} A_{p+1}^T) dE(X) (B_{p-1}^T B_{p-1}), \tag{2.2}$$

where A_p and B_p respectively define the prefactors and postfactors

$$A_p = W_N W_{N-1} \cdots W_p,$$
$$B_p = W_p W_{p-1} \cdots W_1.$$

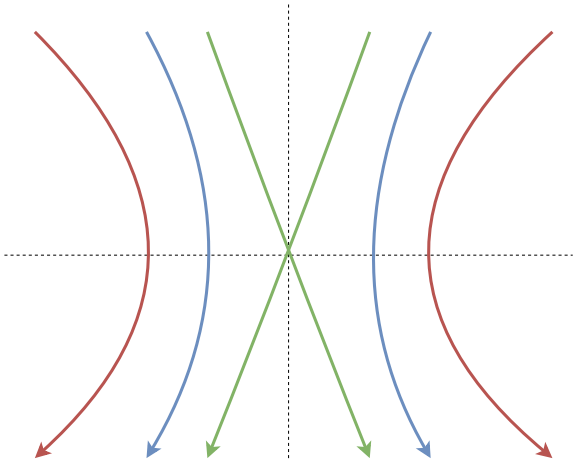


Figure 2.1: Simple illustration of invariant manifolds in DLN geometry. If we consider $d = 1$, the invariant manifolds are exactly hyperbolas. The green manifold is the balanced manifold and passes through 0.

Proof. The proofs of both parts are quite similar and follow from applications of the product rule and Equation 1.6. First, consider

$$\frac{d}{dt}\left(W_{p+1}^T W_{p+1} - W_p W_p^T\right). \tag{2.3}$$

The terms in the product rule come out to be

$$\dot{W}_{p+1}^T W_{p+1} = -(W_p \cdots W_1) dE(X)^T (W_N \cdots W_{p+1}) = W_p \dot{W}_p^T, \tag{2.4}$$

$$W_{p+1}^T \dot{W}_{p+1} = -(W_N \cdots W_{p+1})^T dE(X) (W_p \cdots W_1)^T = \dot{W}_p W_p^T. \tag{2.5}$$

Thus, we conclude that $\frac{d}{dt} G_p = 0$, so these varieties are invariant manifolds. The second part is very similar. We use product rule to expand $\phi(\mathbf{W})$ as

$$\begin{aligned}\frac{d}{dt}\phi(\mathbf{W}) = \dot{W}_N W_{N-1} \cdots W_1 \\ + W_N \dot{W}_{N-1} W_{N-2} \cdots W_1 \\ + \dots \\ + W_N W_{N-1} \cdots W_2 \dot{W}_1.\end{aligned}$$

Plugging in Equation 1.6 into each term gives the expected form of the flow. □

In the special case that $\mathbf{G} = \mathbf{0}$, we find even further structure: all W_p's must be isospectral. Thus, we can foliate $\mathcal{M}_\mathbf{0}$ *by rank* into manifolds $\mathcal{M}_r$ for $0 \le r \le d$. We are particularly interested in the full rank case $\mathcal{M}_d$, called the *balanced manifold* and denoted as just $\mathcal{M}$.

Not only do W_p have the same singular values, their singular vectors must align. Let $W_{p+1} = Q_{p+1}\Lambda Q_p^T$ and $W_p = \tilde{Q}_p \Lambda Q_{p-1}^T$ be the SVDs. Then, we have that

$$W_{p+1}^T W_{p+1} = Q_p \Lambda^2 Q_p^T,$$
$$W_p W_p^T = \tilde{Q}_p \Lambda^2 \tilde{Q}_p^T.$$

So, on the balanced manifold, $Q_p = \tilde{Q}_p$. These symmetries reveal a special parameterization along the balanced manifold that naturally arises through the singular value decomposition (SVD). Explicitly, using the SVD, we can rewrite each coordinate of $\mathbf{W} \in \mathcal{M}$ as

$$W_p = Q_{p+1}\Lambda Q_p^T, 0 \leq p \leq N-1, \tag{2.6}$$

giving the bijective parameterization map

$$\xi : \mathbb{R}_+^d \times O_d^{N+1}, (\Lambda, Q_N, Q_{N-1}, \ldots, Q_0) \mapsto (W_N, W_{N-1}, \ldots, W_1). \tag{2.7}$$

Λ is naturally related to the singular values of $\phi(\mathbf{W})$ by the nth root, and Q_N and Q_0 are bound to the left and right singular vector matrices of $X = \phi(\mathbf{W})$.

We can double check the dimensions: in the full ambient space, we have Nd^2 total dimensions. The G_ps each specify $\frac{d(d+1)}{2}$ equations (they are symmetric). Thus, this leaves

$$Nd^2 - (N-1)\frac{d(d+1)}{2} = \underbrace{d^2}_{\text{end to end}} + \underbrace{(N-1)\frac{d(d-1)}{2}}_{O_d^{N-1}\text{ orbit}}, \tag{2.8}$$

exactly corresponding to the parameterization we found. Furthermore, using this parameterization, we can simplify the downstairs flow in Theorem 2.1.1.

Theorem 2.1.2 (Downstairs Balanced Manifold Flow, Arora et al. [2018b])**.** Let $\mathbf{W}_0 \in \mathcal{M}$. Then, Equation 2.2 simplifies to

$$\dot{X} = -\sum_{p=1}^{N} \left(XX^T\right)^{\frac{N-p}{N}} dE(X) \left(X^T X\right)^{\frac{p-1}{N}}. \tag{2.9}$$

Proof. Because of the parameterization, we notice that the prefactors and postfactors collapse:

$$A_{p+1} = Q_N \Lambda^{\frac{N-p}{N}} Q_p^T \text{ and } B_{p-1} = Q_p \Lambda^{\frac{p-1}{N}} Q_0^T. \tag{2.10}$$

Thus, the flow also simplifies.

$$\dot{X} = -\sum_{p=1}^{N} \left(Q_N \Lambda^{\frac{2(N-p)}{N}} Q_N^T\right) dE(X) \left(Q_0 \Lambda^{\frac{2(p-1)}{N}} Q_0^T\right) \tag{2.11}$$

$$= -\sum_{p=1}^{N} \left(XX^T\right)^{\frac{N-p}{N}} dE(X) \left(X^T X\right)^{\frac{p-1}{N}}. \tag{2.12}$$

□

This final flow on the balanced manifold will be revisited when we compute the geodesics in Section 3.3 and will be the main flow we numerically analyze in Section 4.1.

Remark 2.1.3. *In practice, zero centered initialization schemes with small variance will necessarily begin near the balanced manifold. Indeed, suppose* $(W_p)_{ij} \sim \mathcal{N}(0, \sigma^2)$ *iid for all* $1 \leq p \leq N$ *and* $1 \leq i, j \leq d$*. We will almost surely be full rank. Furthermore,*

$$\mathbb{E}[(G_p)_{ij}] = \mathbb{E}\left[\sum_{k=1}^{d}(W_p)_{ik}(W_p)_{kj}^T - (W_{p+1}^T)_{ik}(W_{p+1})_{kj}\right] \tag{2.13}$$

$$= \sum_{k=1}^{d}\mathbb{E}\left[(W_p)_{ik}(W_p)_{kj}^T\right] - \mathbb{E}\left[(W_{p+1}^T)_{ik}(W_{p+1})_{kj}\right] = 0. \tag{2.14}$$

So, there is practical motivation to be concerned with the balanced manifold - very often at initialization we will be operating close to if not effectively on the balanced manifold.

Remark 2.1.4. *In forthcoming work, Kathryn Lindsey (Boston College) and Govind Menon (Brown University) provide a characterization of the balanced manifold as the solution to a simple optimization problem over the preimage of a downstairs point* X*. In particular, let*

$$\mathcal{F}_X = \{\mathbf{W} | \phi(\mathbf{W}) = X\}. \tag{2.15}$$

Then, the intersection of the balanced manifold with the fiber $\mathcal{M} \cap \mathcal{F}_X$ *is exactly the set of solutions to the optimization problem*

$$\min_{\mathbf{W} \in \mathcal{F}_X} \sum_{p=1}^{N} \operatorname{Tr}\left(W_p W_p^T\right). \tag{2.16}$$

In other words, the balanced manifold is characterized by the minimizers of the Frobenius norm. Naturally, one can consider two follow ups:

1. *Can we construct a gradient flow of the Frobenius norm along the fiber that converges to a point in* $\mathcal{M}$ *using this energy? (This just involves computing the tangent space and appropriate gradient vector* $\operatorname{grad}_\iota \mathcal{M}$*).*

2. *Extend this characterization to the general deep learning case.*

These results are particularly exciting because they are the first steps to extending DLN theory to nonlinear deep learning by generalizing the fundamental notion of balancedness via a characterization that is not specific to linear networks.

2.2 Riemannian Geometry

A clever observation made by Bah et al. [2022] is that Equation 2.9 can be rewritten as a **Riemannian gradient flow** i.e. of the form

$$\dot{X} = -\operatorname{grad}_h E(X) \tag{2.17}$$

for the appropriately defined metric h and manifold $\mathcal{M}$.

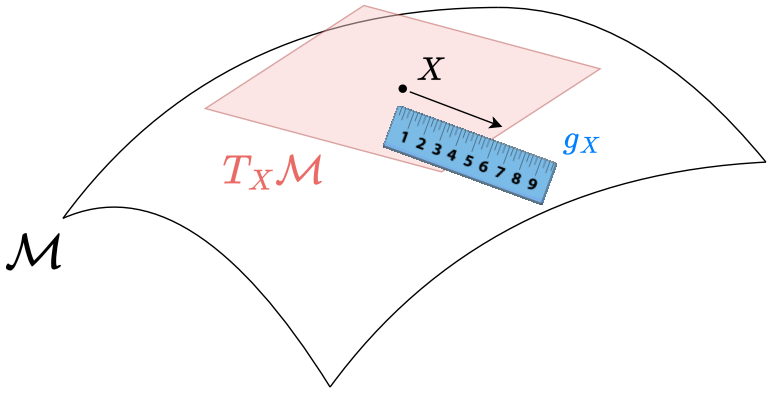


Figure 2.2: An illustration of basic Riemannian geometry. At each point X on a manifold $\mathcal{M}$, we compute the tangent space $T_X\mathcal{M}$, the set of derivatives of smooth paths through X. For each tangent vector in $T_X\mathcal{M}$, the metric g_X specifies the length of the vector, thus redefining the entire geometry in the tangent space.

Specifically, recall a metric assigns a length to each tangent vector $Z \in T_X\mathbb{M}_d$ (Figure 2.2). Let the metric g^N on $\mathbb{M}_d$ be defined as

$$\|Z\|_{g^N}^2 = \mathrm{Tr}\left(Z^T \mathcal{A}_{N,X}^{-1}(Z)\right), \tag{2.18}$$

where $\mathcal{A}_{N,X} : T_X\mathbb{M}_d^* \to T_X\mathbb{M}_d$ is the linear operator that maps from 1-forms to the tangent space with explicit form

$$Z \mapsto \sum_{p=1}^{N} \left(XX^T\right)^{\frac{N-p}{N}} Z \left(X^TX\right)^{\frac{p-1}{N}}. \tag{2.19}$$

The nature of the operator $\mathcal{A}_{N,X}$ is very curious - along the balanced manifold $\mathcal{M}$, it turns out that $\mathcal{A}_{N,X}$ can be diagonalized and is positive definite.

Theorem 2.2.1 (Diagonalization of $\mathcal{A}_{N,X}$)**.** Let $X = U\Sigma V^T$ be the SVD of X. $\mathcal{A}_{N,X}$ is a positive definite and symmetric linear operator with eigenvectors $u_k v_l^T$ and corresponding eigenvalues

$$\lambda_{kl}^N = \begin{cases} \frac{\sigma_k^2 - \sigma_l^2}{\sigma_i^{\frac{2}{N}} - \sigma_j^{\frac{2}{N}}} & k \neq l, \\ N\sigma_k^{\frac{2(N-1)}{N}} & k = l, \end{cases} \tag{2.20}$$

for all $1 \leq k, l \leq d$.

Proof. We first observe that

$$(XX^T)^{\frac{N-p}{N}} = U\Sigma^{\frac{N-p}{N}}U^T \tag{2.21}$$

$$(X^TX)^{\frac{p-1}{N}} = V\Sigma^{\frac{p-1}{N}}V^T \tag{2.22}$$

Thus, the $\mathcal{A}_{N,X}$ operator applied to the rank one matrix $u_k v_l^T$ gives

$$\sum_{p=1}^{N} U\Sigma^{\frac{2(N-p)}{N}} U^T u_k v_l^T V \Sigma^{\frac{2(p-1)}{N}} V^T = \sum_{p=1}^{N} U\Sigma^{\frac{2(N-p)}{N}} e_k e_l^T \Sigma^{\frac{2(p-1)}{N}} V^T$$
$$= \left(\sum_{p=1}^{N} \sigma_k^{\frac{2(N-p)}{N}} \sigma_l^{\frac{2(p-1)}{N}} \right) u_k v_l^T.$$

The summation is a geometric series that, upon simplification, gives the eigenvalue

$$\lambda_{kl}^N = \frac{\sigma_k^2 - \sigma_l^2}{\sigma_k^{2/N} - \sigma_l^{2/N}} \tag{2.23}$$

when $k \neq l$. When $k = l$, the eigenvalue can be seen as (1) the result of the geometric series collapsing or (2) as the limit of $\lim_{\sigma_l \to \sigma_k} \lambda_{kl}^N$. □

With this metric in mind, Equation 2.17 comes into fruition.

Theorem 2.2.2 (DLN Riemannian Gradient Flow, Bah et al. [2022])**.** Equation 2.9 can be rewritten as a Riemannian gradient flow on $(\mathfrak{M}_d, g^N)$:

$$\dot{X} = -\text{grad}_{g^N} E(X), \tag{2.24}$$

where $\mathfrak{M}_d$ is the manifold of all rank d matrices in $\mathbb{M}_d$.

Proof. The core of the proof lies in the computation of the gradient vector through the dual relationship with the differential of E. Recall that by definition, for all $Z \in T_X \mathfrak{M}_d$,

$$dE(X)(Z) = g^N \left(\text{grad}_{g^N} E(X), Z \right). \tag{2.25}$$

We can compute both sides, letting $V = \text{grad}_{g^N} E(X)$ be the vector we are looking for:

$$dE(X)(Z) = \text{Tr}(dE(X)^T Z), \tag{2.26}$$
$$g^N(V, Z) = \text{Tr}\left(V^T \mathcal{A}_{N,X}^{-1}(Z) \right). \tag{2.27}$$

Since this must hold for all Z, we see that

$$dE(X) = \mathcal{A}_{N,X}^{-1}(V) \implies V = \mathcal{A}_{N,X}(dE(X)). \tag{2.28}$$

It is then obvious from the expansion of $\mathcal{A}_{N,X}$ that Equation 2.9 is simply

$$\dot{X} = -\mathcal{A}_{N,X}(dE(X)) = -\text{grad}_{g^N} E(X). \tag{2.29}$$

□

Remark 2.2.3 (Infinite Depth Limit)**.** *$\mathcal{A}_{N,X}$ can also be altered to give a clear infinite depth limit (i.e. $N \to \infty$). In particular, if we add a factor of $\frac{1}{N}$ in front, we find that*

$$\lim_{N \to \infty} \mathcal{A}_{N,X}(Z) = \int_0^1 \left(XX^T \right)^{1-t} Z (X^T X)^t dt. \tag{2.30}$$

The eigenvalues also have clear limits as well after including the adjustment factor.

$$\lim_{N\to\infty} \frac{1}{N}\lambda_{kl}^N = \begin{cases} \frac{\sigma_k^2-\sigma_l^2}{\log\sigma_k^2-\log\sigma_l^2} & k \neq l, \\ \sigma_k^2 & k = l. \end{cases} \tag{2.31}$$

We will use the infinite depth limit often in numerical simulation to reduce the number of hyperparameters to study.

2.2.1 Riemannian Submersion

Recall the following important definitions from Riemannian geometry.

Definition 2.2.4 (Vertical and Horizontal Spaces)**.** Let $(\mathcal{M}, g)$ and $(\mathcal{N}, h)$ be linked by a differentiable map $\phi : \mathcal{M} \to \mathcal{N}$. Then, there exists a direct sum decomposition of the tangent space at any point $m \in \mathcal{M}$ via the differential of ϕ.

$$T_m\mathcal{M} = \ker D\phi[m] \oplus (\ker D\phi[m])^\perp = \mathcal{V}_m \oplus \mathcal{H}_m, \tag{2.32}$$

where $\mathcal{V}_m$ and $\mathcal{H}_m$ denote the **vertical** and **horizontal** spaces at m respectively. A tangent vector is vertical or horizontal if it lies completely within the respective space.

Definition 2.2.5 (Riemannian Submersion)**.** Let $(\mathcal{M}, g)$ and $(\mathcal{N}, h)$ be two Riemannian manifolds. A mapping $\phi : \mathcal{M} \to \mathcal{N}$ is a **Riemannian submersion** iff ϕ satisfies the following.

1. ϕ is a smooth submersion (the differential $D\phi[m]$ at $m \in \mathcal{M}$ is surjective for all m).
2. The differential restricted to the horizontal space at $m \in \mathcal{M}$ is isometric for all m.

As it conveniently turns out, the ϕ mapping defined back in Equation 1.4 is a Riemannian submersion. The upstairs manifold is $\mathcal{M}$ equipped with the **natural metric** ι induced from the Frobenius norm as $\mathcal{M}$ is embedded in $\mathbb{M}_d^N$. The downstairs manifold is exactly $(\mathfrak{M}_d, g^N)$, and g^N is exactly the metric obtained through submersion of ι. More generally, we have the following result.

Theorem 2.2.6 (ϕ Submersion, [Yu and Menon, 2024, Menon, 2024])**.** For all $1 \leq r \leq d$, the metric g^N on $\mathfrak{M}_r$ is exactly obtained from Riemannian submersion via the end-to-end product map $\phi : \mathcal{M}_r \to \mathfrak{M}_r$.

In the proof of this theorem, Yu and Menon [2024] compute the vertical and horizontal spaces of ϕ by computing the differential ϕ^*. The formulas for the particular orthonormal basis vectors are meticulously worked out in their work, but a key observation is that the orthonormal basis for the horizontal space coincides with the eigenvectors in the diagonalization of $\mathcal{A}_{N,X}$ (rank one matrices of the form $u_k v_l^T$ where u_k and v_l are normal).

Lemma 2.2.7 (Orthonormal Basis for Horizontal Space, Yu and Menon [2024])**.** Let

$$\mathbf{M} \in \mathcal{M}, \mathbf{M} = (Q_N \Lambda Q_{N-1}^T, Q_{N-1}\Lambda Q_{N-2}^T, \ldots, Q_1 \Lambda Q_0^T) \tag{2.33}$$

so that $\mathbf{M} = \xi(\Lambda, Q_N, Q_{N-1}, \ldots, Q_1)$. Furthermore, let $q_{p,i}$ denote the ith column of Q_p. Then, define the following special elements of $\mathbb{M}_d^N$ that will form the basis for $\mathcal{H}_\mathbf{M}$ as

$$\mathbf{l}^k = \left(l_N^k, l_{N-1}^k, \ldots, l_1^k\right), \tag{2.34}$$

$$l_s^k = \frac{1}{\sqrt{N}} q_{s,k} q_{s-1,k}^T, \tag{2.35}$$

$$\mathbf{u}^{k,l,0} = \left(u_N^{k,l,0}, u_{N-1}^{k,l,0}, \ldots, u_1^{k,l,0}\right) \tag{2.36}$$

$$u_s^{k,l,0} = C_s^{k,l} q_{s,l} q_{s-1,k}^T, \tag{2.37}$$

and

$$\mathbf{u}^{k,l,N} = \left(u_N^{k,l,N}, u_{N-1}^{k,l,N}, \ldots, u_1^{k,l,N}\right) \tag{2.38}$$

$$u_s^{k,l,N} = D_s^{k,l} q_{s,k} q_{s-1,l}^T, \tag{2.39}$$

where $C_s^{k,l}, D_s^{k,l} \in \mathbb{R}$ are normalizing constants with $s \in [N]$.

The result is that $(\mathbf{l}, \mathbf{u}^0, \mathbf{u}^N)$ where

$$\mathbf{l} = \{l_k | 1 \leq k \leq d\}, \tag{2.40}$$

$$\mathbf{u}^0 = \{u^{k,l,0} | 1 \leq k < l \leq d\}, \tag{2.41}$$

$$\mathbf{u}^N = \{u^{k,l,N} | 1 \leq k < l \leq d\} \tag{2.42}$$

form a basis for $\mathcal{H}_{\mathbf{M}} = (\ker \phi^*[\mathbf{M}])^\perp$.

The proofs of these important results are technical but straightforward so for brevity we refer interested readers to Section 8 of Menon [2024].

Chapter 3

Geodesics

3.1 Background

Given a Riemannian manifold $(\mathcal{M}, g)$, the geodesics (or shortest paths) are the fundamental geometric object to understanding the curvature of the space. We begin with an overview of their definition and state an important result on the behavior of geodesics under Riemannian submersions. A good reference is Do Carmo and Flaherty Francis [1992].

Definition 3.1.1 (Geodesic)**.** A **geodesic** between $A, B \in \mathcal{M}$ on $(\mathcal{M}, g)$ is an extrema of the action

$$\mathcal{A}[x] = \int_0^1 g_{x(t)}(\dot{x}(t), \dot{x}(t))dt \tag{3.1}$$

over the set of paths

$$\mathcal{S} = \{x : [0,1] \to \mathcal{M} | x(0) = A, x(1) = B, x \in \mathcal{C}^1([0,1])\}. \tag{3.2}$$

In essence, the geodesic definition comes from defining the Lagrangian to be

$$L(x, \dot{x}) = g_x(\dot{x}, \dot{x}), \tag{3.3}$$

which is exactly the arc-length of x under the metric g. In this formulation (and more generally for any Lagrangian), the minimizing x is known to satisfy the **Euler-Lagrange** equations (which are equivalent to Newton's laws of motion)

$$\frac{d}{dt}\frac{\partial L(x, \dot{x})}{\partial \dot{x}} = -\frac{\partial L(x, \dot{x})}{\partial x}. \tag{3.4}$$

The optimization can also be seen from a dual perspective, known as the Hamiltonian formulation, and reveals another equivalent set of equations of motion. Given a Lagrangian $L(x, \dot{x})$ over coordinates equipped with an inner product, the Hamiltonian is given from the **Legendre transform** of $L(x, \cdot)$ as

$$H(x, p) = \sup_{\dot{x}}(\langle p, \dot{x}\rangle - L(x, \dot{x})). \tag{3.5}$$

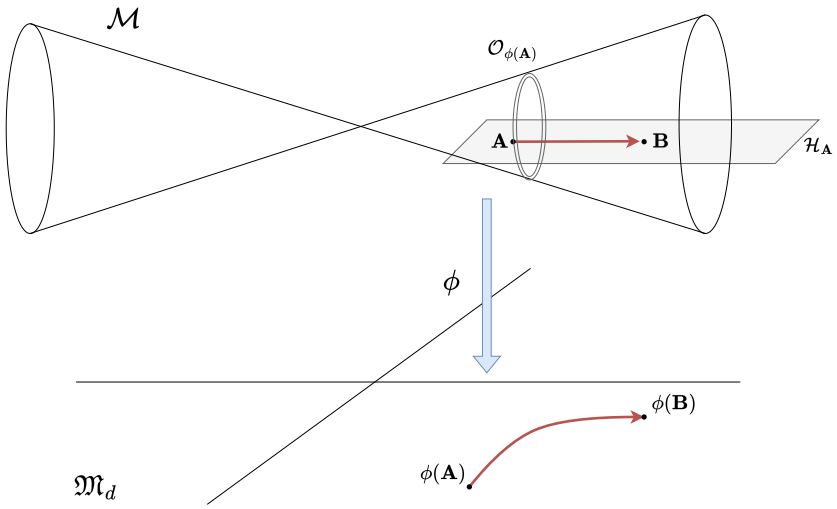


Figure 3.1: Diagram of ϕ acting as a Riemannian submersion from $\mathcal{M}$ to $\mathfrak{M}_d$ and the respective preservation of horizontal geodesics under ϕ (Theorem 3.1.2). This image is the core argument of Theorem 3.4.2.

Differentiating the right hand side with respect to $\dot{x}$ and setting equal to 0 gives

$$0 = \frac{d}{d\dot{x}} \left(\langle p, \dot{x} \rangle - L(x, \dot{x}) \right) \implies p = \frac{\partial L}{\partial \dot{x}}$$

when $\dot{x}$ is maximizing, netting the change of variables formula. The corresponding canonical equations to the Euler-Lagrange equations are given as

$$\dot{x} = \partial_p H(x, p), \tag{3.6}$$
$$\dot{p} = -\partial_x H(x, p). \tag{3.7}$$

3.1.1 Geodesics Under Submersion

In our problem, we have the unique structure given by ϕ and its role as a submersion mapping (Theorem 2.2.6). We would like to leverage this structure when attempting to compute geodesics. We have the following result on a specific case when submersions respect geodesics.

Theorem 3.1.2. Let $(\mathcal{M}, g)$ and $(\mathcal{N}, h)$ be Riemannian manifolds and $\phi : \mathcal{M} \to \mathcal{N}$ be a Riemannian submersion. Let $x : [0, 1] \to \mathcal{M}$ be a geodesic between $A, B \in \mathcal{M}$ in the geometry $(\mathcal{M}, g)$ so that $x'(0) \in \mathcal{H}_A$ i.e. is horizontal. Then,

1. $x'(t)$ is horizontal for all $t \in [0, 1]$.
2. $\phi \circ x$ is a geodesic in $(\mathcal{N}, h)$ between $\phi(A)$ and $\phi(B)$ and is the same length as x.

Theorem 3.1.2 is the core result that will allow us to obtain explicit formulas for the geodesics when specific symmetries exist that result in horizontal straight lines in the upstairs manifold.

We illustrate the procedure in Figure 3.1: as a preview, we will find two points connected by a horizontal geodesic in the upstairs space that under the submersion map to the desired endpoints in the downstairs geometry, and apply Theorem 3.1.2 to derive an explicit formula.

We are now ready to proceed with the computation of various geodesic equations and explicit paths. To motivate our results in the DLN, we present two derivations of the explicit geodesics in the classical Bures-Wasserstein geometry, a special case of the DLN. These arguments will be extended to analogous computations in the DLN, giving rise to a general system of ODEs for general geodesics and explicit formulas in a specific case. The first argument is novel but relies on a classical mechanical computation after a change of variables into the Hamiltonian formulation. The second argument relies on the preservation of horizontal geodesics under Riemannian submersions and offers a more natural geometric perspective [Bhatia et al., 2019]. We will also present a characterization of all straight lines that lie in the balanced manifold, reminiscent of confocal quadrics (Hilbert and Cohn-Vossen [1952], pp. 25-27).

3.2 Bures Wasserstein Geometry

The Bures-Wasserstein (BW) geometry [Bhatia et al., 2019] is a particular special case of the general DLN when $N = 2$ and $W_1 = W_2^T$, and the upstairs space is only invertible matrices (nonzero singular values). Notice then that the parameter space becomes $GL(d)$ and the observable space $\mathbb{P}_d$ of symmetric positive definite $d \times d$ matrices, as the parameter $W \in GL(d)$ maps to the observable space through

$$X = WW^T \in \mathbb{P}_d.$$

BW is still overparameterized: upstairs, we have d^2 dimensions compared to only $\binom{d+1}{2}$ dimensions downstairs as X is symmetric.

The downstairs metric, now denoted g^{BW}, at a point $X \in \mathbb{P}_d$ reduces to a very simple form, given by

$$g^{\mathrm{BW}} : T_X\mathbb{P}_d \times T_X\mathbb{P}_d \to \mathbb{R} \tag{3.8}$$

$$(Z_1, Z_2) \mapsto \frac{1}{2}\operatorname{Tr}(Z_1^T \mathcal{L}_X^{-1}(Z_2)), \tag{3.9}$$

where $\mathcal{L}_X^{-1}(Z)$ is the **Lyapunov operator** that maps Z to the matrix that solves

$$\mathcal{L}_X^{-1}(Z)X + X\mathcal{L}_X^{-1}(Z) = Z. \tag{3.10}$$

Thus it is also clear that

$$\mathcal{L}_X(Y) = XY + YX. \tag{3.11}$$

We will aim to compute the geodesics of the geometry $(\mathbb{P}_d, g^{\mathrm{BW}})$. The Lagrangian in this case is given by (using capital letters for matrix paths)

$$L(X, \dot{X}) = \frac{1}{2}\operatorname{Tr}(\dot{X}\mathcal{L}_X^{-1}(\dot{X})). \tag{3.12}$$

We also quickly define an important quantity.

Definition 3.2.1 (Geometric Mean)**.** Given two positive definite matrices A and B, the **geometric mean** of A and B, denoted by $A\#B$, is given by

$$A\#B = A^{1/2}(A^{-1/2}BA^{-1/2})^{1/2}A^{1/2}. \tag{3.13}$$

Lemma 3.2.2 (Geometric Mean Properties)**.** We will use two key facts about the geometric median, both of which can be proven directly from the definition.

1. $A\#B$ is the unique positive definite solution M to the **Riccati Equation**

$$B = MA^{-1}M. \tag{3.14}$$

2. $A(A^{-1}\#B) = (AB)^{1/2}$ and $(A^{-1}\#B)A = (BA)^{1/2}$.

We will prove the following theorem in two ways.

Theorem 3.2.3 (Geodesics on $(\mathbb{P}_d, g^{\mathrm{BW}})$)**.** The geodesic path $X : [0,1] \to \mathbb{P}_d$ between $A, B \in \mathbb{P}_d$ in the BW geometry is given by

$$X(t) = t^2 B + (1-t)^2 A + t(1-t)\left((AB)^{1/2} + (BA)^{1/2}\right). \tag{3.15}$$

Proof 1 - Hamiltonian Formulation. To begin, recall that $T_X\mathbb{P}_d \cong \mathrm{Symm}_d$. Furthermore, for clarity, denote $Y = \dot{X}$. In order to construct the change of variable into Hamiltonian coordinates, we must first compute $D_Y[L(X,Y)](H)$ where $H \in T_X\mathbb{P}_d$.

$$\begin{aligned} D_Y\left[\frac{1}{2}\operatorname{Tr}(Y^T\mathcal{L}_X^{-1}(Y))\right](H) &= \frac{1}{2}\operatorname{Tr}(D_Y[Y^T\mathcal{L}_X^{-1}(Y)](H)) \\ &= \frac{1}{2}\operatorname{Tr}(H^T\mathcal{L}_X^{-1}(Y) + Y^T\mathcal{L}_X^{-1}(H)) \\ &= \frac{1}{2}(g_X(H,Y) + g_X(Y,H)) \\ &= \operatorname{Tr}(\mathcal{L}_X^{-1}(Y)H) \end{aligned} \tag{3.16}$$

where the second to last equality holds because H and Y are in $T_X\mathbb{P}_d$ and the last equality follows from the cyclic property of trace.

Thus, we obtain that $P = \mathcal{L}_X^{-1}(Y)$. Thus, we also have that $Y = \mathcal{L}_X(P)$. The advantage of working in these dual coordinates now becomes clear: the Hamiltonian is much simpler than the Lagrangian because of the explicit nature of $\mathcal{L}_X$ compared to the implicit nature of $\mathcal{L}_X^{-1}$.

Now, we can compute the Hamiltonian using the cyclic property of trace and the definition of P and $\mathcal{L}_X$ as

$$\begin{aligned} H(X,P) &= \operatorname{Tr}(P^T\mathcal{L}_X(P)) - \frac{1}{2}\operatorname{Tr}(\mathcal{L}_X(P)\mathcal{L}_X^{-1}(\mathcal{L}_X(P))) \\ &= \frac{1}{2}\operatorname{Tr}(P\mathcal{L}_X(P)) \\ &= \operatorname{Tr}(P^2X). \end{aligned} \tag{3.17}$$

We can now state the geodesic equations, which are now essentially trivial from the definition of the Hamiltonian:

$$\begin{cases} \dot{X} = \mathcal{L}_X(P), \\ \dot{P} = -P^2. \end{cases} \tag{3.18}$$

We now find an exact formula for P. In particular, we see that

$$\begin{aligned} \dot{P} = -P^2 \implies -P^{-1}\dot{P}P^{-1} &= I \\ \frac{d}{dt}(P^{-1}) &= I \\ P^{-1}(t) &= tI + P_0^{-1}. \end{aligned} \tag{3.19}$$

Thus, we conclude that $P(t) = (tI + P_0^{-1})^{-1}$. We can then factor out P_0^{-1} from the interior term, revealing an innate structure of $P(t)$ as a time derivative.

$$\begin{aligned} P(t) &= (P_0^{-1}(tP_0 + I))^{-1} \\ &= (tP_0 + I)^{-1}P_0 \qquad (3.20) \\ &= \frac{d}{dt}\log(tP_0 + I). \qquad (3.21) \end{aligned}$$

Using this formula for $P(t)$, we can push on to also derive the explicit formula for $X(t)$. We must now incorporate the Dirichlet boundary conditions of the geodesic:

$$X_0 = A \text{ and } X_1 = B, \tag{3.22}$$

where $A, B \in \mathbb{P}_n$ are the endpoints of the geodesic.

We can expand the first differential equation as

$$\dot{X} = X(t)P(t) + P(t)X(t). \tag{3.23}$$

This equation is both linear and non-autonomous, as it depends on $X(0)$ and $\dot{X}(0)$ through $P(t)$, and is reminiscent of the Lax equation

$$\dot{X} = [X, P] = XP - PX. \tag{3.24}$$

These relationships and the linear of this equation involving both left and right actions motivates the following ansatz for the solution form of X:

$$X(t) = e^{C(t)}X_0e^{D(t)}. \tag{3.25}$$

Plugging in this guess into the original ODE in order to solve for $C(t)$ and $D(t)$ gives that

$$\dot{C}(t) = \dot{D}(t) = P(t), \tag{3.26}$$

so that via matching coefficients both $C(t)$ and $D(t)$ are $\int_0^t P(s)ds$, confirming that our guess is indeed the unique solution to the ODE. Now, using that $P(t) = \frac{d}{dt}\log(tP_0 + I)$, we have that then

$$\begin{aligned} C(t) = D(t) &= \int_0^t \frac{d}{ds}\log(sP_0 + I)ds \qquad (3.27) \\ &= \log(tP_0 + I) - \log(I) \qquad (3.28) \\ &= \log(tP_0 + I). \qquad (3.29) \end{aligned}$$

Returning to our solution form thus leaves just

$$\begin{aligned} X(t) &= \exp\{\log(tP_0 + I)\}X_0 \exp\{\log(tP_0 + I)\} \\ &= (tP_0 + I)X_0(tP_0 + I) \\ &= (tP_0 + I)A(tP_0 + I). \end{aligned} \tag{3.30}$$

Now, we can apply the final unused condition that $X_1 = B$ to find the value for P_0. Plugging this into X gives that

$$B = (P_0 + I)A(P_0 + I), \tag{3.31}$$

which from Lemma 3.2.2 implies that

$$P_0 = (A^{-1}\#B) - I. \tag{3.32}$$

Finally, substituting this into our solution formula for $X(t)$, we conclude that

$$X(t) = (t((A^{-1}\#B) - I) + I)A(t((A^{-1}\#B) - I) + I) \tag{3.33}$$

$$= (t(A^{-1}\#B) + (1-t)I)A(t(A^{-1}\#B) + (1-t)I) \tag{3.34}$$

$$= t^2B + (1-t)^2A + t(1-t)((A^{-1}\#B)A + A(A^{-1}\#B)) \tag{3.35}$$

$$= t^2B + (1-t)^2A + t(1-t)\left((AB)^{1/2} + (BA)^{1/2}\right) \tag{3.36}$$

as desired. □

Proof 2 - Riemannian Submersion. This proof is outlined in detail in Section 4 of Bhatia et al. [2019]. The key steps are

1. Compute the vertical and horizontal spaces of the mapping $\pi : GL(d) \to \mathbb{P}_d$.
2. Show that the line between $A^{\frac{1}{2}}$ and $B^{\frac{1}{2}}U$ is horizontal and is a valid line through $GL(d)$, where U is the orthogonal polar factor of $B^{\frac{1}{2}}A^{\frac{1}{2}}$. Since $\mathbb{M}_d$ is a flat space, the line is also the geodesic between $A^{\frac{1}{2}}$ and $B^{\frac{1}{2}}U$.
3. Apply Theorem 3.1.2 to derive the explicit form of geodesics between $\pi\left(A^{\frac{1}{2}}\right) = A$ and $\pi\left(B^{\frac{1}{2}}U\right) = B$.

□

3.3 General Geodesic Equation

Much like in the BW case, a quick computation reveals that $H(X, P)$ on the manifold $(\mathfrak{M}_d, g^N)$ is exactly

$$H(X, P) = \frac{1}{2}\operatorname{Tr}\left(P^T \mathcal{A}_{N,X}^{-1}(P)\right). \tag{3.37}$$

Using this Hamiltonian formulation, we can obtain similar equations in the general DLN case on the manifold $(\mathfrak{M}_d, g^N)$. However, unfortunately we are unable to extract explicit solutions to this set of ODEs.

We first recall functional calculus on symmetric matrices (reference: [Bhatia, 2009]): if $S = Q\Lambda Q^T$ is a symmetric matrix and $f : \mathbb{R} \to \mathbb{R}$ is a 1D functional, f can be extended to the symmetric matrices as (abusing notation)

$$f(S) = Qf(\Lambda)Q^T = Q\text{diag}(f(\lambda_1), f(\lambda_2), \ldots, f(\lambda_d))Q^T. \tag{3.38}$$

Then, the differential of $f(S)$ at S evaluated in the direction Z is exactly

$$Df[S](Z) = Q\left(f^{[1]}(\Lambda) \circ \left[Q^T Z Q\right]\right)Q^T \tag{3.39}$$

where $f^{[1]}$ is the special matrix given elementwise by the conditional

$$f^{[1]}(\Lambda)_{ij} = \begin{cases} \frac{f(\lambda_i)-f(\lambda_j)}{\lambda_i-\lambda_j} & i \neq j \\ \lim_{\lambda_i \to \lambda_j} \frac{f(\lambda_i)-f(\lambda_j)}{\lambda_i-\lambda_j} & i = j. \end{cases} \tag{3.40}$$

Thus, we have the following lemma.

Lemma 3.3.1 (Derivatives of $(XX^T)^\alpha$)**.** Let $X = U\Sigma V^T$ be the singular value decomposition of X i.e. $U, V \in O_d$ and Σ is a diagonal matrix. Then,

$$D_X\left[\left(XX^T\right)^\alpha\right][Z] = U\left[f_\alpha^{[1]}\left(\Sigma^2\right) \circ \left(\Sigma\tilde{Z}^T + \tilde{Z}\Sigma\right)\right]U^T. \tag{3.41}$$

$$D_X\left[\left(X^TX\right)^\alpha\right][Z] = V\left[f_\alpha^{[1]}\left(\Sigma^2\right) \circ \left(\Sigma\tilde{Z} + \tilde{Z}^T\Sigma\right)\right]V^T. \tag{3.42}$$

where $\tilde{Z} = U^T Z V$ and

$$\left(f_\alpha^{[1]}(\Sigma^2)\right)_{ij} = \begin{cases} \frac{\sigma_i^{2\alpha}-\sigma_j^{2\alpha}}{\sigma_i^2-\sigma_j^2} & i \neq j, \\ \lim_{\sigma_i^2 \to \sigma_j^2} \frac{\sigma_i^{2\alpha}-\sigma_j^{2\alpha}}{\sigma_i^2-\sigma_j^2} & i = j. \end{cases} \tag{3.43}$$

Proof. We write out the proof for the first result - the second is similar.

We notice that XX^T is a symmetric matrix with diagonalization $U\Sigma^2U^T$. Thus, we apply functional calculus along with the chain rule to find that

$$D_X\left[\left(XX^T\right)^\alpha\right](Z) = U\left[f_\alpha^{[1]}\left(\Sigma^2\right) \circ \left(U^T\left(D_X\left[XX^T\right](Z)\right)U\right)\right]U^T. \tag{3.44}$$

Clearly, we know that

$$D_X\left[XX^T\right](Z) = XZ^T + ZX^T. \tag{3.45}$$

Rewriting Z as $\tilde{Z}$ in the singular vector coordinates via the transformation $Z = U\tilde{Z}V^T$ nets the final simplification that gives the form in the Lemma statement. □

Theorem 3.3.2 (DLN Geodesic Equations in Hamiltonian Form.)**.** Let $f_\alpha^{[1]}(\Lambda)$ denote the matrix given by Equation 3.40 with $f(x) = x^\alpha$. The Hamiltonian equations of motion for geodesics on $(\mathfrak{M}_d, g^N)$ are given by

$$\begin{cases} \dot{X} = \mathcal{A}_{N,X}(P), \\ \dot{P} = -U\left(\sum_{p=1}^N \left[f_{\frac{N-p}{N}}^{[1]}\left(\Sigma^2\right) \circ M_1\right]\Sigma + \Sigma\left[f_{\frac{p-1}{N}}^{[1]}\left(\Sigma^2\right) \circ M_2\right]\right)V^T, \end{cases} \tag{3.46}$$

where M_1 and M_2 are the symmetric matrices

$$M_1 = (U^T PV)\Sigma^{\frac{2(p-1)}{N}}(U^T PV)^T \tag{3.47}$$

$$M_2 = (U^T PV)^T\Sigma^{\frac{2(N-p)}{N}}(U^T PV). \tag{3.48}$$

Proof. This first equation follows from exactly the same computation as the BW case but with $\mathcal{L}^{-1}$ replaced with $\mathcal{A}^{-1}_{N,X}$. So the only nontrivial equation is in the second equation. We begin by applying Lemma 3.3.1 to compute $D_X[H(X,P)](Z)$.

$$D_X[H(X,P)](Z) = \frac{1}{2}\sum_{p=1}^{N}\mathrm{Tr}\left(T_U + T_V\right), \tag{3.49}$$

$$T_U = P^T U\left[f^{[1]}_{\frac{N-p}{N}}\left(\Sigma^2\right)\circ\left(\Sigma\tilde{Z}^T + \tilde{Z}\Sigma\right)\right]U^T PV\Sigma^{\frac{2(p-1)}{N}}V^T, \tag{3.50}$$

$$T_V = P^T U\Sigma^{\frac{2(N-p)}{N}}U^T PV\left[f^{[1]}_{\frac{p-1}{N}}\left(\Sigma^2\right)\circ\left(\Sigma\tilde{Z} + \tilde{Z}^T\Sigma\right)\right]V^T. \tag{3.51}$$

Now, to clean up further, we also write P in terms of the singular value coordinates.

$$P = U\tilde{P}V^T \iff U^T PV = \tilde{P}. \tag{3.52}$$

Plugging this in cleans up the equation, leaving just

$$T_U = \tilde{P}^T\left[f^{[1]}_{\frac{N-p}{N}}\left(\Sigma^2\right)\circ\left(\Sigma\tilde{Z}^T + \tilde{Z}\Sigma\right)\right]\tilde{P}\Sigma^{\frac{2(p-1)}{N}} \tag{3.53}$$

$$T_V = \tilde{P}^T\Sigma^{\frac{2(N-p)}{N}}\tilde{P}\left[f^{[1]}_{\frac{p-1}{N}}\left(\Sigma^2\right)\circ\left(\Sigma\tilde{Z} + \tilde{Z}^T\Sigma\right)\right]. \tag{3.54}$$

Lemma 3.3.3. Let K and M be symmetric matrices, Σ be a diagonal matrix, and A be an arbitrary matrix. Then, we have the following identities.

$$\mathrm{Tr}([K\circ(\Sigma A^T)]M) = \mathrm{Tr}([K\circ(A\Sigma)]M) = \mathrm{Tr}(\Sigma[K\circ M]A), \tag{3.55}$$

$$\mathrm{Tr}([K\circ(\Sigma A)]M) = \mathrm{Tr}([K\circ(A^T\Sigma)]M) = \mathrm{Tr}([K\circ M]\Sigma A). \tag{3.56}$$

Proof. 1. Let's first prove that the first expression is equal to the last. We will approach by expanding into coordinates, following Einstein notation convention where repeated indices are summed.

$$\begin{aligned}\mathrm{Tr}([K\circ(\Sigma A^T)]M) &= K_{ij}\sigma_i(A^T)_{ij}M_{ji}\\ &= \sigma_i K_{ij}M_{ji}A_{ji}\\ &= \sigma_i K_{ij}M_{ij}A_{ji}\\ &= \mathrm{Tr}(\Sigma[K\circ M]A).\end{aligned} \tag{3.57}$$

Similarly, we can prove the equality of the second expression and the third expression using the symmetry of K.

$$\begin{aligned}\mathrm{Tr}([K\circ(A\Sigma)]M) &= K_{ij}A_{ij}\sigma_j M_{ji}\\ &= \sigma_j K_{ij}M_{ji}A_{ij}\\ &= \sigma_j K_{ji}M_{ji}A_{ij}\\ &= \mathrm{Tr}(\Sigma[K\circ M]A).\end{aligned} \tag{3.58}$$

2. These follow from extremely similar steps to the proofs in part 1.

□

Applying this lemma to eq. 3.53, where $K = f_\alpha^{[1]}(\Sigma^2)$ for $\alpha = \frac{N-p}{N}$ and $\alpha = \frac{p-1}{N}$ (it is easy to see from definition that these are symmetric) and $M = \tilde{P}^T \Sigma^{\frac{2(N-p)}{N}} \tilde{P}$ and $\tilde{P}\Sigma^{\frac{2(p-1)}{N}}\tilde{P}^T$, $\Sigma = \Sigma$, and $A = \tilde{Z}$, we can clean up even further. For clarity, let

$$M_1 = \tilde{P}\Sigma^{\frac{2(p-1)}{N}}\tilde{P}^T \tag{3.59}$$

$$M_2 = \tilde{P}^T\Sigma^{\frac{2(N-p)}{N}}\tilde{P}. \tag{3.60}$$

Then, we have that

$$D_X[H(X,P)](Z) = \frac{1}{2}\sum_{p=1}^{N} \mathrm{Tr}\left(2\Sigma\left[f^{[1]}_{\frac{N-p}{N}}\left(\Sigma^2\right)\circ M_1\right]\tilde{Z} + 2\left[f^{[1]}_{\frac{p-1}{N}}\left(\Sigma^2\right)\circ M_2\right]\Sigma\tilde{Z}\right) \tag{3.61}$$

$$= \sum_{p=1}^{N} \mathrm{Tr}\left(\left(\Sigma\left[f^{[1]}_{\frac{N-p}{N}}\left(\Sigma^2\right)\circ M_1\right] + \left[f^{[1]}_{\frac{p-1}{N}}\left(\Sigma^2\right)\circ M_2\right]\Sigma\right)\tilde{Z}\right) \tag{3.62}$$

$$= \mathrm{Tr}\left(V\sum_{p=1}^{N}\left(\Sigma\left[f^{[1]}_{\frac{N-p}{N}}\left(\Sigma^2\right)\circ M_1\right] + \left[f^{[1]}_{\frac{p-1}{N}}\left(\Sigma^2\right)\circ M_2\right]\Sigma\right)U^T Z\right) \tag{3.63}$$

This lets us conclude that the second canonical equation in the Hamiltonian formulation is

$$\dot{P} = -U\sum_{p=1}^{N}\left(\left[f^{[1]}_{\frac{N-p}{N}}\left(\Sigma^2\right)\circ M_1\right]\Sigma + \Sigma\left[f^{[1]}_{\frac{p-1}{N}}\left(\Sigma^2\right)\circ M_2\right]\right)V^T. \tag{3.64}$$

□

However, this final form is not particularly amenable to explicit analysis, despite the obvious commutator structure and relation to the singular vector coordinates U and V. Thus, we turn to special cases, where we will be able to find explicit formulas.

3.4 Special Cases of DLN

Since we were unable to get explicit solutions for the geodesics in the general case, we turn to characterizing special cases where explicit formulas are obtainable. We begin with a characterization of straight lines on $(\mathcal{M}, \iota)$ through both a simple sufficient condition on the endpoints and a necessary and sufficient skew-symmetry condition.

Lemma 3.4.1 (Characterization of Straight Lines on $\mathcal{M}$.)**.** Let $\mathbf{A} = \xi(\Lambda, Q_N, \ldots, Q_0)$ and $\mathbf{B} = \xi(\tilde{\Lambda}, \tilde{Q}_N, \ldots, \tilde{Q}_0)$ be the endpoints of interest. Let

$$\mathcal{Q}_p = \tilde{Q}_p^T Q_p \in O_d. \tag{3.65}$$

Then, the line

$$\mathbf{W}(t) = (1-t)\mathbf{A} + t\mathbf{B} \tag{3.66}$$

is balanced for all $t \in [0,1]$:

1. if $\mathcal{Q}_{p+1} = \mathcal{Q}_{p-1}$ for all $1 \leq p \leq N-1$.

2. if and only if
$$(\mathcal{Q}_{p+1} - \mathcal{Q}_{p-1})M^T \tag{3.67}$$
are skew symmetric, where
$$M = \tilde{\Lambda}^{-1}\mathcal{Q}_p\Lambda \tag{3.68}$$
for all $1 \leq p \leq N-1$.

Proof. We wish to find the condition that $\mathbf{W}(t) = (1-t)\mathbf{A} + t\mathbf{B} \in \mathcal{M}$ for all $t \in (0,1)$ (it is by definition balanced at the endpoints). Fix $p \in \{1, 2, \ldots, N-1\}$. We would like to show that
$$W_{p+1}(t)^T W_{p+1}(t) = W_p(t)W_p(t)^T, \tag{3.69}$$
where $W_p(t) = (1-t)W_p(0) + tW_p(1)$ for all p. Expanding the algebra out and using that $W_p(0) = \mathbf{A}_p$ and $W_p(1) = \mathbf{B}_p$ are both in the balanced manifold, we find that $\mathbf{W}(t)$ is in the balanced manifold for all $t \in (0,1)$ if and only if
$$\mathbf{B}_{p+1}^T\mathbf{A}_{p+1} + \mathbf{A}_{p+1}^T\mathbf{B}_{p+1} = \mathbf{B}_p\mathbf{A}_p^T + \mathbf{A}_p\mathbf{B}_p^T. \tag{3.70}$$
Now, we use the parameterizations of $\mathbf{A}$ and $\mathbf{B}$ to expand into the orthogonal matrix components.
$$\begin{aligned}\tilde{Q}_p\tilde{\Lambda}(\tilde{Q}_{p+1}^T Q_{p+1})\Lambda Q_p^T + Q_p\Lambda(Q_{p+1}^T\tilde{Q}_{p+1})\tilde{\Lambda}\tilde{Q}_p^T \\ = \tilde{Q}_p\tilde{\Lambda}(\tilde{Q}_{p-1}^T Q_{p-1})\Lambda Q_p^T + Q_p\Lambda(Q_{p-1}^T\tilde{Q}_{p-1})\tilde{\Lambda}\tilde{Q}_p^T.\end{aligned} \tag{3.71}$$
The sufficient condition is obvious in this form.

We can push further for a cleaner necessary condition by left and right multiplying by $\tilde{Q}^T$ and Q respectively:
$$\tilde{\Lambda}(\mathcal{Q}_{p+1} - \mathcal{Q}_{p-1})\Lambda = -\mathcal{Q}_p\Lambda(\mathcal{Q}_{p+1} - \mathcal{Q}_{p-1})^T\tilde{\Lambda}\mathcal{Q}_p.$$
Then, it is a simple rewriting in terms of M to get the final condition:
$$\mathcal{Q}_{p+1} - \mathcal{Q}_{p-1} = -M(Q_{p+1} - Q_{p-1})^T M^{-T}. \tag{3.72}$$
$$(\mathcal{Q}_{p+1} - \mathcal{Q}_{p-1})M^T = -M(Q_{p+1} - Q_{p-1})^T. \tag{3.73}$$
The structure demonstrates that the difference
$$\mathcal{Q}_{p+1} - \mathcal{Q}_{p-1} \tag{3.74}$$
behaves like a skew symmetric matrix *after* changing into coordinates given by M (or essentially $\mathcal{Q}_p$). □

The importance of straight lines is that they are automatically geodesics on $(\mathcal{M}, \iota)$ because ι is induced from the embedding of $\mathcal{M}$ in the flat ambient space $\mathbb{M}_d^N$. Now, we can put the results together along with Theorem 3.1.2 to get a set of explicit geodesics on the downstairs space. The following argument is depicted in Figure 3.1.

Theorem 3.4.2 (Explicit DLN Geodesics for Rotationally Related Endpoints.)**.** Let $A, B \in \mathfrak{M}_d$ be matrices such that $A = U\Sigma V^T$ and $B = \tilde{U}\tilde{\Sigma}\tilde{V}^T$ are the singular value decompositions. Furthermore, define $\ell : \mathbb{M}_d \times \mathbb{M}_d \to \mathbb{M}_d$ as the linear interpolation between its two arguments:

$$(X, Y) \mapsto (1-t)X + tY. \tag{3.75}$$

Then, if there exists $Q \in O_d$ such that $\tilde{U} = UQ$ and $\tilde{V} = VQ$,

1. there exists $\mathbf{A} \in \phi^{-1}(A)$ and $\mathbf{B} \in \phi^{-1}(B)$ such that $\gamma(t) = \ell(\mathbf{A}, \mathbf{B})$ is a geodesic in $(\mathcal{M}, \iota)$.
2. the geodesic in $(\mathfrak{M}_d, g^N)$ between A and B is given by

$$\gamma(t) = \ell\left(U\Sigma^{\frac{1}{N}}, \tilde{U}\tilde{\Sigma}^{\frac{1}{N}}Q^T\right)\ell\left(\Sigma^{\frac{1}{N}}, Q\tilde{\Sigma}^{\frac{1}{N}}Q^T\right)^{N-2}\ell\left(\Sigma^{\frac{1}{N}}V^T, Q\tilde{\Sigma}^{\frac{1}{N}}\tilde{V}^T\right). \tag{3.76}$$

Proof. First, let's outline the proof. To prove (1), it is sufficient to construct $\mathbf{A} \in \phi^{-1}(A) \subseteq \mathcal{M}$ and $\mathbf{B} \in \phi^{-1}(B) \subseteq \mathcal{M}$ such that

$$\gamma(t) = \ell(\mathbf{A}, \mathbf{B}) \in \mathcal{M} \text{ for all } t \in [0, 1] \tag{3.77}$$

because ι is induced by the flat Frobenius metric. Then, we will show that $\gamma'(t)$ is horizontal at $t = 0$ by proving that

$$\frac{d}{dt}\left((1-t)\mathbf{A} + t\mathbf{B}\right) = -\mathbf{A} + \mathbf{B} \in \mathcal{H}_{\mathbf{A}} \tag{3.78}$$

or that equivalently $\mathbf{A} \in \mathcal{H}_{\mathbf{A}}$ and $\mathbf{B} \in \mathcal{H}_{\mathbf{A}}$ because ϕ^* is linear. Thus, by Theorem 3.1.2, it follows that $\phi \circ \gamma$ is a geodesic in the downstairs space between $\phi(\mathbf{A})$ and $\phi(\mathbf{B})$, giving the explicit formula in (2).

Parameterizations for A and B. Consider the following parameterization for $\mathbf{A}$:

$$\mathbf{A} = \xi\left(\Sigma^{\frac{1}{N}}, U, I, \ldots, I, V^T\right) \tag{3.79}$$

where ξ is the map given in Equation 2.7 since $\mathbf{A} \in \mathcal{M}$. Clearly, $\phi(\mathbf{A}) = A$. Furthermore, Lemma 2.2.7 gives a set of orthonormal basis vectors for the horizontal space $\mathcal{H}_{\mathbf{A}}$:

$$\begin{aligned}
l_s^k &= \frac{1}{\sqrt{N}}\begin{cases} u_k e_k^T & s = N \\ e_k v_k^T & s = 1 \\ e_k e_k^T & \text{o/w} \end{cases}, \\
u_s^{k,l,0} &= C_s^{k,l}\begin{cases} u_l e_k^T & s = N \\ e_l v_k^T & s = 1 \\ e_l e_k^T & \text{o/w} \end{cases}, \\
u_s^{k,l,0} &= D_s^{k,l}\begin{cases} u_k e_l^T & s = N \\ e_k v_l^T & s = 1 \\ e_k e_l^T & \text{o/w} \end{cases}.
\end{aligned} \tag{3.80}$$

So, note that checking if a matrix $X = \xi(\Lambda, Q_N, \ldots, Q_0) \in \mathcal{M}$ is in the span of $\mathbf{l}$, $\mathbf{u}^{k,l,0}$, and $\mathbf{u}^{k,l,N}$ is equivalent to finding a set of coefficients $\alpha \in \mathbb{M}_d$ such that

$$\begin{aligned} U\alpha &= Q_N \Lambda Q_{N-1}^T \\ \alpha &= Q_{p+1} \Lambda Q_p^T, 1 \leq p \leq N-2 \\ \alpha V^T &= Q_1 \Lambda Q_0^T. \end{aligned} \tag{3.81}$$

It is immediately obvious that $\alpha = \Sigma^{\frac{1}{N}}$ gives that $\mathbf{A} \in \mathcal{H}_{\mathbf{A}}$.

Now, also consider

$$\mathbf{B} = \xi\left(\tilde{\Sigma}^{\frac{1}{N}}, \tilde{U}, Q, \ldots, Q, \tilde{V}^T\right), \tag{3.82}$$

where $Q = U^T\tilde{U} = V^T\tilde{V}$ by assumption. Clearly, $\phi(\mathbf{B}) = B$. Furthermore, $\mathbf{A}$ and $\mathbf{B}$ satisfy the sufficient condition of Lemma 3.4.1 since $Q^T = \tilde{U}^T U = \tilde{V}^T V$ and $\mathcal{Q}_p$ is constant for all p. Thus, the line $\ell(\mathbf{A}, \mathbf{B})$ is indeed the geodesic between $\mathbf{A}$ and $\mathbf{B}$ in $(\mathcal{M}, \iota)$.

$\gamma'(0)$ is Horizontal. We have already shown that $\mathbf{A}$ is in $\mathcal{H}_{\mathbf{A}}$. Thus, in order to show the line is horizontal at $t = 0$, we need to show that $\mathbf{B} \in \mathcal{H}_{\mathbf{A}}$ as well. We thus return to solving the system of equations in Equation 3.81, which under the parameterization of $\mathbf{B}$ collapses to

$$U\alpha = \tilde{U}\tilde{\Sigma}^{\frac{1}{N}}Q^T, \tag{3.83}$$

$$\alpha = Q\tilde{\Sigma}^{\frac{1}{N}}Q^T, \tag{3.84}$$

$$\alpha V^T = Q\tilde{\Sigma}^{\frac{1}{N}}\tilde{V}^T, \tag{3.85}$$

and has the obvious solution given by the middle equations of $\alpha = Q\tilde{\Sigma}^{\frac{1}{N}}Q^T$ which is consistent with the others. Thus, $\mathbf{B} \in \mathcal{H}_{\mathbf{A}}$.

Since we have shown that both $\mathbf{B}$ and $\mathbf{A}$ are in $\mathcal{H}_{\mathbf{A}}$ and that the geodesic linking them is a straight line, it follows from Equation 3.78 that $\gamma'(0)$ is horizontal. Applying Theorem 3.1.2, we conclude that $\phi \circ \gamma$ is a geodesic in $(\mathfrak{M}_d, g^N)$ of the same length as γ. The explicit formula comes directly from applying ϕ to γ between our parameterized $\mathbf{A}$ and $\mathbf{B}$. □

Corollary 3.4.2.1 (Diagonal A and B)**.** When A and B are diagonal, we have that the geodesic is

$$\begin{aligned} \gamma(t) &= \ell\left(A^{1/N}, B^{1/N}\right)^N \\ &= \sum_{p=0}^{N} \binom{N}{p} (1-t)^p t^{N-p} A^{p/N} B^{(N-p)/N}. \end{aligned} \tag{3.86}$$

Proof. Apply the theorem and expand the geodesic formula using that diagonal matrices commute. □

Remark 3.4.3. *Other examples of when A and B satisfy this right-left rotation condition is if they are both symmetric or if they have the same singular vectors ($Q = Id$).*

Chapter 4

Numerical Simulations

4.1 Balanced Flow Simulation

We aim to simulate the flow in Equation 2.9

$$\dot{X} = -\sum_{p=1}^{N} \left(XX^T\right)^{\frac{N-p}{N}} dE(X) \left(X^T X\right)^{\frac{p-1}{N}}. \tag{4.1}$$

Numerical schemes require a computation of the tangent vector corresponding to $\dot{X}$ at every integration step. However, naive recomputation of the right hand side at every iteration of the integration would involve recomputing the SVD of X to compute the fractional powers of XX^T and X^TX. This quickly becomes prohibitively expensive as d increases. Furthermore the summation includes $O(N)$ terms, which also becomes expensive as N increases and we cannot explicitly work with the $N \to \infty$ case. Throughout our prior computations, we observed that the seemingly "natural" coordinate system for the balanced manifold flow is in the SVD coordinates, as seen in the parameterization in Equation 2.7 and Theorems 2.1.2, 3.3.2, and 3.4.2 all placing emphasis on the SVD coordinates through assumptions or statements. This gives us the insight that the system should be evolved through the singular value coordinates $U(t)$, $\Sigma(t)$, and $V(t)$.

4.1.1 Simulation Methodology

In order to perform the desired coordinate change, we must first understand how the respective components of the SVD change along a path $X(t)$.

Lemma 4.1.1 (SVD Differentiation, Cohen et al. [2023])**.** Suppose $X(t) = U(t)\Sigma(t)V(t)^T$ and assume that $\sigma_i(t) \neq \sigma_j(t)$ for all t. Then, the time derivatives of the individual components can be

computed as

$$\dot{\sigma}_i(t) = u_i^T(t)\dot{X}(t)v_i(t), \tag{4.2}$$

$$\dot{u}_i(t) = \sum_{j\neq i} \frac{1}{\sigma_i^2 - \sigma_j^2} u_j^T(t)(\dot{X}(t)X^T(t) + X(t)\dot{X}^T(t))u_i(t)u_j(t), \tag{4.3}$$

$$\dot{v}_i(t) = \sum_{j\neq i} \frac{1}{\sigma_i^2 - \sigma_j^2} v_j^T(t)(\dot{X}^T(t)X(t) + X^T(t)\dot{X}(t))v_i(t)v_j(t). \tag{4.4}$$

Proof. Fix $1 \leq i \leq d$. Firstly, notice that

$$U^TU = I \implies U^T\dot{U} \in \mathbb{A}_d, \tag{4.5}$$

which implies that $u_i^T\dot{u}_i = 0$ (similarly for v_i and $\dot{v}_i$). Thus, we know that

$$\dot{u}_i = \sum_{j\neq i} c_{ij}u_j, \dot{v}_i = \sum_{j\neq i} d_{ij}v_j. \tag{4.6}$$

Now, consider differentiating the equation

$$X(t)v_i(t) = \sigma_i(t)u_i(t). \tag{4.7}$$

This gives

$$\dot{X}(t)v_i(t) + X(t)\dot{v}_i(t) = \dot{\sigma}_i(t)u_i(t) + \sigma_i(t)\dot{u}_i(t). \tag{4.8}$$

By left multiplying by $u_i^T(t)$ and using the orthogonality properties and the expansions from equation (4.6), we find that this implies

$$u_i^T(t)\dot{X}(t)v_i(t) = \dot{\sigma}_i(t). \tag{4.9}$$

Now, let's compute $\dot{u}_i(t)$ and $\dot{v}_i(t)$. We can do this by considering differentiating the identities

$$X(t)X^T(t)u_i(t) = \sigma_i^2(t)u_i(t), \tag{4.10}$$

$$X^T(t)X(t)v_i(t) = \sigma_i^2(t)v_i(t). \tag{4.11}$$

We will do the first expression and the second is exactly the same. In particular, we get

$$(\dot{X}(t)X^T(t) + X(t)\dot{X}^T(t))u_i(t) + X(t)X^T(t)\dot{u}_i(t) = \left(\frac{d}{dt}\sigma_i^2(t)\right)u_i(t) + \sigma_i^2(t)\dot{u}_i(t). \tag{4.12}$$

Multiplying this time by u_j^T where $j \neq i$ and plugging in the expansions from equation (4.6),

$$u_j^T(t)(\dot{X}(t)X^T(t) + X(t)\dot{X}^T(t))u_i(t) + \sigma_j^2(t)c_{ij} = \sigma_i^2(t)c_{ij}. \tag{4.13}$$

Collecting terms gives that

$$c_{ij} = \frac{1}{\sigma_i^2 - \sigma_j^2} u_j^T(t)(\dot{X}(t)X^T(t) + X(t)\dot{X}^T(t))u_i(t). \tag{4.14}$$

Similarly we find that

$$d_{ij} = \frac{1}{\sigma_i^2 - \sigma_j^2} v_j^T(t)(\dot{X}^T(t)X(t) + X^T(t)\dot{X}(t))v_i(t). \tag{4.15}$$

These can be plugged into equation (4.6) to get $\dot{u}(t)$. □

We can now use Lemma 4.1.1 to change coordinates and derive the corresponding flows for $U(t)$, $\Sigma(t)$, and $V(t)$.

Lemma 4.1.2 (Balanced Flow in SVD Coordinates, Cohen et al. [2023])**.** Let $\mathfrak{s}(M) = M - M^T$ and let

$$L_N(\Sigma)_{ij} = \begin{cases} \frac{\lambda_{ij}^N}{\sigma_i^2 - \sigma_j^2} & i \neq j \\ 0 & i = j \end{cases} \tag{4.16}$$

where λ_{ij}^N is the eigenvalue for the eigenvector $u_i v_j^T$ of $\mathcal{A}_{N,X}$. Then,

$$\dot{U} = U\mathfrak{s}\left((L_N(\Sigma)\Sigma) \circ \left(U^T dE(X) V\right)\right) \tag{4.17}$$

$$\dot{\Sigma} = -\Sigma^{2-2/N} \operatorname{diag}\left(U^T dE(X) V\right) \tag{4.18}$$

$$\dot{V} = V\mathfrak{s}\left((\Sigma L_N(\Sigma)) \circ \left(U^T dE(X) V\right)\right) \tag{4.19}$$

Remark 4.1.3. *λ_{ij}^N can come from any of the versions of $\mathcal{A}_{N,X}$: infinite depth or finite depth with/without $\frac{1}{N}$ normalization. In our numerical experiments, we will largely use the infinite depth limit for simplicity.*

Proof. These equations can be verified by simply plugging in the $\dot{X}$ dynamics from Equation 2.9 into the equations from Lemma 4.1.1 and cleaning up using the antisymmetric projection $\mathfrak{s}$. □

Corollary 4.1.3.1. Assume X_0 diagonal and $E(X)$ depends only on the singular values of X. Then, for each i, we have that

$$\dot{\sigma_i} = -\sigma_i^{2-2/N} \partial_{\sigma_i} E(X) \tag{4.20}$$

and the matrix X remains diagonal.

By integrating the flow in the above lemma instead of the naive flow in Equation 2.9, we reduce the number of SVD calls from 1 every iteration to just 1 at initialization. Furthermore, the complexity no longer depends on the depth, as previous we would have to compute a summation $\sum_{p=1}^{N}$. We integrate the flow using RK45 [Dormand and Prince, 1980].

4.2 Low Rank Attraction

Much of this section replicates or lightly builds on results from Cohen et al. [2023] and Gunasekar et al. [2017]. In specific situations, Cohen et al. [2023] observed a **low rank attraction** phenomenon, or that solutions to the flow in Equation 2.9 would tend to be biased toward low rank solutions. Gunasekar et al. [2017] found the same phenomenon in a more general energy setup but only restricted to the BW geometry, even characterizing the solution in a special energy setup as the solution minimizing the nuclear norm. As a simple demonstration of this phenomenon, consider the diagonal matrix completion energy defined in $d = 2$ as

$$E(X) = (X_{11} - 1)^2 + (X_{22} - 1)^2. \tag{4.21}$$

The set of rank-1 minimizers of E are given by matrices of the form

$$X = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix} = \begin{bmatrix} 1 & \alpha \\ \frac{1}{\alpha} & 1 \end{bmatrix}. \tag{4.22}$$

Cohen et al. [2023] observe an implicit attraction to this particular set of minimizers, especially when the initial conditions were small. Consider X_0 with elements drawn from $\mathcal{N}(0, 0.01^2)$ in the infinite depth case. We can plot the resulting X_{12} and X_{21} distributions (Figure 4.1a), observing concentration of measure around the expected low rank hyperbola $X_{21} = \frac{1}{X_{12}}$ for various initial conditions drawn from $\mathcal{N}(0, 0.01^2)$.

However, a more important question is how this low rank attraction occurs. The matrix must still be full rank, but it is becoming *effectively* low rank. Therefore, the singular value distribution is becoming skewed, but is it because lower singular values are decreasing or larger singular values increasing? In Figure 4.1b we plot the singular value dynamics for the exact same energy (diagonal matrix completion) but in $d = 5$ with initial condition drawn from $\mathcal{N}(0, 0.001^2)$. We see a clear picture: the small singular values are "lazy" and largely don't move. Instead, the top singular value is moved the most by the energy, meaning the skew in the singular value decomposition is coming from the largest singular value moving up.

Furthermore, this observation implies that as variance of the initial conditions goes to 0, we will observe an increasingly strong low rank attraction effect in the limit as $t \to \infty$, because the lowest singular values will become smaller yet the dynamics will still be dominated by the top singular value(s). This is *exactly* what Gunasekar et al. [2017] find: they find that in order to observe strong low rank attraction effects, it is important to closely simulate the gradient flow (which we do using higher order methods) *and* begin the initialization "close enough" to 0.

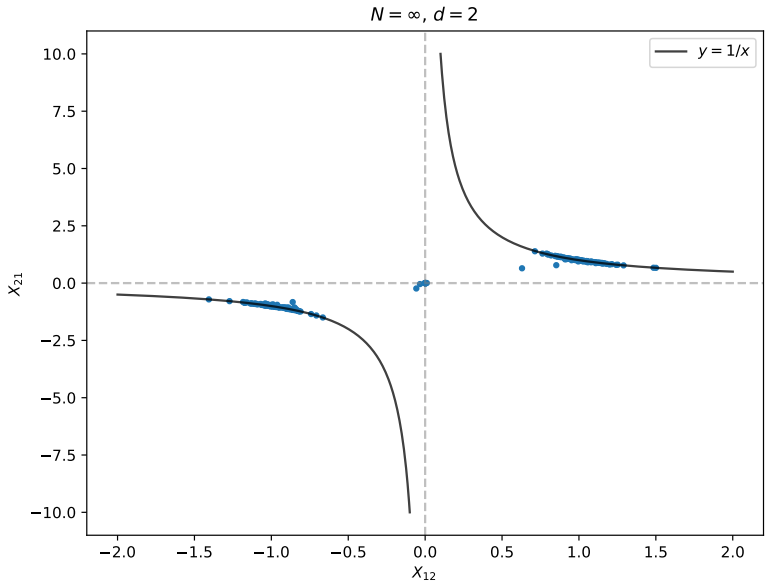
Effective Rank Dynamics. We would like some method to quantify "singular value distribution skew" as a low "effective" rank. That is the role of the following definition:

Definition 4.2.1 (Effective Rank)**.** Let $\{\sigma_i\}_{i=1}^d$ be the singular values and $\{s_i\}_{i=1}^d$ denote the normalized singular values $\frac{\sigma_i}{\sum_{i=1}^d \sigma_i}$. Then, the effective rank is defined as

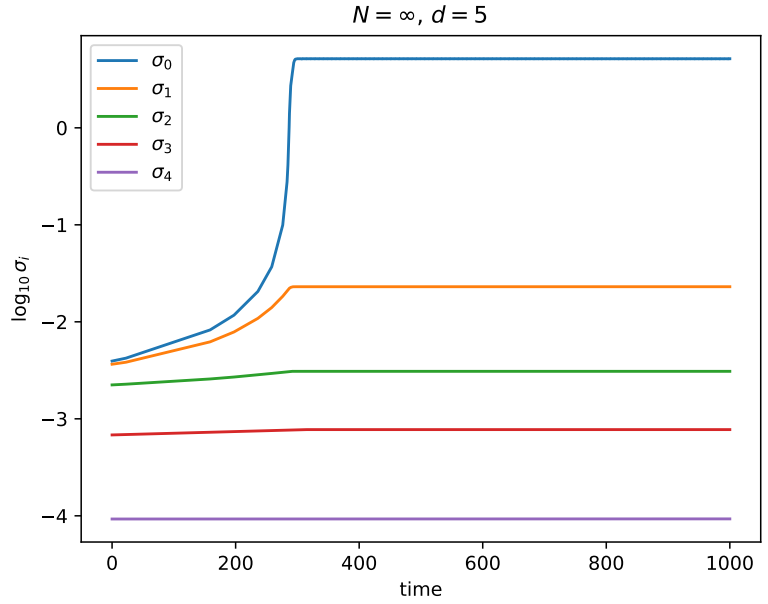
$$\exp\left(-\sum_{i=1}^{d} s_i \log s_i\right). \tag{4.23}$$

Suppose we draw initial conditions from $\mathcal{N}(0, \frac{1}{d})$, so that all eigenvalues are approximately unit and the initialization is more principled. In this case, the dynamics of effective rank are illuminating: for various d's, we observe a drastic gap between the distribution of initial effective ranks and final effective ranks (Figure 4.2), concretely quantifying the low rank phenomenon from earlier. The effect is also robust across various dimensions, indicating the attraction does not occur just in small dimensions.

Rank 1 Attraction Requires Small Initialization. As suspected, we also find a dependence of the low rank attraction on the variance of the initialization. It turns out that the smaller the variance is, the "stronger" the implicit regularization is, in the limit pushing all matrices to effectively rank 1. We simulate the effective rank dynamics for infinite depth diagonal matrix completion similar to Figure 4.2 but vary the initialization variance instead of the dimension d. In Figure 4.3, we observe that as $\sigma^2 \to 0$, the regularization becomes stronger and the final converged solutions essentially all become rank 1. This supports conjectured results from Gunasekar et al. [2017] on the importance of initialization in observing the low rank attraction effects.

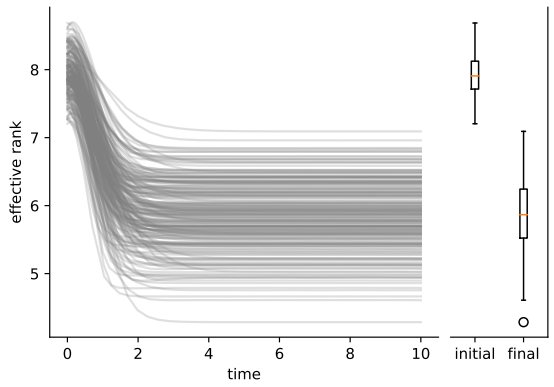


(a) Example of low rank attraction, $d = 2$

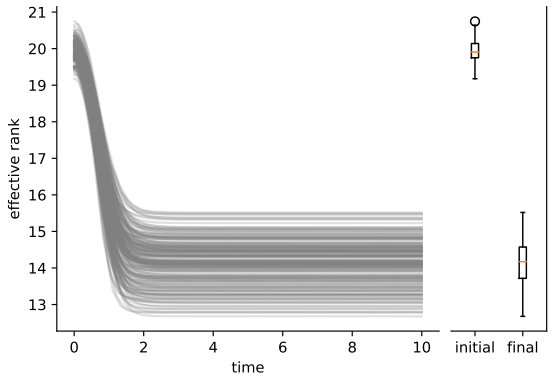


(b) Example of singular value dynamics, $d = 5$

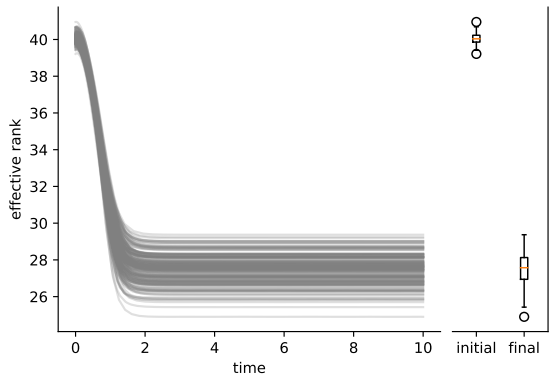
Figure 4.1: (a) Final values of X_{12} and X_{21} for $d = 2$. The converged solutions clearly concentrate around the low rank hyperbola $X_{21} = \frac{1}{X_{12}}$. (b) Singular value dynamics for example run with $d = 5$. We see that the rank attraction is because the top singular values move upward significantly relative to the other singular values. Initial condition was drawn from $\mathcal{N}(0, 0.001^2)$.



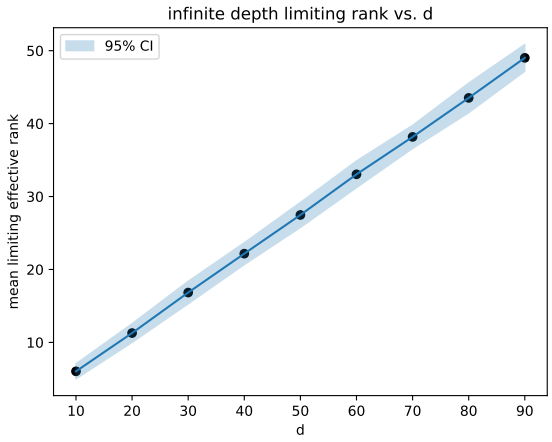
(a) $d = 10$



(b) $d = 25$



(c) $d = 50$



(d) Limiting rank vs. d

Figure 4.2: (a)-(c): Infinite depth effective rank dynamics for various dimensionalities for 200 initial seeds. (d): Mean (over 50 initializations drawn with $X_{ij} \sim \mathcal{N}(0, 1/d)$) limiting rank vs. depth under the infinite depth flow. We see a clear linear scaling with a slope of around 3/5, though we do not have theoretical results.

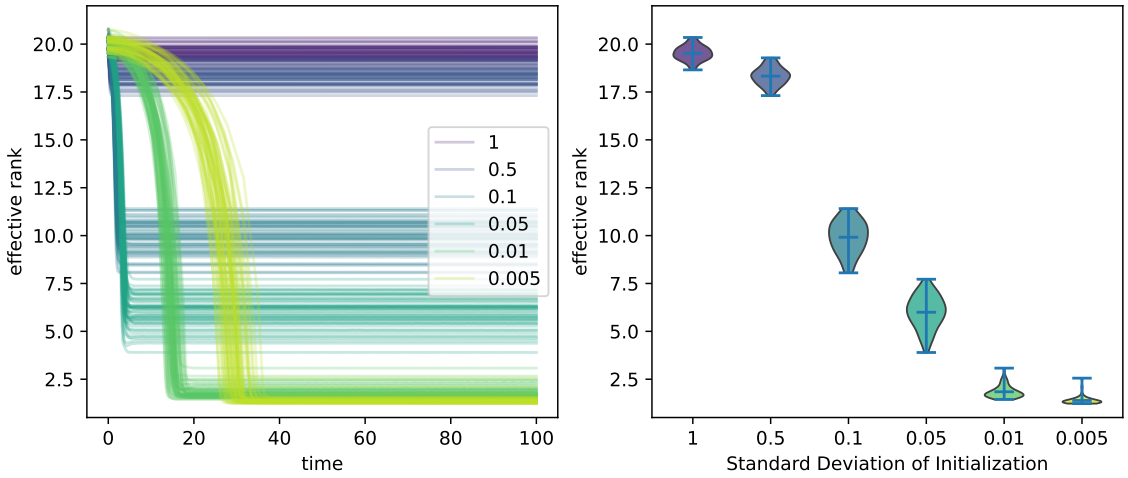


Figure 4.3: (Left) Effective rank dynamics for different initialization variances when $d = 25$ and $N = \infty$. (Right) Violin plot of effective rank distribution of solutions at $T = 100$ as a function of the standard deviation of the initialization.

Arbitrary Matrix Completion. Finally, zooming out a bit, we can notice that Equation 4.21 is a simple case of a more general **matrix completion** problem. We can consider a label matrix ℓ and index set $\mathcal{I}$, and the energy is defined as

$$E(X) = \frac{1}{2} \sum_{(i,j)\in\mathcal{I}} (X_{ij} - \ell_{ij})^2. \tag{4.24}$$

The derivative matrix is clearly

$$\partial_X E(X)_{ij} = 1_{(i,j)\in\mathcal{I}}(X_{ij} - \ell_{ij}). \tag{4.25}$$

Now, let's fix $d = 10$ and remain in the infinite depth situation. We want to more deeply study how this phenomenon is related to the sparsity of the label matrix i.e. $|\mathcal{I}|$. In diagonal matrix completion, we mask d elements to complete. What if we masked more or less - would the low rank attraction phenomenon become more or less pronounced?

To investigate this problem, we randomly generate various label matrices with fixed sparsity K and uniformly random labels ℓ_{ij} in the interval $[-2, 2]$. We conduct 50 runs of each label matrix and compute the mean effective rank at convergence. We then plot the distribution of these mean effective ranks, where the distribution is over randomly sampled label matrices with fixed sparsity.

In Figure 4.4, we find a U shaped pattern as a function of sparsities $K = [1, 10, 25, 50, 99]$ in the distribution of mean end effective rank - when very few elements or almost all elements are masked, the final effective rank tends to not change much. Otherwise, we observe low rank attraction in the middle region. This observation intuitively makes sense: if very few elements are being masked, most elements in the matrix will not be updated, so the effective rank will be similar to initialization. If almost all elements are being masked, then the effective rank will be similar to the label matrix, which in our case is almost always full rank. Otherwise, there is more freedom to "choose," leading us to observe implicit regularization.

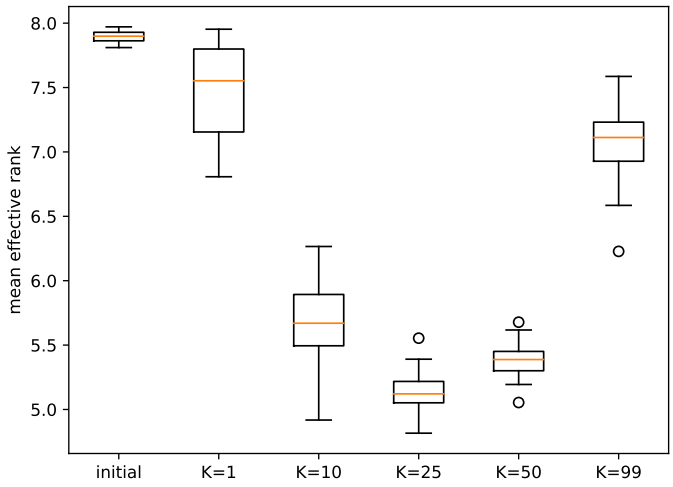


Figure 4.4: Distributions (over randomly drawn label matrices) of mean initial and end effective rank for various sparsities at $d = 10$ and the same initial conditions. We find a U-shaped pattern as sparsity increases.

4.3 Regularization by Entropy

So far, our numerics have focused solely on the downstairs flow without consideration for the upstairs space. However, it turns out that the upstairs space can introduce interesting modifications to the class of energy functions we can consider. The geometric intuition is to consider a definition of the *volume* of the group sitting above each point $X \in \mathbb{M}_d$.

Formally, let $\mathcal{F}_X$ denote the preimage fiber above a value $X \in \mathbb{M}_d$ (Equation 2.15). Then, the volume at each point X gives a sense of how "large" the upstairs fiber $\mathcal{F}_X$ is or, in the language of physics, how many microstates are associated with each observable/macrostate X. If $X \in \mathfrak{M}_d$, then we know from the parameterization in Equation 2.7 that above each point X sits an orthogonal group O_d^{N-1} (let this be denoted $\mathcal{O}_X$). One can compute [Menon, 2024] the volume of $\mathcal{O}_X$ as

$$\mathrm{vol}\,(\mathcal{O}_X) \propto \prod_{1\le j<k\le d} \sqrt{\frac{\sigma_j^2 - \sigma_k^2}{\sigma_j^{2/N} - \sigma_k^{2/N}}}. \tag{4.26}$$

Normally, the log of the volume is considered, giving the entropy $S_N(X)$

$$S_N(X) = \log \mathrm{vol}(\mathcal{O}_X) \propto \sum_{1\le j<k\le d} \frac{1}{2} \log\left(\frac{\sigma_j^2 - \sigma_k^2}{\sigma_j^{2/N} - \sigma_k^{2/N}}\right). \tag{4.27}$$

This also gives the infinite depth limit

$$S_\infty(X) \propto \sum_{1\le j<k\le d} \frac{1}{2} \log\left(\frac{\sigma_j^2 - \sigma_k^2}{\log \sigma_j^2 - \log \sigma_k^2}\right). \tag{4.28}$$

Remark 4.3.1. *These log volumes can also be understood as ratios of Vandermonde determinants. It can be shown that*

$$S_\infty(X) = \log \sqrt{\frac{\text{van}(\Sigma^2)}{\text{van}(\log \Sigma^2)}}, \tag{4.29}$$

where if $A = diag(a_1, \ldots, a_d)$, $\text{van}(A)$ *is the determinant of the matrix*

$$\begin{bmatrix} 1 & 1 & \ldots & 1 \\ a_1 & a_2 & \ldots & a_d \\ a_1^2 & a_2^2 & \ldots & a_d^2 \\ \vdots & & & \vdots \\ a_1^{d-1} & a_2^{d-1} & \ldots & a_d^{d-1} \end{bmatrix}. \tag{4.30}$$

These forms have deep connections to random matrix theory and this perspective should be explored further in future work.

Intuitively, this log volume is large when the singular values are very far apart and small when the singular values are all the same (one can minimize just this energy and observe this behavior). Cohen et al. [2023] noted that the rank 1 hyperbolas we observed in Figure 4.1a are *exactly* the high volume regions of the space. This observation inspires using the volume as a regularizer. However, we need to be careful to confirm some important properties. As further intuition on the nature of the log volume, we note the following properties of the entropy: a scaling identity, the first derivatives, and the strict concavity of S with respect to Σ.

Lemma 4.3.2 (Scaling of $S(X)$)**.** For $a \in \mathbb{R}$, we have that

$$S_N(aX) = \binom{d}{2}\left(1 - \frac{1}{N}\right)\log a + S_N(X), \tag{4.31}$$

$$S_\infty(aX) = \binom{d}{2}\log a + S_\infty(X). \tag{4.32}$$

Proof. These identities can be verified by direct computation with log rules. The main observation is that the singular values of aX are $|a|\sigma_i$ for $i \in \{1, 2, \ldots, d\}$. We will do S_∞ - the proof for S_N is the same.

$$S_\infty(aX) = \sum_{1 \le i < j \le d} \frac{1}{2}\log\left(\frac{(a\sigma_i)^2 - (a\sigma_j)^2}{\log(a\sigma_i)^2 - \log(a\sigma_j)^2}\right) \tag{4.33}$$

$$= \sum_{1 \le i < j \le d} \frac{1}{2}\log a^2 + S_\infty(X) \tag{4.34}$$

$$= \binom{d}{2}\log a + S_\infty(X). \tag{4.35}$$

Thus, we can always make $S_N(X)$ or $S_\infty(X)$ arbitrarily large by scaling the matrix and singular values. □

Lemma 4.3.3 (Derivative of Entropy)**.**

$$\frac{\partial S_N}{\partial \sigma_i} = \sum_{j \neq i} \frac{\sigma_i}{\sigma_i^2 - \sigma_j^2} - \frac{\sigma^{\frac{2-N}{N}}}{N\left(\sigma_i^{\frac{2}{N}} - \sigma_j^{\frac{2}{N}}\right)}, \tag{4.36}$$

$$\frac{\partial S_\infty}{\partial \sigma_i} = \sum_{j \neq i} \frac{\sigma_i}{\sigma_i^2 - \sigma_j^2} - \frac{\sigma_i^{-1}}{\log \sigma_i^2 - \log \sigma_j^2}. \tag{4.37}$$

Proof. Again, we will prove the result just for S_∞: S_N is largely the same. First, by log rules, we know that

$$S_\infty(X) \propto \sum_{1 \leq j < k \leq d} \frac{1}{2} \log(\sigma_j^2 - \sigma_k^2) - \frac{1}{2} \log(\log \sigma_j^2 - \log \sigma_k^2). \tag{4.38}$$

We can just directly differentiate the two terms separately, "selecting" out the terms in the summation that depend on σ_i.

$$\sum_{1 \leq j < k \leq d} \frac{\partial}{\partial \sigma_i} \frac{1}{2} \log(\sigma_j^2 - \sigma_k^2) = \frac{1}{2} \sum_{1 \leq j < i} \frac{-2\sigma_i}{\sigma_j^2 - \sigma_i^2} + \frac{1}{2} \sum_{i < k \leq d} \frac{2\sigma_i}{\sigma_i^2 - \sigma_k^2} \tag{4.39}$$

$$= \sum_{j \neq i} \frac{\sigma_i}{\sigma_i^2 - \sigma_j^2}. \tag{4.40}$$

$$\sum_{1 \leq j < k \leq d} \frac{\partial}{\partial \sigma_i} \frac{1}{2} \log(\log \sigma_j^2 - \log \sigma_k^2) = \frac{1}{2} \sum_{1 \leq j < i} \frac{-2\sigma_i/\sigma_i^2}{\log \sigma_j^2 - \log \sigma_i^2} + \frac{1}{2} \sum_{i < k \leq d} \frac{2\sigma_i/\sigma_i^2}{\log \sigma_i^2 - \log \sigma_k^2} \tag{4.41}$$

$$= \sum_{j \neq i} \frac{\sigma_i^{-1}}{\log \sigma_i^2 - \log \sigma_j^2}. \tag{4.42}$$

Finally, differencing these two terms gives the desired derivative form. □

We now recall some important theorems from linear algebra as sufficient properties to determine if matrices are positive definite. The following results, unless otherwise specified, are proved on S_∞ but are also true for S_N.

Lemma 4.3.4 (Positive Definite Diagonally Dominant Condition)**.** Let A be a $d \times d$ symmetric matrix. If

1. $A_{ii} > 0$ for all $1 \leq i \leq d$ and
2. $A_{ii} > \sum_{j \neq i} |A_{ij}|$,

Then, A is positive definite.

Corollary 4.3.4.1. Again, let A be a $d \times d$ symmmetric matrix. If

1. $A_{ii} < 0$ for all $1 \leq i \leq d$ and
2. $|A_{ii}| > \sum_{j \neq i} |A_{ij}|$,

Then, A is negative definite.

Proof. These results can be argued from the Gershgorin circle theorem [Horn and Johnson, 2012], which states that every eigenvalue of a matrix A lies in a *Gershgorin disc* $D\left(A_{ii}, \sum_{j\neq i} A_{ij}\right)$. $\square$

Lemma 4.3.5 (S_∞ Concavity)**.** S_∞ is strictly concave in σ.

Proof. For brevity, we omit the computation of many of the final inequalities that determine the signs of quantities. They all are the same proof - reparameterize in terms of $u = \log(\sigma_i/\sigma_j)$ and argue that the inequality holds when $u > 0$ using the inequalities $e^x \geq 1 + x$ and $e^{-x} \leq 1$ for all $x \in \mathbb{R}$.

We've already computed first derivatives in Lemma 4.3.3, so we just differentiate further and prove that the Hessian is negative definite. Recall that we assume $\sigma_i \neq \sigma_j$ for all $1 \leq i, j \leq d$. Then, we can compute the next derivative as

$$\frac{\partial^2 S_\infty}{\partial \sigma_i^2} = \sum_{j\neq i} \frac{(\sigma_i^2 - \sigma_j^2) - \sigma_i(2\sigma_i)}{(\sigma_i^2 - \sigma_j^2)^2} - \frac{-\sigma_i^{-2}(\log(\sigma_i/\sigma_j)) - \sigma_i^{-1}\sigma_i^{-1}}{2\left(\log(\sigma_i/\sigma_j)\right)^2} \tag{4.43}$$

$$= \sum_{j\neq i} \frac{-\sigma_i^2 - \sigma_j^2}{(\sigma_i^2 - \sigma_j^2)^2} - \frac{-\sigma_i^{-2}\log(\sigma_i/\sigma_j) - \sigma_i^{-2}}{2(\log(\sigma_i/\sigma_j))^2} < 0, \tag{4.44}$$

where the inequality follows from the mentioned procedure applied to each term in the summation (as in, each term is also strictly negative when $u > 0$). We also compute the cross terms.

$$\frac{\partial^2 S_\infty}{\partial \sigma_i \partial \sigma_j} = \frac{2\sigma_i\sigma_j}{(\sigma_i^2 - \sigma_j^2)^2} - \frac{\sigma_i^{-1}\sigma_j^{-1}}{2(\log(\sigma_i/\sigma_j))^2} < 0. \tag{4.45}$$

Now, we check the conditions of Corollary 4.3.4.1 on the Hessian. We've already noted that all elements $\frac{\partial^2 S_\infty}{\partial \sigma_i \partial \sigma_j}$ are negative for all $1 \leq i, j \leq d$. The domination condition is more subtle. We would like to show that

$$\sum_{j\neq i} \left| \frac{\partial^2 S_\infty}{\partial \sigma_i \partial \sigma_j} \right| < \left| \frac{\partial S_\infty}{\partial \sigma_i^2} \right| \tag{4.46}$$

for all $1 \leq i \leq d$. Because all cross derivatives *and* all terms in the summation that make up $\frac{\partial S_\infty}{\partial \sigma_i^2}$ are negative, we can account for the absolute value bars accordingly. Let $\frac{\partial S_\infty}{\partial \sigma_i^2} = \sum_{j\neq i} s_{ij}$ for clarity ($s_{ij} < 0$ when $\sigma_i > \sigma_j$). Namely, the want to show becomes

$$\sum_{j\neq i} -\frac{\partial S_\infty}{\partial \sigma_i \partial \sigma_j} < \sum_{j\neq i} -s_{ij}. \tag{4.47}$$

To show this inequality, it is sufficient to show that $-\frac{\partial S_\infty}{\partial \sigma_i \partial \sigma_j} < -s_{ij}$ for all i, j where $\sigma_i > \sigma_j$. Direct computation finds that

$$\frac{\sigma_i^2 + \sigma_j^2}{(\sigma_i^2 - \sigma_j^2)^2} - \frac{\log(\sigma_i/\sigma_j) + 1}{2\sigma_i^2(\log(\sigma_i/\sigma_j))^2} > \frac{\sigma_i/\sigma_j}{2\sigma_i^2(\log(\sigma_i/\sigma_j))^2} - \frac{2\sigma_i\sigma_j}{(\sigma_i^2 - \sigma_j^2)^2} \tag{4.48}$$

$$\frac{(\sigma_i + \sigma_j)^2}{(\sigma_i^2 - \sigma_j^2)^2} - \frac{\log(\sigma_i/\sigma_j) + 1 + \sigma_i/\sigma_j}{2\sigma_i^2(\log(\sigma_i/\sigma_j))^2} > 0 \tag{4.49}$$

$$\frac{1}{(\sigma_i - \sigma_j)^2} - \frac{\log(\sigma_i/\sigma_j) + 1 + \sigma_i/\sigma_j}{2\sigma_i^2(\log(\sigma_i/\sigma_j))^2} > 0. \tag{4.50}$$

which again follows from the aforementioned reparameterization trick. Thus, we conclude that the domination condition is true, and thus that the Hessian is negative definite and S_∞ is strictly concave

□

Equipped with these observations, the usual setup is to consider the *free energy* landscape at temperature parameter $\beta \in (0, \infty)$

$$F_\beta(X) = E(X) - \frac{1}{\beta} S(X), \tag{4.51}$$

extending the gradient flow to

$$\dot{X} = -\text{grad}_{g^N} F_\beta(X). \tag{4.52}$$

In some sense, we are using $-S(X)$ as a strictly convex regularizer to $E(X)$ with relative strength given by the inverse temperature $\frac{1}{\beta}$. Changing from the energy function $E(X)$ to the free energy $F_\beta(X)$ does not change the geometry of overparameterization. In order to numerically simulate $F_\beta(X)$ we use Lemma 4.1.1 to see that

$$\frac{\partial S_{[N,\infty]}}{\partial X} = U \text{diag}\left(\nabla_\sigma S_{[N,\infty]}\right) V^T. \tag{4.53}$$

We now investigate the convergence behavior of $F_\beta(X)$ with respect to special class of energies.

Convergence to Isotropic Matrices. The simplest example of a free energy is if we take

$$E(X) = \frac{1}{2} \text{Tr}(X^T X) = \frac{1}{2} \sum_{i=1}^{d} \sigma_i^2. \tag{4.54}$$

In this case, the equilibrium SVD distribution can be computed via differentiation and equating to 0 as a solution to the nonlinear system of equations

$$\sigma_i = \frac{1}{\beta} \sum_{j \neq i} \frac{\sigma_i}{\sigma_i^2 - \sigma_j^2} - \frac{\sigma_i^{-1}}{\log \sigma_i^2 - \log \sigma_j^2}. \tag{4.55}$$

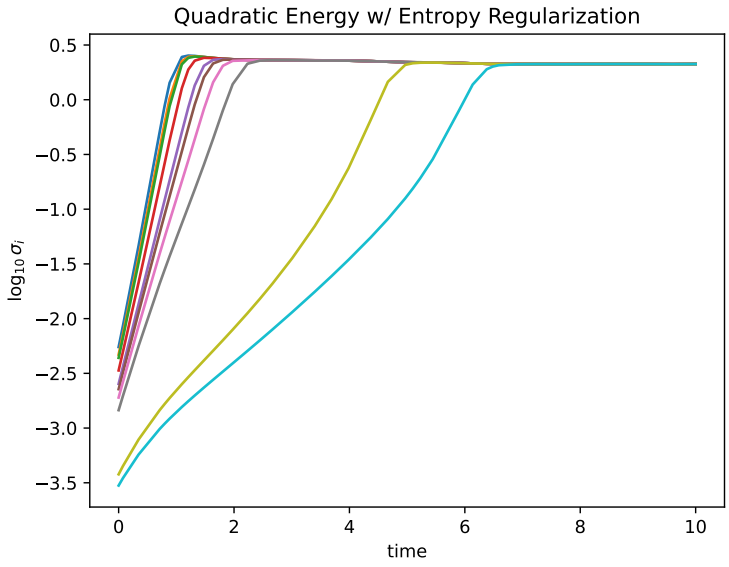
If we simulate the gradient flow, we find that in fact all the singular values converge to the same value (Figure 4.5a). Therefore, despite the fact that we are minimizing the L^2 norm of the singular values (which pushes the singular values towards 0) and maximizing $S(X)$ (which causes the singular values to spread out), when combined they "cancel" each other out and result in robust convergence to a full rank isotropic matrix with one eigenvalue and multiplicity d.

We make these observations rigorous through the following series of results for a more general form of energy functions $E_p(X) = \frac{1}{p} \sum_{i=1}^{d} \sigma_i^p$.

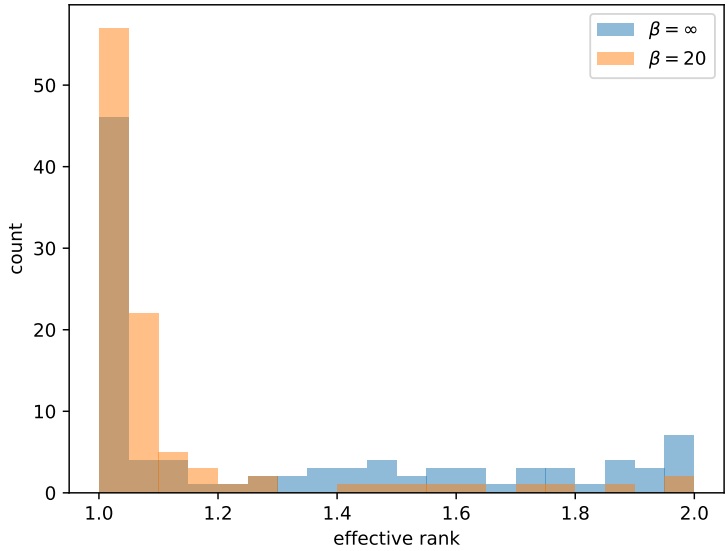
Lemma 4.3.6. We note a few properties about the expression

$$g(\sigma_i, \sigma_j) := \frac{\sigma_i}{\sigma_i^2 - \sigma_j^2} - \frac{\sigma_i^{-1}}{\log \sigma_i^2 - \log \sigma_j^2}. \tag{4.56}$$

1. $g(\sigma_i, \sigma_j)$ is monotonically decreasing in its first argument.



(a) Singular value dynamics of quadratic energy with regularization



(b) Effective ranks with and without specially chosen regularization.

Figure 4.5: (a) Singular value dynamics with $F_\beta(X) = \frac{1}{2}\operatorname{Tr}(X^T X) - S_\infty(X)$ (i.e. $p = 2$) and $d = 10$ in the infinite depth. We see convergence to an isotropic matrix, verifying Theorem 4.3.7. The limiting singular value is also exactly as expected: $\log_{10}\sqrt{\frac{d-1}{2}} \approx 0.33$. (b) $d = 2$, $N = \infty$ simulation of diagonal matrix completion with and without explicit regularization via entropy. When regularization is included, the solutions are generally more tightly distributed at effective rank 1.

2. If $\sigma_i > \sigma_j$, $g(\sigma_i, \sigma_j) > g(\sigma_j, \sigma_i)$.

3. g has a clean limit as $\sigma_i \to \sigma_j$, namely

$$\lim_{\sigma_i \to \sigma_j} g(\sigma_i, \sigma_j) = \lim_{\sigma_i \to \sigma_j} \frac{\sigma_i}{\sigma_i^2 - \sigma_j^2} - \frac{\sigma_i^{-1}}{\log \sigma_i^2 - \log \sigma_j^2} = \frac{1}{2\sigma_j}. \tag{4.57}$$

Proof. For the first part, notice that $\sum_{j \neq i} g(\sigma_i, \sigma_j) = \frac{\partial S_\infty}{\partial \sigma_i}$. Thus,

$$\sum_{j \neq i} \partial_{\sigma_i} g(\sigma_i, \sigma_j) = \frac{\partial^2 S_\infty}{\partial \sigma_i^2}. \tag{4.58}$$

When proving the diagonal elements of the Hessian were negative, we proved each individual term in the summation was negative, so we conclude that $\partial_{\sigma_i} g(\sigma_i, \sigma_j) < 0$. The second argument is similar - rewrite $g(\sigma_i, \sigma_j)$ in terms of the ratio $r = \sigma_i/\sigma_j > 1$. The expression will be 0 when $r = 1$ but monotonically decreasing when $r > 1$.

To prove the limit, begin by letting $h = \sigma_i - \sigma_j$. Then, with some algebra we find that

$$\frac{\sigma_i}{\sigma_i^2 - \sigma_j^2} = \frac{h + \sigma_j}{h(h + 2\sigma_j)} = \frac{1}{2h}\left(1 + \frac{h}{\sigma_j}\right)\left(1 + \frac{h}{2\sigma_j}\right)^{-1}. \tag{4.59}$$

For $h \ll 1$, we have that the final term is the geometric series

$$\left(1 + \frac{h}{2\sigma_j}\right)^{-1} = 1 - \frac{h}{2\sigma_j} + \left(\frac{-h}{2\sigma_j}\right)^2 + \cdots = 1 - \frac{h}{2\sigma_j} + o(h^2). \tag{4.60}$$

Distributing gives that

$$\frac{\sigma_i}{\sigma_i^2 - \sigma_j^2} = \frac{1}{2h}\left(1 + \frac{h}{2\sigma_j} + o(h^2)\right). \tag{4.61}$$

Similarly, we can expand the second term.

$$\frac{\sigma_i^{-1}}{\log \sigma_i^2 - \log \sigma_j^2} = \frac{1}{2\sigma_i \log(\sigma_i/\sigma_j)} \tag{4.62}$$

$$= \frac{1}{2(h + \sigma_j)\left(\frac{h}{\sigma_j} - \frac{1}{2}\left(\frac{h}{\sigma_j}\right)^2 + \frac{1}{3}\left(\frac{h}{\sigma_j}\right)^3 + \dots\right)} \tag{4.63}$$

using that $\log(1+z) = z - z^2/2 + z^3/3 + o(z^4)$. Distributing and using that the sequences absolutely converge, we can intertwine the terms to collapse to

$$\frac{\sigma_i^{-1}}{\log \sigma_i^2 - \log \sigma_j^2} = \frac{1}{2h}\left(\left(\frac{h}{\sigma_j} - \frac{h^2}{2\sigma_j^2} + \dots\right) + \left(1 - \frac{h}{2\sigma_j} + \frac{h^2}{3\sigma_j^2} + \dots\right)\right)^{-1} \tag{4.64}$$

$$= \frac{1}{2h}\left(1 + \frac{h}{2\sigma_j} + o(h^2)\right)^{-1} \tag{4.65}$$

$$= \frac{1}{2h}\left(1 - \frac{h}{2\sigma_j} + o(h^2)\right) \tag{4.66}$$

where the final line follows from similar geometric series logic to before. Combining the two terms now reveals the clean limit.

$$\frac{\sigma_i}{\sigma_i^2 - \sigma_j^2} - \frac{\sigma_i^{-1}}{\log \sigma_i^2 - \log \sigma_j^2} = \left(\frac{1}{2h} + \frac{1}{4\sigma_j} + o(h)\right) - \left(\frac{1}{2h} - \frac{1}{4\sigma_j} + o(h)\right) \tag{4.67}$$

$$= \frac{1}{2\sigma_j} + o(h). \tag{4.68}$$

So as $h \to 0$, the limit is just $\frac{1}{2\sigma_j}$. One could also prove this by combining the fractions and applying L'Hopital's rule twice. □

Theorem 4.3.7 (Isotropic Equlibrium for Entropy Regularized p-Energies)**.** Fix $p > 1$ and consider the energy

$$E_p(X) = \frac{1}{p}\sum_{i=1}^{d} \sigma_i^p. \tag{4.69}$$

The equilibria of $F_\beta(X) = E(X) - \frac{1}{\beta} S_\infty(X)$ are the class of matrices

$$\left\{ X \text{ such that } \sigma_1 = \sigma_2 = \cdots = \sigma_d = \left(\frac{d-1}{\beta}\right)^{\frac{1}{p}} \right\}. \tag{4.70}$$

Proof. Assume for the sake of contradiction that at an equilibrium point there exist i, j such that $\sigma_i > \sigma_j$. Then, we can subtract the two corresponding equilibrium conditions from each other, giving

$$\sigma_i^{p-1} - \sigma_j^{p-1} = \frac{1}{\beta}\sum_{k\neq i} \frac{\sigma_i}{\sigma_i^2 - \sigma_k^2} + \frac{\sigma_i^{-1}}{\log \sigma_i^2 - \log \sigma_k^2} - \frac{1}{\beta}\sum_{k\neq j} \frac{\sigma_j}{\sigma_j^2 - \sigma_k^2} + \frac{\sigma_j^{-1}}{\log \sigma_j^2 - \log \sigma_k^2} \tag{4.71}$$

$$= g(\sigma_i, \sigma_j) - g(\sigma_j, \sigma_i) + \sum_{k\neq i,j} g(\sigma_i, \sigma_k) - g(\sigma_j, \sigma_k). \tag{4.72}$$

From Lemma 4.3.6, we see that every single difference on the right side is negative when the singular values are distinct, but the difference on the left hand side is positive because x^{p-1} is monotonically increasing when $p > 1$. This implies that at an equilibrium, it cannot be the case that $\sigma_i > \sigma_j$ for any $1 \leq i, j \leq d$.

Thus, we see that $\sigma_1 = \sigma_2 = \ldots = s$ at equilibrium. The equilibrium relationship reduces to

$$s^{p-1} = \frac{1}{\beta}\sum_{j\neq i} \lim_{\sigma_i \to \sigma_j} \left(\frac{\sigma_i}{\sigma_i^2 - \sigma_j^2} - \frac{\sigma_i^{-1}}{\log \sigma_i^2 - \log \sigma_j^2}\right) \tag{4.73}$$

$$= \frac{1}{\beta}\sum_{j\neq i} \frac{1}{2s} \tag{4.74}$$

$$s = \left(\frac{d-1}{2\beta}\right)^{\frac{1}{p}}. \tag{4.75}$$

□

Remark 4.3.8. *In fact, this theorem is still true for the wider class of symmetric energies*

$$E_p(X) = \sum_{i=1}^{d} f(\sigma_i) \tag{4.76}$$

where $f : \mathbb{R} \to \mathbb{R}$ is convex. The proof is exactly the same contradiction proof, except the left hand side of the equilibrium condition is nonnegative as justified by the convexity of f. Indeed,

$$f'(\sigma_i) - f'(\sigma_j) = (\sigma_i - \sigma_j) f''(\xi) \tag{4.77}$$

for $\xi \in (\sigma_j, \sigma_i)$ by mean value theorem. Since f convex and $\sigma_i > \sigma_j$ the right hand side is nonnegative. Obviously we are unable to obtain an explicit equilibrium value but the equilibrium points are thus still exactly

$$\{X | \sigma_1 = \sigma_2 = \cdots = \sigma_d\}. \tag{4.78}$$

Matrix Completion Regularization. Finally, we find that including the entropy in a matrix completion energy can indeed act as an explicit regularizer. Figure 4.5b displays the distribution of final effective ranks with and without including the explicit regularization with strength $\beta = 20$. However, generally the effects are very sensitive to β and initialization parameters, so we mark these results as mixed. We conjecture that the interaction between an energy that does not have the same symmetries as the log volume (which is invariant with respect to transformations of the form $X \mapsto Q_1 X Q_2$ with $Q_1, Q_2 \in O_d$) leads to a messy loss landscape that has peaks and valleys that are heavily dependent on how strong the regularization is.

Chapter 5

Conclusion and Discussion

We began with a distilled expository overview of the key results in the DLN focusing on explicit formulas for flows and invariant manifolds that we later use in numerical simulations, culminating in the equation for the flow

$$\dot{X} = -\sum_{p=1}^{N} \left(XX^T\right)^{\frac{N-p}{N}} dE(X) \left(X^T X\right)^{\frac{p-1}{N}}. \tag{5.1}$$

The emergence of a natural Riemannian geometry defined by the positive definite linear operator $\mathcal{A}_{N,X}$ led to new computations of the geodesics. We first studied the two proofs of the geodesics in the BW geometry subcase: one based on a mechanical Hamiltonian coordinates computation and another based on Riemannian submersion. We extended the proofs to the DLN case, finding a general geodesic equation and explicit formulas in a special case given by a rotational symmetry condition. To obtain the explicit formulas, we also characterized all straight lines that run through the balanced manifold in the process.

We concluded with a variety of numerical simulations that expand on existing intuition and understanding of low rank attraction in the training dynamics of matrix factorizations, such as experimenting with the dependence of the phenomenon on initialization and sparsity of the energy function. Finally, we construct a regularizer based on the log volume of an orthogonal group. We can derive explicit characterizations of the equilibrium states of the class of energies

$$F_\beta(X) = E_p(X) - \frac{1}{\beta} S_\infty(X) \tag{5.2}$$

When we attempt to use the regularizer for more general energies (matrix completion), we find mixed results likely because of the delicate interplay of invariances, as $S_\infty(Q_1 X Q_2) = S_\infty(X)$ for $Q_1, Q_2 \in O_d$.

5.1 Open Problems

Extending to True Deep Learning. The most obvious open problem is extending the theory to nonlinear neural networks. As a starting point, if the activation function is just the ReLU

$$\text{ReLU}(x) = \begin{cases} x & x > 0 \\ 0 & \text{o/w} \end{cases} \tag{5.3}$$

one can envision a neural network as *locally* a DLN, where locally is defined in the sense of the input space, as e.g. for all inputs where all the ReLU's activate, the function is trivially a DLN.

Another important limitation of the presented theory is that our energy functions and flows are completely deterministic. In empirical model training, the system incurs numerous sources of noise from the data. In general, one only has an approximation of the true energy function and true gradient at each discretization step, usually given as an average over a minibatch of noisy data samples, thus transforming the flow into a stochastic flow. Yu and Menon [2024] and Menon [2024] discuss a Riemannian Langevin equation (RLE) extension of the DLN that introduces sources of stochastic noise, but the idea of incorporating and studying the effects of noise on the gradient flow is still relatively preliminary. However, there is evidence to believe that noise could have benefits such as preventing local minima entrapment and further biasing solutions to flat/stable regions of the loss landscape [Sutskever et al., 2013, Smith et al., 2021], inspiring additional empirical tricks like Dropout layers [Srivastava et al., 2014].

If we desire even further applicability of our results to modern empirical deep learning, we also need to extend beyond just naive gradient descent, which is only guaranteed to find a global minima if the loss is convex. Modern deep learning models are largely trained with momentum-based optimizers like Adam/AdamW to help escape local minima [Kingma and Ba, 2014, Loshchilov and Hutter, 2017]. However, our results are limited to just the gradient flow setting.

Geodesics and Convergence Rates. On the more technical side of results, the computation of explicit geodesics in the general DLN case remains open, as we only compute the general ODEs but do not solve them. Geodesics are always fundamentally of interest in any Riemannian geometry, but in algorithms research they can provide further intuition on fundamental limitations of convergence rates of algorithms that operate in the geometry [Karmarkar, 1990]. Energies that result in gradient flow paths that remain close to the geodesic path between X_0 and the optimizer X^* are ideal to speed up convergence.

Low Rank Manifolds. The current work focuses on the full rank downstairs manifold $\mathfrak{M}_d$. However, as is clear from our numerical results, for particular energies, the matrix appears to have limiting behavior that gets asymptotically close to $\mathfrak{M}_r$ for $r < d$. It is likely important to also understand the behavior on $\mathfrak{M}_r$ more generally (Menon [2024] have parameterizations along lower rank manifolds).

General Energies. Restricting ourselves to matrix completion energies in the numerical experiments was also an arbitrary choice. We can generalize matrix completion even further to the general class of quadratic energies given by

$$E(X) = \sum_{i=1}^{M} \|\langle A_i, X\rangle - y_i\|_2^2, \tag{5.4}$$

where $A_i \in \mathbb{R}^{d\times d}$ are a series of M *measurement matrices* and $y \in \mathbb{R}^M$ are the *measurement labels*. M is usually $\ll d$. Under this setup $dE(X)$ is just

$$dE(X) = \sum_{i=1}^{M} (\langle A_i, X\rangle - y_i)A_i. \tag{5.5}$$

Gunasekar et al. [2017] works in this setup with the BW geometry and is able to derive some preliminary theoretical results on characterization of optima so it could be fruitful to extend their characterization to also apply when $X = \phi(\mathbf{W})$, as numerical simulations seem to imply it is the case.

We would also like to still more deeply understand the entropy form introduced in Section 4.1. For example, an interesting theoretical result would be to characterize the limit X_∞ for some free energy $F_\beta(X)$ either explicitly or implicitly as the solution to another optimization problem, such as the minimal nuclear norm solution [Gunasekar et al., 2017]. We also do not fully understand if there is limiting behavior as $d \to \infty$ in analogy to random matrix theory, which one might expect because of the appearance of Vandermonde determinants.

5.2 Concluding Thoughts

Understanding training dynamics of deep learning models is crucial as "post-mortem" evaluations (analyses done on fully trained models) are largely insufficient on their own to formulate a clear picture. The DLN is just one model of the training dynamics of deep learning models - one that is limited in scope on its own but simple enough to be amenable to seemingly endless explicit analysis. As increasing amounts of energy and resources is invested into training large scale AI models in various domains, understanding *how* and *why* the models work also becomes increasingly important not only for developing better models but also social responsibility. The DLN among other evidence gives us hope that this problem is not as impossible as previously imagined and the framework of isolating individual components like overparameterization under the umbrella of deep learning could be critical in efforts to understand large models.

Bibliography

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pages 242–252. PMLR, 2019.

Raman Arora, Sanjeev Arora, Joan Bruna, Nadav Cohen, Simon Du, Rong Ge, Suriya Gunasekar, Chi Jin, Jason Lee, Tengyu Ma, et al. Theory of deep learning, 2020.

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. *arXiv preprint arXiv:1810.02281*, 2018a.

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International conference on machine learning*, pages 244–253. PMLR, 2018b.

Bubacarr Bah, Holger Rauhut, Ulrich Terstiege, and Michael Westdickenberg. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *Information and Inference: A Journal of the IMA*, 11(1):307–353, 2022.

Rajendra Bhatia. *Positive definite matrices*. Princeton university press, 2009.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the bures–wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

Nadav Cohen, Govind Menon, and Zsolt Veraszto. Deep linear networks for matrix completion—an infinite depth limit. *SIAM Journal on Applied Dynamical Systems*, 22(4):3208–3232, 2023.

Manfredo Perdigao Do Carmo and J Flaherty Francis. *Riemannian geometry*, volume 2. Springer, 1992.

J.R. Dormand and P.J. Prince. A family of embedded runge-kutta formulae. *Journal of Computational and Applied Mathematics*, 6(1):19–26, 1980. ISSN 0377-0427. doi: https://doi.org/10.1016/0771-050X(80)90013-3. URL `https://www.sciencedirect.com/science/article/pii/0771050X80900133`.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli,

Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. *Advances in neural information processing systems*, 30, 2017.

David R. Hilbert and S. Cohn-Vossen. *Geometry and the Imagination*. Chelsea Publishing, 1952.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Narendra Karmarkar. Riemannian geometry underlying interior-point methods for linear programming. *Contemp. Math.*, 114:51–75, 1990.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Govind Menon. The geometry of the deep linear network. *arXiv preprint arXiv:2411.09004*, 2024.

David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning internal representations by error propagation, 1985.

Samuel L Smith, Benoit Dherin, David GT Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. *arXiv preprint arXiv:2101.12176*, 2021.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147. PMLR, 2013.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Tianmin Yu and Govind Menon. *Riemannian Langevin Equation and Its Applications in Random Matrix Theory and Gibbs Sampling Problems*. Doctoral dissertation, Brown University, 2024.