The Deep Linear Network – Dynamics, Riemannian Geometry and Overparametrization

by

Zsolt Veraszto

B.Sc., Budapest University of Technology and Economics; Budapest, Hungary, 2016

M.Sc., ETH Zürich; Zürich, Switzerland, 2018

A dissertation submitted in partial fulfillment of the

requirements for the degree of Doctor of Philosophy

in The Division of Applied Mathematics at Brown University

PROVIDENCE, RHODE ISLAND

May 2023

This dissertation by Zsolt Veraszto is accepted in its present form
by The Division of Applied Mathematics as satisfying the
dissertation requirement for the degree of Doctor of Philosophy.

Date_____ _____

Govind Menon, Ph.D., Advisor

Recommended to the Graduate Council

Date_____ _____

Nadav Cohen, Ph.D., Reader

Date_____ _____

Jérôme Darbon, Ph.D., Reader

Approved by the Graduate Council

Date_____ _____

Thomas A. Lewis, Dean of the Graduate School

Curriculum Vitae

Zsolt Veraszto graduated from Budapest University of Technology where he received a B.Sc. in Mechatronical Engineering. He also graduated from ETH Zürich where he received a M.Sc. in Mechanical Engineering.

Acknowledgments

To Nadav Cohen, Jérôme Darbon, Patrick Flynn, George Haller, Florian Kogelbauer, Govind Menon, Eszter Rónai, Gábor Stépán, Brigitta Szilágyi, Csilla Szonderik, Gyula Szonderik, Anita Varga, László Verasztó, Petra Verasztó and Charlene Wooten: according to mathematical tradition, the names are listed in alphabetical order. I trust all of you to know your respective contributions to this journey. Thank you all.

CONTENTS

LIST OF FIGURES

Chapter One

Introduction

Theoretical research in deep learning concerns itself with three primary questions: expressivity, convergence and generalization. Expressivity pertains to the characterization of the space of functions that a particular network architecture can represent. The issue of convergence pertains to the asymptotic behavior of training. For instance, does the training process converge to a (local or global) minimum? Is the training process able to escape saddle points efficiently? Finally, the question of generalization pertains to how well the model performs on test data, and in particular its potential for overfitting.

Overfitting generally occurs when the ratio of parameters to data points exceeds a certain threshold. When a model has too many parameters, it attempts to reproduce the statistical noise in the observations, rather than meaningful features of the data. While overfitted models perform very well on the training data, they tend to show poor performance on new, unseen data. In classical regression and machine learning, this is normally mitigated using some explicit regularization technique, e.g. modifying

the loss function by an explicit regularizer term.

Implicit regularization, on the other hand, is the phenomenon where certain optimization algorithms produce simpler models, which are less prone to overfitting. There is a longstanding belief that deep learning architectures avoid overfitting better than their wide and shallow counterparts. The deep linear network (DLN) is a model for implicit regularization in gradient based optimization of deep learning architectures.

The DLN consists of compositions of linear functions, and therefore the network implements the linear function represented by the product of the weight matrices of individual layers. This renders the question of expressivity trivial, since the space of functions represented by matrices is linear maps. From the deep learning perspective, a simple way to think about the DLN is to consider a fully connected feedforward network, stripped from its activation functions. While expressivity is solely a feature of the architecture, the study of convergence and generalization also requires a specific choice of loss function.

Our choice of loss function corresponds to a relaxed version of the matrix completion problem. The optimization objective is to minimize quadratic distance from a target matrix with only a subset of its elements represented in the loss function. This setting is a model for an extreme case of overparametrization. A single observation of some matrix elements is available, while the number of learnable parameters can be arbitrarily large depending on the number of layers in the network.

A simple model of training is gradient flow on the space of weight matrices (referred to as "upstairs") given the above described loss function. Previous work on the DLN also showed that training admits more structure. The product of weight

matrices, referred to as "downstairs", satisfies a Riemannian gradient flow during training. The Riemannian metric downstairs is defined by the architecture of the network (i.e. products of matrices). The main idea is to equip the downstairs space with a metric such that the upstairs and downstairs spaces are related by a Riemannian submersion. This construction makes the pushforward of the gradient flow upstairs a gradient flow downstairs.

We extend this geometric framework to obtain a matrix representation for the Riemannian metric under varying depth, including the limit case of infinite depth. We derive explicit formulas for the volume forms in both natural matrix coordinates and singular value decomposition coordinates. We also give formulas of a moving frame that can aid computations of higher order geometric quantities.

In the infinite depth limit, one loses correspondence to a practical, finite network upstairs. Nevertheless, we get a far simpler model that captures the main qualitative features of finite depth. We compare the infinite depth DLN to its finite depth counterparts using numerical simulations of training downstairs. Upon analyzing various examples, we conclude that the infinite depth model exhibits the same implicit regularization as was previously observed in case of finite depth.

It is an intuitive idea that rank measures the "simplicity" of a linear function, and therefore implicit regularization can be understood using rank. Bias towards low rank minimizers has been observed in the literature. Increasing depth strengthens implicit bias towards low rank matrices. We provide more examples of this phenomenon, proposing a new, geometric understanding of implicit regularization.

We show that volume of the state space blows up at low rank matrices. We observe that the set of low rank solutions to the matrix completion problem is larger

than the set of minimizers observed upon training. Moreover, we observe that certain low rank completions are favored over others. Using analysis and numerics, the blowup of volume at these observed minimizers is demonstrated to be faster than at unobserved low rank minimizers. Based on our findings, we propose that implicit regularization is a result of bias towards high state space volume, with rank merely serving as a proxy for this phenomenon.

Chapter Two

Deep Linear Networks for Matrix Completion – An Infinite Depth Limit

2.1 Introduction

2.1.1 The deep linear network

Deep learning has proven its general applicability in several fields of applied science in the past decade [21]. While neural networks are structurally simple function approximators (e.g. [20]), several questions around them remain unanswered. Two fundamental problems are to obtain first principles explanations for generalizability and implicit regularization. Generalizability refers to a network's performance on new, previously unseen data. Implicit regularization is the feature of deep networks to avoid overfitting despite overparametrization. In the context of classical regression problems, overfitting is mitigated by explicit regularization. Deep learning architec-

tures are observed not to overfit despite the lack of explicit regularization. This is referred to as implicit regularization.

A simple model in which the effect of overparametrization in deep networks can be studied is the deep linear network (DLN) [1, 2]. The DLN is simple enough to serve as a minimal model for neural networks. It may also be applied directly to optimization problems, such as matrix completion [8], autoencoders [3] and multi-task training. The most common approach to the training process is gradient based minimization of a loss function, a process known as empirical risk minimization.

Recent work has shown that the DLN is also of intrinsic mathematical interest. Training the DLN corresponds to a gradient flow of a loss function with a Riemannian structure determined by the architecture of the network. The purpose of this paper is to further develop the mathematical theory of DLN. The main new contributions are explicit formulas for volume forms, including the infinite depth limit, that parallel corresponding expressions in random matrix theory. A numerical investigation that interprets some aspects of these formulas is also presented. We define the DLN below and then state these results more precisely.

2.1.2 The dynamical systems

Let $\mathbb{M}_d$ denote the space of real $d \times d$ matrices. The state space of the DLN consists of N weight matrices, $W_i \in \mathbb{M}_d$, $1 \leq i \leq N$, which we denote by

$$\mathbf{W} = (W_N, W_{N-1}, \ldots, W_1)\,. \tag{2.1.1}$$

The number of weight matrices, N, is referred to as the *depth* of the network. We have assumed that all matrices have the same size for simplicity. In fact, all that is required is that the matrix sizes be such that the product W in equation (2.1.2) is well-defined. Properties of the DLN in such generality have been studied in [2, 3]. We build on these results in our work, but restrict ourselves to $W_i \in \mathbb{M}_d$ for ease of computation. Some aspects of our analysis extend to a manifold of fixed rank matrices as in [3] (see Section 2.3.4.3).

The training process is an optimization problem for the weight matrices based on a lower dimensional observable. This observable, referred to as the *end-to-end matrix*, is the product of the weight matrices

$$W = \pi(\mathbf{W}) := \prod_{i=N}^{1} W_i. \tag{2.1.2}$$

Notice that the dimension of W is d^2 regardless of the network depth.

The gradient flow for a *loss function* $E(W)$

$$\frac{d}{dt} W_j = -\partial_{W_j} E(W), \quad 1 \le j \le N, \tag{2.1.3}$$

is used as our training model. Equation (2.1.3) is an idealization for the manner in which training is implemented in practice. Since W is defined through equation (2.1.2), we also find that

$$\dot{W}_j = -W_{j+1}^T \dots W_N^T \, \partial_W E(W) \, W_1^T \dots W_{j-1}^T, \quad j = 1, \dots, N. \tag{2.1.4}$$

Here and below, the notation $\partial_A f(A)$ for a differentiable function $f : \mathbb{M}_d \to \mathbb{R}$ denotes the Euclidean gradient, that is, for every $B \in \mathbb{M}_d$, $df(B) = \operatorname{trace}\left(\partial_A f(A)^T B\right)$.

Our interest lies in the long-time behavior of the observable $W(t)$. The limit $\lim_{t\to\infty} W(t)$ is called the *training outcome*. Since the dynamics is governed by a gradient flow, this limit exists when the loss function $E(W)$ is bounded below and has compact sublevel sets. However, natural loss functions, such as those for certain matrix completion problems, do *not* have compact sublevel sets. In fact, the minima of these loss functions are submanifolds of $\mathbb{M}_d$. Thus, the prediction of training outcomes is a subtle problem. The existence of $\lim_{t\to\infty} W(t)$ when each of the W_i has full rank was established using Lojasiewicz's theorem [3, Thm 10]. In order to use this convergence theorem as a foundation for the analysis in this paper, we restrict ourselves to $W_i \in GL(d)$ for most of the analysis. The space $GL(d)$ is the set of bijective linear transformations of $\mathbb{R}^d$, which is isomorphic to the set of d-by-d invertible matrices. Despite this assumption, training outcomes for many matrix completion problems have low rank. The repeated appearance of low rank matrices as training outcomes has stimulated the use of several heuristics for the prediction of training outcomes (see the discussion in [27]). The goal of this paper is to shed light on this question using the Riemannian geometry of the DLN and numerical simulations.

Let us now review the Riemannian geometry underlying the Euclidean gradient flow defined in equation (2.1.3). An important concept, see [1, Defn.1], is the notion of *balancedness*. We say that the weight matrices W_j and W_{j+1} are G_j-balanced if

$$G_j := W_{j+1}^T W_{j+1} - W_j W_j^T. \tag{2.1.5}$$

For fixed G_j, $1 \leq j \leq N$, equation (2.1.5) defines an algebraic variety $\mathbb{M}_d^N$ that is stratified by the rank. When $G_j = 0$, for each value of the rank r, the gradient flow (2.1.3) leaves the manifold $\mathcal{M}_r$ of rank-r matrices solving (2.1.5) invariant [3, Cor.6]. These are called the *balanced manifolds*. It is immediate from equation (2.1.5)

that the singular values of the weight matrices $\{W_j\}_{j=1}^N$ are equal on the balanced manifolds.

In order to use existing convergence theory, we restrict attention to full-rank matrices, and we denote $\mathcal{M}_d$ by $\mathcal{M}$ in what follows. This balanced manifold allows to separate the dynamics into a flow 'upstairs' in $\mathbb{M}_d^N$, described by equation (2.1.3), and a flow 'downstairs' for the end-to-end matrix $W(t)$ in $GL(d)$. The flow downstairs is a Riemannian gradient flow with a metric computed in [3] that may be described as follows. We define the linear map $\mathcal{A}_{N,W} : T_W GL(d) \simeq \mathbb{M}_d \to \mathbb{M}_d$

$$\mathcal{A}_{N,W}(Z) := \frac{1}{N} \sum_{j=1}^{N} \left(WW^T\right)^{\frac{N-j}{N}} Z \left(W^T W\right)^{\frac{j-1}{N}}. \tag{2.1.6}$$

On the balanced manifold the end-to-end matrix satisfies the Riemannian gradient flow

$$\dot{W} = -\mathrm{grad}_{g^N} E(W), \tag{2.1.7}$$

under the metric

$$g^N(Z_1, Z_2) = \mathrm{trace}\left(\mathcal{A}_N^{-1}\left(Z_1\right)^T Z_2\right). \tag{2.1.8}$$

This structure allows us to extend the DLN geometry the infinite depth limit, with the linear operator

$$\mathcal{A}_\infty\left(Z\right) = \lim_{N\to\infty} \mathcal{A}_N(Z) = \int_0^1 \left(WW^T\right)^{(1-\tau)} Z \left(W^T W\right)^{\tau} d\tau, \tag{2.1.9}$$

replacing $\mathcal{A}_N$ in equation (2.1.8) to define a limiting metric g^∞. When N is finite, the dynamical system (2.1.7) corresponds to a flow upstairs in $\mathbb{M}_d^N$. In the limit of infinite depth, equation (2.1.7) continues to hold, even though there is no longer a well-defined flow upstairs.

The existence of the infinite depth metric, in particular its resemblance to the Bogoliubov inner product in quantum statistical mechanics, was noted in [3, Remark 8]. We develop some properties of this metric below. The appearance of such elegant Riemannian structures in the DLN are surprising at first sight. However, it is helpful to note that fundamental interior point methods for conic programs also have a surprising gradient structure (see [4, 5, 15]).

The above structure applies to an arbitrary loss function. In practice, the loss function is determined by the learning task and the ease of computation. We focus on matrix completion in this paper, choosing the loss function $E(W)$ to be the quadratic distance from a fixed matrix Φ termed the *optimization objective* [11]. Using $\circ$ for element-wise (Hadamard) product, the family of loss functions we study takes the form

$$E_{\mathcal{B}}(W) = \frac{1}{2}\|\mathcal{B} \circ (\Phi - W)\|_2^2. \tag{2.1.10}$$

Here $\mathcal{B}$ is a matrix whose entries are either zero or one. This notation allows us to include several forms of matrix completion. An important example is the following: Let $\mathcal{B} = I$ select the diagonal elements, and consider

$$E_I(W) = \frac{1}{2}\|\mathrm{diag}\,(\Phi - W)\|_2^2 \tag{2.1.11}$$

Here $\mathrm{diag}(\cdot)$ returns a diagonal matrix constructed from the diagonal elements of its arguments.

The nature of the loss function is determined by the matrix $\mathcal{B}$. For example, the energy defined by (2.1.11) has a submanifold of global minima. Indeed, given a diagonal matrix Φ, $E_I(W)$ vanishes when W is chosen to be any matrix whose diagonal entries are Φ. Moreover, since the matrix can be completed to any rank

from 1 to d, some of these minima are noninvertible matrices. We consider numerical examples with several such $\mathcal{B}$.

2.1.3 Statement of Results

2.1.3.1 The metric and volume forms

The singular value decomposition (SVD) of W is denoted $W = U\Sigma V^T$, where $U, V \in O(d)$ and Σ denotes the diagonal matrix of singular values. We write $\Sigma_{ii} = \sigma_i$ and order the singular values in decreasing order: $\sigma_i \geq \sigma_j$ if $i < j$. The metric g^N may be expressed in a simple manner using the SVD. We find (see equation (2.2.17) below) that

$$g^N = (V \otimes U) D^N(\Sigma)(V \otimes U)^T. \tag{2.1.12}$$

Here $V \otimes U$ is the Kronecker product of V and U and $D_N \in \mathbb{R}^{d^2 \times d^2}$ is a diagonal operator with nonzero entries

$$D^N_{il} = \frac{N}{\sum_{j=1}^N (\sigma_i^2)^{N-j/N} (\sigma_l^2)^{j/N}}, \quad 1 \leq i, l \leq d. \tag{2.1.13}$$

Here the subscript il denotes a double index on the diagonal elements of D^N. The expression is unambiguous due to the symmetry of (2.1.13) in i and l. The expression (2.1.12) holds in the limit $N = \infty$, with

$$D^\infty_{il} = \frac{2\log(\sigma_i/\sigma_l)}{\sigma_i^2 - \sigma_l^2} \quad i \neq l, \quad D_{ii} = \frac{1}{\sigma_i^2}. \tag{2.1.14}$$

These expressions for the metric are used to compute the associated volume forms in $GL(d)$. Let $\mathrm{van}(\Lambda)$ denote the Vandermonde determinant of a diagonal matrix Λ,

let $d\Sigma$ denote Lebesgue measure on $\mathbb{R}^N$ and let dU and dV denote Haar measure on $O(d)$.

Theorem 2.1.1. *The volume form of g^N is given by*

$$\sqrt{\det g^N} dW = N^{\frac{d(d-1)}{2}} \det(\Sigma^2)^{\frac{1-N}{2N}} \operatorname{van}\left(\Sigma^{2/N}\right) d\Sigma dU dV. \tag{2.1.15}$$

In the limit $N \to \infty$, the volume form of g^∞ is

$$\sqrt{\det g^\infty} dW = \frac{\operatorname{van}(\log \Sigma^2)}{\sqrt{\det(\Sigma^2)}} d\Sigma dU dV. \tag{2.1.16}$$

The volume forms allow us to quantify the importance of regions of high volume that correspond to empirical observations of training outcomes. Notice that the volume density blows up when passing to the limit $\sigma_i \to 0$, showing a clear relationship between low rank and high volume. Formulas of this kind are also of interest in random matrix theory and suggest interesting asymptotics in DLN in the limits $N \to \infty$ and $d \to \infty$ (see [12] for similar investigations).

2.1.3.2 Normal hyperbolicity

The spectral decomposition of g^∞ given in equation (2.1.12) (with $N = \infty$) and (2.1.14) has important consequences for the dynamics given by equation (2.1.7). For different choices of $\mathcal{B}$, let $\mathcal{N}_\mathcal{B}$ denote the set of global minima of the energy function (2.1.10),

$$\mathcal{N}_\mathcal{B} = \{W : W \in \mathbb{M}_d,\, \mathcal{B} \circ (\Phi - W) = 0\}. \tag{2.1.17}$$

Recall that $\mathcal{B}$ is a matrix whose entries are either zero or one. Let K denote the number of zeros in $\mathcal{B}$,

$$K = d^2 - \sum_{i,j=1}^{d} \mathcal{B}_{ij}. \tag{2.1.18}$$

The set $\mathcal{N}_{\mathcal{B}}$ is an open submanifold of $GL(d)$ with dimension K.

Theorem 2.1.2. *Any K-dimensional compact submanifold of $\mathcal{N}_{\mathcal{B}}$ is normally hyperbolic under the dynamical system (2.1.7) for $N = \infty$.*

The statement is a consequence of Lemma 2.3.4. The proof is presented in Section 2.3.2.

2.1.3.3 Are the training outcomes low rank matrices?

The origin of implicit regularization was proposed to arise from (quasi-)norms in [25]. This idea is motivated by classical regression, where overfitting effects are often mitigated by adding explicit regularization terms to the optimization problem. Previous work in this direction tried to explicitly construct the regularizers in the form of (quasi-)norms. This explanation was challenged in [27], where the training outcomes were observed to be low rank matrices, even though matrix norms blew up. We develop this idea in Section 2.3. On one hand, we view the bias towards low rank matrices as an entropic effect determined by the volume forms above (the volume forms diverge as the singular values approach zero). On the other hand, we construct numerical examples where using low rank as a criterion does not accurately predict the training outcome under small initial conditions. The numerical examples show that withing the set of low rank minimizers, the actually observed ones are also the ones with the largest concentration of volume around them.

2.1.3.4 Numerical simulations for finite and infinite depth

Section 2.3 details numerical simulations of system (2.1.7) for $N \leq \infty$. As equation (2.1.6) is very costly computationally for large N and (2.1.9) can only be explicitly evaluated in terms of SVD coordinates, (2.1.7) is inefficient to simulate directly. To keep track of the evolution of SVD coordinates, we derive their dynamics directly under the flow of (2.1.7).

In the first few examples, we make use of energy function (2.1.11). In Section 2.3.3.1, we demonstrate that for large N and diagonal matrix completion, bias towards low rank should be viewed as bias towards *minimal* rank. This example also motivates the study of smaller matrices, where the matrix size and the minimal rank of minimizers are comparable. The example in Section 2.3.3.2 is on 2×2 matrices; yet it illustrates the effect of depth, demonstrating the validity of the infinite depth limit.

The examples of Section 2.3.4.1 show simulation outcomes for different choices of energy functions in the family (2.1.10). For these examples we chose $\mathcal{B}$ such that $E_{\mathcal{B}}$ has a finite number of rank deficient minimizers. We see that not all of the rank deficient minimizers are actually observed as training outcomes, contradicting the idea that low rank accurately predicts simulation outcomes. Instead, we find that the observed minimizers are the minimum rank minimizers with maximal volume.

Section 2.3.4.3 details an example where the entire state space has minimal rank. In this case, it is impossible to to predict simulation outcomes in terms of bias towards low rank. We demonstrate that high state space volume remains predictive, and the simulation outcomes are clustered in the region of state space where the volume is maximal.

2.1.4 Organization of the paper

The rest of this paper is organized as follows. In Section 2, we review the Riemannian geometry of the DLN and establish equations (2.1.12)–(2.1.13). We then extend this geometry to the infinite depth limit and prove Theorem 2.1.1. In Section 3 we use numerical experiments to show the correspondence between state space volume and implicit regularization. We also present several comparisons between finite depth networks and the infinite depth limit. Section 2.3.5 contains a collection of minor results specific to the energy function

$$E_1(W) = \frac{1}{2}\|W - \Phi\|^2. \tag{2.1.19}$$

Our conclusions are summarized in Section 2.4.

2.2 The Riemannian Geometry of DLN

This section contains several results on the state space geometry for the DLN. We first review the balanced manifold as well as the Riemannian submersion that determines the Riemannian metric for the DLN, basing our discussion on past work [2, 3]. This is followed by computations in SVD coordinates that provide matrix representations of the metric, including the infinite depth limit $N \to \infty$ (equations (2.1.12)–(2.1.13)). Finally, we prove Theorem 2.1.1 on the volume forms.

2.2.1 Riemannian submersion of the balanced manifold

The geometry on $GL(d)$ is defined by a Riemannian submersion of the balanced manifold, introduced in [3] and illustrated in Figure 2.1. This observation effectively reduces the dynamics of the DLN to a space of dimension d^2, independent of the depth N.

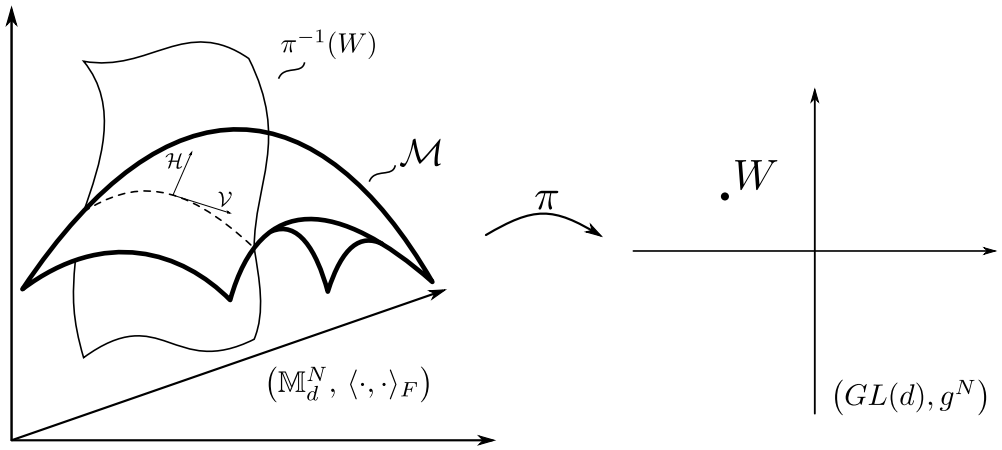


Figure 2.1: The left side is the optimization space and shows the balanced manifold as an immersed manifold. The right side represents the space of trained networks equipped with a Riemannian metric given by the Riemannian submersion of the balanced manifold.

Recall that $\mathbf{W} = (W_N, W_{N-1}, \ldots, W_1) \in \mathbb{M}_d^N$ (see equation (2.1.1)).

Definition 2.2.1 (Balanced manifold)**.** The balanced manifold of full rank matrices is

$$\mathcal{M} := \{\mathbf{W} | W_j \in GL(d), \text{ and } G_j = 0 \text{ for } j \in 1 \ldots N-1\} \tag{2.2.1}$$

For more on the interpretation of balancedness in machine learning, see [1]. The following lemma is included for completeness; it has already been established in [3].

Lemma 2.2.2 (Invariant manifold)**.** *The balanced manifold is invariant under the gradient flow* (2.1.3).

Proof. The invariance of G_j (not just $G_j = 0$) along trajectories can be checked by direct computation of the derivative along trajectories of (2.1.3):

$$\dot{G}_j = \dot{W}_{j+1}^T W_{j+1} + W_{j+1}^T \dot{W}_{j+1} - \dot{W}_j W_j^T - W_j \dot{W}_j^T = 0. \tag{2.2.2}$$

We need to show that (2.2.1) implicitly defines a manifold. For that we need to check if its differential is surjective as a linear map from $\mathbb{M}_d \times \mathbb{M}_d$ to the space of $d \times d$ symmetric matrices. In other words we need to show that for $W_j, W_{j+1} \in GL(d)$ and arbitrary symmetric S the equation

$$\dot{W}_{j+1}^T W_{j+1} + W_{j+1}^T \dot{W}_{j+1} - \dot{W}_j W_j^T - W_j \dot{W}_j^T = S \tag{2.2.3}$$

has a solution for $\dot{W}_j, \dot{W}_{j+1}$. It is easy to see that $\dot{W}_{j+1}^T W_{j+1} + W_{j+1}^T \dot{W}_{j+1}$ and $-\dot{W}_j W_j^T - W_j \dot{W}_j^T$ are arbitrary symmetric matrices. Thus the problem reduces to writing S as a sum of two symmetric matrices. □

The invariance of G_j in the above proof does not require that $G_j = 0$, $1 \leq j \leq N$. In practice, $\mathcal{M}$ is particularly important because the DLN is initialized with small initial conditions. Thus, all singular values are small, and the dynamical system begins close to $\mathcal{M}$.

Lemma 2.2.3. *(Symmetries) For $Q \in O(d)$, let $L_i(Q)$ denote the linear map*

$$L_i(Q)(\mathbf{W}) = \left(W_N, W_{N-1}, \ldots, W_{i+1}Q, Q^T W_i, \ldots, W_1\right). \tag{2.2.4}$$

Then $\pi\left(L_i\left(Q\right)\left(\mathbf{W}\right)\right) = \pi(\mathbf{W})$ and $L_i\left(Q\right)\left(\mathcal{M}\right) = \mathcal{M}$.

Proof. Both statements are obtained through direct computations. In order to see

that $\pi\left(L_i\left(Q\right)\left(\mathbf{W}\right)\right) = \pi\left(\mathbf{W}\right)$, we compute

$$W_N W_{N-1} \cdots W_{i+1} Q^T Q W_i \cdots W_1 = W_N W_{N-1} \cdots W_{i+1} W_i \cdots W_1,$$

since $QQ^T = I$.

Next assume $\mathbf{W}$ is balanced and observe that under the action of $L_i(Q)$

$$W_{i+1} Q Q^T W_{i+1}^T = W_{i+1} W_{i+1}^T = W_{i+2}^T W_{i+2}, \tag{2.2.5}$$

$$Q^T W_{i+1}^T W_{i+1} Q = Q^T W_i W_i^T Q \tag{2.2.6}$$

and

$$W_i^T Q Q^T W_i Q = W_i^T W_i = W_{i-1} W_{i-1}^T. \tag{2.2.7}$$

□

2.2.2 The operator $\mathcal{A}_{N,W}$ and the metric g^N

The metric g^N was defined in [3] using the linear operator $\mathcal{A}_{N,W} : T_W GL(d) \to \mathbb{M}_d$ (see equations (2.1.6)–(2.1.8) and [3, Defn.3]). We find it convenient to represent $\mathcal{A}_N$ and g^N in SVD coordinates since this yields explicit formulas for the metric and volume form. In all that follows we write the SVD of a matrix W as

$$W = U \Sigma V^T, \quad \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_d), \quad \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_d. \tag{2.2.8}$$

The standard basis of $\mathbb{M}_d$ is denoted by E_{ij}, $1 \leq i, j \leq d$.

Lemma 2.2.4 (Spectral decomposition of $\mathcal{A}_{N,W}$)**.** *For finite N, the d^2 number of*

eigenvalues of $\mathcal{A}_{N,W}$ *are*

$$\lambda_{il}^N = \frac{1}{N}\sum_{j=1}^{N}\left(\sigma_i^2\right)^{\frac{N-j}{N}}\left(\sigma_l^2\right)^{\frac{j-1}{N}} = \frac{\left(\sigma_i^2\right)^{\frac{N-1}{N}}}{N}\frac{1-\frac{\sigma_l^2}{\sigma_i^2}}{1-\left(\frac{\sigma_l^2}{\sigma_i^2}\right)^{\frac{1}{N}}},\ i,l\in 1,\ldots,d. \tag{2.2.9}$$

In the limit $N=\infty$*, the eigenvalues are*

$$\lambda_{il}^{\infty} = \frac{\sigma_i^2-\sigma_l^2}{2\log\left(\sigma_i/\sigma_l\right)},\ i,l\in 1,\ldots,d. \tag{2.2.10}$$

The corresponding eigenvectors are independent of N *and are given by*

$$T_{il} = UE_{il}V^T,\quad 1\le i,l\le d. \tag{2.2.11}$$

Proof. Let

$$Y=\mathcal{A}_N(X), \tag{2.2.12}$$

and introduce new variables $\tilde{X}=U^TXV$ and $\tilde{Y}=U^TYV$. By the definition of $\mathcal{A}_N$

$$\tilde{Y} = \frac{1}{N}\sum_{j=1}^{N}\left(\Sigma^2\right)^{\frac{N-j}{N}}\tilde{X}\left(\Sigma^2\right)^{\frac{j-1}{N}} \tag{2.2.13}$$

Notice that we have obtained a diagonal form in coordinates:

$$\tilde{y}_{il} = \tilde{x}_{il}\frac{1}{N}\sum_{j=1}^{N}\left(\sigma_i^2\right)^{\frac{N-j}{N}}\left(\sigma_l^2\right)^{\frac{j-1}{N}}, \tag{2.2.14}$$

which proves equation (2.2.9).

The proof of equation (2.2.10) is similar. Fix $i\neq l$ and take the limit $N\to\infty$ in

equation (2.2.14) to obtain

$$\lim_{N\to\infty} \lambda_{il}^N = \lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} \left(\sigma_i^2\right)^{\frac{N-j}{N}} \left(\sigma_l^2\right)^{\frac{j-1}{N}} = \int_0^1 \sigma_i^{2(1-t)} \sigma_l^{2t} dt = \frac{\sigma_i^2 - \sigma_l^2}{\log\left(\sigma_i^2/\sigma_l^2\right)}. \quad (2.2.15)$$

Finally, when $i = l$

$$\lim_{N\to\infty} \lambda_{il}^N = \lim_{N\to\infty} \left(\sigma_i^2\right)^{\frac{N-1}{N}} = \sigma_i^2. \quad (2.2.16)$$

□

As a direct consequence of Lemma 2.2.4, we obtain an explicit matrix representation for the DLN metric (2.1.8). Let the vectorization of the matrix coordinate one-forms be denoted dw^α, $\alpha = 1, \ldots, d^2$. Then the metric $g^N = g^N_{\alpha\beta} dw^\alpha \otimes dw^\beta$, $N \leq \infty$ at a point $W = U\Sigma V^T$ is given by

$$g^N = (V \otimes U)\, D^N(\Sigma)\, (V \otimes U)^T, \quad (2.2.17)$$

where D^N is the diagonal operator whose non-zero entries are

$$\frac{1}{\lambda_{il}^N}, \quad 1 \leq i, l \leq d. \quad (2.2.18)$$

The above calculations prove equations (2.1.12)–(2.1.13).

The following inequalities are used to establish normal hyperbolicity of an invariant manifold in Theorem 2.1.2 below.

Lemma 2.2.5 (Inequalities on the spectrum of $\mathcal{A}_{\infty,W}$)**.** *Let $1 \leq i < j \leq k < l \leq d$, then $\lambda_{jj}^\infty \leq \lambda_{ij}^\infty \leq \lambda_{ii}^\infty$ and $\lambda_{ij}^\infty \leq \lambda_{kl}^\infty$.*

Proof. $\lambda^{\infty}_{jj} = \sigma_j^2 \leq \sigma_i^2 = \lambda^{\infty}_{ii}$ since $\sigma_j \leq \sigma_i$.

Given a concave differentiable function f and $x < y$,

$$f'(y) \leq \frac{f(y) - f(x)}{y - x} \leq f'(x). \tag{2.2.19}$$

Thus

$$\frac{d}{dx}\log(x)|_{\sigma_j^2} = \frac{1}{\sigma_j^2} \geq \frac{\log\sigma_j^2 - \log\sigma_i^2}{\sigma_j^2 - \sigma_i^2} \geq \frac{d}{dx}\log(x)|_{\sigma_i^2} = \frac{1}{\sigma_i^2}, \tag{2.2.20}$$

where the middle term is

$$\frac{\log\sigma_j^2 - \log\sigma_i^2}{\sigma_j^2 - \sigma_i^2} = \frac{1}{\lambda^{\infty}_{ij}}. \tag{2.2.21}$$

The remaining inequality also follows from the concavity of $\log(x)$. □

2.2.3 Volume forms and the proof of Theorem 2.1.1

Vandermonde determinants appear often in random matrix theory as the Jacobians for diagonalization (see for example [10, §5.3], [23, §2.2] and Lemma 2.2.6 below). Theorem 2.1.1 reflects an analogous feature of the DLN geometry. The proof is an easy consequence of Lemma 2.2.4.

Proof of Theorem 2.1.1. We compute the determinants of the matrices g^N and g^{∞} as a product of the reciprocal of the eigenvalues of $\mathcal{A}_{N,W}$ given in Lemma 2.2.4. For

g^{∞} we find

$$\begin{aligned}\sqrt{\det g^{\infty}}dW &= \sqrt{\frac{1}{\sigma_1^2\dots\sigma_d^2}\prod_{i\neq j}\frac{\log\left(\sigma_i^2\right)-\log\left(\sigma_j^2\right)}{\sigma_i^2-\sigma_j^2}}dW \\ &= \frac{1}{\sqrt{\det\left(\Sigma^2\right)}}\prod_{i<j}\frac{\log\left(\sigma_i^2\right)-\log\left(\sigma_j^2\right)}{\sigma_i^2-\sigma_j^2}dW \\ &= \frac{\mathrm{van}\left(\log\Sigma^2\right)}{\sqrt{\det\left(\Sigma^2\right)}\mathrm{van}\left(\Sigma^2\right)}dW.\end{aligned} \tag{2.2.22}$$

Similarly, for $N<\infty$, when $i\neq l$ we use the eigenvalues of $\mathcal{A}_N$ to find

$$\begin{aligned}\lambda_{il}^N = \frac{1}{N}\sum_{j=1}^{N}\left(\sigma_i^2\right)^{\frac{N-j}{N}}\left(\sigma_l^2\right)^{\frac{j-1}{N}} &= \frac{\left(\sigma_i^2\right)^{\frac{N-1}{N}}}{N}\left(1+\left(\frac{\sigma_l^2}{\sigma_i^2}\right)^{\frac{1}{N}}+\dots+\left(\frac{\sigma_l^2}{\sigma_i^2}\right)^{\frac{N-1}{N}}\right) \\ &= \frac{\left(\sigma_i^2\right)^{\frac{N-1}{N}}}{N}\frac{1-\frac{\sigma_l^2}{\sigma_i^2}}{1-\left(\frac{\sigma_l^2}{\sigma_i^2}\right)^{\frac{1}{N}}}.\end{aligned} \tag{2.2.23}$$

In the case $i=l$ we obtain instead

$$\lambda_{il}^N = \left(\sigma_i^2\right)^{\frac{N-1}{N}}. \tag{2.2.24}$$

The volume form is given by the product of the reciprocals of all the eigenvalues λ_{il}^N, $1\leq i,l\leq d$. Thus,

$$\sqrt{\det g^N}dW = \sqrt{\prod_{i=1}^{d} \frac{1}{(\sigma_i^2)^{\frac{N-1}{N}}} \prod_{i\neq l} N\left(\sigma_i^2\right)^{\frac{1-N}{N}} \frac{1-\left(\frac{\sigma_l^2}{\sigma_i^2}\right)^{\frac{1}{N}}}{1-\frac{\sigma_l^2}{\sigma_i^2}}} dW \tag{2.2.25}$$
$$= \frac{N^{\frac{d(d-1)}{2}} \det(\Sigma^2)^{\frac{1-N}{2N}}}{V(\Sigma^2)} V\left(\Sigma^{2/N}\right) dW.$$

Finally, we use Lemma 2.2.6 below to express dW in SVD coordinates completing the proof of Theorem 2.1.1.

□

Remark (Volume form in new coordinates). Notice that using the coordinates $\Lambda = \log(\Sigma)$, we have $\frac{d\lambda_i}{d\sigma_i} = \frac{1}{\sigma_i}$ and thus

$$\sqrt{\det g^\infty}dW = 2^{\frac{d(d-1)}{2}} \text{van}\Lambda \, d\Lambda dU dV. \tag{2.2.26}$$

This shows a strong formal similarity between the DLN and the Gaussian Orthogonal Ensemble of random matrix theory, suggesting the study of probability distributions and large d asymptotics.

2.2.4 The Jacobian of singular value decomposition

The following Lemma is included for completeness since we were unable to find a convenient reference.

Lemma 2.2.6. *The Jacobian determinant of the SVD map:* $W \mapsto (U, \Sigma, V) \in O(d) \times \mathbb{R}^d \times O(d)$ *is given by the Vandermonde determinant* $\text{van}(\Sigma^2)$.

Proof. Assume W is a matrix with distinct singular values. Consider a C^1 curve $W(t)$, $t \in [-1, 1]$, with $W(0) = W$ such that the SVD coordinates of $W(t)$, written $U(t)\Sigma(t)V(t)^T$ are also C^1. Let $\dot{W}(0)$ be denoted $\dot{W}$ and similarly for U,Σ and V. We compute

$$\dot{W} = \dot{U}\Sigma V^T + U\dot{\Sigma}V^T + U\Sigma\dot{V}^T. \tag{2.2.27}$$

Since U, V are orthogonal and Σ is diagonal,

$$\dot{U} = UA^U, \quad \dot{V} = VA^V, \quad \dot{\Sigma} = D$$

where A^U and A^V are skew-symmetric and D is diagonal. Thus, we may write

$$\dot{W} = U(A^U\Sigma + D + \Sigma A^V)V^T. \tag{2.2.28}$$

This expression defines $\dot{W}$ as the image of a linear map from $TO(d) \times TO(d) \times T\mathbb{R}^d$ to $TGL(d)$. The Jacobian we are seeking is the determinant of this map. We compute this determinant, by first simplifying the expression above with the isometry $\dot{W} \mapsto U^T\dot{W}V$, and defining the linear transformation

$$\mathcal{L} : TO(d) \times TO(d) \times T\mathbb{R}^d \to \mathbb{R}^{d\times d}, \quad (A^U, A^V, D) \mapsto A^U\Sigma + D + \Sigma A^V. \tag{2.2.29}$$

Finally, we compute $\det(\mathcal{L})$ as follows. First observe that the action of $\mathcal{L}$ on diagonal matrices has eigenvalue 1 with multiplicity d. The action of $\mathcal{L}$ on $TO(d) \times TO(d)$ is given in coordinates by

$$\sum_j a^U_{ij}\delta_{jk}\sigma_j + \delta_{ij}\sigma_i a^V_{jk}. \tag{2.2.30}$$

Introducing the row-major half-vectorization for the skew-symmetric matrices (example for $d = 3$: $\mathrm{vec}(A^U) = [a^U_{12}\, a^U_{13}\, a^U_{23}]^T$) the action of $\mathcal{L}$ is now represented by

the block matrix with diagonal blocks:

$$\begin{bmatrix} D_1 & D_2 \\ -D_2 & -D_1 \end{bmatrix} \begin{bmatrix} \mathrm{vec}(A^U) \\ \mathrm{vec}(A^V) \end{bmatrix}, \tag{2.2.31}$$

where D_1 is

$$D_1 = \mathrm{diag}(\underbrace{\sigma_1, \ldots, \sigma_1}_{d-1}, \underbrace{\sigma_2, \ldots, \sigma_2}_{d-2}, \ldots, \underbrace{\sigma_{d-1}}_{1}), \tag{2.2.32}$$

and D_2 is

$$D_2 = \mathrm{diag}(\underbrace{\sigma_2, \ldots, \sigma_d}_{d-1}, \underbrace{\sigma_3, \ldots, \sigma_d}_{d-2}, \ldots, \underbrace{\sigma_{d-1}, \sigma_d}_{2}, \underbrace{\sigma_d}_{1}). \tag{2.2.33}$$

The matrices D_1 and D_2 commute because they are diagonal. Thus, the absolute value of the determinant of the block matrix above is

$$|\det(D_2^2 - D_1^2)| = |\det \mathrm{diag}(\sigma_i^2 - \sigma_j^2,\ i > j)| = \prod_{i<j} (\sigma_i^2 - \sigma_j^2) = \mathrm{van}(\Sigma^2). \tag{2.2.34}$$

□

2.3 Dynamics of matrix completion

This section is devoted to the gradient dynamics of the DLN for matrix completion. We consider quadratic energy functions as in equation (2.1.10) and equation (2.1.11). As $\mathcal{B}$ varies, the energy $E_{\mathcal{B}}$ may have a unique minimum or a submanifold of minima. By equation (2.1.7) the dynamics of the DLN is determined by an interplay between the Riemannian geometry and the nature of $E_{\mathcal{B}}(W)$.

This section primarily focuses on describing numerical experiments. However, in

order to generate efficial numerical simulations, it is necessary to first express the dynamics in SVD coordinates. These expressions are summarized in Theorem 2.3.2. We also prove Theorem 2.1.2 on normal hyperbolicity of submanifolds of equilibria to shed light on the attraction to the balanced manifold.

The main themes in the numerical experiments are as follows. First, we consider the variation with depth N, in order to demonstrate the utility of the infinite depth limit. Second, we construct energies where low rank heuristics are insufficient to explain the accumulation of training outcomes. Instead, we demonstrate that state space volume (as measured by the intrinsic Riemannian metric g^N) is a better predictor of the training outcome.

2.3.1 Numerical integration of equation (2.1.7) using SVD

Numerically integrating (2.1.7) gets very costly with increasing depth, due to the large number of matrix powers needed. Alternatively, one can use the factorized formula (2.2.17) to compute the Riemannian gradient through the dual metric:

$$\dot{w} = -g^{N*}(w)\partial E(w). \tag{2.3.1}$$

This formulation, however, requires the SVD of W at every time step. Instead, we compute the SVD of the initial condition and directly evolve the singular coordinates using smooth singular value decomposition. All numerical results in this paper are obtained using the fourth order, fixed timestep Runge-Kutta method. This formulation is stated in Theorem 2.3.2, building on the well known result Lemma 2.3.1, that we review for completeness.

Lemma 2.3.1 (Smooth SVD)**.** *Given a smooth curve $W(t) : (t_1, t_2) \to GL(d)$, $W(t)$ having distinct singular values for all $t \in (t_1, t_2)$, a smooth singular value decomposition $W(t) = U(t)\Sigma(t)V(t)^T$ exists satisfying the following system of differential equations:*

$$\dot{\sigma}_i = u_i^T \dot{W} v_i \tag{2.3.2}$$

$$\dot{u}_i = \sum_{j \neq i} \frac{1}{\sigma_i^2 - \sigma_j^2} \langle (\dot{W}W^T + W\dot{W}^T)u_i, u_j \rangle u_j \tag{2.3.3}$$

$$\dot{v}_i = \sum_{j \neq i} \frac{1}{\sigma_i^2 - \sigma_j^2} \langle (\dot{W}^T W + W^T \dot{W})v_i, v_j \rangle v_j \tag{2.3.4}$$

Proof. Under our assumptions these formulas can be verified by direct computation. For a more careful treatment for dealing with repeated singular values see [7]. □

Using Lemma 2.3.1 and equation (2.3.1), we can write down the evolution equations for the singular coordinates $U(t)\Sigma(t)V^T(t) = W(t)$ directly. For $N < \infty$, this result is equivalent to statements in [2](see Thm 3 and Lemma 2). Let $\mathfrak{s}(M) := M - M^T$, $\alpha = 1 - 1/N$ and $\circ$ denote the Hadamard (elementwise) product. For $N \leq \infty$ introduce the matrix:

$$L_N^{il} = \begin{cases} \frac{\lambda_{il}^N}{\sigma_i^2 - \sigma_l^2}, & \text{for } i \neq l \\ 0 \text{ otherwise.} \end{cases} \tag{2.3.5}$$

Theorem 2.3.2. *Under the assumptions of Lemma 2.3.1, and differentiable loss function E, the singular value decomposition of the end-to-end matrix evolve accord-*

ing to the differential equations:

$$\begin{aligned}\dot{U} &= U\mathfrak{s}\left(\left(L_N\left(\Sigma\right)\Sigma\right)\circ\left(U^T\partial_W E\, V\right)\right)\\ \dot{\Sigma} &= -\Sigma^{2\alpha}\mathrm{diag}\left(U^T\partial_W E\, V\right)\\ \dot{V} &= V\mathfrak{s}\left(\left(\Sigma L_N\left(\Sigma\right)\right)\circ\left(U^T\partial_W E\, V\right)\right).\end{aligned} \tag{2.3.6}$$

Proof. For $N<\infty$, see Theorem 3 and Lemma 2 in [2]. Under infinite depth, direct computation using 2.3.1 and Lemma 2.3.1 verifies the result. □

Recall that the state space of the Riemannian gradient flow is $GL(d)$. Since this is a dense, open subset of the space of all quadratic matrices of given size, any lower rank matrix can be approximated in this space. An appropriate way to quantify the rank of a matrix that is stable under numerical approximations is the following.

Definition 2.3.3 (Effective rank)**.** Let $W\in GL(d)$ with singular values σ_i and s_i be its normalized singular values

$$s_i=\frac{\sigma_i}{\sum_j\sigma_j}. \tag{2.3.7}$$

Then the effective rank of W is

$$r_e\left(W\right)=\exp\left(-\sum_i s_i\log\left(s_i\right)\right). \tag{2.3.8}$$

2.3.2 Attraction rates and the proof of Theorem 2.1.2

We derive bounds on the normal attraction rates for the submanifold of equilibria $\mathcal{N}_{\mathcal{B}}$. Given a choice of $\mathcal{B}$, let $\mathcal{I}$ denote the vectorized index set of observed elements,

that is

$$\operatorname{vec}(\mathcal{B})_i = \begin{cases} 1 \text{ if } i \in \mathcal{I} \\ 0 \text{ otherwise.} \end{cases} \tag{2.3.9}$$

In the following, we compute the linearization of equation (2.1.7) in the normal direction of $\mathcal{B}_{\mathcal{N}}$. We use coordinates $W = \mathcal{B} \circ \Phi + X + Y$, where $\mathcal{B} \circ X = X$ and $\mathcal{B} \circ Y = 0$. Lowercase x, y and w denote the vectorization of these coordinates. The dual metric tensor (the matrix of which is represented by the inverse of g^N) is denoted g^{N*}. Let $W_0 \in \mathcal{N}_{\mathcal{B}}$, be a fixed point.

Lemma 2.3.4. *At W_0, the linearization of equation (2.1.7) in the normal direction x is given by*

$$\dot{x} = -Ax, \tag{2.3.10}$$

where A is a principle submatrix of g^{N*}, consisting of rows and columns of indices observed by $\mathcal{B}$:

$$A = \left[g^{N*}\big|_{W_0} \right]_{i,j \in \mathcal{I}}. \tag{2.3.11}$$

Proof. Using the general form of the Riemannian gradient flow (2.3.1) and taking a derivative

$$\frac{d}{d\epsilon} g^{N*}(\epsilon x, y) \partial E_I((\epsilon x, y)) = \left(\partial_x g^{N*}\right) \partial E_I(w_0) + g^{N*}(w_0) \partial^2 E(w_0) x, \tag{2.3.12}$$

we notice that the first term is zero since $\partial_W E(W_0) = 0$ by W_0 being an equilibrium of the gradient flow. Computing

$$\partial^2 E_i(W_0)_{ij} = \begin{cases} 1, \text{ if } i = j, i \in \mathcal{I} \\ 0 \text{ otherwise,} \end{cases} \tag{2.3.13}$$

we see that multiplication by this matrix results in the above principle submatrix of g^{N*}. □

Proof of Theorem 2.1.2. We show a nonzero lower bound on the attraction rates. Let $\{\alpha_i\}_1^d$, $\alpha_i \leq \alpha_j$ if $j \leq i$ denote the eigenvalues of A. Then by the Cauchy interlacing theorem and Lemma 2.2.5,

$$\sigma_d^2 \leq \alpha_d \leq \alpha_{d-1} \leq \cdots \leq \alpha_1, \tag{2.3.14}$$

and note that this lower bound is nonzero on any compact subset on $\mathcal{N}_{\mathcal{B}}$. □

Notice that for Lemma 2.2.5 and the Cauchy interlacing theorem completely characterizes the order of λ_{ij}^{∞} and the characteristic exponents of the $d = 2$ case:

$$\sigma_2^2 \leq \alpha_2 \leq \frac{\sigma_1^2 - \sigma_2^2}{\log \sigma_1^2 - \log \sigma_2^2} \leq \alpha_1 \leq \sigma_1^2. \tag{2.3.15}$$

2.3.3 Diagonal matrix completion

In this section we show numerical simulations for the energy function E_I under variable N and d. For this choice of energy function, the rank of possible completions ranges from 1 to d, and so it constitutes one of the important cases for studying bias towards low rank in matrix completion problems.

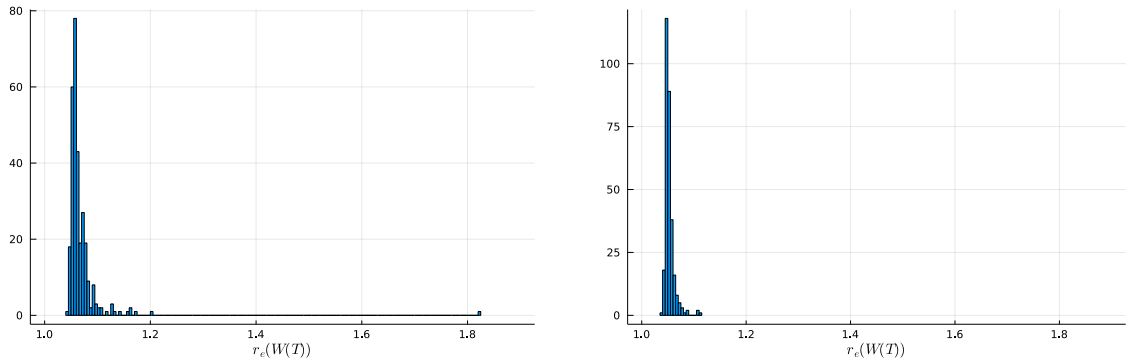
2.3.3.1 Example: Diagonal matrix completion, $d = 20$

We start with simulations of a larger, $d = 20$, example. Figures 2.2-2.3 show histograms of effective rank of optimization outcomes. Note that the rank and thus the effective rank here can be as large as 20, the size of the matrices and therefore bias towards low rank could mean convergence to a matrix of any effective rank smaller than 20.

Notice that even the shallowest, $N = 3$ case shows strong bias towards low rank completions. In case of $N = 10$ and $N = \infty$, the obtained histograms are nearly identical, showing strong bias towards minimal, rank one outcomes. This suggests that building intuition around the dynamics under sufficient depth is possible using small, $d = 2$ or $d = 3$ examples, in which case the rank deficient cases (2.1.3) are easier to characterize.

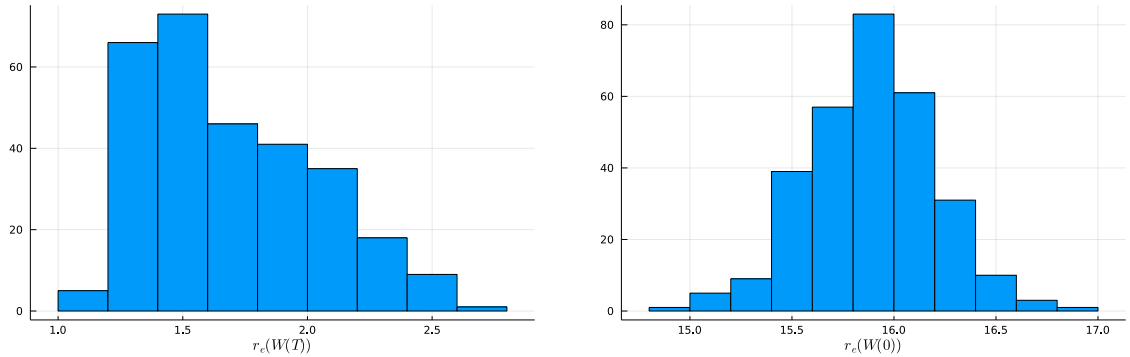
Approximately 300 optimization outcomes are included for each N. Panel (b) of Figure 2.3 shows the distribution of effective rank upon initialization. The small random initial conditions are drawn from a Wigner ensemble, specifically they consist of matrices of independent normal entries of mean zero and standard deviation 0.001, which distribution is denoted Wigner$(0, 0.001)$).

Numerical convergence criterion used for these examples is $E_I(W) < 10^{-6}$.



(a) $N = 10$ (b) $N = \infty$

Figure 2.2: Empirical distributions of effective rank in 20×20 matrix completion simulations.



(a) $N = 3$ (b) Initialization

Figure 2.3: (a) Effective rank distributions of shallow matrix completion simulations. (b) Sample distribution of effective rank at initialization for all above examples.

2.3.3.2 Example: Diagonal matrix completion, $d = 2$

In the following example for matrix completion, we illustrate how an increase in depth influences training outcome and at what depth do the results start to closely resemble our infinite depth limit. For easier visualization and building on the intuition developed in the previous section, we keep the width at the minimum $d = 2$. The results shown in Figure 2.4 are obtained by approximately 3000 runs for $N = 5, 10, 20$ and infinity each. Each of these figures show the outcome of these batch simulations.

For all of these batch runs, the optimization objective is a fixed invertible matrix of diagonal elements $[0.58724; 1.447]$. The small random initial conditions are drawn from Wigner$(0, 0.001)$[1]. Note that initializing the end-to-end matrix directly does not lead to the same distribution as initializing individual layers a similar way and then taking a product. However, for the purposes of this example, these two approaches lead to identical results.

Notice that here any matrix with fixed diagonal elements

$$\begin{bmatrix} \Phi_1 & w_{12} \\ w_{21} & \Phi_2 \end{bmatrix} \tag{2.3.16}$$

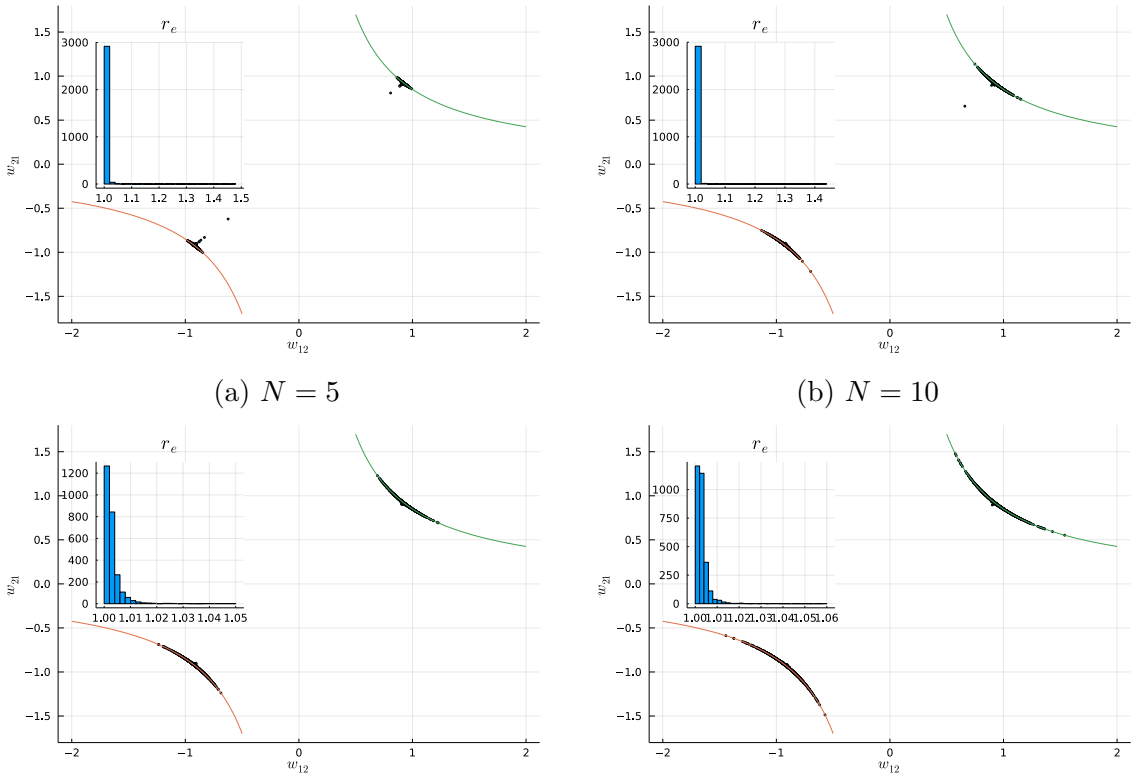
is a global minimizer. Batch simulations allow us to see whether there is a concentration of training outcomes on the w_{12}, w_{21} plane of possible global minimizers. Figure 2.4 show the results of these simulations, and Figure 2.5 show the clear correspondence to high phase space volume. The hyperbola on the left pane correspond to the rank one singularities on the manifold defined by

$$\frac{\Phi_1}{w_{21}} = \frac{w_{12}}{\Phi_2}. \tag{2.3.17}$$

At these singularities phase space volume blows up for any depth, and the probability of landing in this high volume region is expected to be high. The logarithmic volume can be interpreted as an entropic quantity [22], suggesting the use of statistical mechanical tools in our future analysis.

Numerical convergence criterion is $E(W) < 10^{-15}$ which is achieved in no more than $T = 1200$ simulation time in all included examples.

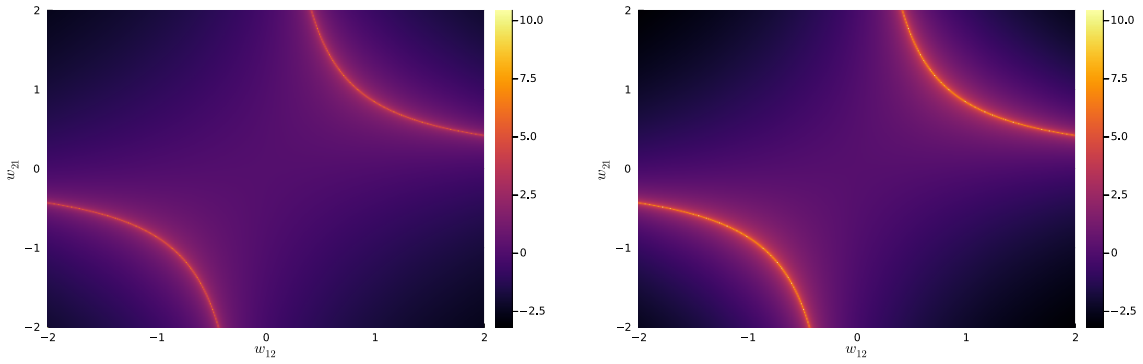
[1]The Wigner ensemble Wigner(μ, σ) is a distribution of random matrices with independent normal elements of mean μ and standard deviation σ.



(a) $N = 5$ (b) $N = 10$

(c) $N = 20$ (d) $N = \infty$

Figure 2.4: Outcomes of batch simulations under $N = 5, 10, 20$ and ∞, respectively. The black dots indicate training outcomes, while the red and green hyperbola lobes visualize rank one minimizers. Sample distributions of effective rank are also included in the embedded graphs.



(a) (b)

Figure 2.5: Heatmap of the logarithmic volume on the plane of global minimizers, for $N = 5$ and $N = \infty$, respectively. Note that the volume itself does not depend on the loss function, only the plane where the section was taken.

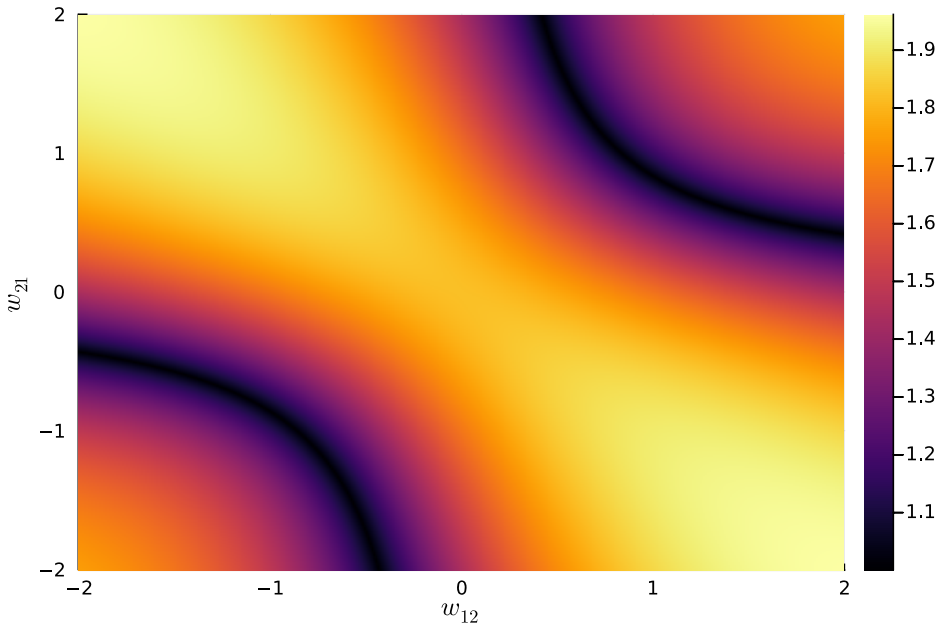


Figure 2.6: Effective rank of global minimizers.

Based on the example in section 2.3.3.1 and the results in Figure 2.4 we conclude that the infinite depth limit shows very similar behavior qualitatively to finite, sufficiently large depth models. Thus it is worth studying and a predictive theory of implicit regularization must be consistent with the infinite depth model.

Comparing Figure 2.4 and Figure 2.6, we see that there is no clustering of simulation outcomes in high effective rank regions. However, effective rank does not take the effect of depth into consideration. As seen in Figure 2.5, state space volume shows higher concentration at higher depth, just as simulation outcomes in higher depth show more concentration in these regions. While we are not ready to make this relation quantitative, these results suggest that state space volume is a good candidate to rely on for a quantitative theory of implicit regularization.

2.3.4 Other configurations

2.3.4.1 Example: Finite number of rank deficient minimizers

In the previous examples, both global minimizers and rank deficient global minimizers were nonunique. Here we provide an example, where minimizers are nonunique, but there is only one rank deficient (and high state space volume) minimizer. The setup is the following: the energy function is $E_{\mathcal{T}}(W)$,

$$\mathcal{T} = \begin{bmatrix} 1 & 1 \\ & \\ 0 & 1 \end{bmatrix}, \tag{2.3.18}$$

and $N = \infty$. In this case, the minimizers are of the form

$$\begin{bmatrix} \Phi_{11} & \Phi_{12} \\ w_{21} & \Phi_{11} \end{bmatrix}, \tag{2.3.19}$$

for arbitrary w_{21}, while imposing rank deficiency gives one solution, $w_{21} = \Phi_{11}\Phi_{22}/\Phi_{12}$. All simulation outcomes of 1000 random initial conditions drawn from Wigner$(0, 0.001)$ showed convergence to this minimizer. As illustration, 5 sample trajectories are shown in Figure 2.7.

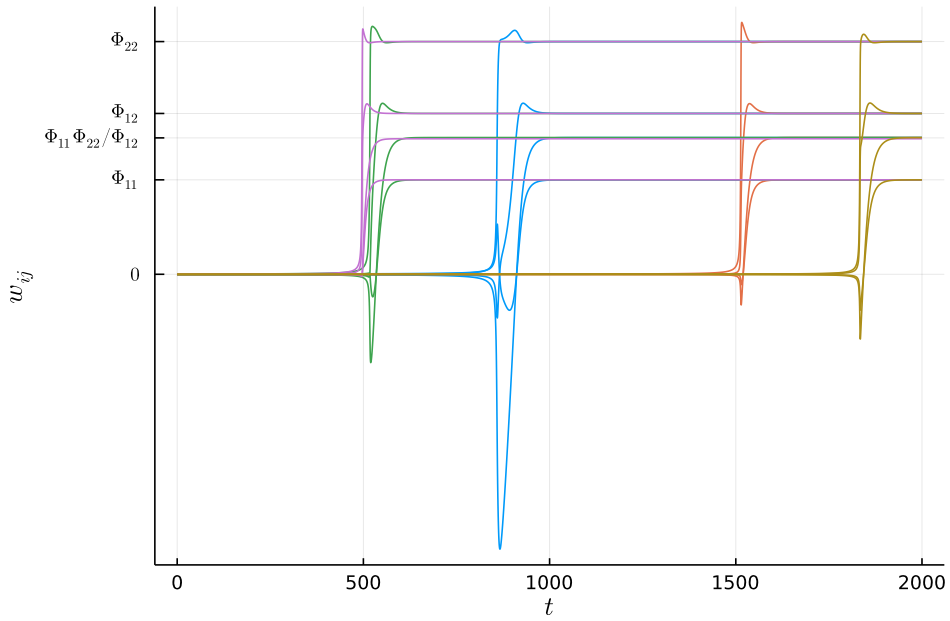


Figure 2.7: Five sample trajectories of the matrix elements under training are shown. All of them converging to the rank deficient minimizer.

2.3.4.2 A 3-by-3 example

We continue with a more complicated, $N = \infty$, $d = 3$ example. Computing the minimal rank to which a partially known matrix can be completed is in general nontrivial. For a detailed discussion of the computational complexity of matrix completion, see [13]). We covered cases before, where either the small matrix size, or the symmetry of the configuration in the observed elements (diagonal matrix completion) makes this problem trivial.

The equations for the missing elements can be written out by (symbolic) LU factorization under the assumption of a given rank. If the equations have a solution for the unknown elements, the matrix has a completion to the given rank. Solving the resulting system of multivariate polynomials may not be possible.

We show this process for an easily solvable case, take for example the first step of LU factorization for the following configuration of $\Phi_{ij} \neq 0$ and arbitrary w_{ij} we are trying to solve for:

$$M(w_{31}, w_{12}, w_{23}) = \begin{bmatrix} \Phi_{11} & w_{12} & \Phi_{13} \\ \Phi_{21} & \Phi_{22} & w_{23} \\ w_{31} & \Phi_{32} & \Phi_{33} \end{bmatrix} = \begin{bmatrix} \Phi_{11} \\ \Phi_{21} \\ w_{31} \end{bmatrix} \begin{bmatrix} 1 & \frac{w_{12}}{\Phi_{11}} & \frac{\Phi_{13}}{\Phi_{11}} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \Phi_{22} - \frac{\Phi_{21} w_{12}}{\Phi_{11}} & w_{23} - \frac{\Phi_{13}\Phi_{21}}{\Phi_{11}} \\ 0 & \Phi_{32} - \frac{w_{31} w_{12}}{\Phi_{11}} & \Phi_{33} - \frac{w_{31}\Phi_{13}}{\Phi_{11}} \end{bmatrix} \tag{2.3.20}$$

We see that there is no solution for the rank one completion and that the rank-two completions are implicitly given by

$$0 = \Phi_{33} - \frac{\Phi_{13}}{\Phi_{11}} w_{31} - \frac{\left(\Phi_{32} - \frac{w_{31} w_{12}}{\Phi_{11}}\right)\left(w_{23} - \frac{\Phi_{13}\Phi_{21}}{\Phi_{11}}\right)}{\Phi_{22} - \frac{\Phi_{21} w_{12}}{\Phi_{11}}}. \tag{2.3.21}$$

In particular, an easily identifiable rank-two completion is given by the choice

$$w_{12}^1 = \frac{\Phi_{32}\Phi_{13}}{\Phi_{33}},\ w_{23}^1 = \frac{\Phi_{13}\Phi_{21}}{\Phi_{11}},\ w_{31}^1 = \frac{\Phi_{33}\Phi_{11}}{\Phi_{13}}. \tag{2.3.22}$$

M denotes the zero energy matrix of with coordinates (2.3.22).

In order to run numerical simulations, we pick arbitrary matrix elements for the

target matrix:

$$\Phi = \begin{bmatrix} -1.55795 & \cdot & 1.58397 \\ 0.212869 & 0.0337805 & \cdot \\ \cdot & 1.32488 & 1.92653 \end{bmatrix}. \tag{2.3.23}$$

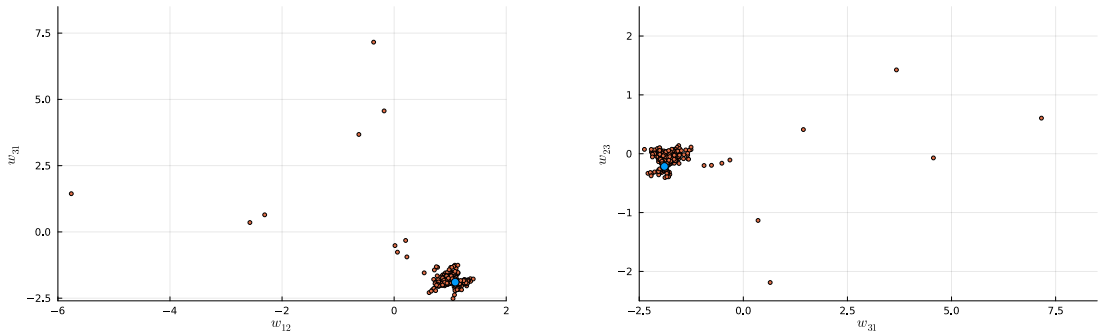


Figure 2.8: Clustering of training outcomes obtained in the setup of equation (2.3.20). The majority of the 500 outputs cluster near one particular rank-two minimizer indicated by the blue dot. Random initial conditions were drawn from Wigner$(0, 0.02)$.

Figure 2.8 shows the outcome of strong clustering in a region near M. We interpret this as a shortcoming of the rank based predictions, as the entire two parameter family of minimizers defined by (2.3.21) has rank-two. We attempt to explain the observation using phase space volume. We approximate phase space volume within the zero energy plane spanned by coordinates (w_{12}, w_{31}, w_{23}) upon simple Monte Carlo integration. We integrate the volume form (2.1.16) in a small cube around some rank-two minimizers.

Since these points are singular, volume form (2.1.16) blows up. As a result of this, the volume in a domain containing these points can be infinity. The volume, however, is locally finite, so working with domains bounded away from singular points allows us to make quantitative comparison between regions of the state space. Given a rank-two minimizer, we can study the concentration of volume around it by simple monte carlo integration of the volume density (2.2.22).

Define the set of three-by-three matrices with smallest singular value strictly larger than h,

$$GL_3^h = \{X \in GL(3) : \sigma_3(X) > h\}, \tag{2.3.24}$$

and the three dimensional H-cube around a point X

$$\mathcal{C}_H(X) = \{W \in \mathbb{R}^{d\times d} : \|x_{31} - w_{31}\|_1 < H/2, \|x_{12} - w_{12}\|_1 < H/2, \|x_{23} - w_{23}\|_1 < H/2\}. \tag{2.3.25}$$

We can generate uniform random points in $GL_3^h \cap \mathcal{C}_H(X)$ by drawing from a uniform distribution on $\mathcal{C}_H(X)$ and discarding the point when its smallest singular value is smaller than h.

To carry out these integrations, random rank-two minimizers with undetermined elements (w_{12}, w_{31}, w_{23}) solving (2.3.21) are also required. We obtain these matrices by sampling w_{12} and w_{31} from a centered normal distribution of standard deviation 10. Parameters $H = 0.001$ and $h = 0.00001$ were chosen.

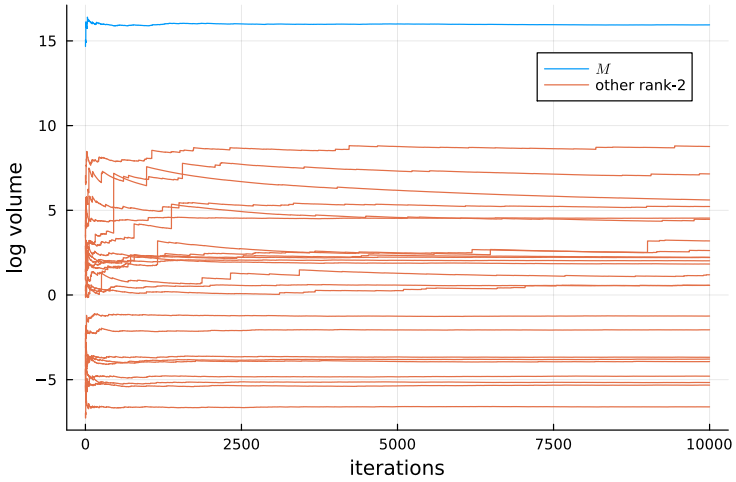


Figure 2.9: Monte Carlo integration of volume around rank-two minimizers.

The outcome of simulations conducted on M and 24 other randomly selected rank-two minimizers is shown in Figure 2.9. M is the point of the highest volume

vicinity, the difference to other equal rank minimizers spanning several orders of magnitude. This is in alignment with the clustering of training outcomes.

This example motivates the rigorous analysis of the asymptotics of the volume form around singular matrices. While we observe large quantitative differences of volume in the cluster of training outcomes, notice that M was a point arbitrarily selected in this cluster. Based on these results, we cannot conclude that there are no points with even larger volume around them, or even points with higher blowup rates. Our results nonetheless strongly suggest that state space volume is the right tool to make predictions on training outcomes, when rank based predictions are inconclusive.

2.3.4.3 Example: Minimal Rank State Space

Our last example is a pathological case showing an explicit shortcoming of the rank-hypothesis of implicit regularization. Let $\mathcal{R}^1 = \{X \in \mathbb{M}_2 : \operatorname{rank}(X) = 1\}$. It is shown in Chapter 1 in [3] that the DLN geometry for $N < \infty$ can be defined on the manifold of fixed rank matrices using equation (2.1.8) for the metric.

It is easy to show that the volume density of this metric is

$$\sqrt{G_1^N(W)} = \frac{N}{\sigma^{3\frac{N-1}{N}}}, \tag{2.3.26}$$

where σ denotes the single nonvanishing singular value of $W \in \mathcal{R}^1$. This volume form does not blow up at any point other than points near the origin.

For numerical simulations, take the setup of section 2.3.3.2, save for setting $N = 20$ and the change in the generation of random initial conditions. We generate

samples $u\sigma v^T = W_0' \sim \text{Wigner}(0, 0.001)$, then use

$$W_0 = u \begin{bmatrix} \sigma_1 & 0 \\ 0 & 0 \end{bmatrix} v^T \tag{2.3.27}$$

as initial conditions. The set of minimizers (same as minimum rank minimizers) consists of matrices

$$\begin{bmatrix} \Phi_1 & w_{12} \\ \frac{\Phi_2\Phi_1}{w_{12}} & \Phi_2 \end{bmatrix} = \begin{bmatrix} 0.58724 & w_{12} \\ \frac{0.8497}{w_{12}} & 1.447 \end{bmatrix} \tag{2.3.28}$$

Volume form (2.3.26) shows that highest volume is obtained at smallest σ, but notice that in this case the volume form does not blow up. Nonetheless, we can predict highest concentration of training outcomes at the minimal singular value completions. Simple computation verifies that these are the same as the symmetric completions,

$$\begin{bmatrix} 0.58724 & \pm 0.921811 \\ \pm 0.921811 & 1.447 \end{bmatrix}, \tag{2.3.29}$$

which both admit $\sigma_{\min} = 2.03424$. Thus, based on phase space volume, matrices with singular value at $\sigma_{\min}$ and slightly above are expected to have high representation among training outcomes. Figure 2.10 shows the outcome of 1072 simulations using random initial conditions verifying the predictions.

Notice furthermore that $r_e(W) = 1$ for all $W \in \mathcal{R}^1$, therefore it is impossible to make any predictions based on effective rank.

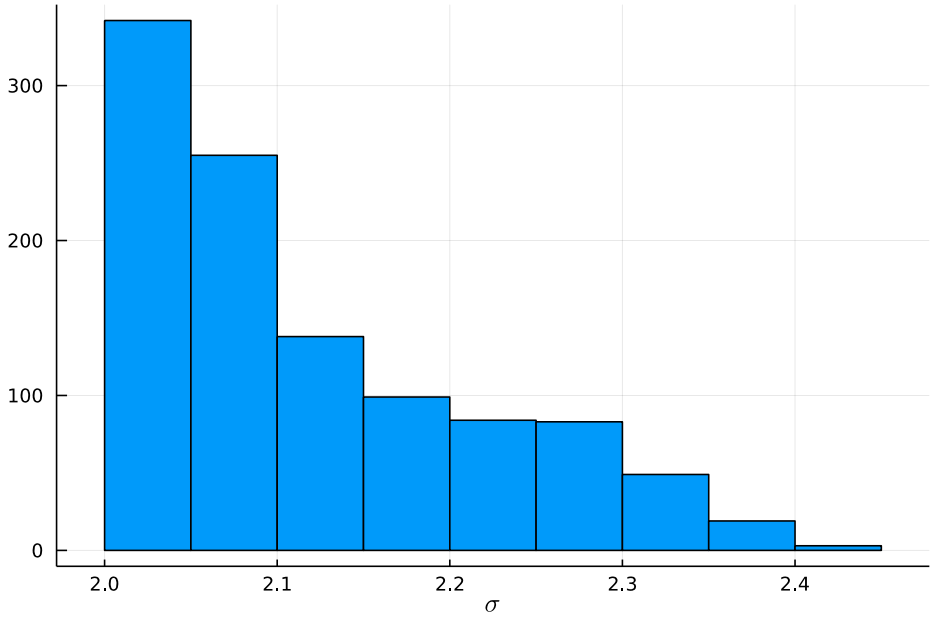


Figure 2.10: Sample distribution of 1072 simulation outcomes of rank one, diagonal matrix completion.

2.3.5 Results on dynamics under energy function $E_1(W)$

Let the energy function be

$$E_1(W) = \frac{1}{2}\|W - \Phi\|^2, \tag{2.3.30}$$

corresponding to a degenerate case of matrix completion in which all matrix elements are observed. Let $N = \infty$ and introduce $P = U^T \Phi V$ as a new dynamical variable. Then (2.3.6) takes the form

$$\begin{aligned} \dot{\Sigma} &= -\Sigma^2 \mathrm{diag}\left(\Sigma - P\right), \\ \dot{P} &= \mathfrak{s}\left(\left(L\left(\Sigma\right)\Sigma\right) \circ P\right) P - P\mathfrak{s}\left(\left(\Sigma L\left(\Sigma\right)\right) \circ P\right). \end{aligned} \tag{2.3.31}$$

The second equation in this form shows similarities to the double bracket flow in [6]. In particular, its application to dynamical singular value decomposition, see [14]. This isn't surprising. Upon successful convergence of training, $\lim_{t\to\infty} W(t) = \Phi$, thus the left and right singular vectors of W must diagonalize Φ as well. In this sense, we can think of equation (2.3.31) as a dynamical system computing the the singular value decomposition of P as long as $W \to \Phi$.

We notice that the original dynamical system (2.1.7) or the SVD dynamics (2.3.6) were d^2 dimensional, while the number of new coordinates sum up to $d + d^2$. The reason for this is the d new conserved quantities introduced by the singular values of $P(t) = U(t)\Phi V(t)^T$ being constant. Notice furthermore that the original energy function can be expressed with the new coordinates as

$$E(\Sigma, P) = \|\Sigma - P\|^2. \tag{2.3.32}$$

Obviously, due to the equivalence to (2.1.7), this is still a monotonically decreasing quantity. Another monotonic quantity is trace(P). Let $\hat{L}_{ij}$ denote the symmetric matrix quantity

$$\hat{L}_{ij} = \begin{cases} \sqrt{L_{ij}(\Sigma_i - \Sigma_j)} & \text{if } i \neq j, \\ 0 & \text{if } i = j. \end{cases} \tag{2.3.33}$$

Lemma 2.3.5. *Along the trajectories of* (2.3.31), $\frac{d}{dt}\text{trace}(P(t)) = \frac{1}{2}\|\hat{L} \circ (P - P^T)\|^2$.

Proof.

$$\begin{aligned}
\frac{d}{dt}\text{trace}(P(t)) &= \sum_{k,i} \Sigma_k P_{ik} P_{ki}(L_{ik} - L_{ki}) + L_{ik}\Sigma_i P_{ki} P_{ki} - L_{ki}\Sigma_i P_{ik} P_{ik} \\
&= \sum_{i,k} 2L_{ik} P_{ik} P_{ki} \Sigma_k + L_{ik}(\Sigma_i - \Sigma_k) P_{ki}^2 \\
&= \sum_i 2L_{ii} P_{ii}^2 \Sigma_i + 2\sum_{i<k} L_{ik} P_{ik} P_{ki} \Sigma_k + L_{ki} P_{ki} P_{ik} \Sigma_i \\
&\quad + \sum_{i<k} L_{ik}(\Sigma_i - \Sigma_k) P_{ki}^2 + L_{ki}(\Sigma_k - \Sigma_i) P_{ik}^2 \\
&= \sum_{i<k} -2L_{ik}(\Sigma_i - \Sigma_k) P_{ik} P_{ki} + L_{ik}(\Sigma_i - \Sigma_k)(P_{ik}^2 + P_{ki}^2) \\
&= \sum_{i<k} L_{ik}(\Sigma_i - \Sigma_k)(P_{ik} - P_{ki})^2 \\
&= \frac{1}{2}\|\hat{L} \circ (P - P^T)\|^2 \\
&> 0.
\end{aligned} \tag{2.3.34}$$

□

Since the norm of P is conserved and

$$\text{trace}(P) = \langle P, Id \rangle \leq \sqrt{d}\|P\|, \tag{2.3.35}$$

we find that trace(P) is monotonically increasing and bounded from above. As a consequence, we see that

$$\lim_{t\to\infty} \|\hat{L} \circ (P - P^T)\| = 0, \tag{2.3.36}$$

or explicitly

$$\frac{\Sigma_i - \Sigma_j}{\log(\Sigma_i^2) - \log(\Sigma_j^2)}(P_{ij} - P_{ji}) \to 0, \tag{2.3.37}$$

for $i \neq j$ as $t \to \infty$. However, more is true. Our next result shows that either P converges to a diagonal matrix, or σ_i converge to zero.

Theorem 2.3.6. *Along the trajectories of system (2.3.31), $\Sigma^2 - \Sigma \circ P \to 0$, as well as $\tilde{L}P \to 0$, as $t \to \infty$.*

Proof. Define the symmetric matrix

$$\tilde{L} = \left(\sqrt{L_{ik}(\Sigma_i^2 - \Sigma_k^2)}\right)_{ij}, \tag{2.3.38}$$

and arrive at

$$\begin{aligned} \frac{d}{dt}\|\Sigma - P\| &= -2\|\Sigma^2 - \operatorname{diag}(\Sigma P)\|^2 - 2\sum_{i<k}(\Sigma_i^2 - \Sigma_k^2)L_{ik}(P_{ki}^2 + P_{ik}^2) \\ &= -2\|\Sigma^2 - \operatorname{diag}(\Sigma P)\|^2 - 2\|\tilde{L} \circ P\|^2 \end{aligned} \tag{2.3.39}$$

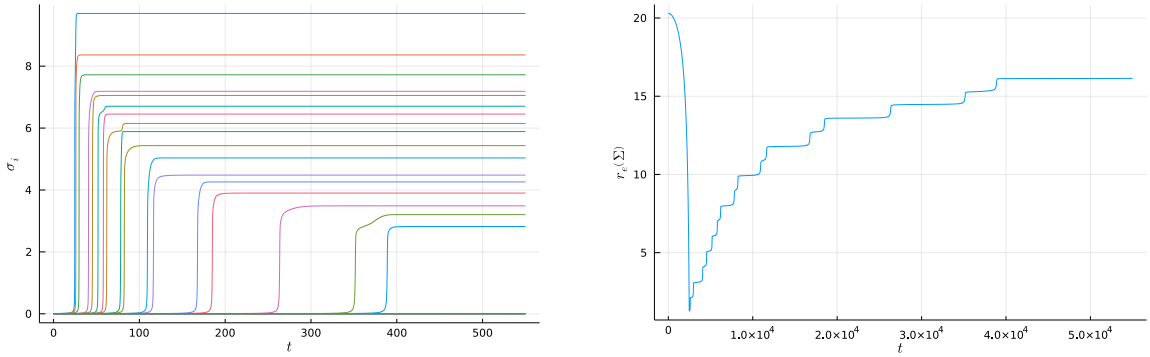
Since $\|\Sigma - P\|$ is monotonically decreasing and bounded from below, it is convergent, completing the proof. □

2.3.5.1 Results on singular value non-crossing

The direct SVD dynamics (2.3.6) and the new formulation (2.3.31) depend on the same assumption. In order to use smooth singular value decomposition (Lemma 2.3.1) we assume that the singular values cannot cross along the trajectories of the gradient flow (2.1.7).

So far we have not addressed this assumption. During the extensive numerical experiments, crossing of singular values was never observed. In fact, for matrix completion problems under small initial conditions, singular values tend to "fan out" after being clustered close to zero upon initialization. However, interesting near-collisions can happen under the present choice of energy function (2.3.30).

Examples of these near-collisions can be seen in Figure 2.11. This picture motivated us trying to isolate the dynamics of singular values from the rest of the system, in an attempt to identify the observed strong repulsion between singular values. For $d = 2$, this idea is made rigorous by the following theorem, characterizing the time evolution of Σ as a closed second order dynamical system.



(a) (b)

Figure 2.11: Generic trajectories of the singular values and effective rank throughout training under small sufficiently large depth ($N = \infty$ shown). Notice the near collisions $t \approx 60$ and $t \approx 80$.

Notice that $\det P$ and $\|P\|$ only depend on the singular values of P, thus $C_1 := \|P\|^2 + 2\det P$ and $C_2 := \|P\|^2 - 2\det P$ are conserved quantities of (2.3.31). Let f and g denote functions

$$f(\sigma_i, \dot{\sigma}_i) := \frac{\sigma_1 - \sigma_2}{2\log \sigma_1/\sigma_2}\left(C_1 - \left(\sigma_1 - \frac{\dot{\sigma}_1}{\sigma_1^2} + \sigma_2 - \frac{\dot{\sigma}_2}{\sigma_2^2}\right)^2\right), \tag{2.3.40}$$

$$g(\sigma_i, \dot{\sigma}_i) := \frac{\sigma_1 + \sigma_2}{2\log \sigma_1/\sigma_2}\left(C_2 - \left(\sigma_1 - \frac{\dot{\sigma}_1}{\sigma_1^2} - \sigma_2 + \frac{\dot{\sigma}_2}{\sigma_2^2}\right)^2\right). \tag{2.3.41}$$

$$\tag{2.3.42}$$

Theorem 2.3.7. *Let $d = 2$. Under the flow of (2.3.31), σ_1 and σ_2 solve the system*

$$\ddot{\sigma}_1 = \frac{\sigma_1^2}{2}\left(f(\sigma_i, \dot{\sigma}_i) + g(\sigma_i, \dot{\sigma}_i)\right) + \frac{2\dot{\sigma}_1^2}{\sigma_1} - \dot{\sigma}_1\sigma_1^2, \tag{2.3.43}$$

$$\ddot{\sigma}_2 = \frac{\sigma_2^2}{2}\left(f(\sigma_i, \dot{\sigma}_i) - g(\sigma_i, \dot{\sigma}_i)\right) + \frac{2\dot{\sigma}_2^2}{\sigma_2} - \dot{\sigma}_2\sigma_1^2. \tag{2.3.44}$$

Proof. Let us introduce notation

$$l := \left(\log\left(\sigma_1^2\right) - \log\left(\sigma_2^2\right)\right)^{-1}, \tag{2.3.45}$$

$$\tau_1 := \operatorname{trace}(P) = P_{11} + P_{22}, \tag{2.3.46}$$

$$\tau_2 := P_{11} - P_{22}. \tag{2.3.47}$$

Simple computation verifies that

$$\dot{\tau_1} = l\left(\sigma_1 - \sigma_2\right)\left(C_1 - \tau_1^2\right) \text{ and} \tag{2.3.48}$$

$$\dot{\tau_2} = l\left(\sigma_1 + \sigma_2\right)\left(C_2 - \tau_2^2\right). \tag{2.3.49}$$

Notice that the RHS is only dependent on σ_i, while from the second equation of (2.3.31), one can express

$$\tau_1 = \sigma_1 + \frac{\dot{\sigma}_1}{\sigma_1^2} + \sigma_2 + \frac{\dot{\sigma}_2}{\sigma_2^2}, \tag{2.3.50}$$

$$\tau_2 = \sigma_1 + \frac{\dot{\sigma}_1}{\sigma_1^2} - \sigma_2 - \frac{\dot{\sigma}_2}{\sigma_2^2}. \tag{2.3.51}$$

Differentiating in time and solving the equations for $\ddot{\sigma}_i$ yields the result.

□

Corollary 2.3.8 (Singular values cannot cross)**.** *Let $d = 2$ and let $\sigma_i(t)$ be trajectories of 2.3.31. Then $\sigma_1(t) \neq \sigma_2(t)$, $\forall t$.*

Proof. Notice that

$$\lim_{\sigma_2 \to \sigma_1} g(\sigma_i, \dot{\sigma}_i) = \infty. \tag{2.3.52}$$

Thus

$$\lim_{\sigma_2 \to \sigma_1} \ddot{\sigma}_1 = \infty \tag{2.3.53}$$

$$\lim_{\sigma_2 \to \sigma_1} \ddot{\sigma}_2 = -\infty. \tag{2.3.54}$$

Moreover

$$\int_{\sigma_1 - \epsilon}^{\sigma_1} \frac{\sigma_1 + \sigma_2}{2\log\sigma_1 - 2\log\sigma_2} d\sigma_2 \geq (\sigma_1 - \epsilon) \int_{\sigma_1 - \epsilon}^{\sigma_1} \frac{1}{\sigma_1 - \sigma_2} d\sigma_2 = \infty. \tag{2.3.55}$$

□

Unfortunately, the above way of proving singular value non-crossing does not immediately generalize to $d > 2$. The algebra of isolating Σ from P using conserved quantities is increasingly more complicated under $d > 2$. Nonetheless, one can still identify the leading order behavior

$$\ddot{\sigma}_i + \sigma_i^2 \dot{\sigma}_i - 2\frac{\dot{\sigma}_i^2}{\sigma_i} = \sigma_i^2 \frac{\sigma_{i-1}}{\log(\sigma_i^2) - \log(\sigma_{i-1}^2)}) P_{i,i-1} P_{i-1,i} + \mathcal{O}(1), \tag{2.3.56}$$

as $\sigma_{i-1} \to \sigma_i$. This formula is promising, showing the same non-integrable blowup in the repulsion force as Theorem 2.3.7. Unfortunately, we have no reason to assume, and we were unable to prove that the coefficient $P_{i,i-1}P_{i-1,i}$ cannot become zero in finite time.

2.4 Conclusion

We have used Riemannian geometry to improve the rank-based understanding of implicit regularization in DLN. We derived new representations for the DLN metric and formalized the infinite depth limit. The most salient feature of our analysis is the derivation for the volume forms under depths of all positive integers, as well as infinity.

Using our results on the geometry of the DLN, we also improved upon the understanding of training dynamics. We found formulas for linear attraction rates using the eigendecomposition of the DLN metric. We then proved local normal hyperbolicity for the critical manifold of training dynamics under a relatively general family of loss functions.

Looking forward, we would like to continue our study on the DLN geometry. The simplicity of the formulas presented in this paper suggest that further intrinsic quantities could be expressed using explicit formulas. The derivation of higher order geometric quantities could give rise to new results on the dynamics. The precise formulation of stochastic training dynamics requires the notion of Brownian motion on a Riemannian manifold. The stochastic differential equation describing intrinsic Brownian motion relies on computations of the Levi-Civita connection and curvatures [17, Chapter 3].

We engineered low dimensional examples to verify our idea, that implicit regularization is explained by high state space volume. We showed that training in the case of $N = \infty$ shows implicit regularization the same way as has been established for $N < \infty$. To summarize, the DLN at $N = \infty$ provides a novel model of implicit regularization which is simpler than the $N < \infty$ counterpart.

Chapter Three

Riemannian Langevin Equation for DLN

3.1 Introduction

The analytical results and numerical examples of Chapter 2 demonstrate the predictions that can be made based on understanding the volume form and volume in the DLN. In this chapter we further develop the intuition on the importance of volume and use computational tools to explore the zero energy surface. Throughout this chapter we restrict our attention to $N = \infty$.

While for the purposes of numerical simulations, it is practical to approximate the coordinates of the connection using finite differences, in Section 3.3.1 we detail another way of obtaining second order geometric information.

In Section 3.3.2 we review basic stochastic differential geometry and the derivation of equation (3.3.2). Sections 3.3.1 and 3.3.2 use the summation convention over repeated indices, unless it is explicitly suspended. Numerical simulations results on

the RLE are shown in Section 3.3.3.

3.2 Volume density and blowup rates

The results of Section 2.3.4.2 show that the volume around rank deficient minimizer is larger when there is observed clustering of training outcomes. The natural question arises whether this difference is merely quantitative or the blowup rate of the volume density qualitatively differs. In this Section we revisit the examples of Section 2.3.3.2 and 2.3.4.2 to investigate this question.

Exercise 3.2.1 (Explicit computation of volume perturbation)**.** *Let us start by revisiting the example given in Section 2.3.3.2, the case of* 2×2 *diagonal matrix completion. We saw that in this example, the accumulation of training outcomes is strong near the corners of the hyperbola of rank one completions.*

We compute the perturbation of singular values and the volume density in the normal direction of the hyperbola. Fixing the diagonal elements to 1, we get that the one-parameter family of rank-one completions is given by

$$W = \begin{bmatrix} 1 & \gamma \\ \frac{1}{\gamma} & 1 \end{bmatrix}. \tag{3.2.1}$$

To simplify some of the formulas, we assume $\gamma > 0$*, restricting the analysis on the positive lobe of the hyperbola. The singular value decomposition* $W(\gamma) =$

$U(\gamma)\Sigma(\gamma)V^T(\gamma)$ can be explicitly computed:

$$U = \frac{1}{\sqrt{1+\gamma^2}}\begin{bmatrix}\gamma & -1\\ 1 & \gamma\end{bmatrix},\ \Sigma = \begin{bmatrix}\gamma+\gamma^{-1} & 0\\ 0 & 0\end{bmatrix},\ V = \frac{1}{\sqrt{1+\gamma^2}}\begin{bmatrix}1 & \gamma\\ \gamma & -1\end{bmatrix}. \tag{3.2.2}$$

We will consider perturbations of size η in the normal direction of the hyperbola $1/\gamma$,

$$\dot{W} = \frac{\eta}{\sqrt{1+\gamma^{-4}}}\begin{bmatrix}0 & \gamma^{-2}\\ 1 & 0\end{bmatrix}. \tag{3.2.3}$$

Using Lemma 2.3.1, the perturbed singular values can be computed,

$$\Sigma(\gamma,\eta) = \begin{bmatrix}\gamma+\gamma^{-1}+\frac{2\eta}{(1+\gamma^2)\sqrt{1+\gamma^{-4}}} & 0\\ 0 & \eta\frac{\sqrt{\gamma^4+1}}{\gamma^2+1}\end{bmatrix} + \mathcal{O}\left(\eta^2\right). \tag{3.2.4}$$

And consequently, up to leading order, the determinant is

$$\det\left(\Sigma(\gamma,\eta)\right) = \eta\frac{\sqrt{\gamma^4+1}}{\gamma} + \mathcal{O}(\eta^2). \tag{3.2.5}$$

Plugging this back into the volume density function from equation (2.2.22),

$$\begin{aligned}\frac{\operatorname{van}\left(\log\Sigma^2\right)}{\det(\Sigma)\operatorname{van}\left(\Sigma^2\right)} &= \frac{2\gamma\left(\log\left(\gamma+\gamma^{-1}+\frac{2\eta}{(\gamma^2+1)\sqrt{1+\gamma^{-4}}}\right)-\log\left(\eta\frac{\sqrt{\gamma^4+1}}{\gamma^2+1}\right)\right)}{\left(\eta\sqrt{\gamma^4+1}+\mathcal{O}\left(\eta^2\right)\right)\left(\left(\gamma+\gamma^{-1}+\frac{2\eta}{(1+\gamma^2)\sqrt{1+\gamma^{-4}}}\right)^2-\left(\eta\frac{\sqrt{\gamma^4+1}}{\gamma^2+1}\right)^2\right)}\\ &= \mathcal{O}\left(\frac{|\log\eta|}{\eta}\right)\end{aligned} \tag{3.2.6}$$

as $\eta \to 0$. This shows that the volume is infinite in regions containing points on the hyperbola even within the zero energy surface. As a secondary result, we also gain quantitative understanding of how the singular values perturb. Specifically, we learn that the leading order coefficient of the smaller singular value, σ_2, is $\frac{\sqrt{\gamma^4+1}}{\gamma^2+1}$. This function has a minimum at $\gamma = 1$, corresponding to the quickest blow up of the volume density.

The above exercise gives a complete description of our simplest model, but it also shows the limitation of this approach. Any higher dimensional case would likely be impossible to compute due the number of variables and lack of explicit singular value decomposition. However, we conjecture that for a perturbation of size η, normal to the rank-r manifold, the volume density blows up at rate $\mathcal{O}\left(\frac{|\log \eta|^{d-1}}{\eta^{d-r}}\right)$.

To test this conjecture numerically, we revisit the three-by-three example of Section 2.3.4.2. In this setting, $d = 3$, and the lowest rank minimizers have rank 2. Let $\mathcal{R}^2$ denote the eight dimensional submanifold of rank-two matrices in $\mathbb{R}^{3\times 3}$. In this codimension-one setup, given a point $X \in \mathcal{R}^2$ with its singular value decomposition $X = U_X \Sigma_X V_X^T$, the unique unit normal is given by

$$N_X = U_X \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} V_X^T. \tag{3.2.7}$$

Let $v_X(\eta)$ denote the volume density is restricted to the normal direction

$$v_X(\eta) := \sqrt{\det g^{\infty}(X + \eta N_X)}. \tag{3.2.8}$$

Informed by (3.2.6), we expect the affine relationship

$$\log \frac{v_X(\eta)}{|\log(\eta)|^2} = \log(c) - a\log(\eta) \tag{3.2.9}$$

to hold with good accuracy for small η. We sample points $X_i \in E^{-1}(0) \cap \mathcal{R}^2$ according to the procedure laid out in Section 2.3.4.2.

The results of these experiments are summarized in Figures 3.1-3.2. We see that the volume density seems to have the same blowup rate inside the cluster (point M) and for random points outside. The conjectured blowup rate $\mathcal{O}\left(\frac{|\log\eta|^2}{\eta}\right)$ holds with good accuracy. The right pane of Figure 3.2 shows that the multiplicative constant c is several orders of magnitude larger for point M than the sampled points.

We found no evidence of qualitative difference between blowup rates within the cluster and outside. However, the arbitrary choice of M prevents us from drawing a definitive conclusion.

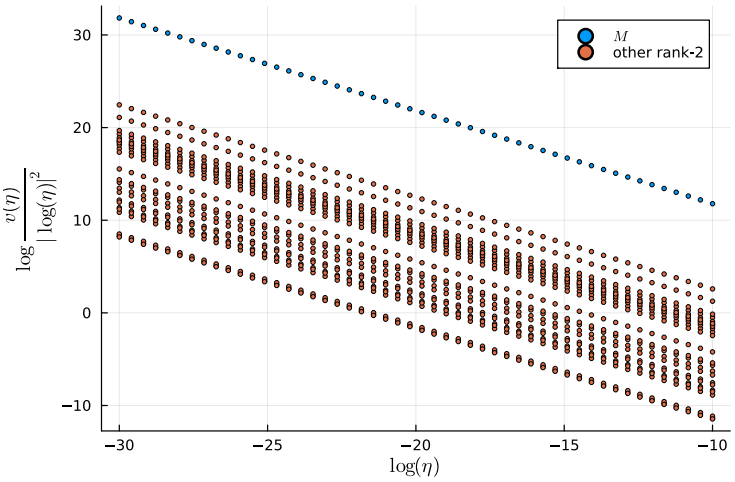


Figure 3.1: Scaling of the volume form near random rank-two minimizers, and the point M.

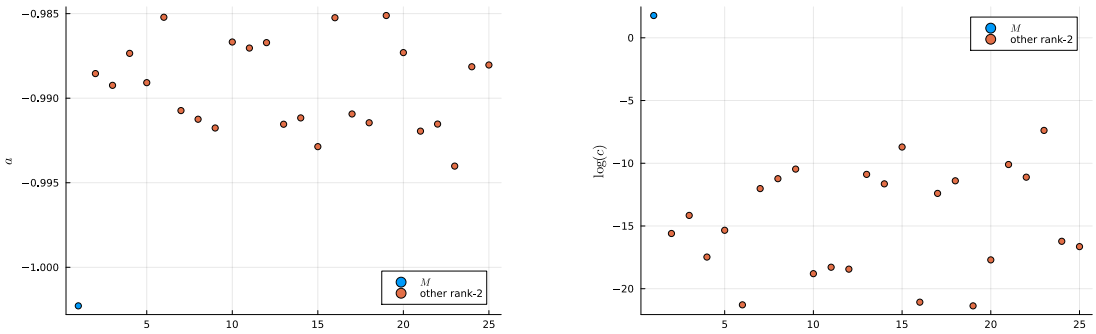


Figure 3.2: Slope and additive constant parameters obtained by fitting straight lines on data shown in Figure 3.1.

3.3 Riemannian Langevin Equation

The numerical observations of Section 2.3 show accumulation of training outcomes in high volume regions under random initialization. In an attempt to control for the effect initialization, we use deterministic initialization paired with stochastic dynamics. A natural way to add noise to the Riemannian gradient flow (2.1.7) is to write down the Ito equation of combined Riemannian gradient dynamics and intrinsic Brownian motion, referred to as the Riemannian Langevin Equation (RLE):

$$dW_t = -\text{grad}_{g^\infty} E(W)dt + \frac{1}{\sqrt{\beta}} dX_t. \tag{3.3.1}$$

The scaling parameter $\beta > 0$ determines the intensity of the noise, and is referred to as the inverse temperature. The Ito stochastic differential equation for intrinsic Brownian motion is given by

$$dX_t^i = \sigma_j^i\left(X_t\right) dB_t^j - \frac{1}{2} g^{lk}(X_t)\Gamma_{kl}^i(X_t)dt, \tag{3.3.2}$$

where σ is the symmetric square root of the dual metric, Γ_{kl}^i denote the Christoffel symbols and B_t denotes standard Brownian motion on $\mathbb{R}^{d^2}$.

The Riemannian gradient and the diffusion term depend on the dual metric we already gave explicit representation for in Chapter 2, using the diagonalization of the metric (2.1.14). The only remaining piece is the drift term, which requires second order geometric information.

3.3.1 Moving frames and the basis one-forms of DLN

Given an m-dimensional differentiable manifold $\mathcal{M}$, a smooth choice of basis vectors on a neighborhood is called a frame field, or (moving) frame.

Definition 3.3.1 (moving frame)**.** A (moving) frame at $x \in \mathcal{M}$ is a local linear isomorphism $u : \mathbb{R}^d \to T_x\mathcal{M}$.

We assume that $\mathcal{M}$ is Riemannian, i.e. equipped with metric g and Levi-Civita connection ∇. The canonical projection from the tangent bundle to the manifold is denoted $p : T\mathcal{M} \to \mathcal{M}$, while e_i denote the usual basis of $\mathbb{R}^m$.

Given $A_i = ue_i$, a moving frame, the usual way of representing the Levi-Civita connection is using the Christoffel symbols

$$\nabla_{A_i} A_j = \Gamma_{ij}^k A_k. \tag{3.3.3}$$

Then, for two vector fields $X = X^i A_i$, $Y = Y^i A_i$ their covariant derivative (Levi-Civita connection) writes:

$$\nabla_X Y = \left(dY\left(X^k\right) + \sum_{i,j} \Gamma_{ij}^k X^i Y^j \right) A_k. \tag{3.3.4}$$

Note that the above formula holds for an arbitrary frame field A_i, but most commonly

local coordinates are used, i.e. $A_i = \frac{\partial}{\partial \phi^i}$ for a chart (ϕ, U).

Computation of the Christoffel symbols in matrix geometries can be very long and is often not feasible, due to their large dimension and involved formulas. [9] developed another way of representing the connection using differential one-forms. Cartan's moving frame machinery sometimes leads to simpler computations, specifically when metric has some symmetries. One example is when a coordinate system in which the matrix of the metric takes a diagonal form is known.

Definition 3.3.2 (one-form)**.** A one-form on a differentiable manifold $\mathcal{M}$ is a smooth covector field $\theta_p : T_p\mathcal{M} \to \mathbb{R}$ for $p \in \mathcal{M}$.

One forms can be expressed in coordinates given a basis of the cotangent space, usually using differentials of the coordinates:

$$\theta = f_i(x)dx^i. \tag{3.3.5}$$

Example: $df(v) = \langle \text{grad} f, v \rangle_g$, the differential of a function f on a Riemannian manifold. Curvature calculations also require 2-forms, which are tensors taking two tangent vectors as arguments to return a scalar and are skew-symmetric in their arguments.

Let a collection of one-forms $(\alpha^1, \ldots, \alpha^m)$ be dual to $(A_1, \ldots, A_m)$, i.e. $\alpha^j(A_i) = \delta^i_j$ for all i, j. Such a collection of 1-forms is said to form a coframe field dual to the frame field $(A_1, \ldots, A_m)$. Then we have the one forms $\omega^k_j = \sum_i \Gamma^k_{ij}\alpha^i$ that also characterize the connection:

$$\nabla_X Y = \left(dY\left(X^k\right) + \omega^k_j(X)Y^j\right) A_k. \tag{3.3.6}$$

Remark. Easy calculation shows the skew-symmetry of these one forms $\omega_i^j = -\omega_j^i$.

We will use two different operations to generate two-forms from one-forms, the exterior product and the exterior derivative. The definitions sufficing for our setting can be simply stated.

Definition 3.3.3 (Exterior product)**.** The exterior product of one-forms α, β is the two-form

$$(\alpha \wedge \beta)_p (v_1, v_2) = \alpha_p(v_1)\beta_p(v_2) - \alpha_p(v_2)\beta_p(v_1) \tag{3.3.7}$$

for $v_1, v_2 = T_p\mathcal{M}$.

Definition 3.3.4 (Exterior derivative)**.** The exterior derivative of a one-form $\alpha = f_i dx^i$ is the two-form defined by

$$d\alpha = \frac{\partial f_j}{\partial x^i} dx^i \wedge dx^j. \tag{3.3.8}$$

For further calculations we need an orthonormal coframe field in which the Riemannian metric writes $g = \theta^i \otimes \theta^i$. We call α^i basis one-forms.

Remark. Obtaining the values of θ^i is often impossible, since Riemannian metrics usually lack a "sum of squares" structure that enables these computations.

Two classical theorems of Cartan are stated below without proof. The first one, the Cartan structure equations give explicit ways of computing the connection and curvature forms in terms of basis one-forms, the second gives the representation of curvature in terms of the curvature forms.

Theorem 3.3.5 (Cartan's Structure equations)**.** *Two statements:*

1. *The basis one-forms determine the connection one-forms:*

$$d\theta^i = -\omega^i_j \wedge \theta^j. \tag{3.3.9}$$

2. *The connection one forms ω^i_j determine the curvature two-forms:*

$$\Omega^i_j = d\omega^i_j + \omega^i_k \wedge \omega^k_j \tag{3.3.10}$$

Theorem 3.3.6. *The Riemannian curvature tensor can be expressed using the connection forms*

$$R^l_{ijk} = \Omega^l_i(A_j, A_k) \tag{3.3.11}$$

In particular, the Ricci curvature writes

$$R_{ik} = \Omega^l_i(A_l, A_k). \tag{3.3.12}$$

We were unable to find the following obvious result expressing the connection form in terms of a general basis coframe $\theta^i = F^i_j dw^j$ and its inverse $H := F^{-1}$.

Lemma 3.3.7 (Connection forms, general formula)**.**

$$\omega^i_j = \left(F^i_a H^k_j - g_{jc} g^{im} F^c_a H^k_m\right) \frac{\partial H^a_l}{\partial w_k} \theta^l \tag{3.3.13}$$

Proof. Taking an exterior derivative of θ^i and expressing the coordinate one-forms with basis one-forms gives:

$$d\theta^i = \frac{\partial F^i_j}{\partial w_k} (H\theta)^k \wedge (H\theta)^j. \tag{3.3.14}$$

Using the derivative of the inverse metric $\frac{\partial F}{\partial x} = -F\frac{\partial H}{\partial x}F$, we get

$$d\theta^i = -F^i_a \frac{\partial H^a_b}{\partial w_k} F^b_j H^k_l H^j_m \theta^l \wedge \theta^m = -F^i_a \frac{\partial H^a_b}{\partial w_k} \delta^b_m H^k_l \theta^l \wedge \theta^m = -F^i_a \frac{\partial H^a_j}{\partial w_k} H^k_l \theta^l \wedge \theta^j. \tag{3.3.15}$$

To compute the connection forms, one can use the first structure equation (Theorem 3.3.5) and antisymmetrize in indices i and j via index gymnastics. □

The DLN metric is significantly more complicated than most examples in which these calculations can be carried out.[1] In the following statement we show that basis one-forms are available in for the DLN geometry under arbitrary depth.

Let dw^j denote the coordinate one-forms of the column-major vectorization of the standard matrix coordinates.

Theorem 3.3.8 (Basis one-forms of DLN)**.** *The basis one forms of g^N are given by*

$$\theta^i = F_{ij} dw^j, \tag{3.3.16}$$

where

$$F = (V \otimes U)\sqrt{D^N(\Sigma)}. \tag{3.3.17}$$

Proof. Notice that for basis one-forms, $\delta^{ij} = g^*(\theta^i, \theta^j)$. Thus θ^i can be identified by inspection of equation (2.2.17). □

Combining Lemma 3.3.7 and Theorem 3.3.8 gives the connection forms for the DLN geometry, under arbitrary N. Unfortunately, the resulting formulas cannot be

[1] We would like to point the interested reader's attention to the valuable examples provided in Fred Akalin's blog post.

expressed in a compact way fitting a page. The utility of these results lies in the potential usage of computational investigations involving the connection.

Examples include numerical integration of geodesic, or parallel transport equations and the RLE (3.3.1). Notice that the connection forms (3.3.13) only require matrix multiplications and partial derivatives that can be efficiently computed using a automatic differentiation.

We remark that this methodology may be effective to study other matrix geometries. The Bures-Wasserstein metric and the trace metric of positive semidefinite matrices admit similar diagonalizing frames (see Section 3 in [22]). Based on these diagonalizations, the equivalents of Theorem 3.3.8 for these geometries can be proven. The frames of these metrics are given by simpler formulas, projecting meaningful potential contributions to the understanding of these geometries.

3.3.2 Brownian motion on a Riemannian manifold

In this section we review some tools of differential geometry. These tools enable the definition of (stochastic) differential equations on manifolds, given their Euclidean counterparts. This machinery is known as "(stochastic) rolling without slipping".

This section summarizes chapters 1-3 of [16], with the goal of arriving at the Ito differential equation describing intrinsic Brownian motion on a manifold (3.3.29).

We assume an m-dimensional differentiable manifold $\mathcal{M}$ is given with connection ∇. The canonical projection from the tangent bundle to the manifold is denoted $p : T\mathcal{M} \to \mathcal{M}$.

This general setup is assumed throughout the section, until we write out Brownian motion in coordinates. Then we restrict our attention to Riemannian manifolds, specifying the connection to be the Levi-Civita connection.

Definition 3.3.9 (moving frame, frame bundle)**.** The frame bundle is the disjoint union

$$\mathcal{F}(\mathcal{M}) = \bigsqcup_x \mathcal{F}(\mathcal{M})_x, \tag{3.3.18}$$

where $\mathcal{F}(\mathcal{M})_x$ denotes the space of all frames at point x.

Recall that a (moving) frame is a linear map that maps $e_1, \ldots, e_m$, a basis of $\mathbb{R}^m$ to $ue_1, \ldots, ue_m$, a basis of the tangent space in a neighborhood of x. Thus, the frame bundle consists of all the possible smooth choices of frames along the manifold $\mathcal{M}$, constituting a $m + m^2$ dimensional manifold. A curve u_t in $\mathcal{F}(\mathcal{M})$ is a smooth choice of frames along a curve πu_t.

Remark. In the Riemannian case, we further specify the frame bundle to be the orthonormal frame bundle $\mathcal{O}(\mathcal{M})$. This includes all smooth choices of orthonormal bases. The orthogonality of the frame is preserved using the natural, Levi-Civita connection on the manifold.

We review the correspondence between vector fields of the manifold, and vector fields of its frame bundle, which is done by defining horizontal directions on $\mathcal{F}(\mathcal{M})$.

Definition 3.3.10. A curve $u_t \in \mathcal{F}(\mathcal{M})$ is called horizontal if for any $e \in \mathbb{R}^m$, the vector field $u_t e$ is parallel along $p\, u_t$.

Clearly, not every tangent vector $X \in T_{(x,u)}\mathcal{F}(\mathcal{M})$ can be the tangent vector of a horizontal curve emanating from u. Such vectors are called horizontal, the space

of which is denoted $H_u\mathcal{F}((M))$, giving the decomposition of the tangent space

$$T_u\mathcal{F}(\mathcal{M}) = V_u\mathcal{F}(\mathcal{M}) \oplus H_u\mathcal{F}(\mathcal{M}). \tag{3.3.19}$$

Based on this, an isomorphism can be defined by the pushforward of the canonical projection $p_* : H_u\mathcal{F}(\mathcal{M}) \mapsto T_{\pi u}\mathcal{F}(\mathcal{M})$. Using this isomorphism, the horizontal vector X^* is unique given the choice $X \in T_x\mathcal{M}$, and is called the horizontal lift of X.

The horizontal lift can be extended pointwise on a vector field $X \in \Gamma(\mathcal{M})$ along a curve, giving a unique horizontal vector field X^* on $\mathcal{F}(\mathcal{M})$. Specifically, using the horizontal lift of vector field, the unique horizontal lift $u_t \in \mathcal{F}(\mathcal{M})$ of a curve $x_t \in \mathcal{M}$ given u_0 can be defined.

Remark (Local chart on $\mathcal{F}(\mathcal{M})$). Let x^i be a local chart, then locally, for a frame u, we have $ue_i = e_i^j \partial_{x^j}$ for some matrix $e_i^j \in GL(d)$. Then (x^i, e_j^i) is a local chart on the frame bundle.

The usual notion of a parallel transport is directly defined by the connection. Given an initial point $(x, u) \in \mathcal{F}(\mathcal{M})$, and a direction on the manifold $X \in T\mathcal{M}$, the last m^2 coordinates of the tangent vector $X^* \in H_u\mathcal{F}(\mathcal{M})$ are determined by the parallel transport equation.

Definition 3.3.11 (Parallel transport)**.** Given a curve x_t and a frame u_0 at x_0, denote u_t the unique horizontal lift of x_t from u_0. The linear map

$$\tau_{t_0 t_1} = u_{t_1} u_{t_0}^{-1} \tag{3.3.20}$$

is independent of the initial frame u_0 and is called the parallel transport along x_t.

Remark. The above reasoning suggests that the development of these ideas can be turned around, and parallel transport itself can be defined using the notion of the horizontal lift. This turns out to be the case: a unique connection can be defined after a smooth assignment of m dimensional subspaces that we want to keep horizontal.

The next step in developing stochastic processes (semimartingales) given their Euclidean equivalents are the fundamental horizontal fields $\{H_i\}_1^m$. These are defined to be the horizontal lifts of the coordinate unit vectors $\{e_i\}_1^m$ of $\mathbb{R}^m$. These can be computed explicitly given that the connection is represented in a local chart by the following proposition:

Proposition 3.3.12. *In terms of a local chart, the fundamental horizontal fields can be expressed*

$$H_i(u) = e_i^j \partial_{x^i} - e_i^j e_n^l \Gamma_{jl}^k(x) \partial_{e_n^k}. \tag{3.3.21}$$

The horizontal lift establishes the way of passing data (vectors, vector fields, curves) from the manifold to the frame bundle, while the frame itself was a linear isomorphism between the tangent space and $\mathbb{R}^m$. Putting these pieces together we find a way of mapping smooth curves of $\mathbb{R}^m$ to the manifold such that horizontality is preserved.

Let u_t and x_t be curves in the frame bundle and on the manifold respecitvely, linked by horizontal lift. We can define the curve in $\mathbb{R}^m$ called the anti-development (of x_t) by

$$w_t = \int_0^t u_s^{-1} \dot{x}_s ds. \tag{3.3.22}$$

Differentiating this expression and using horizontality we get

$$H_{\dot{w}_t}(u_t) = \dot{u}_t, \tag{3.3.23}$$

which can be expressed in terms of the fundamental horizontal fields as

$$\dot{u}_t = H_i(u_t)\dot{w}_t^i, \tag{3.3.24}$$

defining a differential equation to map a smooth curve w_t in $\mathbb{R}^m$ to the frame bundle. This is the construction known as "rolling without slipping", and can be further generalized in a principled way to define a way of mapping stochastic evolutions in $\mathbb{R}^m$ to $\mathcal{M}$.

The key idea here is to replace Equation (3.3.24) with its stochastic version

$$dU_t = H(U_t) \circ dY_t, \tag{3.3.25}$$

where Y is a semimartingale on $\mathbb{R}^m$. This machinery enables the definition of the horizontal Brownian motion, that is, the horizontal lift of Brownian motion of a Riemannian manifold. Given B, a m-dimensional Euclidean Brownian motion, the solution of

$$dU_t = H_i(U_t) \circ dB_t^i \tag{3.3.26}$$

is called horizontal Brownian motion on the orthonormal frame bundle $\mathcal{O}(\mathcal{M})$. Combining this with Equation (3.3.21), we get the explicit Stratonovich equation for the evolution on the frame bundle:

$$dX_t^i = e_j^i(t) \circ dB_t^j \tag{3.3.27}$$

$$de_j^i(t) = -\Gamma_{kl}^i(X_t)e_j^l(t)e_m^k(t) \circ dB_t^m. \tag{3.3.28}$$

By comptuation of the drift term, the Ito SDE of intrinsic Brownian motion on a Riemannian manifold can be written down.

$$dX_t^i = \sigma_j^i\left(X_t\right) dB_t^j - \frac{1}{2} g^{lk}(X_t)\Gamma_{kl}^i(X_t)dt. \tag{3.3.29}$$

The diffusion matrix is the positive definite square root of the matrix defining the dual metric, $\sigma_j^i = \sqrt{g^*}^i{}_j$.

3.3.3 Numerical simulations of the RLE on DLN

In this section we present simulation results obtained by numerical integration of the RLE (3.3.1). Setting up the correct simulation framework itself poses significant challenges. In the deterministic case, instead of evolving the matrix coordinates W, we pulled back the dynamics to the evolution of the triple (U, Σ, V). This approach was detailed in Section 2.3.1.

The replication of this idea for the stochastic evolution is very cumbersome. A coordinate change of stochastic differential equations like the RLE (3.3.1) requires a second order perturbation of the coordinate transformations. On one hand, the second order equivalent of Lemma 2.3.1 is certainly available. On the other hand, the RLE (3.3.1) was already not explicitly written down due to no simple closed form formula being available for the Christoffel symbols.

To get the simplest setup that generates sample trajectories for the stochastic model of training, we opt for a numerical approximation of the Christoffel symbols using their definition

$$\Gamma_{ij}^k = \frac{1}{2}\, g^{kl}\left(\partial_i g_{jl} + \partial_j g_{il} - \partial_l g_{ij}\right) \tag{3.3.30}$$

and a simple centered finite difference scheme. We integrate this explicit numerical representation of the RLE (3.3.1), using the Euler-Maruyama approximation (see Chapter 9.1 [19]).

The goal of these simulations is twofold. First, setting up a stochastic model for DLN training is of individual interest. When studying the deterministic training dynamics, we neglected that in real-world applications, training is always stochastic. The RLE is a principled way of putting back the randomness into the dynamics. Second, it also serves as an attempt to abstract away the effects of initialization on the training outcome.

The main thesis of Chapter 2 was that accumulation of training outcomes can be predicted by understanding state space volume on the zero energy surface. Intrinsic Brownian motion, as a scaling limit of geodesic random walks is an intuitive candidate for a purely numerical exploration of the volume landscape.

To fulfill this purpose, we initialize on the zero energy surface. This minimizes the energetic effects on the trajectories, since the zero energy surface consists of stable fixed points of the gradient dynamics. By adjusting the inverse temperature β, we can ensure that the trajectories do not drift away from the zero energy surface.

The results shown in Figure 3.3 support our intuition. Stochastic trajectories seem to accumulate in the same regions as their deterministic counterparts. This happens despite using completely different initialization, reinforcing that the accumulation of training outcomes in the DLN is primarily a result of volumetric ("entropic") effects as compared to a feature of the energy landscape.

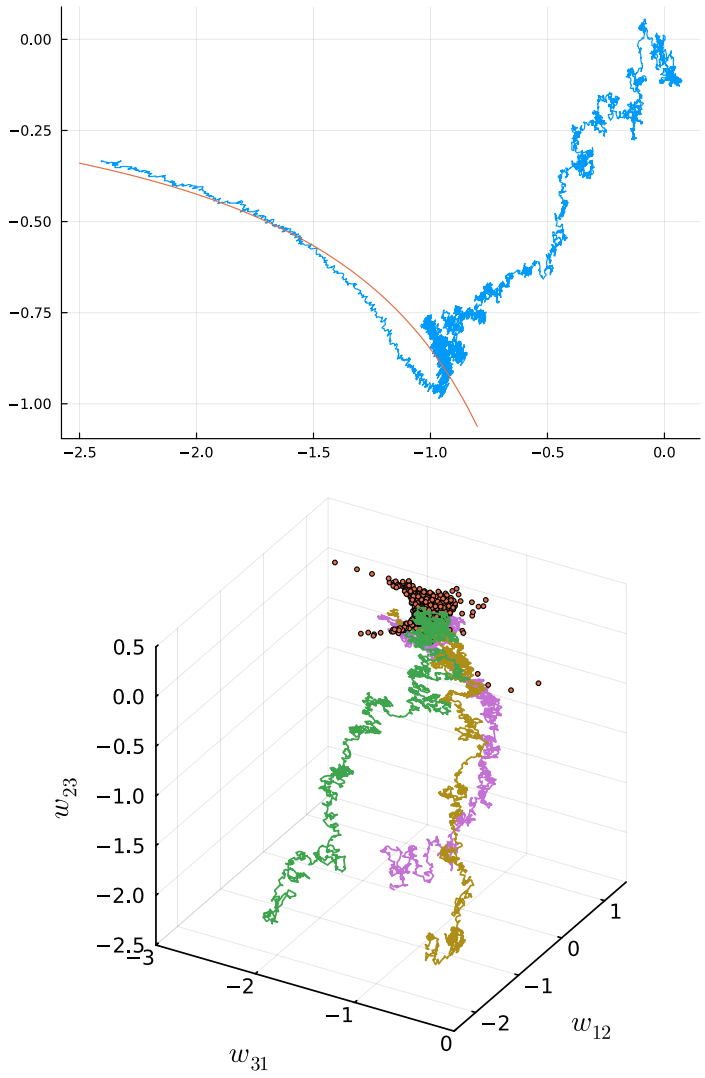


Figure 3.3: Typical trajectories of the RLE for the example of Sections 2.3.3.2 and 2.3.4.2.

Chapter Four

Discussion and Future Work

This dissertation presented contributions to the understanding of the DLN geometry and its training dynamics. In particular, we extended the geometry by providing the $N \to \infty$ limit. We also gave an explicit matrix representation to the metric and computed its volume form. Basis one-forms, which allow for easier computations of second order geometric quantities were also identified. It would be useful to continue these calculations in an attempt to obtain simple formulas. Upon successful simplification, Theorem 3.3.5 gives a way of computing curvatures. Based on the observed large volume near singularities, we conjecture that the DLN geometry is at least in part negatively curved.

The infinite depth limit is a theoretical construction that does not immediately correspond to real learning applications. Beyond its striking formal simplicity, the $N \to \infty$ limit has also proven to be relevant. In our analysis and experiments, it demonstrated equivalent behavior when compared to its finite N counterparts. When studying the $N = \infty$ limit, the obvious notion of upstairs is lost. It would

be very interesting to see whether an appropriate definition of infinite dimensional upstairs, admitting the equivalent of the balanced manifold exists.

The understanding of training dynamics have been extended by computing linear attraction rates and establishing normal hyperbolicity for matrix completion. We proved partial results on the question of singular value non-crossing, the completion of which offers obvious ways for further contributions. We pointed out formal similarities between DLN training dynamics and dynamical systems computing the SVD of a fixed matrix. Making these connections more explicit can insert the DLN dynamics in a greater context of well studied matrix dynamical systems.

We extended the deterministic training dynamics by offering a canonical way of adding noise to the DLN training. We set up numerical simulations that showed that this model is consistent with previous deterministic training experiments. Under deterministic dynamics, source of randomness was coming merely from initialization, which is now replaced by the noise in the dynamics.

Brownian motion on Riemannian submersions, the gradient of logarithmic volume of the associated group orbits and mean curvature flows are related. The interested reader is referred to [24], [26] and [18] for an introduction. While mean curvature for the DLN is not available at the moment, calculations of the volume of group orbits can be feasible, as demonstrated by the following computation for $N = 2$.

Take a balanced pair $W_2 = U_2\sqrt{\Sigma}U_1^T$ $W_1 = U_1\sqrt{\Sigma}V_1^T$. The group orbit at a point is parametrized by $\mathcal{O}(Q) = (W_2Q^T, QW_1)$. The tangent space at $Q = I$ on the orbit is given by

$$T_{(W_2,W_1)}O(I) = \{(-W_2A, AW_1)|A \in \text{Skew}(d)\}. \tag{4.0.1}$$

Using e_1 to denote the standard basis of $\mathbb{R}^d$, define

$$e_{ij} = \frac{1}{\sqrt{2}}(e_i e_j^T - e_j e_i^T). \tag{4.0.2}$$

Notice that

$$\alpha_{ij} = (U_1 e_{ij} U_1^T, -U_1 e_{ij} U_1^T) \tag{4.0.3}$$

forms an orthogonal basis of the tangent space. The length of these basis vectors is

$$\|\alpha_{ij}\|^2 = \text{trace}(\alpha_{ij}^T W_2^T W_2 \alpha_{ij} + W_1^T \alpha_{ij}^T \alpha_{ij} W_1) = 2(\sigma_i + \sigma_j). \tag{4.0.4}$$

Thus the volume of the group orbit only needs to be known up to a multiplicative constant, which is given by the volume spanned by α_{ij}:

$$\sqrt{\prod_{i<j}(\sigma_i + \sigma_j)} = \sqrt{\frac{\text{van}(\Sigma^2)}{\text{van}(\Sigma)}}. \tag{4.0.5}$$

This calculation may be generalized to $N > 2$, and it could give rise to contributions to undestanding o stochastic training dynamics.

We demonstrated the explanatory power of volume in predicting training outcomes. In specific examples we showed that the blowup of volume is at least quantitatively faster near observed training outcomes. In these examples, we also obtained analytical and numerical results on the blowup rate of volume density. While we expect the computed blowup rates to be "generic", the exact sense in which such a statement holds remains unknown at the time of writing this dissertation.

In a greater scope, the main motivation for studying matrix models is the idea that they may generalize to a larger class of learning architectures. The geometric construction of Riemannian submersion allows to write down dynamical systems

directly on the space of functions on which the architecture is expressive. We expect this picture to survive for the setting of (nonlinear) artificial neural networks, and we are looking forward to exciting new work being completed in this space.

BIBLIOGRAPHY

[1] S. Arora, N. Cohen, N. Golowich, and W. Hu. A convergence analysis of gradient descent for deep linear neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[2] S. Arora, N. Cohen, W. Hu, and Y. Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7411–7422, 2019.

[3] B. Bah, H. Rauhut, U. Terstiege, and M. Westdickenberg. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *Inf. Inference*, 11(1):307–353, 2022.

[4] D. A. Bayer and J. C. Lagarias. The nonlinear geometry of linear programming. i. affine and projective scaling trajectories. *Transactions of the American Mathematical Society*, 314(2):499–526, 1989.

[5] D. A. Bayer and J. C. Lagarias. The nonlinear geometry of linear programming. ii. legendre transform coordinates and central trajectories. *Transactions of the American Mathematical Society*, 314(2):527–581, 1989.

[6] R. W. Brockett. Dynamical systems that sort lists, diagonalize matrices, and solve linear equations. *Linear algebra and its applications*, 79:155–169, 1986.

[7] A. Bunse-Gerstner, R. Byers, V. Mehrmann, and N. K. Nichols. Numerical computation of an analytic singular value decomposition of a matrix valued function. *Numer. Math.*, 60(1):1–39, 1991.

[8] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[9] É. Cartan. Les groupes de transformations continus, infinis, simples. *Annales scientifiques de l'École normale supérieure*, 26(3):93–161, 1909.

[10] P. A. Deift. *Orthogonal polynomials and random matrices: a Riemann-Hilbert approach*, volume 3 of *Courant Lecture Notes in Mathematics*. New York University, Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence, RI, 1999.

[11] S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6151–6159, 2017.

[12] B. Hanin and M. Nica. Products of many large random matrices and gradients in deep neural networks. *Comm. Math. Phys.*, 376(1):287–322, 2020.

[13] M. Hardt, R. Meka, P. Raghavendra, and B. Weitz. Computational limits for matrix completion. In *COLT*, 2014.

[14] U. Helmke and M. J. B. Singular-value decomposition via gradient and self-equivalent flows. *Linear Algebra and its Applications*, 169:223–248, 1992.

[15] R. Hildebrand. Canonical barriers on convex cones. *Mathematics of Operations Research*, 39(3):841–850, Feb. 2014.

[16] E. Hsu. *stochastic analysis on manifolds*. American Mathematical Society, 2001.

[17] E. P. Hsu. *Stochastic analysis on manifolds*, volume 38 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2002.

[18] C.-P. Huang. A model of invariant control system using mean curvature drift from brownian motion under submersions, 2022.

[19] P. E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, Berlin, 1992.

[20] A. Lapedes and R. Farber. How neural nets work. In *Evolution, learning and cognition*, pages 331–346. World Sci. Publ., Teaneck, NJ, 1988.

[21] Y. LeCun, Y. Bengio, and G. E. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[22] G. Menon. Gibbs measures for semidefinite programming. *Manuscript, Brown University*, 2020.

[23] G. Menon and T. Trogdon. Random matrix theory and numerical linear algebra. *Manuscript, Brown University*, 2020.

[24] G. Menon and T. Yu. The riemannian langevin equation and conic programs, 2023.

[25] B. Neyshabur, R. Tomioka, and N. Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning, 2014.

[26] T. Pacini. Mean curvature flow, orbits, moment maps. *Transactions of the American Mathematical Society*, 355(8):3343–3357, 2003.

[27] N. Razin and N. Cohen. Implicit regularization in deep learning may not be explainable by norms, 2020.