THE GEOMETRY OF THE DEEP LINEAR NETWORK

GOVIND MENON

Abstract. This article provides an expository account of training dynamics in the Deep Linear Network (DLN) from the perspective of the geometric theory of dynamical systems. Rigorous results by several authors are unified into a thermodynamic framework for deep learning.

The analysis begins with a characterization of the invariant manifolds and Riemannian geometry in the DLN. This is followed by exact formulas for a Boltzmann entropy, as well as stochastic gradient descent of free energy using a Riemannian Langevin Equation. Several links between the DLN and other areas of mathematics are discussed, along with some open questions.

Contents

This work was supported by the National Science Foundation (DMS grants 171487, 2107205 and 2407055), the Simons Foundation (Award 561041) and the Charles Simonyi Foundation. The author thanks the School of Mathematics at the Institute for Advanced Study, Princeton for partial support during the completion of this work.

1. Introduction

In its simplest form, deep learning is a version of function approximation by neural networks, where the parameters of the network are determined by given data. The best fit is determined, or the network is trained, by minimizing an empirical cost function using gradient descent. Many of the mysteries of deep learning concern training dynamics: Do we have convergence? If so, how fast and to what? How do we make training more efficient? How do the training dynamics depend on the network architecture or on the size of data?

This article focuses on mathematical foundations. Our goal is to illustrate the utility of the geometric theory of dynamical systems for the study of these questions. While gradient flows have been studied in mathematics since the 1930s, gradient flows arising in deep learning have two subtle aspects –*overparametrization* and *degenerate loss functions* – that prevent naive applications of the standard theory of gradient flows (see Section 3.2) . We present a geometric framework for a simplified model, the Deep Linear Network (DLN), where these aspects can be studied with complete rigor.

The DLN is deep learning for *linear* functions. This reduces the training dynamics to gradient flows on spaces of matrices. Despite its apparent simplicity, the

model has a rich mathematical structure. Overparametrization provides a foliation of phase space by invariant manifolds. Of these, there is a fundamental class of invariant manifolds, the *balanced manifolds*, which are themselves foliated by group orbits. This geometric structure allows us to define a natural Boltzmann entropy (the logarithm of the volume of a group orbit) that may be computed explicitly. Microscopic fluctuations that underlie the entropy may be described by Riemannian Langevin equations. This approach unifies the work of several authors into a thermodynamic framework. In particular, it suggests an entropic origin for implicit regularization.

My view is that the DLN is a gift to mathematics from computer science. On one hand, it is subtle, but tractable, providing a rich set of practical questions and insights. On the other hand, the study of the DLN is filled with sharp theorems, exact formulas and unexpected mathematical structure. While the gradient dynamics of the DLN are different from the standard theory, we also see familiar aspects in surprising combinations. This gives the analysis a classical feel, even though all the results here were obtained in the very recent past. There is plenty more to be discovered and the real purpose of this article is to explain why.

In order to apply the methods of dynamical systems theory, what is of most value is to understand the dynamicists 'way of seeing'. This is not so much a collection of theorems, as a systematic use of geometry, and particular examples, to figure out what questions one should ask. Geometric methods provide a powerful intuition that is often a source of new discoveries.

This is the approach we use in this article. All the theorems we prove are guided by the work of computer scientists in the area. While this article does include some advanced mathematics, especially Riemannian geometry and random matrix theory, we stress explicit calculations, heuristic insights and representative examples. We hope this approach reveals the conceptual power of dynamical systems theory while remaining of interest to practitioners. The references are representative, not exhaustive, since our aim in this article is to provide a pedagogical treatment. We include a brief discussion of the literature on the DLN in Section 13.2. For broader surveys of deep learning that include related mathematical ideas, the reader is referred to the recent books [1, 37].

The article concludes with open questions that emerge from this perspective. These include specific mathematical questions on the DLN, as well as the extension of our entropy formula to gauge groups arising in deep learning.

2. A caricature of deep learning

In its simplest mathematical variant, the purpose of deep learning is to create a function $f : \mathbb{R}^{d_x} \to \mathbb{R}^{d_y}$ given a collection of training data $\{(X_k, Y_k)\}_{k=1}^n$, with $X_k \in \mathbb{R}^{d_x}$, $Y_k \in \mathbb{R}^{d_y}$ and n denoting the size of the training dat set. In contrast with approximation by a linear combination of basis functions, such as the use of Fourier series or polynomials, in deep learning a function is approximated using neural networks. The neural network consists of individual function elements (neurons) and a hierarchical architecture (the network) so that complicated functions can be constructed by scaling, translating and composing sums of the basic function element. The parameters, denoted for convenience $\mathbf{W} \in \mathbb{R}^M$, of the neural network describe the underlying scaling and translation (see [15, p.5] for explicit descriptions).

The parameters $\mathbf{W} \in \mathbb{R}^M$ are determined from the data by minimizing an empirical loss function. A natural example is the quadratic loss function

$$L(\mathbf{W}) = \frac{1}{n}\sum_{k=1}^{n} |f(X_k;\mathbf{W}) - Y_k|^2. \tag{2.1}$$

The parameters $\mathbf{W}$ are then obtained by minimizing the loss function using gradient descent. We model this process with the differential equation

$$\dot{\mathbf{W}} = -\nabla_{\mathbf{W}} L(\mathbf{W}), \tag{2.2}$$

where $\nabla_{\mathbf{W}}$ denotes the gradient with respect to the Euclidean norm in $\mathbb{R}^M$. In practice, the gradient descent is discrete in time, relies on the back-propagation algorithm, and includes random batch processing. These dynamics involve subtle corrections to the gradient flow (2.2), but for the purposes of rigorous analysis, it is is first necessary to have an understanding of equation (2.2).

Our aim in this article will be to foster a geometric understanding of training dynamics for deep networks. In order to do so, we first form a geometric picture of the underlying approximation theory. We will then use the DLN to make precise several aspects of this geometric picture. Here are the important ideas in our work.

(1) *Riemannian submersion.* The creation of a function $\mathbf{W} \mapsto f(\cdot,\cdot;\mathbf{W})$ by the neural network defines a map $\mathcal{F} : \mathbb{R}^M \to C(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$. The image of this map, $\mathcal{G}$, is the class of functions that can be created with the network. We will visualize this by thinking of the parameter space $\mathbb{R}^M$ as living 'upstairs' and its image $\mathcal{G}$ living 'downstairs'. The natural geometric relation between these spaces is provided by Riemannian submersion.

(2) *Scale-by-scale composition.* Simple architectural motifs define the architecture for bump functions [28, 33]. Thus, intuitively deep networks form complicated functions through the scale-by-scale composition of bump functions. The expressivity of the deep network – its ability to approximate complicated functions – is quantified by the 'size' of $\mathcal{G} \subset C(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$. The word size is in quotes, because of the complexity of the map $\mathcal{F}$ and the fact that $C(\mathbb{R}^{d_x}, \mathbb{R}^{d_y})$ is infinite-dimensional. However, this idea may be explored with approximation theory and numerical experiments [15].

(3) *The geometry of overparametrization.* The problem is overparametrized when the same function $h \in \mathcal{G}$ may be created with different choices of weights $\mathbf{W} \in \mathbb{R}^M$. Precisely, given $h \in \mathcal{G}$ we must understand a 'large' inverse image $\mathcal{F}^{-1}(h) := \mathcal{H} \subset \mathbb{R}^M$. It is natural to assume at first that $\mathcal{H}$ is a manifold. Then $T_{\mathbf{W}}\mathcal{H}$ defines a set of 'null directions' for the gradient descent. (Since the loss function E depends on $\mathbf{W}$ through $h = \mathcal{F}(\mathbf{W})$ alone it does not change when $\mathbf{W}$ is varied along the directions in $T_{\mathbf{W}}\mathcal{H}$.) If $\mathcal{H}$ is large, then the gradient descent is restricted to a 'thin' set of directions in $\mathbb{R}^M$. We will quantify these ideas precisely for the DLN with a Boltzmann entropy. Further, we show that noise in the null directions, gives rise to motion by curvature.

(4) *Degenerate loss functions.* While the loss function is quadratic, there may be many choices of h such that the empirical loss function is minimized for such h (for example, when n is sufficiently small). Naively, this corresponds to *overfitting*. For example, sufficiently high-degree polynomials will fit the data, but also generate spurious oscillations in between the data points. The

functions created by deep learning do not overfit. Heuristically, this seems to be due to the efficient coding of functions made possible by scale-by-scale composition. However, a rigorous analysis must explain this behavior as a consequence of the training dynamics.

We revisit these ideas at the conclusion of this article. This allows us to compare the DLN and deep learning equipped with a precise geometric understanding of training dynamics in the DLN.

3. The deep linear network

3.1. **The model.** Given a positive integer m, we denote by $\mathbb{M}_m$ the space of $m \times m$ real matrices; $\mathrm{Symm}_m \subset \mathbb{M}_m$ the space of symmetric matrices; $\mathbb{A}_m$ the space of anti-symmetric matrices; and $\mathbb{P}_m \subset \mathbb{M}_m$ the space of positive semidefinite (psd) matrices. Finally, O_m denotes the orthogonal group of dimension m and $\mathrm{St}_{r,m}$ denotes the Stiefel manifold of r-frames in $\mathbb{R}^m$.

We fix two positive integer d and N referred to as the width and depth of the network. The state space for the DLN is $\mathbb{M}_d^N$. Each point $\mathbf{W} \in \mathbb{M}_d^N$ is denoted by

$$\mathbf{W} = (W_N, W_{N-1}, \dots, W_1). \tag{3.1}$$

We equip $\mathbb{M}_d$ with the Frobenius norm so that $\mathbb{M}_d^N$ is Euclidean with the norm

$$\|\mathbf{W}\|^2 = \sum_{p=1}^{N} \mathrm{Tr}\left(W_p^T W_p\right). \tag{3.2}$$

We define the projection $\phi : \mathbb{M}_d^N \to \mathbb{M}_d$ and the *end-to-end* matrix W by[1]

$$\phi(\mathbf{W}) = W_N W_{N-1} \cdots W_1 =: W. \tag{3.3}$$

We assume given an energy $E : \mathbb{M}_d \to \mathbb{R}$. The training dynamics are described by the gradient flow in $\mathbb{M}_d^N$ with respect to the Frobenius norm of the 'lifted' loss function $L = E \circ \phi$

$$\dot{\mathbf{W}} = -\nabla_{\mathbf{W}} L(\mathbf{W}). \tag{3.4}$$

This is a collection of N equations in $\mathbb{M}_d$

$$\dot{W}_p = -\nabla_{W_p} E(W_N \cdots W_1), \quad p = 1, \dots, N. \tag{3.5}$$

A computation using equation (3.3) (see Section 7.1.1) simplifies equation (3.5) to

$$\dot{W}_p = -(W_N \cdots W_{p+1})^T E'(W)(W_{p-1} \cdots W_1)^T, \quad p = 1, \dots, N. \tag{3.6}$$

Here $E'(W)$ denotes the $d \times d$ matrix with entries

$$E'(W)_{jk} = \frac{\partial E}{\partial W_{jk}}, \quad 1 \le j, k \le d. \tag{3.7}$$

This article focuses on the analysis of this gradient flow.

[1] It is not necessary to assume that all the matrices are $d \times d$. All that is required is that the matrix multiplication in equation (3.3) is well-defined. However, we restrict attention to square matrices to illustrate the main ideas.

3.2. **A comparison with deep learning.** Let $d = d_x = d_y$ and let $f_p : \mathbb{R}^d \to \mathbb{R}^d$ define the linear function $f_p(x) = W_p x$ for $1 \leq p \leq n$. Then the linear function $f : \mathbb{R}^d \to \mathbb{R}^d$ corresponding to the matrix $f(x) = Wx$ is $f = f_N \circ f_{N-1} \ldots \circ f_1$. The output function f depends only on the end-to-end matrix W. However, the same function f may be represented by Nd^2 choices of training parameters $\mathbf{W}$. Thus, despite the absence of the nonlinear activation element and shifts as in a neural network, the choice of variables in the DLN models *overparametrization*.

Natural learning tasks for the DLN, such as matrix completion, give rise to *degenerate loss functions*. Assume given a subset $S \subset \{(i,j)\}_{1 \leq i,j \leq d}$ and assume given the values of W_{ij}, for $(i,j) \in S$, say $W_{ij} = a_{ij}$. The task in matrix completion task is to obtain a principled answer to the question: how do we reconstruct W from the partial observations a_{ij} for $(i,j) \in S$?

The DLN may be used to study this question as follows. We define the quadratic loss function

$$E(W) = \frac{1}{2} \sum_{(i,j) \in S} |W_{ij} - a_{ij}|^2, \tag{3.8}$$

and seek the limit of $W(t) = \phi(\mathbf{W}(t))$ as $t \to \infty$ when $\mathbf{W}(t)$ solves the gradient flow (3.6).

This loss function is degenerate because it does not depend on the values M_{ij} when the indices (i,j) do not lie in S. Thus, the loss function is minimized on the affine subspace

$$\mathcal{S} = \{W \in \mathbb{M}_d : W_{ij} = a_{ij}, \quad (i,j) \in S\}.$$

The analysis of the gradient flow (3.6) lies beyond the standard theory because of the interplay between overparametrization and the degeneracy of the loss function. Here by standard theory, we mean the analysis of gradient flows using the main results on convergence of the gradient flow such as La Salle's invariance principle and the closely related Barbashin-Krasovskii criterion [9, 29], the use of the Lojasiewicz convergence criterion [27, 40] for analytic loss functions, and the Morse-Smale decomposition of the gradient flow [35] when the loss function is Morse. These results provide the typical framework for the analysis of gradient flows, but they cannot be naively applied to the DLN. The use of La Salle's invariance principle is not valid when the loss function $E(W)$ does not grow sufficiently fast as $|W| \to \infty$ (for example, for matrix completion). The Morse-Smale decomposition does not hold since $\phi \circ E(\mathbf{W})$ is not a Morse function due to overparametrization.

3.3. **Main results.** The theorems in this article are a composite of results obtained by several authors. They have been chosen to illustrate three geometric aspects.

(1) *Foliation by invariant varieties; the* $\mathbf{G}$*-balanced varieties*. The phase space $\mathbb{M}_d^N$ is foliated by invariant varieties described by quadratic matrix equations parametrized by $\mathbf{G} \in \mathrm{Symm}_d^{N-1}$. Theorem 1 explains this structure. It is a geometric consequence of 'thin gradients' and holds for all E.

(2) *Riemannian gradient flows*. The dynamics on the invariant varieties may be explicitly described when $\mathbf{G} = \mathbf{0}$. The variety $\mathcal{M}_\mathbf{0}$ is itself foliated by manifolds corresponding to W with rank r, $1 \leq r \leq d$. We refer to these as *balanced manifolds*, denoted $\mathcal{M}_r$, or simply $\mathcal{M}$, when $r = d$. Each balanced manifold $\mathcal{M}_r$ is invariant and the dynamics on the balanced manifolds is

described by a Riemannian gradient flow, with a metric that may be computed exactly using Riemannian submersion. This metric has an obvious infinite-depth limit (though the Riemannian submersion itself does not).

(3) *Group orbits, entropy and stochastic dynamics*. We define a Boltzmann entropy of the form $\log \mathrm{vol}(\mathcal{O}_W)$ for group orbits on $\mathcal{M}_r$. This entropy may be computed explicitly (see Theorem 10). The microscopic dynamics associated to the entropy are described using a Riemannian Langevin Equation (RLE) which reveals the role of curvature in the dynamics.

The rigorous results are only part of the story. Each of the main theorems formalizes a different form of geometric intuition and was guided by heuristics, numerical experiments and connections with other areas of mathematics. These connections are truly surprising. For example, the balanced varieties have a subtle relation with the theory of minimal surfaces; a special case of the Riemannian geometry of the DLN is the Bures-Wasserstein geometry on $\mathbb{P}_d$; and the entropy formula is obtained by analogy with Dyson Brownian motion in random matrix theory.

For these reasons, we first state these results along with some background. This allows the reader to obtain an overview of the geometry of the DLN without much baggage. Most proofs are presented in the sections that follow.

4. $\mathbf{G}$-balanced varieties are invariant

4.1. **Definitions.** Denote the coordinates of $\mathbf{G} \in \mathrm{Symm}_d^{N-1}$ by

$$\mathbf{G} = (G_{N-1}, \cdots, G_1). \tag{4.1}$$

Given $\mathbf{G}$, consider the system of $N-1$ quadratic equations

$$W_{p+1}^T W_{p+1} = W_p W_p^T - G_p, \quad 1 \le p \le N-1. \tag{4.2}$$

The solution set defines an algebraic variety that is termed the $\mathbf{G}$-balanced variety. We denote it by

$$\mathcal{M}_{\mathbf{G}} = \{\mathbf{W} \in \mathbb{M}_d^N \,|\, W_{p+1}^T W_{p+1} = W_p W_p^T - G_p,\ 1 \le p \le N-1.\} \tag{4.3}$$

Not all values of $\mathbf{G}$ give rise to non-empty $\mathcal{M}_{\mathbf{G}}$. However, given a point $\mathbf{W} \in \mathbb{M}_d^N$, we may use equation (4.2) to define $\mathbf{G}$. Thus, the space $\mathbb{M}_d^N$ is fibered by the varieties $\mathcal{M}_{\mathbf{G}}$.

When $\mathbf{G} = \mathbf{0}$, equation (4.2) reduces to the important special case

$$W_{p+1}^T W_{p+1} = W_p W_p^T, \quad 1 \le p \le N-1. \tag{4.4}$$

These equations define the *balanced variety*

$$\mathcal{M}_{\mathbf{0}} = \{\mathbf{W} \in \mathbb{M}_d^N \,|\, W_{p+1}^T W_{p+1} = W_p W_p^T,\ 1 \le p \le N-1.\} \tag{4.5}$$

If $\mathbf{W} \in \mathcal{M}_{\mathbf{0}}$, the singular values, and thus rank, of each W_p are the same. In Section 5, we use this observation to construct a parametrization which shows that $\mathcal{M}_{\mathbf{0}}$ is foliated by manifolds $\mathcal{M}_r$ corresponding to the rank r, $1 \le r \le d$. We refer to the leaf of $\mathcal{M}_{\mathbf{0}}$ with rank $r = d$ as the *balanced manifold* and denote it by $\mathcal{M}$. Our main theorems concern the behavior on $\mathcal{M}_r$.

We lack a complete understanding of the singularities of the variety $\mathcal{M}_{\mathbf{G}}$. However, it is easy to check that $\mathcal{M}$ is a manifold. For each p, equation (4.2) is Symm_d-valued; thus, there are $(N-1)d(d+1)/2$ scalar equations. Since we have Nd^2

parameters, we find that

$$\dim(\mathcal{M}) = d^2 + (N-1)\frac{d(d-1)}{2}. \tag{4.6}$$

Of these, d^2 degrees of freedom correspond to an end-to-end matrix $W \in \mathbb{M}_d$ and the remaining $(N-1)d(d-1)/2$ degrees of freedom correspond to an O_d^{N-1} group orbit. We parametrize $\mathcal{M}_{\mathbf{G}}$ in Section 5 to make this explicit. We focus mainly on $\mathcal{M}$ for simplicity. When we consider the rank-deficient cases, $\mathcal{M}_r$ with $r < d$, the O_d^{N-1} orbits have to be replaced by $\mathrm{St}_{r,d}^{N-1}$ orbits.

4.2. **Dynamics on invariant manifolds.** Each variety $\mathcal{M}_{\mathbf{G}}$ is invariant under the dynamics. Let us state this assertion precisely.

Assume that $E \in C^2$ and consider the initial value problem

$$\dot{\mathbf{W}} = -\nabla_{\mathbf{W}} E \circ \phi(W), \quad \mathbf{W}(0) = \mathbf{W}_0. \tag{4.7}$$

This is a differential equation with a locally Lipschitz vector field. Thus, Picard's theorem guarantees the existence of a unique solution on a maximal time interval $(T_{\min}, T_{\max})$ containing $t = 0$. Let $\mathbf{G}$ be given by the initial condition

$$G_p = W_{p+1}^T(0)W_{p+1}(0) - W_p(0)W_p^T(0), \quad 1 \le p \le n. \tag{4.8}$$

Theorem 1 (Arora, Cohen, Hazan [3]). *The following hold on the maximal interval of existence of solutions to equation* (4.7).

(a) *The solution* $\mathbf{W}(t)$ *lies on the variety* $\mathcal{M}_{\mathbf{G}}$.
(b) *The end-to-end matrix* $W(t)$ *satisfies*

$$\dot{W} = -\sum_{p=1}^{N} (A_{p+1}A_{p+1}^T)\, E'(W)\, (B_{p-1}^T B_{p-1}), \tag{4.9}$$

where $A_{N+1} = B_0 = 1$ *and we have defined the partial products*

$$A_p = W_N \cdots W_p, \quad B_p = W_p \cdots W_1, \quad 1 \le p \le N. \tag{4.10}$$

The specific nature of E is irrelevant to Theorem 1. Instead, it reflects a fundamental geometric restriction forced by overparametrization. The gradient vector fields $\nabla_{\mathbf{W}} L(\mathbf{W})$ always lie along a 'thin' space of dimension d^2 in $\mathbb{M}_d^N$. We explain this point further in Remark 5 below. Theorem 1 tells us that each variety $\mathcal{M}_{\mathbf{G}}$ is invariant under the dynamics, but it does not provide a closed description of the reduced dynamics. However, on the balanced manifold $\mathcal{M}$ we have

Theorem 2 (Arora, Cohen, Hazan [3]). *Assume* $\mathbf{W}(0) \in \mathcal{M}$. *The end-to-end matrix* $W(t) = \phi(\mathbf{W}(t))$ *satisfies the differential equation*

$$\dot{W} = -\sum_{k=1}^{N} (WW^T)^{\frac{N-k}{N}}\, E'(W)\, (W^T W)^{\frac{k-1}{N}}. \tag{4.11}$$

Theorem 1 and Theorem 2 are easy to establish (see Section 7). However, despite their simplicity, both Theorems have a fascinating character. The most important structural feature is that equation (4.11) is a *Riemannian* gradient flow in disguise. We explain this idea below and then conclude with some remarks on these theorems.

4.3. **Riemannian gradient descent.** We recommend [30] as an introduction to Riemannian geometry. The reader willing to take some definitions on faith can also understand the main ideas in this work without a detailed understanding of Riemannian geometry.

A metric g on $\mathbb{M}_d$ assigns a length to each tangent vector $Z \in T\mathbb{M}_d$. Since the tangent space to $\mathbb{M}_d$ at any point is itself isomorphic to $\mathbb{M}_d$, a metric is an assignment of lengths of the form

$$g(W)(Z,Z) := \|Z\|^2_{g(W)}, \quad Z \in T_W\mathbb{M}_d. \tag{4.12}$$

The inner-product $g(W)(Z_1,Z_2)$ may be recovered from the polarization identity

$$g(W)(Z_1,Z_2) = \frac{1}{4}\left(\|Z_1+Z_2\|^2_{g(W)} - \|Z_1-Z_2\|^2_{g(W)}\right). \tag{4.13}$$

We define a metric g^N on $\mathbb{M}_d$ as follows. Given the depth N for every $W \in \mathbb{M}_d$ we define the linear map

$$\mathcal{A}_{N,W} : T_W\mathbb{M}_d \to T_W\mathbb{M}_d, \quad Z \mapsto \sum_{k=1}^{N} (WW^T)^{\frac{N-k}{N}} Z (W^TW)^{\frac{k-1}{N}}. \tag{4.14}$$

This linear map allows us to rewrite equation (4.11) as

$$\dot{W} = -\mathcal{A}_{N,W}\left(E'(W)\right). \tag{4.15}$$

We will show that the operator $\mathcal{A}_{N,W}$ is always invertible when W has full rank. In this setting, the structure of equation (4.15) may be understood better. Define

$$g^N(W)(Z,Z) = \mathrm{Tr}(Z^T\mathcal{A}^{-1}_{N,W}Z). \tag{4.16}$$

It then follows from the polarization identity and the identity $(\mathcal{A}_{N,W}(Z))^T = \mathcal{A}_{N,W}(Z^T)$ that

$$g^N(W)(Z_1,Z_2) = \mathrm{Tr}(Z_1^T\mathcal{A}^{-1}_{N,W}Z_2). \tag{4.17}$$

Given a Riemannian manifold $(\mathcal{N},h)$ and a differentiable function $E : \mathcal{N} \to \mathbb{R}$, the Riemannian gradient of E is a vector in $T_x\mathcal{N}$ obtained from the duality pairing

$$dE(x)(z) = h(\mathrm{grad}_h E(x), z), \quad z \in T_x\mathbb{N}. \tag{4.18}$$

In our setting, $(\mathcal{N},h) = (\mathbb{M}_d, g^N)$ where g^N is given by equation (4.16) for $Z \in T_M\mathbb{M}_d$. Then the left and right hand sides of this equation are

$$dE(W)(Z) = \mathrm{Tr}(Z^TE'(W)), \quad g^N(\mathrm{grad}_{g^N}E(W), Z) = \mathrm{Tr}(Z^T\mathcal{A}^{-1}_{N,W}E'(W)). \tag{4.19}$$

A more careful analysis [5] reveals that the assumption that W have full rank is not necessary. Following [5] we foliate $\mathbb{M}_d$ by rank, defining

$$\mathfrak{M}_r = \{W \in \mathbb{M}_d \,|\mathrm{rank}(W) = r\}. \tag{4.20}$$

The balanced manifolds $\mathcal{M}_r$ are naturally related to $\mathfrak{M}_r$: $\mathbf{W} \in \mathcal{M}_0$ lies in $\mathcal{M}_r$ if and only if $W = \phi(\mathbf{W}) \in \mathfrak{M}_r$. Further, it is shown in [5, §3] that the metric g^N restricts naturally to $\mathfrak{M}_r$ for $r < d$. We then have

Theorem 3 (Bah, Rauhut, Terstiege, Westdickenberg [5])**.** *Equation* (4.11) *is equivalent to the Riemannian gradient flow on* $(\mathfrak{M}_r, g^N)$

$$\dot{W} = -\mathrm{grad}_{g^N}E(W). \tag{4.21}$$

In particular, when $\mathbf{W}(0) \in \mathcal{M}_r$*, the end-to-end matrix* $W(t) = \phi(\mathbf{W}(t))$ *evolves in* $\mathfrak{M}_r$ *according to this Riemannian gradient flow.*

When $W(t)$ evolves according to (4.21) we have the fundamental estimate

$$\frac{dE(W(t))}{dt} = -\|\mathrm{grad}_{g^N} E\|^2_{g^N} \leq 0. \tag{4.22}$$

It follows that when $W(t)$ converges to a critical point of E, then $\mathbf{W}(t)$ converges to a critical point of $L = E \circ \phi$ at exactly the same rate. Overparametrization does not change the speed of convergence.

4.4. **Remarks on the invariant manifold theorems.**

Remark 4. Equation (4.11) was first obtained by using the gradient flow (3.5) and the identity $W = W_N W_{N-1} \cdots W_1$ [3]. The underlying metric g^N was then obtained by noting the properties of the map $\mathcal{A}_{N,W}$, including its restriction to rank r in [5]. These calculations may be explained geometrically using Riemannian submersion as discussed in Section 6.

Remark 5 (Thin gradients)**.** The space of gradients of loss functions of the form $L = E \circ \phi$ at a point $\mathbf{W} \in \mathbb{M}_d^N$ has at most d^2 dimensions. To see this, fix l, m and choose linear energies $E_{lm}(W) := W_{lm}$. There are d^2 such energies; thus, their gradients form a basis for the space of gradients of energies E at $W \in \mathbb{M}_d$. Equation (3.6) then shows that the space of gradients of $L = E \circ \phi$ at $\mathbf{W}$ has d^2 dimensions. However, the dimension of the space $\mathbb{M}_d^N$ is Nd^2.

Remark 6 (Group orbits)**.** Since the gradient flow can explore at most d^2 directions we have $(N-1)d^2$ free parameters. Of these, $(N-1)d(d+1)/2$ parameters are fixed by the constants $(G_N, \ldots, G_1)$. We use this observation to parametrize $\mathbf{W}$ by $\mathbb{M}_d \times O_d^{N-1}$ using the polar factorization in Section 5. In particular, for each $W \in \mathfrak{M}_r$ the inverse image $\phi^{-1}(W) \cap \mathcal{M}$ is a $\mathrm{St}_{r,d}^{N-1}$ orbit $\mathcal{O}_W$.

Remark 7 (Global existence and convergence)**.** The following method for establishing convergence was introduced in [5].

Picard's theorem provides local existence on an interval $(T_{\min}, T_{\max})$ with $T_{\min} < 0 < T_{\max}$. In order to obtain global existence for $\mathbf{W}(t)$, it is only necessary to show that $W(t) = \phi(\mathbf{W}(t))$ is bounded in time. If so, we have the uniform bounds

$$\|W_{p+1}\|^2 = \mathrm{Tr}(W_{p+1}^T W_{p+1}) = \mathrm{Tr}(W_p W_p^T - G_p) = \|W_p\|^2 - \mathrm{Tr}(G_p). \tag{4.23}$$

Thus, we may inductively control $\|W_N(t)\|, \ldots, \|W_2(t)\|$ in terms of $\|W_1(t)\|$. Since $W = W_N \cdots W_1$, the bound on $\|W_1(t)\|$ is equivalent to a bound on $\|W\|$. It then follows from a continuation argument that $T_{\min} = -\infty$ and $T_{\max} = +\infty$ if $\|W(t)\|$ is uniformly bounded on its interval of existence.

Once we have established that $\mathbf{W}(t)$ remains in a compact set, we may use standard criterion to establish convergence as $t \to \infty$. When E is analytic, the Lojasiewicz criterion implies convergence to a critical point. For $E \in C^2$, we may use the La Salle invariance principle to classify the ω-limit set $\omega(\mathbf{W}(0))$.

Remark 8 (Relation to the Simons cone)**.** The variety $\mathcal{M}_{\mathbf{G}}$ is a conic section in $\mathbb{M}_d^N$ with a parametrization discussed in Section 5. We may understand the balanced manifold $\mathcal{M}$ explicitly when $d = 2$ and $N = 2$. Let

$$W_1 = \begin{pmatrix} x_1 & x_2 \\ x_3 & x_4 \end{pmatrix}, \quad W_2 = \begin{pmatrix} x_5 & x_6 \\ x_7 & x_8 \end{pmatrix}. \tag{4.24}$$

Then equation (4.4) reduces to a system of three quadratic equations

$$\begin{aligned} x_5^2 + x_7^2 &= x_1^2 + x_2^2, && (4.25)\\ x_5x_6 + x_7x_8 &= x_1x_3 + x_2x_4 && (4.26)\\ x_6^2 + x_8^2 &= x_3^2 + x_4^2. && (4.27) \end{aligned}$$

When we add equations (4.25) and (4.27) we obtain the *Simons cone*

$$\mathcal{C} = \{x \in \mathbb{R}^8 \,|\, x_1^2 + x_2^2 + x_3^2 + x_4^2 = x_5^2 + x_6^2 + x_7^2 + x_8^2\} \subset \mathbb{R}^8. \tag{4.28}$$

The Simons cone is a fundamental counterexample in the calculus of variations and geometric measure theory. The variety $\mathcal{C}$ defined by (4.28) has zero mean curvature at all points $x \neq 0$ and is a local minimizer of the area functional. However, it is not smooth because of the singularity at $x = 0$. On the other hand, there are no such singularities for minimal surfaces in $\mathbb{R}^n$ when $n < 7$. Thus, the Simons cone is the simplest stable minimal surface with a singularity, and it manifests only in $\mathbb{R}^n$ with $n > 8$.

We see that the balanced manifold $\mathcal{M}$ is a five-dimensional variety contained within the Simons cone $\mathcal{C}$. We will use a stochastic extension of equation (3.5) to shed some light on this unexpected connection.

Remark 9. The space of positive definite matrices $\mathbb{P}_d$ may be equipped with the Bures-Wasserstein metric g^{BW} (see [10] for an exposition). Gradient flows on $(\mathbb{P}_d, g^{BW})$ can be obtained from the DLN as follows. We set $N = 2$ and further restrict $\mathbf{W}$ to the subspace

$$\mathcal{V} = \{\mathbf{W} \in \mathbb{M}_d^2 \,|\, W_2 = W_1^T\}. \tag{4.29}$$

Then $W = \phi(\mathbf{W}) = W_1^T W_1$ is a psd matrix and the metric g^2 given by Riemannian submersion of $\mathcal{V}$ is the Bures-Wasserstein metric g^{BW}.

5. Parametrization of $\mathcal{M}_{\mathbf{G}}$ and $\mathcal{M}$

5.1. **Polar factorization and the SVD.** Recall that a matrix $W \in \mathbb{M}_d$ admits left and right polar decompositions

$$W = QP, \quad W = RU^T \tag{5.1}$$

where $P, R \in \mathbb{P}_d$ and $Q, U \in O_d$. Further, $P = \sqrt{W^T W}$ and $R = \sqrt{WW^T}$ are the unique psd square roots of these matrices.

The polar factorization is related to the singular value decomposition (SVD) of W as follows. Let us denote the SVD by

$$W = V_R \Sigma V_P^T, \tag{5.2}$$

where $V_R, V_P \in O_d$ and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_d)$ with each $\sigma_i \geq 0$. Then we have immediately

$$P = \sqrt{W^T W} = V_P \Sigma V_P^T, \quad R = \sqrt{WW^T} = V_R \Sigma V_R^T. \tag{5.3}$$

Therefore, V_P and V_R provide an orthonormal basis of eigenvectors for P and R respectively. Finally, we have the relation between the O_d factors

$$Q = V_R V_P^T, \quad U = V_P V_R^T. \tag{5.4}$$

5.2. **Parametrization by the polar factorization.** We now parametrize each $\mathbf{G}$-balanced variety $\mathcal{M}_{\mathbf{G}}$ using the polar factorization, defining a map

$$\mathfrak{r} : O_d^{N-1} \times \mathbb{M}_d \to \mathbb{M}_d^N, \quad (Q_N, \dots Q_2, W_1) \mapsto (W_N, \dots, W_1). \tag{5.5}$$

An interesting feature of equation (4.2) is the manner in which the left and right polar factors alternate along the network as the index p varies. We use the polar factors of W_p to rewrite equation (4.2) in the form

$$P_{p+1}^2 = R_p^2 - G_p, \quad 1 \le p \le N-1. \tag{5.6}$$

We define the map $\mathfrak{r}$ as follows. Compute

$$R_1^2 = W_1 W_1^T, \quad P_2 = \sqrt{R_1^2 - G_1}, \quad W_2 = Q_2 P_2. \tag{5.7}$$

Now proceed algorithmically: as p increases, we find sequentially

$$R_p^2 = W_p W_p^T, \quad P_{p+1} = \sqrt{R_p^2 - G_p}, \quad W_{p+1} = Q_{p+1} P_{p+1}. \tag{5.8}$$

These equations may also be combined into an iterative sequence for P_p. We have

$$P_1 = W_1^T W_1, \quad P_{p+1} = \sqrt{Q_p P_p^2 Q_p^T - G_p}. \tag{5.9}$$

Each square-root is a smooth map when G_p is a negative definite matrix. Thus, the variety $\mathcal{M}_{\mathbf{G}}$ is a manifold with dimension $d^2 + (N-1)d(d-1)/2$ when each G_p is negative definite.

The balanced varieties are matrix-valued conic sections. A useful geometric caricature is obtained by considering the case $d = 1$.[2] Assume given $(v_{N-1}, \dots, v_1, w_1)$ where each $v_p = \pm 1$. Then given $\mathbf{g} := (g_{N-1}, \dots, g_1)$, we see that $\mathcal{M}_{\mathbf{g}}$ is a conic section in $\mathbb{R}^N$ described by the hyperbolas

$$w_{p+1} = v_{p+1} \sqrt{w_p^2 - g_p}, \quad 1 \le p \le N-1. \tag{5.10}$$

Finally, note that the parametrization goes from right to left. We have an analogous parametrization from left to right

$$\mathfrak{y} : \mathbb{M}_d \times O_d^{N-1} \to \mathbb{M}_d^N, \quad (W_N, U_{N-1}, \dots U_1) \mapsto (W_N, \dots, W_1), \tag{5.11}$$

given for $N - 1 \ge p \ge 1$ by the sequential polar factorizations

$$P_{p+1}^2 = W_{p+1}^T W_{p+1}, \quad R_p = \sqrt{P_{p+1}^2 + G_p}, \quad W_p = R_p U_p^T. \tag{5.12}$$

5.3. **Parametrization by the SVD.** The parametrization may be simplified further on the balanced variety $\mathcal{M}_{\mathbf{0}}$. We now have

$$W_{p+1}^T W_{p+1} = W_p W_p^T \tag{5.13}$$

Let us first study the case when W_N has full rank. Then P_N is positive definite and the parametrization $\mathfrak{r}$ continues to be smooth because equation (5.8) reduces to the chain of equalities

$$R_p = P_{p+1}, \quad W_p = R_p V_p^T, \quad P_p = \sqrt{W_p^T W_p} = V_p R_p V_p^T. \tag{5.14}$$

It follows that each P_p is positive definite and that $\mathcal{M}_{\mathbf{0}}$ is locally a manifold. We denote this branch of $\mathcal{M}_{\mathbf{0}}$ as $\mathcal{M}$ and call it the *balanced manifold*.

[2] We switch notation to lower case letters to prevent any confusion with the case $d \ge 2$.

The singular values of each W_p are identical on $\mathcal{M}$. This allows us to simplify the parametrization $\mathfrak{x}$ further. We define the coordinate map

$$\mathfrak{z} : \mathbb{R}_+^d \times O_d^{N+1} \to \mathcal{M}, \quad (\Lambda, Q_N, \ldots, Q_0) \to (W_N, \ldots, W_1), \tag{5.15}$$

where for each $1 \leq p \leq N$ we set

$$W_p = Q_p \Lambda Q_{p-1}^T, \quad \text{and} \quad \Lambda = \Sigma^{1/N}. \tag{5.16}$$

Here Λ denotes the singular values of each W_p and Σ denotes the singular values of $W = W_N \cdots W_1$. Indeed, it follows from equation (5.16) that

$$W = Q_N \Sigma Q_0^T. \tag{5.17}$$

This parametrization is a local bijection when the singular values are distinct. However, it can fail to be smooth when Σ has repeated singular values. We will compute the metric on $\mathcal{M}$ in Section 8.

5.4. **The rank-deficient case.** When W has rank $r < d$, we may extend each of the above parametrizations in a natural way. First, we fix the action of the permutation group on Σ by ordering the singular values

$$\sigma_1 \geq \sigma_2 \ldots \geq \sigma_d. \tag{5.18}$$

Thus, when W has rank $r < d$,

$$\sigma_{d-r+1} = \ldots = \sigma_d = 0. \tag{5.19}$$

Taking a limit of full-rank matrices in equation (5.16), we see that it is only the first r orthonormal vectors in each Q_p that affect $\mathbf{W}$ in the rank-deficient limit. Thus, in order to extend $\mathfrak{z}$ to $\mathcal{M}_r$ we must replace the action of O_d in the parametrization with the action of $\mathrm{St}_{r,d}$. Then (5.15) extends to the family of parametrizations indexed by r, $1 \leq r \leq d$

$$\mathfrak{z}_r : \mathbb{R}_+^d \times \mathrm{St}_{r,d}^{N+1} \to \mathcal{M}_r, \quad (\Lambda, Q_N, \ldots, Q_0) \to (W_N, \ldots, W_1), \tag{5.20}$$

Equations (5.16)–(5.17) continue to hold true and we see immediately that $\mathcal{M}_r$ is also a manifold. The image $\phi(\mathcal{M}_r) = \mathfrak{M}_r$, where $\mathfrak{M}_r$ was defined in equation (4.20).

We define the parametrizations $\mathfrak{x}_r$ and $\mathfrak{y}_r$ by extending (5.11) and (5.20) in an analogous manner to $r < d$.

6. Entropy of group orbits and Riemannian submersion

6.1. **Overview.** In this section, we use Riemannian submersion to develop a thermodynamic framework for the DLN. We restrict ourselves to full rank matrices. The generalization to rank r is natural, but it requires that O_d be replaced by $\mathrm{St}_{r,d}$ and a careful treatment of zero singular values.

6.2. **The entropy formula.** The parametrization (5.15)–(5.17) allows us to foliate $\mathcal{M}$ with group orbits as follows. For each $W \in \mathfrak{M}_d$ consider its preimage in $\mathcal{M}$

$$\mathcal{O}_W = \phi^{-1}(W) = \{\mathbf{W} \in \mathcal{M} \,|\, W_N \cdots W_1 = W\}. \tag{6.1}$$

We use equations (5.15)–(5.17) to find that $\mathcal{O}_W$ is an orbit of O_d^{N-1}. Indeed, the SVD of W fixes Q_0, Σ and Q_N, leaving $(Q_{N-1}, \ldots, Q_1)$ free.

We interpret this geometric picture in the language of thermodynamics. Matrices $W \in \mathfrak{M}_d$ downstairs are macrostates, or *observables*, whereas matrices $\mathbf{W} \in \mathcal{M}$ upstairs are microstates. Conceptually, the entropy enumerates the number of microstates corresponding to a given macrostate, with the enumeration given by

Boltzmann's formula, $S = \log(\#)$, where $(\#)$ denote the number of microstates. In our setting, since $\mathcal{M}$ inherits a metric from its embedding in $\mathbb{M}_d^N$, the number of microstates associated to $W \in \mathfrak{M}_d$ is the volume of the group orbit $\mathcal{O}_W$ with respect to this metric. This reduces the computation of the entropy to an evaluation of a matrix integral and we find

Theorem 10 (Menon, Yu [34]). *For each $W \in (\mathfrak{M}_d, g^N)$ with SVD given by* (5.17)

$$\operatorname{vol}(\mathcal{O}_W) = c_d^{N-1} \sqrt{\frac{\operatorname{van}(\Sigma^2)}{\operatorname{van}(\Sigma^{\frac{2}{N}})}} = c_d^{N-1} \prod_{1 \le i < j \le d} \sqrt{\frac{\sigma_i^2 - \sigma_j^2}{\sigma_i^{\frac{2}{N}} - \sigma_j^{\frac{2}{N}}}} \tag{6.2}$$

where c_d is the volume of the orthogonal group O_d with the standard Haar measure.

Here $\operatorname{van}(A)$ denoted the Vandermonde determinant associated to the matrix $A = \operatorname{diag}(a_1, \dots, a_d)$ as follows:

$$\operatorname{van}(A) = \begin{vmatrix} 1 & 1 & & 1 \\ a_1 & a_2 & & a_d \\ a_1^2 & a_2^2 & \cdots & a_d^2 \\ \vdots & & & \vdots \\ a_1^{d-1} & a_2^{d-1} & \cdots & a_d^{d-1} \end{vmatrix} = \prod_{1 \le j < k \le d} (a_k - a_j). \tag{6.3}$$

Similar factors appear in a different way in the following

Theorem 11 (Cohen, Menon, Veraszto [14]). *The volume form on $(\mathfrak{M}_d, g^N)$ is*

$$\sqrt{\det g^N}\, dW = \det(\Sigma^2)^{\frac{N-1}{2N}} \operatorname{van}(\Sigma^{\frac{2}{N}})\, d\Sigma\, dQ_0\, dQ_N. \tag{6.4}$$

Both theorems follow from explicit computations of metrics, but the two theorems correspond to different metrics. Theorem 11 concerns the metric downstairs and follows easily from Lemma 1. On the other hand, Theorem 10 concerns the metric upstairs $(\mathcal{M}, \iota)$ as explained in Section 8. It follows from a computation of the pullback metric $\mathfrak{z}^\sharp \iota$ on the parameter space $\mathbb{R}_+^d \times O_d^{N+1}$.

6.3. Equilibrium thermodynamics.

Definition 12. Given $W \in (\mathfrak{M}_d, g^N)$ we define the Boltzmann entropy

$$S(W) = \log \operatorname{vol}(\mathcal{O}_W). \tag{6.5}$$

At the inverse temperature $\beta \in (0, \infty)$ we define the *free energy*

$$F_\beta(W) = E(W) - \frac{1}{\beta} S(W). \tag{6.6}$$

The introduction of the entropy allows us to extend gradient descent of the energy $E(W)$ to gradient descent of the free energy on $(\mathfrak{M}_d, g^N)$

$$\dot{W} = -\operatorname{grad}_{g^N} F_\beta(W). \tag{6.7}$$

The geometry of group orbits also allows us to introduce natural microscopic dynamics that capture the notion of 'fluctuations in the gauge'. In Section 10 we extend equation (6.7) to Riemannian Langevin Equations of the form

$$dW_t = -\operatorname{grad}_{g^N} F_\beta(W_t)\, dt + dB_t^{\beta, g^N}, \tag{6.8}$$

where B_t^{β, g^N} is Brownian motion at inverse temperature β on $(\mathfrak{M}_d, g^N)$.

The inclusion of the entropy provides a selection principle when $E(W)$ is degenerate. We expect that a typical solution to equation (6.7) is attracted to the minimizing set

$$\mathcal{S}_\beta = \operatorname{argmin}_{W\in\mathfrak{M}_d} F_\beta(W). \tag{6.9}$$

At present, we do not understand the structure of this set completely even for the simple energy corresponding to matrix completion (see Section 14). In particular, it is natural to conjecture that $\mathcal{S}_\beta$ consists of a single point $A_\beta \in \mathfrak{M}_d$ and that $\lim_{\beta\to\infty} A_\beta$ is selected by the training dynamics due to noise introduced through round-off errors. However, we do not have rigorous results in this direction.

6.4. **Riemannian submersion.** The metric g^N was first introduced in [5, §3] using an analysis of $\mathfrak{M}_r$. The parametrization (5.15)-(5.16) provides a simple geometric explanation for the origin of this metric.

Theorem 13 (Menon, Yu [34])**.** *For each rank r, $1 \le r \le d$, the metric g^N on $\mathfrak{M}_r$ is obtained from the map $\phi : \mathcal{M}_r \to \mathfrak{M}_r$ by Riemannian submersion.*

The main ideas in the proof of this theorem are as follows:

(1) Each manifold $\mathcal{M}_r$ is Riemannian with the metric ι induced by its embedding in $\mathbb{M}_d^N$. We present the essential ideas in the computation of this metric in Theorem 14 in Section 8.
(2) The O_d^{N-1} action discussed in Section 5 is an isometry of $(\mathcal{M}, \iota)$. This idea is implicit in the entropy formula and is made explicit in Theorem 14.
(3) The differential $\phi_* : T\mathcal{M}_r \to T\mathfrak{M}_r$, its kernel and orthogonal complement may be computed explicitly using Theorem 14. We then find an orthonormal basis of $(\operatorname{Ker}\phi_*)^\perp$ and show that the projection of this basis to $T\mathfrak{M}_r$ provides an orthonormal basis for $(T_W\mathfrak{M}_r, g^N)$.

7. Proof of Theorem 1–Theorem 3

In this section, we provide the essential steps in the proofs of Theorem 1–Theorem 3. The proofs involve direct matrix calculations. We typically assume the Einstein convention and sum over repeated indices.

7.1. Balanced varieties.

7.1.1. *Overparametrization and equation* (3.6)*.* Fix p, $1 \le p \le N$. We must compute the matrix $\nabla_{W_p} E(W)$ with $W = \phi(\mathbf{W})$. By the chain rule, this is the matrix with (j,k) entries

$$\frac{\partial E(W)}{\partial W_{p,jk}} = \frac{\partial E(W)}{\partial W_{lm}}\frac{\partial W_{lm}}{\partial W_{p,jk}} := E'(W)_{lm}\frac{\partial W_{lm}}{\partial W_{p,jk}} \tag{7.1}$$

Clearly, the form of E is not that important, what really matters is the application of the chain rule to the product $W = W_N W_{N_1}\cdots W_1$. Let us write this as

$$W_{lm} = W_{N,li_{N-1}} W_{N-1,i_{N-1}i_{N-2}}\cdots W_{p,i_p i_{p-1}}\cdots W_{1,i_1 m}. \tag{7.2}$$

Therefore,

$$\frac{\partial W_{lm}}{\partial W_{p,jk}} = W_{N,li_{N-1}} W_{N-1,i_{N-1}i_{N-2}}\cdots \delta_{i_p j}\delta_{i_{p-1}k}\cdots W_{1,i_1 m}, \tag{7.3}$$

which may be rewritten as

$$\frac{\partial W_{lm}}{\partial W_{p,jk}} = (W_N \cdots W_{p+1})_{lj}(W_{p-1} \cdots W_1)_{km}. \tag{7.4}$$

It then follows from equation (7.1) and (7.4) that

$$\dot{W}_p = -\nabla_{W_p} E(W) = -(W_N \cdots W_{p+1})^T E'(W)(W_{p-1} \cdots W_1)^T. \tag{7.5}$$

7.1.2. *Proof of Theorem 1.* By the product rule we have

$$\frac{d}{dt} W_p W_p^T = \dot{W}_p W_p^T + W_p \dot{W}_p^T. \tag{7.6}$$

We now find from equation (7.5) that

$$\dot{W}_p W_p^T = -(W_N \cdots W_{p+1})^T E'(W)(W_p W_{p-1} \cdots W_1)^T, \tag{7.7}$$

where we have observed that

$$(W_{p-1} \cdots W_1)^T W_p^T = (W_p W_{p-1} \cdots W_1)^T.$$

In a similar manner, using equation (7.5) with p replaced by $p+1$ we also have

$$W_{p+1}^T \dot{W}_{p+1} = -(W_N \cdots W_{p+1})^T E'(W)(W_p W_{p-1} \cdots W_1)^T. \tag{7.8}$$

Part (a) of Theorem 1 now follows immediately from the identity.

$$\dot{W}_p W_p^T = W_{p+1}^T \dot{W}_{p+1}. \tag{7.9}$$

Part (b) of Theorem 1 is similar. The factors A_p and B_p defined in equation (4.10) appear naturally in the calculation. We use the chain rule to see that

$$\dot{W} = \dot{W}_N W_{N-1} \cdots W_1 + W_N \dot{W}_{N-1} W_{N-2} \cdots W_1 + W_N \cdots W_2 \dot{W}_1. \tag{7.10}$$

We then apply equation (7.5) to each term, observing that in each case the product above has factors that complement the products $(W_N \cdots W_{p+1})^T$ and $(W_{p-1} \cdots W_1)^T$. For example, the first term simplifies to

$$\dot{W}_N W_{N-1} \cdots W_1 = -E'(W)(W_{N-1} \cdots W_1)^T W_{N-1} \cdots W_1 = -E'(W) B_{N-1}^T B_{N-1}.$$

The general term is given by

$$W_N \cdots \dot{W}_p \cdots W_1 = -A_{p+1} A_{p+1}^T E'(W) B_{p-1}^T B_{p-1}.$$

We sum over the depth index p from 1 to N to obtain equation (4.9).

7.1.3. *Proof of Theorem 2.* Theorem 2 is a specialization of Theorem 1(b) to the balanced variety. Since

$$W_{p+1}^T W_{p+1} = W_p W_p^T$$

on the balanced variety we may simplify the prefactors $A_{p+1} A_{p+1}^T$ and $B_{p-1} B_{p-1}^T$ appearing in equation (4.9). It follows immediately from equation (4.9) and the parametrization (5.16) that

$$A_p = W_N \cdots W_p = Q_N \Sigma^{\frac{N-p}{n}} Q_{p-1}^T, \quad B_p = W_p \cdots W_1 = Q_p \Sigma^{\frac{p}{n}} Q_0^T. \tag{7.11}$$

We then have

$$A_p A_p^T = Q_N \Sigma^{\frac{2(N-p)}{n}} Q_N^T = (W W^T)^{\frac{N-p}{n}}, \quad B_p^T B_p = Q_0 \Sigma^{\frac{2p}{n}} Q_0^T = (W^T W)^{\frac{p}{n}}.$$

We substitute in equation (4.9) to complete the proof.

7.2. **The Riemannian manifold $(\mathcal{M}, g^N)$ and Theorem 3.** We hold N and W fixed in this subsection. To simplify notation, we refer to $\mathcal{A}_{N,W}$ as $\mathcal{A}$.

The main observation in this section is that $\mathcal{A}$ is diagonalized in singular value coordinates. Precisely, let $\{u_1, \cdots, u_d\}$ and $\{v_1, \cdots, v_d\}$ be the column vectors of Q_N and Q_0 respectively. Then we may write the SVD (5.17) as

$$W = Q_N \Sigma Q_0^T = \sum_{j=1}^{d} \sigma_j u_j v_j^T. \tag{7.12}$$

Lemma 1. *The operator $\mathcal{A} : T_W \mathfrak{M}_d \to T_W \mathfrak{M}_d$ is symmetric and positive definite with respect to the Frobenius inner-product. It has the spectral decomposition*

$$\mathcal{A}\, u_k v_l^T = \frac{\sigma_k^2 - \sigma_l^2}{\sigma_k^{\frac{2}{N}} - \sigma_l^{\frac{2}{N}}}\, u_k v_l^T, \quad 1 \le k, l \le d, \tag{7.13}$$

when $k \neq l$ and

$$\mathcal{A}\, u_k v_k^T = N \sigma_k^{2-\frac{2}{N}}\, u_k v_k^T, \quad 1 \le k \le d. \tag{7.14}$$

Proof. We have

$$W W^T = \sum_{i=1}^{d} \sigma_i^2 u_i u_i^T, \quad W^T W = \sum_{j=1}^{d} \sigma_j^2 v_j v_j^T.$$

We then find from equation (4.14) that for each pair k,l

$$\mathcal{A}\, u_k v_l^T = \sum_{p=1}^{N} (W W^T)^{\frac{N-p}{N}}\, u_k v_l^T\, (W^T W)^{\frac{p-1}{N}} = \left(\sum_{p=1}^{N} \sigma_k^{\frac{2(N-p)}{N}} \sigma_l^{\frac{2(p-1)}{N}} \right) u_k v_l^T,$$

yielding (7.13) and (7.14).

Observe that distinct eigenvectors are orthogonal with respect to the Frobenius metric. Indeed, for each pair of indices $\{k, l\}$ and $\{m, n\}$

$$\mathrm{Tr}((u_k v_l^T)^T u_m v_n) = \delta_{km} \delta_{ln},$$

which vanishes unless the pairs agree. It then follows from the eigendecomposition that $\mathcal{A}$ is a symmetric and positive definite operator from $T_W \mathfrak{M}_d \to T_W \mathfrak{M}_d$. □

We may represent an arbitrary matrix $Z \in T_W \mathfrak{M}_d$ in this basis as a sum $Z = Z_{kl} u_k v_l^T$. Then we use equation (4.16) and Lemma 1 to obtain

$$g^N(Z, Z) = N \sum_{1 \le k \le d} \sigma_k^{2(1-\frac{1}{N})} Z_{kk}^2 + \sum_{1 \le k, l \le d, k \neq l} \frac{\sigma_k^{\frac{2}{N}} - \sigma_l^{\frac{2}{N}}}{\sigma_k^2 - \sigma_l^2} Z_{kl}^2. \tag{7.15}$$

Theorem 3 follows directly from the explicit diagonalization of the metric.

8. Embedding and the metric on $\mathcal{M}$

The balanced manifold $\mathcal{M}$ is a Riemannian manifold since it is locally embedded in $\mathbb{M}_d^N$ and inherits the Frobenius metric, denoted ι, from $\mathbb{M}_d^N$. In order to use our parametrization $\mathfrak{Z}$ to understand the Riemannian manifold $(\mathcal{M}, \iota)$, we must pull back the metric ι onto the parameter space $\mathbb{R}_+^d \times O_d^{N-1}$. In this section, we compute an orthonormal basis for $(T_{\mathbf{W}} \mathcal{M}, \iota)$ by computing an orthonormal basis for the pullback metric $\mathfrak{Z}^\sharp \iota$. This is the technical core of [34].

8.1. **The tangent space $T_{\mathbf{W}}\mathcal{M}$.** First, at each point $\mathbf{W} \in \mathcal{M} \subset \mathbb{M}_d^N$, we compute the tangent space $T_{\mathbf{W}}\mathcal{M}$ by differentiating the parametrization (5.16) as follows. The tangent space at the identity to the orthogonal group O_d is the space of anti-symmetric matrices, denoted $\mathbb{A}_d$. Thus, if the singular values Σ of W are distinct, the parametrization is smooth and for each $(\theta, \mathbf{A}) \in \mathbb{R}^d \times \mathbb{A}_d^{N+1}$ we may define a smooth curve in $\mathbf{W}(t) \in \mathcal{M}$ using

$$\Lambda(t) = \Lambda + t\theta, \quad Q_p(t) = e^{tA_p} Q_p, \quad \mathbf{W}(t) = \mathfrak{z}(\Lambda(t), \mathbf{Q}(t)).$$

where θ is the diagonal matrix $\mathrm{diag}(\theta_1, \ldots, \theta_d)$ and $\Lambda(t) = \Sigma(t)^{1/N}$.

Then we obtain a tangent vector in $T_{\mathbf{W}}\mathcal{M}$ by differentiating in t.[3]

$$\left.\frac{d\mathbf{W}(t)}{dt}\right|_{t=0} = D\mathfrak{z}(\mathbf{W})\,(\theta, \mathbf{A})\,. \tag{8.1}$$

Explicitly, the p-th matrix in $D\mathfrak{z}(\mathbf{W})\,(\theta, \mathbf{A})$ is

$$D\mathfrak{z}(\mathbf{W})\,(\theta, \mathbf{A})_p = A_p W_p + Q_p \theta Q_{p-1}^T - W_p A_{p-1}, \quad 1 \le p \le N. \tag{8.2}$$

8.2. **Computing $\mathfrak{z}^\sharp\iota$ in the standard basis.** Since the metric on $\mathcal{M}$ is inherited from its embedding in $\mathbb{M}_d^N$, the length of each vector

$$\mathbf{V} := D\mathfrak{z}(\mathbf{W})\,(\theta, \mathbf{A}) \in T_{\mathbf{W}}\mathcal{M} \tag{8.3}$$

is given by the Frobenius norm

$$\|\mathbf{V}\|^2 = \sum_{k=1}^{N} \mathrm{Tr}(V_p^T V_p), \quad \mathbf{V} = (V_N, \ldots, V_1). \tag{8.4}$$

Similarly, the inner product between two vectors $\mathbf{V}^{(i)} \in T_{\mathbf{W}}\mathcal{M}$, $i = 1, 2$ is given by

$$\langle \mathbf{V}^{(1)}, \mathbf{V}^{(2)} \rangle = \sum_{k=1}^{N} \mathrm{Tr}\left((V_p^{(1)})^T, V_p^{(2)}\right). \tag{8.5}$$

By linearity, we may reduce the computation of the pullback metric to inner products between the action of $D\mathfrak{z}$ on the basis vectors on $\mathbb{R}^d \times \mathbb{A}_d^{N+1}$. We arrange these basis vectors as follows. First, let $\{e_i\}_{i=1}^d$ be the standard basis on $\mathbb{R}^d$ and define the image of diagonal matrices

$$\mathbf{l}_i = D\mathfrak{z}(\mathbf{W})\,(e_i, \mathbf{0})\,, \quad i = 1, \ldots, d. \tag{8.6}$$

Next, form the standard basis for $\mathbb{A}_d$

$$\alpha_{ij} = \frac{1}{\sqrt{2}}(e_i e_j^T - e_j e_i^T), \quad 1 \le i < j \le d, \tag{8.7}$$

and for each $0 \le p \le N$ use it to define a tangent vector in $T_{\mathbf{W}}\mathcal{M}$ by setting

$$\mathbf{A}_{kl}^p = (\ldots, 0, \alpha_{kl}^p, 0, \ldots), \quad \mathbf{a}_{kl}^p = D\mathfrak{z}(\mathbf{W})\,(0, \mathbf{A}_{kl}^p)\,, \quad 1 \le k < l \le d. \tag{8.8}$$

Here $\alpha_{kl}^p = \alpha_{kl}$ and we have used the superscript p to index depth p.

We may now express the pullback metric $\mathfrak{z}^\sharp\iota$ in the standard basis on $\mathbb{R}^d \times \mathbb{A}_d^{N+1}$ by computing the Frobenius inner product between the d tangent vectors $\mathbf{l}_i$ and the $(N+1)d(d-1)/2$ tangent vectors $\mathbf{a}_{kl}^p$. Let I_d denote the $d \times d$ identity matrix. We then have

[3] Note that the term 'vector' here is used to mean 'vector in the sense of vector space'. Each vector in $T_{\mathbf{W}}\mathcal{M}$ 'sits in' $\mathbb{M}_d^N$.

Lemma 2. *The standard basis on* $\mathbb{R}^d \times \mathbb{A}_d^{N+1}$ *may be ordered such that the pullback metric* $\mathfrak{Z}^\sharp l$ *has the block diagonal structure*

$$h = \begin{pmatrix} NI_d & & & \\ & h_a^{1,2} & & \\ & & \ddots & \\ & & & h_a^{d,d-1} \end{pmatrix}, \tag{8.9}$$

where $h_a^{k,l}$ *is the* $(N+1)\times(N+1)$ *symmetric tridiagonal matrix*

$$h_a^{k,l} = \begin{pmatrix} \frac{1}{2}(\lambda_k^2+\lambda_l^2) & -\lambda_k\lambda_l & & & & \\ -\lambda_k\lambda_l & \lambda_k^2+\lambda_l^2 & -\lambda_k\lambda_l & & & \\ & -\lambda_k\lambda_l & \lambda_k^2+\lambda_l^2 & -\lambda_k\lambda_l & & \\ & & \ddots & \ddots & & \\ & & -\lambda_k\lambda_l & \lambda_k^2+\lambda_l^2 & -\lambda_k\lambda_l & \\ & & & -\lambda_k\lambda_l & \frac{1}{2}(\lambda_k^2+\lambda_l^2) \end{pmatrix}. \tag{8.10}$$

There are $d(d-1)/2$ blocks, each indexed by a basis matrix $\alpha_{kl} \in \mathbb{A}_d$. This ordering is less intuitive than arranging the metric by depth into $N+1$ blocks, each of size $d(d-1)/2 \times d(d-1)/2$. However, this structure appears naturally in the computation of inner products. It reflects a subtle non-local coupling along the depth of the network that is due to balancedness.

Sketch of the proof. First, rewrite equation (8.2) as follows

$$Q_p^T D\mathfrak{Z}(\mathbf{W})\,(\theta,\mathbf{A})_p\, Q_{p-1} = Q_p^T A_p Q_p \Lambda + \theta - \Lambda Q_{p-1}^T A_{p-1} Q_{p-1}. \tag{8.11}$$

Now observe that the linear transformation of the p-th tangent space $\mathbb{A}_d$ defined by $A_p \mapsto Q_p^T A_p Q_p := B_p$ is an isometry for the Frobenius norm. We see immediately that when $\mathbf{A} = \mathbf{0}$ the image

$$Q_p^T D\mathfrak{Z}(\mathbf{W})\,(\theta,\mathbf{0})_p\, Q_{p-1} = \theta, \tag{8.12}$$

is a collection of N diagonal matrices. On the other hand, when $\theta = 0$,

$$Q_p^T D\mathfrak{Z}(\mathbf{W})\,(0,\mathbf{A})_p\, Q_{p-1} = B_p\Lambda - \Lambda B_{p-1} \tag{8.13}$$

always vanishes on the diagonal. Thus, the inner-products

$$\langle D\mathfrak{Z}(\mathbf{W})\,(\theta,\mathbf{0}),\, D\mathfrak{Z}(\mathbf{W})\,(0,\mathbf{A})\rangle = 0. \tag{8.14}$$

This gives rise to the block NI_d in the upper left corner.

The only non-zero inner products concern $\mathbf{V}_i = D\mathfrak{Z}(0,\mathbf{A}_i)$, $i = 1,2$ where

$$\mathbf{A}_1 = (\ldots,0,\alpha^{p+1},0,\ldots), \quad \mathbf{A}_2 = (\ldots,0,\alpha^{p},0,\ldots), \tag{8.15}$$

have overlapping indices $p+1$ and p. In this case, we find that the only non-zero inner product corresponds to $\alpha^{p+1} = \alpha^p$. The main step involves the identity

$$\mathrm{Tr}(\alpha_{ij}\Lambda\alpha_{kl}\Lambda) = \lambda_i\lambda_j\left(\delta_{il}\delta_{jk} - \delta_{ik}\delta_{jl}\right). \tag{8.16}$$

□

8.3. **An orthonormal basis for $\mathfrak{z}^\sharp\iota$.** We may compute an orthonormal basis for $T_{\mathbf{W}}\mathcal{M}$ by diagonalizing the block tridiagonal matrix of Lemma 2. The diagonalization procedure uses the theory of Jacobi matrices, in particular the fact that the eigenbasis of each block $h_a^{k,l}$ may be computed using Chebyshev polynomials.

The results of this approach may be summarized as follows. Assume that $\mathbf{W} = \mathfrak{z}(\Lambda, Q_N, \dots, Q_0)$ and let us denote the columns of each Q_p by

$$Q_p = \begin{pmatrix} \uparrow & \uparrow & \cdots & \uparrow \\ q_{p,1} & q_{p,2} & \cdots & q_{p,d} \\ \downarrow & \downarrow & \cdots & \downarrow \end{pmatrix}. \tag{8.17}$$

We will define a collection of vectors denoted by

$$\mathbf{l}^k = (l_N^k, \cdots, l_1^k), \quad 1 \le k \le d; \tag{8.18}$$

$$\mathbf{u}^{k,l,p} = (u_N^{k,l,p}, \cdots, u_1^{k,l,p}), \quad 1 \le k < l \le d, \quad 0 \le p \le N. \tag{8.19}$$

There are d vectors of type $\mathbf{l}$ and $(N+1)d(d-1)/2$ vectors of type $\mathbf{u}$. The coordinates of these vectors are as follows. First, for the $\mathbf{l}$ vectors

$$l_s^k = \frac{1}{\sqrt{N}} q_{s,k} q_{s-1,k}^T, \quad 1 \le s \le N. \tag{8.20}$$

We next consider the $\mathbf{u}$ vectors for $1 \le p \le N-1$. This range for p corresponds to the overparametrization freedom O_d^{N-1}. We define

$$u_s^{k,l,p} = a^{k,l,p,s} q_{s,k} q_{s-1,l}^T + a^{l,k,p,s} q_{s,l} q_{s-1,k}^T, \quad 1 \le s \le N. \tag{8.21}$$

Here we have denoted for brevity

$$a^{k,l,p,s} = \sqrt{\frac{1}{N\left(\lambda_k^2 + \lambda_l^2 - 2\lambda_k\lambda_l \cos\frac{p\pi}{N}\right)}} \left(\lambda_k \sin\frac{(s-1)p\pi}{N} - \lambda_l \sin\frac{sp\pi}{N}\right). \tag{8.22}$$

Finally, we consider the action of the matrices Q_0 and Q_N that generate the SVD of W. First, for $p = 0$ define

$$u_s^{k,l,0} = \sqrt{\frac{\lambda_k^2 - \lambda_l^2}{\lambda_k^{2N} - \lambda_l^{2N}}} \lambda_k^{s-1} \lambda_l^{N-s} q_{s,l} q_{s-1,k}^T. \tag{8.23}$$

Similarly, define for $p = N$

$$u_s^{k,l,N} = \sqrt{\frac{\lambda_k^2 - \lambda_l^2}{\lambda_k^{2N} - \lambda_l^{2N}}} \lambda_k^{N-s} \lambda_l^{s-1} q_{s,k} q_{s-1,l}^T. \tag{8.24}$$

We then have

Theorem 14 (Menon,Yu [34])**.** *The vectors $(\mathbf{l}, \mathbf{u})$ defined in equations (8.20)–(8.24) form an orthonormal basis for $(T_{\mathbf{W}}\mathcal{M}, \iota)$.*

Theorem 14 is the main tool in the proofs of Theorem 10 and Theorem 13. Let us first explain how Theorem 13 is obtained from it.

In order to show that the metric g^N on $\mathfrak{M}_d$ is obtained by Riemannian submersion of $\mathcal{M}_r$ under ϕ we must compute $\operatorname{Ker}\phi_*$ and find an isometry between $(\operatorname{Ker}\phi_*)^\perp$ and $T_W\mathfrak{M}_d$. The kernel $\operatorname{Ker}\phi_*$ corresponds to the group action of O_d^{N-1} and it is the span $\mathbf{u}^{k,l,p}$ for $1 \le p \le N-1$. Thus, $(\operatorname{Ker}\phi_*)^\perp$ is spanned by the vectors $\mathbf{l}$, $\mathbf{u}^{k,l,0}$, and $\mathbf{u}^{k,l,N}$. We see immediately that these correspond exactly to the eigendecomposition of A computed in Lemma 1. This establishes that $(\mathfrak{M}_d, g^N)$ is obtained by Riemannian submersion from $(\mathcal{M}, \iota)$.

Theorem 10 is obtained as follows. We must compute the volume of the O_d^{N-1} orbit $\phi^{-1}(W) \subset \mathcal{M}$. We work in local coordinates given by our parametrization, and observe that fixing W corresponds to fixing Λ, Q_0 and Q_N. The pullback metric $\mathfrak{z}^\sharp \iota$ is invariant under the O_d^{N-1} action. This reduces the computation of the group volume to the computation of the determinant of $\mathfrak{z}^\sharp \iota$ restricted to the subspace spanned by $\mathbf{u}^{k,l,p}$, $1 \leq p \leq N-1$. This determinant may be computed explicitly using Theorem 14, yielding Theorem 10.

The proof for the rank-deficient case requires more care, but the above arguments captures the essence of these theorems. Further details may be found in [34].

9. Cartoons

The way dynamicists actually work is by drawing pictures that capture different forms of geometric intuition. The purpose of this section is to illustrate this way of thinking, complementing several theorems with an impressionistic visualization of the theorems. A good picture can tell us what a theorem means; or better yet, help us guess what theorems we should be proving.

The DLN is an attractive model because it provides us with a rich set of pictures that suggest new questions for study. We organize our cartoons visually to bring out these aspects, focusing on the following themes:

(1) *Riemannian submersion.* We often organize images into a $\mathbf{W}$-space upstairs and W-space downstairs. Riemannian submersion arises in several areas of analysis, especially mass transportation, and it is helpful to study the effects of overparametrization in analogy with these areas.

(2) *Conic sections.* The equations (4.2) that define the balanced varieties are *quadratic* equations. Thus, when drawing a two or three dimensional sketch, we caricature the balanced varieties by conic sections. This caricature then immediately focuses our attention on the balanced variety $\mathcal{M}_{\mathbf{0}}$.

(3) *Foliation by rank.* The balanced variety is itself foliated by rank. We have only studied the case $d = r$ in depth, but the rank-deficient case $r < d$ is of great interest. This is harder to visualize, so we have demonstrated the effect of rank within the zero energy set for an example discussed in equations (14.1)–(14.4).

(4) *Foliation by group action.* The parametrizations by the polar factorization and SVD provide a different way of exploring balanced manifolds with group action. The computation of the pullback metric illustrates the subtle nature of the coupling by depth along the network. We visualize this by slicing the balanced manifold.

(5) *Mean curvature is an Itô correction.* We often use the fact that the mean curvature from of an embedded manifold is the 'deterministic push' that arises from isotropic tangential stochastic fluctuations. This cartoon allows us to introduce Riemannian Langevin equations for the DLN in an intuitive manner. This can be illustrated rather simply and then generalized to more sophisticated foliations, such as in Section 11.1.

10. The Riemannian Langevin equation (RLE)

10.1. **Overview.** The next three sections address the following question: can we use the DLN to develop the thermodynamics of deep learning?

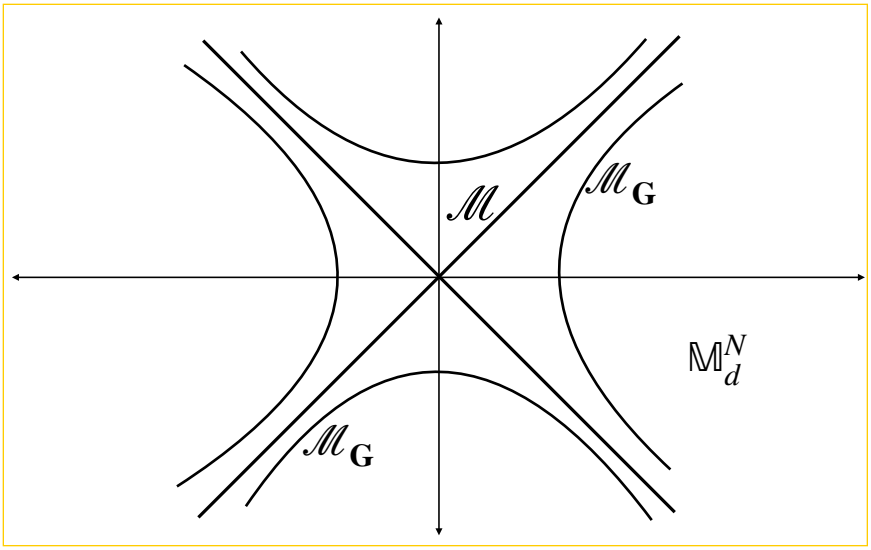

FIGURE 9.1. The foliation of $\mathbb{M}_d^N$ by the balanced varieties $\mathcal{M}_{\mathbf{G}}$ may be visualized as a foliation by conic sections (see equation (5.10)).

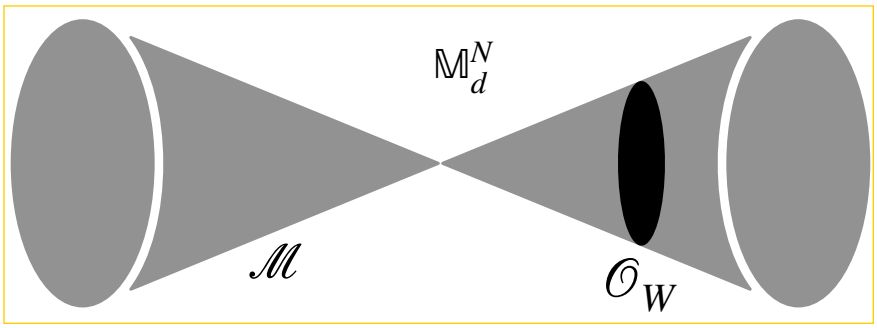

FIGURE 9.2. In comparison with Figure 9.1, we now blow up the balanced manifold $\mathcal{M}$ and visualize the foliation into group orbits $\mathcal{O}_W$ by slicing $\mathcal{M}$.

In the spirit of this article, such a framework must be rooted in the geometry of overparametrization. We have seen that the analysis of the DLN leads naturally to the Riemannian manifolds $(\mathcal{M}_r, \iota)$ and $(\mathfrak{M}_r, g^N)$. Further, we have used this geometry to define a Boltzmann entropy as well as the gradient descent of free energy (see Section 6). In the physical analogy with thermodynamics, what we are now seeking is an explicit understanding of microscopic fluctuations and Gibbs measures. Our task, therefore, is to develop the underlying 'statistical physics' that corresponds to the entropy formula, based on strictly geometric foundations.

In this section, we introduce the Riemannian Langevin Equation (RLE) as the natural geometric model for fluctuations. Our argument is by mathematical analogy

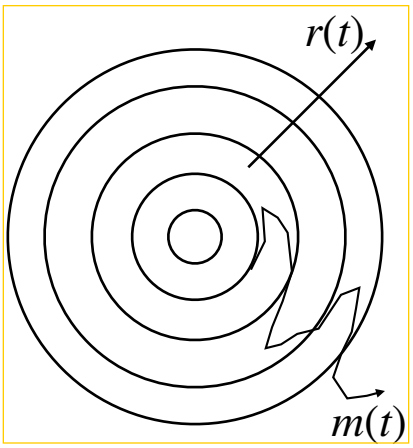

FIGURE 9.3. Motion by (minus one half) curvature arising from tangential Brownian fluctuations as discussed in Section 11.2.

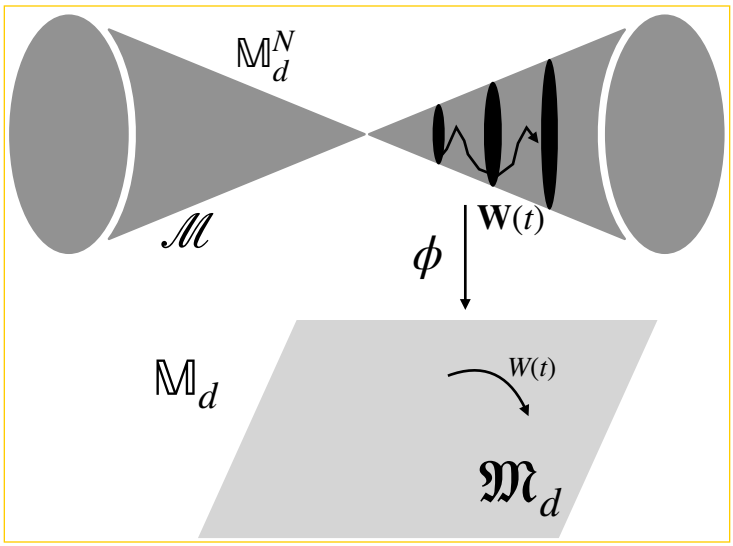

FIGURE 9.4. Riemannian submersion $\phi : \mathcal{M} \to \mathfrak{M}_d$ and the dynamics upstairs and downstairs. In this image, we illustrate the RLE with stochastic dynamics upstairs and deterministic gradient descent of free energy downstairs. See equations (12.6) and (12.7).

and it takes the following form. First, we review the Langevin equation on $\mathbb{R}^d$. We then explain how the RLE is its natural extension to Riemannian manifolds. This is followed by the description of an important example from random matrix theory (Dyson Brownian motion). We then abstract the underlying structure combining Riemannian submersion with the theory of Brownian motion on manifolds. Finally, in Section 12.2, we describe the explicit nature of the RLE in the DLN.

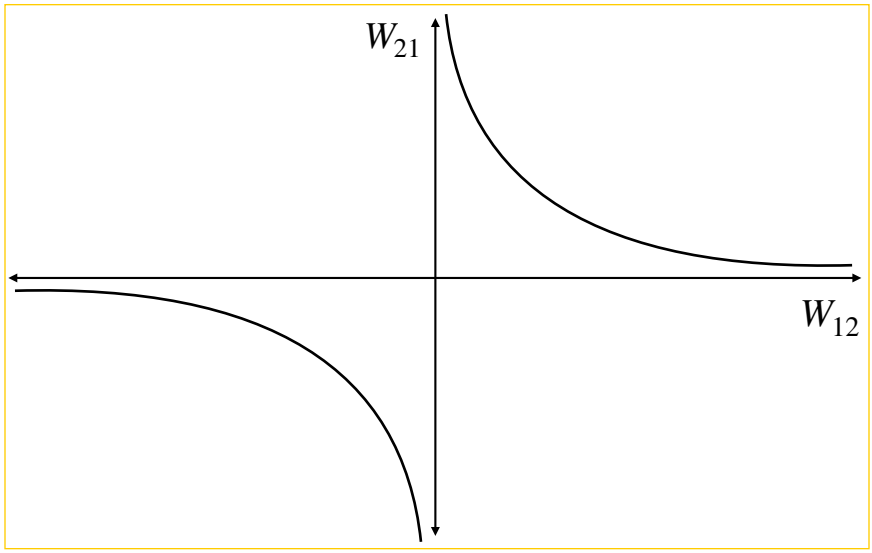

FIGURE 9.5. A rank-one variety within the zero energy set for matrix completion. See equations (14.3)–(14.4).

The correspondence between the RLE for the DLN and Dyson Brownian motion is a powerful tool. It provides a way to use ideas from random matrix theory to analyze training dynamics of the DLN. It immediately leads to open questions of intrinsic mathematical interest which are discussed in Section 14. Algorithmically, the RLE replaces the problem of minimizing the free energy with Gibbs sampling.

The RLE is a generalization of the Langevin equation based on the intrinsic Riemannian geometry. We introduce it here as a phenomenological model that incorporates the effect of noise, revealing the interplay of noise and curvature. Whether a mathematical theory works in pratice is a different matter altogether. Our RLE loosely corresponds to how noise due to round-off errors may interact with the geometry of overparametrization. It does not account for the many different ways in which noise actually arises in training dynamics (discretization, batch processing, random initialization,...). Nevertheless, we hope the structure of the RLE will provide a useful framework for practitioners, since it is a reference model which is a natural stochastic analogue of gradient descent (though it is *not* stochastic gradient descent (SGD) corresponding to batch processing).

10.2. **The Langevin equation.** First, let us recall the Langevin equation on $\mathbb{R}^d$. Assume given $\beta > 0$ and a potential $V : \mathbb{R}^d \to \mathbb{R}$. The Gibbs measure μ_β with respect to this potential is the probability measure with density

$$\rho_\beta(x) = \frac{e^{-\beta V(x)}}{Z_\beta}, \quad Z_\beta = \int_{\mathbb{R}^d} e^{-\beta V(x')}\, dx'. \tag{10.1}$$

The Langevin equation describes microscopic stochastic fluctuations in and out of equilibrium. It is the Itô SDE [4]

$$dx = -\nabla V(x)\,dt + \sqrt{\frac{2}{\beta}}\,dW \tag{10.2}$$

where W_t denotes the standard Wiener process in $\mathbb{R}^d$.

Assume the law of x_t has a density $\rho(t,\cdot)$ and assume that the law of x_0 is given. The evolution of the probability density ρ is then given by the Fokker-Planck (or forward Kolmogorov) equation

$$\partial_t \rho = \frac{1}{\beta}\triangle\rho + \nabla\cdot(\rho\,\nabla V(x_t))\,. \tag{10.3}$$

Under suitable assumptions on V, $\lim_{t\to\infty}\rho(t,\cdot)$ is the Gibbs density ρ_β defined in equation (10.1).

10.3. **RLE: general principles.** The Langevin equation and the Fokker-Planck equation have been widely studied in mathematical physics. Our interest here lies in the generalization of these equation to Riemannian manifolds. The main issue is to understand how to replace the noise in equation (10.2).

Assume given a Riemannian manifold $(\mathcal{N}, h)$. We define the Laplace-Beltrami operator $\triangle_h$ by its action on a smooth function $f:\mathcal{N}\to\mathbb{R}$ in coordinates by

$$\triangle_h f = \frac{1}{\sqrt{|h|}}\partial_{x^i}\left(\sqrt{|h|}h^{ij}\partial_{x^j}f\right), \tag{10.4}$$

where $|h|$ denotes the determinant of h_{ij} and h^{ij} denotes its inverse. Brownian motion on $(\mathcal{N}, h)$ at inverse temperature $\beta>0$ is a diffusion on $\mathcal{N}$ whose generator is $\frac{1}{\beta}\triangle_h$. While the parameter β can be included within the metric, we will include it separately, since there are many situations where it is necessary to hold h fixed and rescale the strength of the noise.

The abstract characterization of Brownian motion as a diffusion process with a generator is powerful. However, in practice (for example, for simulation) we seek 'hands-on' constructions of Brownian motion as the limit of suitable random walks. There are several such constructions of Brownian motion using Stratonovich SDE (see for example [21, §3.2], [23, §V.4], [24, Thm 3]). Once Brownian motion on $(\mathcal{N}, h)$ has been constructed with SDE, it is easy to (formally) extend the Langevin equation (10.2) to the Riemannian setting.

Let $B_t^{\beta,h}$ denote Brownian motion on $(\mathcal{N}, h)$ at inverse temperature $\beta>0$ and consider the (formal) Itô SDE

$$dx = -\mathrm{grad}_h V(x)\,dt + dB^{\beta,h}. \tag{10.5}$$

This equation is only formal, because SDE on manifolds must be defined using the Stratonovich formulation to ensure coordinate independence. However, equation (10.5) makes the analogy with (10.2) transparent and one may recover the 'true' Stratonovich SDE by the inclusion of a drift term computed using the Itô-Stratonovich correction formula. The important geometric aspect for matrix manifolds is that this yields a curvature correction that is often explicitly computable.

[4] We suppress the usual subscript t (such as $dx_t = -\nabla V(x_t)\,dt + \sqrt{2/\beta}\,dW_t$) in the notation to simplify the SDE for coordinates and matrices.

11. Curvature and entropy: examples

11.1. **Dyson Brownian motion via Riemannian submersion.** Dyson Brownian motion at inverse temperature $\beta > 0$ is the interacting particle system described by the Itô SDE

$$dx_i = \sum_{j \neq i} \frac{1}{x_i - x_j}\, dt + \sqrt{\frac{2}{\beta}} dW_i, \quad 1 \leq i \leq d. \tag{11.1}$$

Here $W_t = (W_1, \ldots, W_d)_t$ denotes the standard Wiener process in $\mathbb{R}^d$. This equation has a unique strong solution when $\beta \geq 1$ that remains confined to the Weyl chamber

$$\mathcal{W}_d = \{x \in \mathbb{R}^d \,|\, x_1 < x_2 < \ldots < x_d\}. \tag{11.2}$$

In [22] we presented a geometric construction of Dyson Brownian motion using Riemannian submersion. This construction has a natural extension to the DLN, but we must first introduce some notation to explain these ideas.

Let Her_d denote the space of $d \times d$ Hermitian matrices equipped with the norm $\|M\|^2 = \mathrm{Tr}(M^* M)$ and let U_d denote the unitary group. Given $x \in \mathcal{W}_d$, let $X = \mathrm{diag}(x)$, and let $\mathcal{O}_x$ denote the isospectral orbit

$$\mathcal{O}_x = \{M \in \mathrm{Her}_d \,|\, M = UXU^*,\ U \in U_d\}. \tag{11.3}$$

We observe that the space Her_d is foliated by U_d orbits, each of which is an isospectral set, and an isospectral manifold when the elements of X are distinct. The main insight in [22] is the interplay between mean curvature, tangential noise, and the entropy that is captured in the following 'lift upstairs' of equation (11.1).

Let $M \in \mathcal{O}_x$ and let P_M and $P_M^\perp$ denote the orthonormal projections onto $T_M \mathcal{O}_x$ and $(T_M \mathcal{O}_x)^\perp$ respectively with inner-products computed according to $\|\cdot\|^2$. Let H_t denote the standard Wiener process on $(\mathrm{Her}_d, \|\cdot\|^2)$. For every $\beta > 0$ we define the Itô SDE

$$dM = P_M dH + \sqrt{\frac{2}{\beta}} P_M^\perp dH. \tag{11.4}$$

Assume that x_t is the unique strong solution to equation (11.1) for its maximal interval of existence $[0, T_{\max})$ (this is $[0, +\infty)$ when $\beta \geq 1$). We then have the following

Theorem 15 (Huang, Inauen, Menon [22])**.** (a) *The eigenvalues of M_t have the same law as the solution x_t to (11.1) for $t \in [0, T_{\max})$.*
(b) *When $\beta = +\infty$, the group orbits $\mathcal{O}_{x_t}$ evolve by motion by (minus one half) mean curvature.*

The main insight in the theorem is captured in the case $\beta = +\infty$. In this setting, the stochastic fluctuation $P_M dH$ is *tangential* to the orbit $\mathcal{O}_x$. However, the Itô correction is minus a half times the mean curvature. In particular, it is deterministic and *normal* to $\mathcal{O}_x$.

Remark 16. The entropy formula in Theorem 10 is inspired by the analogous formula for $\mathcal{O}_x$:

$$S(x) = \log \mathrm{vol}(\mathcal{O}_x) = 2 \log \mathrm{van}(x) + C_d, \tag{11.5}$$

where C_d is a universal constant. We then see that we may rewrite equation (11.1) as follows [5]

$$dx = \frac{1}{2}\nabla S(x)dt + \sqrt{\frac{2}{\beta}}dW. \tag{11.6}$$

Remark 17. The role of β in equation (11.1) reflects the standard terminology of random matrix theory. In equation (11.4), however, the parameter β reflects an anisotropic splitting of the inner-product in Her_d between the tangent space $T_M\mathcal{O}_x$ and $T_M\mathcal{O}_x^{\perp}$. For these reasons, we will introduce a separate parameter $\kappa > 0$ to describe the anisotropy, and use $\beta > 0$ for the 'true' inverse temperature when augmenting equation (11.4) with a loss function.

In order to define stochastic gradient descent by the RLE, we include a loss function in equation (11.4) as follows. Given $E : \mathcal{W} \to \mathbb{R}$, set $L(M) = E(\mathrm{eig}(M))$ and consider

$$dM = -\nabla_M L(M)\, dt + \sqrt{\frac{2}{\beta}}\left(P_M dH + \sqrt{\kappa}P_M^{\perp}dH\right). \tag{11.7}$$

The corresponding equation downstairs is

$$dx = -\nabla_x F_\beta(x)dt + \sqrt{\frac{2\kappa}{\beta}}dW, \tag{11.8}$$

where W_t is the standard Wiener process in $\mathbb{R}^d$ and

$$F_\beta(x) = E(x) - \frac{1}{\beta}S(x) \tag{11.9}$$

is the free energy. When $\kappa = 0$, we have gradient descent of free energy. In equations (11.7) and (11.8) the gradient operator ∇ is with respect to the standard inner products on Her_d and $\mathbb{R}^d$ respectively. In fact, these metrics are related by Riemannian submersion.

In what follows, we will use equation (11.7) as a model for stochastic gradient descent by the RLE of a loss function, which accounts separately for anisotropic fluctuations in the gauge group and the observable.

11.2. **The stochastic origin of motion by mean curvature.** The appearance of mean curvature in Theorem 15(b) may be understood through a simpler example. Let $m \in \mathbb{R}^d$, let $r = |m| = \sqrt{m^T m}$ denote the length of m, and let S_r^{d-1} denote the sphere of radius r in $\mathbb{R}^d$. Let P_m denote the orthonormal projection onto $T_m S_r^{d-1}$; explicitly, we have

$$P_m = I_d - \frac{mm^T}{|m|^2}.$$

Let W_t denote the standard Wiener process in $\mathbb{R}^d$ and consider the Itô SDE

$$dm = P_m\, dW. \tag{11.10}$$

This SDE is the continuum limit of a random walk which may be intuitively described as 'at each time step, take an isotropically distributed random spatial step in the tangent space'.

[5] In [22], we work with Hermitian matrices, which is the most natural setting for Dyson Brownian motions. The factors of 2 that cancel are included for consistency with the general theory of RLE.

Stochastic calculus makes it easy to capture the intuitive content of such clumsy verbal descriptions. In particular, since the SDE is in the Itô form, it gives rise to a correction in r_t. The reader is invited to apply Itô's formula to see that while m_t always evolves by *tangential* stochastic fluctuations governed by equation (11.10), the radius $r_t = |m_t|$ satisfies

$$\dot{r} = \frac{d-1}{2r}. \tag{11.11}$$

The standard normalization of the mean curvature of the sphere S_r^{d-1} is such that it has magnitude $(d-1)/r$ and points inwards. Consequently, equation (11.11) tells us that the concentric spheres S_r^{d-1} evolve by motion by minus a half times mean curvature.

Theorem 15(b) is a more sophisticated instance of the same interplay. The Itô correction for the eigenvalues of M_t are computed using the first and second-order variation formulas for eigenvalues (see [22]).

12. RLE for stochastic gradient descent

12.1. **Riemannian submersion with a group action.** In this section, we introduce Riemannian Langevin equations to model stochastic gradient descent of free energy. We formalize the insights obtained in the examples of Sections 11.1– 11.2 in a general geometric framework. We also include 'lifted' loss functions as in deep learning. These RLE provide stochastic extensions of gradient flows such as (3.4), which are consistent with the underlying geometry. The tunable parameter $\kappa > 0$ modulates the relative strength of the noise in null directions, or equivalently the gauge group, with the noise in the observable. When $\kappa = 0$, we recover Riemannian gradient descent of free energy for the observable.

The geometric assumptions are as follows. We assume given a reference Riemannian manifold $(\mathcal{M}, g)$ and a Lie group $\mathcal{G}$ of isometries. We then consider the quotient space $\mathcal{X} = \mathcal{M}/\mathcal{G}$ equipped with the metric h given via Riemannian submersion. Let ϕ: $\mathcal{M} \to \mathcal{X}$ denote the submersion map and let $\mathcal{O}_x$ denote the inverse image $\phi^{-1}(x)$ for each $x \in \mathcal{X}$. Then the following general principles hold:

(1) There is always a natural Boltzmann entropy $S(x) = \log \mathrm{vol}(\mathcal{O}_x)$.
(2) The gradient in $(\mathcal{M}, g)$ of the entropy $S(\phi(m))$ is the mean curvature at each point $m \in \mathcal{O}_x$.
(3) Brownian motion on $(\mathcal{M}, g)$ projects to Brownian motion on $(\mathcal{X}, h)$. Precisely, if M_t^β is a diffusion in $\mathcal{M}$ with generator $\beta^{-1}\triangle_g$ then $\phi(M_t^\beta)$ is a diffusion on $\mathcal{X}$ with generator $\beta^{-1}\triangle_h$.

The tangent space $T_m\mathcal{M}$ at each point $m \in \mathcal{O}_x$ splits into $T_m\mathcal{O}_x$ and its orthogonal complement $T_m\mathcal{O}_x^\perp$. Let P_m and $P_m^\perp$ be the orthonormal projections onto these subspaces of $T_m\mathcal{O}_x$. The formal analog of the anisotropic splitting in equation (11.7) is

$$dm^{\beta,\kappa} = P_m dM^\beta + \sqrt{\kappa} P_m^\perp dM^\beta. \tag{12.1}$$

The use of stochastic calculus makes it easy to describe the intuitive basis of the anisotropic decomposition and we will use this expression below for its simplicity and suggestive power. Such Itô calculus expressions may be made rigorous in two ways. First, we may use the equivalent Stratonovich SDE, computing the required Itô-Stratonovich correction. Second, it is also possible to avoid stochastic calculus by replacing the standard Laplacian with an anisotropic Laplacian as follows.

Rigorously, the decomposition (12.1) can be expressed using the generator of the process m_t^β. The orthogonal decomposition $T_m\mathcal{M} = \mathbb{T}_m\mathcal{O}_x \oplus T_m\mathcal{O}_x^\perp$ allows us to split the Laplacian $\triangle_g$ into 'angular' and 'radial' Laplacians, denoted by $\triangle_\mathcal{O}$ and $\triangle_{\mathcal{O}^\perp}$ respectively. Then the anisotropic process m_t^β is the diffusion with generator

$$\frac{1}{\beta}\left(\triangle_\mathcal{O} + \kappa\triangle_{\mathcal{O}^\perp}\right). \tag{12.2}$$

An explicit description of an orthonormal basis for $T_m\mathcal{M}$ is very useful when we apply this idea in practice. In particular, the basis in Section 8.3 is used when we apply these ideas to the DLN.

The energetics are as follows. We assume given a loss function $E : \mathcal{X} \to \mathbb{R}$ downstairs that we lift to a loss function $L : \mathcal{M} \to \mathbb{R}$, $L = E \circ \phi$.

We then define *stochastic gradient descent by the RLE* of the loss function through the (formal) Itô SDE 'upstairs'

$$dm^{\beta,\kappa} = -\mathrm{grad}_g L(m)\, dt + P_m dM^\beta + \sqrt{\kappa}\, P_m^\perp dM^\beta. \tag{12.3}$$

This equation is an analogue of (11.7). We then expect in analogy with equation (11.8) that the corresponding SDE 'downstairs' is

$$dx = -\mathrm{grad}_h F_\beta(x)\, dt + dX^{\beta/\kappa}, \tag{12.4}$$

with free energy

$$F_\beta(x) = L(x) - \frac{1}{\beta}S(x). \tag{12.5}$$

In the limit $\kappa = 0$, we have the *stochastic* flow upstairs

$$dm = -\mathrm{grad}_g L(m)\, dt + P_m dM^\beta. \tag{12.6}$$

The flow of m_t projects to the *deterministic* gradient flow downstairs

$$\dot{x} = -\mathrm{grad}_h F_\beta(x). \tag{12.7}$$

We have thus obtained a strictly geometric model for the thermodynamic concept of quasistatic equilibration. The observable x_t has no fluctuations and evolves according to the gradient descent of free energy. However, m_t upstairs stochastically 'rolls without slipping' along a mean path with drift $\mathrm{grad}_g L$. [6]

12.2. **RLE for the DLN.** We now apply these general principles to the DLN. The reference Riemannian manifold $(\mathcal{M}, g)$ is the balanced manifold $(\mathcal{M}, \iota)$. The Lie group of isometries is O_d^{N-1}, and the quotient manifold is $(\mathfrak{M}_d, g^N)$.

Let $\mathbf{M}_t$ denote Brownian motion on $(\mathcal{M}, \iota)$. We may construct $\mathbf{M}_t$ by projecting standard Brownian motion on $\mathbb{M}_d^N$ onto $\mathcal{M}$, or by pushing forward Brownian motion on the parameter space under the parametrization $\mathfrak{z}$. The inverse temperature β may be included by a trivial scaling so that we have the process $\mathbf{M}_t^\beta$. We further split $\mathbf{M}^\beta$ into two processes using the orthogonal projection onto the group orbit $\mathcal{O}_W \subset \mathcal{M}$, obtaining the process $\mathbf{M}_t^{\beta,\kappa} \in \mathcal{M}$. This process is analogous to $m^{\beta,\kappa}$ defined in equation (12.1), though we use slightly different notation for convenience.

Theorem 18 below provides an explicit description of Brownian motion downstairs on $(\mathfrak{M}_d, g^N)$. We must first introduce some notation to make the theorem

[6] The terminology 'rolling without slipping' refers to the kinematics of the wheel. When the center of a wheel moves at steady velocity, the point of contact with the ground always has instantaneous velocity zero (and thus does not 'slip'). This idea can be used to define Brownian motion on any manifold with a connection [21, 23].

transparent. Recall from equation (5.17) that $W = Q_N \Sigma Q_0^T$ is the SVD of W and that $\Lambda = \Sigma^{1/N}$. We also define two diagonal matrices obtained by differentiation from Σ. Let Σ' be the diagonal matrix with k-th entry

$$\Sigma'_{kk} = \sum_{l \neq k} \left(\frac{N\lambda_k^{2N-1}}{\lambda_k^{2N} - \lambda_l^{2N}} - \frac{\lambda_k}{\lambda_k^2 - \lambda_l^2} \right) \lambda_k^{N-1}, \quad 1 \leq k \leq d. \tag{12.8}$$

Similarly, define the matrix Σ'' to be the diagonal matrix with entries

$$\Sigma''_{kk} = (N-1)\lambda_k^{N-2}, \quad 1 \leq k \leq d. \tag{12.9}$$

Assume given a matrix B_t valued Brownian motion in $\mathbb{M}_d$, or equivalently, d^2 standard independent Wiener processes labeled $\{B_t^{k,l}\}_{1 \leq k,l \leq d}$. Finally, to fix the Brownian motion, we assume given an initial condition $W_0 \in \mathfrak{M}_d$.

Theorem 18 (Menon, Yu [34])**.** *The solution X_t^β to the following Itô SDE with initial condition $X_0^\beta = W_0$ is Brownian motion on $(\mathfrak{M}_d, g^N)$ started at W_0:*

$$dX_t^\beta = \sqrt{\frac{2}{\beta}} \begin{pmatrix} \sqrt{N}\lambda_1^{N-1} dB_t^{1,1} & \sqrt{\frac{\lambda_1^{2N} - \lambda_2^{2N}}{\lambda_1^2 - \lambda_2^2}} dB_t^{1,2} & \dots \\ \sqrt{\frac{\lambda_2^{2N} - \lambda_1^{2N}}{\lambda_2^2 - \lambda_1^2}} dB_t^{2,1} & \sqrt{N}\lambda_2^{N-1} dB_t^{2,2} & \dots \\ \vdots & \ddots & \vdots \end{pmatrix} + \frac{1}{\beta} Q_N \Sigma'' Q_0^T \, dt. \tag{12.10}$$

We may now state the RLE for the DLN in analogy with equations (12.3) and (12.4). The RLE for $\mathbf{W}^{\beta,\kappa}$ is (cf. equation (3.4))

$$d\mathbf{W}^{\beta,\kappa} = -\nabla_{\mathbf{W}} E(\phi(\mathbf{W}))\, dt + d\mathbf{M}^{\beta,\kappa}. \tag{12.11}$$

Then we show in [34] that the law of the end-to-end matrix $W_t = \phi(\mathbf{W}_t)$ is given by the SDE

$$dW^{\beta,\kappa} = -\mathrm{grad}_{g^N} F_\beta(W^{\beta,\kappa})\, dt + dX^{\beta/\kappa}, \tag{12.12}$$

where $dX^{\beta/\kappa}$ is described by Theorem 18 and $F_\beta(W) = E(W) - \beta^{-1} S(W)$ where the entropy $S(W)$ is given by Theorem 14. The gradient may be computed explicitly and we find

$$\mathrm{grad}_{g^N} F_\beta(W^{\beta,\kappa}) = \mathcal{A}_{N,W}(E'(W)) - \frac{1}{\beta} Q_N \Sigma' Q_0^T. \tag{12.13}$$

The second term on the right is the gradient of $-\beta^{-1} S(W)$. The expression for Σ' in equation (12.8) provides the analogue of Coulombic repulsion in the DLN. This 'physical effect' is strictly due to the geometry of the DLN.

13. Discussion

13.1. **Summary.** This article has used the geometric theory of dynamical systems to study training dynamics in the DLN. Let us review the main ideas.

Theorem 1 and Theorem 2 allow us to understand the foliation of $\mathbb{M}_d^N$ by invariant manifolds, identify the important concept of balancedness and the dynamics on the balanced manifolds. Theorem 3 identifies the important role of Riemannian geometry. Our presentation of these results differs from the original sources in that we use Riemannian geometry as the foundation for our analysis.

Let us explain this shift more more carefully. By viewing Riemannian submersion as the fundamental paradigm, we discover an entropy formula for the DLN (Theorem 10) and obtain an explicit parametric description of the Riemannian

manifolds $(\mathcal{M}_r, \iota)$ and $(\mathfrak{M}_r, g^N)$ (Theorem 13 and Theorem 14). We also obtain a thermodynamic framework, including an understanding of the underlying microscopic fluctuations as follows. The minimal intrinsic description of stochastic fluctuations is provided by the theory of Brownian motion on Riemannian manifolds. Our understanding of this abstract idea is guided by Dyson Brownian motion, a fundamental model in random matrix theory. In particular, the entropy formula is tied to Brownian motion on group orbits in analogy with Dyson Brownian motion. Finally, in order to extend gradient descent to stochastic gradient descent in a geometrically faithful manner, we introduce and compute Riemannian Langevin equations for the DLN. We stress the importance of an anisotropic splitting of the noise between 'noise in the gauge' and 'noise in the observable', in order to provide a geometric framework for the thermodynamics of deep learning.

As noted at the outset, while all the Theorems in this article have a classical feel, the results presented here have been obtained very recently. As a dynamical system, the DLN is fascinating for the range of concrete, tractable questions that arise from a careful examination of overparametrization. In particular, the connections with the theory of minimal surfaces and random matrix theory, suggest new questions on motion by curvature and the large d and large N asymptotics of the DLN. Some of these are summarized in Section 14.

13.2. **Linear networks and deep learning.** A good model must provide insights of practical importance, while at the same time being tractable to to a detailed mathematical analysis. As we have seen, the analysis of training dynamics in the DLN can be studied in rich detail. But how good is it as a guide to deep learning in practice? Let us now present some ideas explored in the literature on linear networks. We then return to the heuristics of deep learning listed in Section 2, contrasting these with the DLN.

Phenomenological linear models have been explored in the neural network literature since the 1990s. The energy landscape for linear networks was studied carefully by Baldi and Hornik [7, 8]. Since the advent of deep learning, a central concern has been to use the DLN to understand implicit bias, to understand the role of depth, and to understand the energy landscape for several matrix learning tasks. This has given rise to a large literature on linear networks, of which we can only present a few samples.

An important result in a similar spirit to [2, 3], especially a characterization of the minimizer when the depth $N = 2$ and $W_2 = W_1^T$ (the Bures-Wasserstein reduction of the DLN) is given in [19, 20]. There have also been several studies of the nature of critical points for matrix learning tasks and the related convergence theory in the DLN for different choices of loss functions [12, 18, 36]. An early use of exact solutions in DLN, as well as the use of the DLN to develop a theory for semantic development is presented in [16, 38, 39].

Linear *convolutional* networks (LCNs) were introduced to shed light on deep learning with convolutional neural networks. An algebraic geometric analysis of LCNs has provided a detailed understanding of the nature of the singularities of the function spaces as a function of the network architecture [25, 26]. In these studies, linear networks are used to shed light on the expressivity of neural networks. A more recent direction is to include nonlinearity that respects the matrix geometry by depending only on the singular values [13].

13.3. **Numerical experiments.** Like most models in machine learning, our understanding of the DLN relies heavily on numerical experiments. These experiments reveal subtle phenomena, such as the importance of low-rank matrices as $\lim_{t\to\infty} W(t)$, and the importance of the balanced manifolds. Let us explain these in turn.

Numerical experiments in [2, 3, 4, 14] explore several aspects of the DLN, including the attraction to low-rank matrices. For example, in a numerical study of matrix completion in [14] we observe that the effective rank of $W(t)$ decreases in time through sudden jumps; roughly the effective rank stays constant for a long period and suddenly drops by 1. We also find an asymptotic distribution of limits $\lim_{t\to\infty} W(t)$ on the low-rank manifolds.

We may design similar numerical experiments for deep learning as follows. We fix a family of random functions $\mathcal{F}$ (e.g. random Fourier series on the line with a fixed decay rate) and then numerically observe the training dynamics for a neural network approximation to a randomly chosen $f \in \mathcal{F}$. It is easy to vary the network architecture (depth, width, neural unit) and study the resulting variation in the training dynamics. Experiments of this nature have been carried out in classroom projects designed by the author. They reveal that typical neural networks fit functions scale-by-scale, with jumps in scale (i.e. sudden increases in oscillations) that are analogous to the jumps in rank observed in the DLN. These experiments lead us to conjecture that the fidelity between the DLN and deep learning is much closer than what may be naively expected.

These are numerical experiments, not theorems. But the nature of the transient dynamics, not just the time asymptotics, is a fertile area of enquiry where we believe it is possible to transfer insights between the DLN and deep learning. In particular, the training dynamics in deep learning also suggest an interplay between curvature and entropy, corresponding to fluctuations in the null directions. We hope that the results in this article will lead to methods to quantify this entropy using functional analytic tools. Any such analysis must be matched with numerics; it is necessary to continuously design numerical experiments, in parallel with the refinement of the theory of DLN, to create an intuition for deep learning.

13.4. **Balancedness.** Let us now discuss aspects of the balanced manifolds that remains somewhat mysterious. That the balanced variety has a beautiful mathematical structure is not in doubt, but the invariant manifold theorems don't really explain why the balanced manifolds should be as important in practice as they are. Since the balanced variety is just the $\mathbf{G} = \mathbf{0}$ special case of the $\mathbf{G}$-balanced varieties, the odds that a randomly chosen initial condition will lie on $\mathcal{M}$ is zero!

There have been extensions of the idea of balancedness to networks that vary linearity locally [17]. More recently, the geometry of the vector fields in the determination of conservation laws has been considered in [31, 32]. These results are promising, but do not yet constitute a satisfactory understanding of balancedness.

Our introduction of the RLE in Section 10, especially the framework of quasistatic thermodynamic equilibration and the RLE for the DLN is intended to shed light on this question. While these theorems are restricted to the balanced manifold $\mathcal{M}$ of full rank, we may define analogous RLE 'upstairs' on any $\mathcal{M}_{\mathbf{G}}$. The underlying conjecture is that small noise, for example due to round-off error, may be modeled by such RLE and that in the $\beta > 0$, $\kappa = 0$ regime, we may observe

motion by curvature of $\mathcal{M}_{\mathbf{G}}$ towards $\mathcal{M}_{\mathbf{0}}$. We hope that this approach will explain the mysterious appearance of the Simons cone in the DLN.

In Section 2 we noted that formally Riemannian submersion also applies to deep learning. What is still missing in this analogy is the concept of balancedness. That is, while there is no ambiguity about the importance of balancedness for the DLN, we don't yet know what form this idea must take for deep learning. The parametrization by β and Theorem 14 reveal a subtle coupling along the depth of the network on $\mathcal{M}$. A natural conjecture is that balancedness must correspond to local equilibrium, i.e. *detailed balance*, along the network. However, the specific mathematical form of this idea remains elusive.

14. Open problems

We begin with some concrete questions that do not require much machinery. This is followed by a broader discussion of perspectives to explore.

14.1. **Convergence to a low-rank matrix.** Here is a simply stated problem. Fix $d = 2$ and consider matrix completion for W when given that $W_{11} = W_{22} = 1$. We study this problem with the DLN as follows: we use equation (3.8) to define the quadratic energy function

$$E(W) = \frac{1}{2}\left((W_{11}-1)^2 + (W_{22}-1)^2\right). \tag{14.1}$$

Observe that E vanishes for any matrix of the form

$$W = \begin{pmatrix} 1 & * \\ * & 1 \end{pmatrix}. \tag{14.2}$$

Further, there is a variety of rank-one solutions of the form

$$W = \begin{pmatrix} 1 & a \\ \frac{1}{a} & 1 \end{pmatrix}, \quad a \in \mathbb{R}\backslash\{0\}. \tag{14.3}$$

This situation is illustrated in Figure 9.1. Numerical experiments (see [14, Figures 3–6]) reveal that the solution to (4.21) with random initial condition $W_0 \in \mathbb{M}_2$ is attracted to the low-rank solutions. Further, these solutions cluster around the matrix

$$W = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad a \in \mathbb{R}\backslash\{0\}. \tag{14.4}$$

A rigorous analysis of this example, including the N dependence, would provide considerable new insight. Specifically, we seek the resolution of the following question. Consider the dynamical system

$$\dot{W} = -\sum_{k=1}^{N} (WW^T)^{\frac{N-k}{N}} E'(W) (W^T W)^{\frac{k-1}{N}}, \tag{14.5}$$

with $E(W)$ given by equation (14.1). Can one establish convergence of the solution $W(t)$ for suitable initial conditions to the low-rank energy minimizer in equation (14.2) as $t \to \infty$? Are there initial conditions with $E(W) > 0$ whose solutions converge to other minimizers of E?

This problem illustrates a fundamental issue in dynamical systems theory. While one may have theorems that guarantee convergence, the precise identification of the limit is not straightforward even in apparently simple problems.

14.2. **The free energy landscape.** We return to a theme discussed in Section 6.3. Does the inclusion of entropy provide a selection principle, especially for degenerate loss functions?

For each depth N, we may consider the entropy defined in Theorem 10, and the free energy defined in equation (6.6). An interesting aspect of this free energy is that while the entropy depends on the singular values Σ alone, the energy depends on all of W. This makes the computation of the set of minimizers an interesting problem, even when $d = 2$.

The simplest concrete problem in this class is as follows. Consider again the energy in equation (14.1). What is the nature of the minimizing set $\mathcal{S}_\beta$? Specifically, how does it vary with β and N?

14.3. **Large d and large N asymptotics.** An important feature of the DLN is that it admits a natural $N \to \infty$ limit. This limit requires that we rescale $\mathcal{A}_{N,W}$ defined in equation (4.14) as follows

$$\frac{1}{N}\sum_{p=1}^{N}(WW^T)^{\frac{N-p}{N}}\, Z\, (W^TW)^{\frac{p-1}{N}}, \tag{14.6}$$

and let $N \to \infty$ to obtain the limiting operator

$$\mathcal{A}_{\infty,W}(Z) = \int_0^1 (WW^T)^{1-s}\, Z\, (W^TW)^s\, ds. \tag{14.7}$$

This operator and the resulting metric g^∞ were studied in [14]. In particular, the metric may be diagonalized as in Section 7.2, the crucial lemma being

Lemma 3. *The operator $\mathcal{A}_{\infty,W} : T_W\mathfrak{M}_d \to T_W\mathfrak{M}_d$ is symmetric and positive definite with respect to the Frobenius inner-product. It has the d^2 eigenvalues and eigenvectors*

$$\mathcal{A}_{\infty,W}\, u_k v_l^T = \frac{\sigma_k^2 - \sigma_l^2}{2\log \sigma_k/\sigma_l}\, u_k v_l^T, \quad 1 \le k, l \le d, \tag{14.8}$$

when $k \neq l$ and

$$\mathcal{A}\, u_k v_k^T = \sigma_k^2\, u_k v_k^T, \quad 1 \le k \le d. \tag{14.9}$$

Recall that u_k and v_l define the normalized left and right singular vectors of W. The analog of equation (7.15) is

$$g^\infty(Z,Z) = \sum_{1\le k\le d} \frac{1}{\sigma_k^2} Z_{kk}^2 + \sum_{1\le k,l\le d, k\neq l} \frac{2\log \sigma_k/\sigma_l}{\sigma_k^2 - \sigma_l^2} Z_{kl}^2, \tag{14.10}$$

where we have written $Z \in T_M\mathfrak{M}_d$ as $Z = Z_{kl}u_k v_l^T$.

The rescaling of the metric affects the entropy by a constant factor, but since it is only the *gradient* of the entropy that matters we also have the limit

$$S_\infty(W) = \sum_{1\le k<l\le d} \log\left(\frac{\sigma_k^2 - \sigma_l^2}{2\log \sigma_k/\sigma_l}\right). \tag{14.11}$$

These formulas are reminiscent of determinantal formulas in random matrix theory (see [6] for examples). Their large d asymptotics have not been studied. Further, the analogy between Dyson Brownian motion and the RLE for DLN in Section 12.2 suggest that the large d asymptotics for the equilibrium distribution should be described by universal fluctuations. In order to make such a study precise, we suggest the reader first examine the equilibrium distribution of equation (12.12)

for the simplest quadratic energy $E(W) = \frac{1}{2}\mathrm{Tr}(W^T W)$, establishing the analog of the semicircle law for the DLN. This may be followed by studies of fluctuations.

An interesting facet of the large N limit is that while the formulas for the metric and entropy simplify since these rely only on W downstairs, there is no longer a space upstairs! Is there a limiting framework for such Riemannian submersion?

14.4. **Low-rank varieties.** The study of the DLN would benefit greatly from a more careful examination of the underlying algebraic and differential geometry. Let us focus on the influence of rank to illustrate this point.

All the main ideas in this article were first established for matrices of full rank. While this is a convenient choice, a careful analysis of the foliations by rank is necessary for both theoretical and practical purposes.

First, as we have remarked above, in practice low-rank matrices seem to appear as $t \to \infty$ limits, even when the dynamics is restricted to $W \in \mathfrak{M}_a$. Even for simple energies, such as in matrix completion, it would be very useful to understand how the set of minimizers may be arranged by rank.

Second, our results on Riemannian submersion and RLE, typically require smoothness of the submersion map ϕ and the assumption of distinct singular values. At present, these are convenient assumptions that allow us to prove interesting theorems. However, an understanding of the singularities that may arise at repeated singular values, especially repeated zero singular values as in the low-rank setting, is necessary for a comprehensive understanding of the model.

14.5. **Coexistence of gradient and Hamiltonian structures.** Gradient flows typically have a character that is complementary to Hamiltonian systems. There are, however, rare cases of dynamical systems that are both gradient-like and completely integrable Hamiltonian [11]. We expect that a deeper examination of the DLN will reveal similar structure for the following reasons.

Training dynamics in the DLN are given by a gradient flow. However, the theorems in this article also suggest a complementary symplectic geometry. Here we note that the natural geometric setting for a Hamiltonian system is that of a symplectic manifold. Further, co-adjoint orbits of Lie groups constitute one of the fundamental examples of symplectic manifolds, including most known completely integrable systems. In the analysis of the DLN, we see that the existence of the conserved quantities $\mathbf{G}$ – a typical feature of integrable Hamiltonian systems – is tied to the symmetries of overparametrization as in the theory of Hamiltonian systems. Further, the appearance of determinantal formulas as in random matrix theory is also suggestive of completely integrability.

15. Acknowledgements

Sanjeev Arora introduced me to the DLN (and Nadav Cohen!) at IAS in 2019. Little did I realize at the time how rewarding the study of the DLN would be, so many thanks Sanjeev. The results presented here are all joint work with Nadav Cohen, Zsolt Veraszto, and Tianmin Yu. They also build on related projects with Dominik Inauen, Ching-Peng Huang, Tejas Kotwal, Jake Mundo and Lulabel Ruiz-Seitz. I am particularly grateful to Nadav for many patient explanations about the nature and promise of deep learning and the pleasure of joint work.

This article was developed on the basis of minicourses at Kyushu University, CIMAT, the National University of Singapore and the Palazzone di Cortona of the

Scuola Normale Superiore, Pisa. Each lecture provided an opportunity to synthesize several results on the DLN for a diverse audience, improving the exposition in each iteration. I express my thanks to Andrea Agazzi (Pisa), Octavio Arizmendi Echegaray (CIMAT), Carlos Pacheco (Cinvestav), Subhro Ghosh (NUS), Philippe Rigollet (MIT) and Tomoyuki Shirai (Kyushu) for providing me with the opportunity to share these ideas and for many stimulating discussions. I am especially grateful to the students at CIMAT and Cortona for their warmth and enthusiasm.

Related discussions with José Luis Pérez Garmendia (CIMAT), Alex Dunlap, Jian-Guo Liu and Jonathan Mattingly (Duke), Guido Montufar (UCLA), Austin Stromme (ENSAE/CREST), Praneeth Netrapalli (Google), Courtney and Elliott Paquette (Montreal), Qianxiao Li and Xin Tong (NUS), Boris Hanin, Pier Beneventano and Mufan Li (Princeton), Jamie Mingo (Queens University), Eulalia Nualart (UPF, Barcelona), Noriyoshi Sakuma (Nagoya) and Sinho Chewi and Zhou Fan (Yale) have improved this work.

I am a dynamicist, not an expert in deep learning, and it has been a real thrill to learn from computer scientists and statisticians. I hope that this article communicates the spirit of the geometric theory of dynamical systems in a way that is useful to practitioners, reciprocating the gift of a beautiful model.

References

[1] R. Arora, S. Arora, J. Bruna, N. Cohen, S. Du, R. Ge, S. Gunasekar, C. Jin, J. Lee, T. Ma, et al., *Theory of deep learning*, 2020.

[2] 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.

[3] S. Arora, N. Cohen, and E. Hazan, *On the optimization of deep networks: Implicit acceleration by overparameterization*, in International Conference on Machine Learning, PMLR, 2018, pp. 244–253.

[4] 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, 2019, pp. 7411–7422.

[5] B. Bah, H. Rauhut, U. Terstiege, and M. Westdickenberg, *Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers*, Information and Inference: A Journal of the IMA, 11 (2022), pp. 307–353.

[6] J. Baik, P. Deift, and T. Suidan, *Combinatorics and random matrix theory*, vol. 172 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, 2016.

[7] P. Baldi and K. Hornik, *Neural networks and principal component analysis: Learning from examples without local minima*, Neural networks, 2 (1989), pp. 53–58.

[8] P. F. Baldi and K. Hornik, *Learning in linear neural networks: A survey*, IEEE Transactions on neural networks, 6 (1995), pp. 837–858.

[9] E. A. Barbaˇsin and N. N. Krasovskiˇi, *On stability of motion in the large*, Doklady Akad. Nauk SSSR (N.S.), 86 (1952), pp. 453–456.

[10] R. Bhatia, T. Jain, and Y. Lim, *On the Bures–Wasserstein distance between positive definite matrices*, Expositiones Mathematicae, 37 (2019), pp. 165–191.

[11] A. M. Bloch, R. W. Brockett, and T. S. Ratiu, *Completely integrable gradient flows*, Communications in Mathematical Physics, 147 (1992), pp. 57–74.

[12] P. Bréchet, K. Papagiannouli, J. An, and G. Montúfar, *Critical points and convergence analysis of generative deep linear networks trained with Bures-Wasserstein loss*, in International Conference on Machine Learning, PMLR, 2023, pp. 3106–3147.

[13] H. T. M. Chu, S. Ghosh, C. T. Lam, and S. S. Mukherjee, *Implicit regularization via spectral neural networks and non-linear matrix sensing*, arXiv:2402.17595, (2024).

[14] N. Cohen, G. Menon, and Z. Veraszto, *Deep linear networks for matrix completion—an infinite depth limit*, SIAM Journal on Applied Dynamical Systems, 22 (2023), pp. 3208–3232.

[15] R. DeVore, B. Hanin, and G. Petrova, *Neural network approximation*, Acta Numerica, 30 (2021), pp. 327–444.
[16] C. C. J. Dominé, N. Anguita, A. M. Proca, L. Braun, D. Kunin, P. A. M. Mediano, and A. M. Saxe, *From lazy to rich: Exact learning dynamics in deep linear networks*, arXiv:2409.14623, (2024).
[17] S. S. Du, W. Hu, and J. D. Lee, *Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced*, Advances in Neural Information Processing Systems, 31 (2018).
[18] R. Ge, C. Jin, and Y. Zheng, *No spurious local minima in nonconvex low rank problems: A unified geometric analysis*, in International Conference on Machine Learning, PMLR, 2017, pp. 1233–1242.
[19] S. Gunasekar, J. D. Lee, D. Soudry, and N. Srebro, *Implicit bias of gradient descent on linear convolutional networks*, Advances in Neural Information Processing Systems, 31 (2018).
[20] 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, 2017, pp. 6151–6159.
[21] E. P. Hsu, *Stochastic analysis on manifolds*, vol. 38 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, 2002.
[22] C.-P. Huang, D. Inauen, and G. Menon, *Motion by mean curvature and Dyson Brownian motion*, Electronic Communications in Probability, 28 (2023), pp. 1–10.
[23] N. Ikeda and S. Watanabe, *Stochastic differential equations and diffusion processes*, vol. 24 of North-Holland Mathematical Library, North-Holland Publishing Co., Amsterdam; Kodansha, Ltd., Tokyo, second ed., 1989.
[24] D. Inauen and G. Menon, *Stochastic Nash evolution*, arXiv preprint arXiv:2312.06541, (2023).
[25] K. Kohn, T. Merkh, G. Montúfar, and M. Trager, *Geometry of linear convolutional networks*, SIAM J. Appl. Algebra Geom., 6 (2022), pp. 368–406.
[26] K. Kohn, G. Montúfar, V. Shahverdi, and M. Trager, *Function space and critical points of linear convolutional networks*, SIAM J. Appl. Algebra Geom., 8 (2024), pp. 333–362.
[27] S. Ł ojasiewicz, *Sur les trajectoires du gradient d'une fonction analytique*, in Geometry seminars, 1982–1983 (Bologna, 1982/1983), Univ. Stud. Bologna, Bologna, 1984, pp. 115–117.
[28] A. Lapedes and R. Farber, *How neural nets work*, in Evolution, learning and cognition, World Sci. Publ., Teaneck, NJ, 1988, pp. 331–346.
[29] J. P. LaSalle, *Stability theory and invariance principles*, in Dynamical systems (Proc. Internat. Sympos., Brown Univ., Providence, R.I., 1974), Vol. I, Academic Press, New York-London, 1976, pp. 211–222.
[30] J. M. Lee, *Introduction to Riemannian manifolds*, vol. 176 of Graduate Texts in Mathematics, Springer, Cham, second ed., 2018.
[31] S. Marcotte, R. Gribonval, and G. Peyré, *Abide by the law and follow the flow: Conservation laws for gradient flows*, Advances in Neural Information Processing Systems, 36 (2024).
[32] ———, *Keep the momentum: Conservation laws beyond euclidean gradient flows*, arXiv preprint arXiv:2405.12888, (2024).
[33] G. Menon, *Pattern theory: old and new*, Lecture notes, Brown University, June 2023.
[34] G. Menon and T. Yu, *An entropy formula for the deep linear network*, 2024.
[35] J. Milnor, *Morse theory*, vol. No. 51 of Annals of Mathematics Studies, Princeton University Press, Princeton, NJ, 1963. Based on lecture notes by M. Spivak and R. Wells.
[36] G. M. Nguegnang, H. Rauhut, and U. Terstiege, *Convergence of gradient descent for learning linear neural networks*, 2021.
[37] D. A. Roberts, S. Yaida, and B. Hanin, *The principles of deep learning theory*, vol. 46, Cambridge University Press, Cambridge, MA, USA, 2022.
[38] A. M. Saxe, J. L. McClelland, and S. Ganguli, *Exact solutions to the nonlinear dynamics of learning in deep linear neural networks*, arXiv preprint arXiv:1312.6120, (2013).
[39] ———, *A mathematical theory of semantic development in deep neural networks*, Proceedings of the National Academy of Sciences, 116 (2019), pp. 11537–11546.

[40] L. Simon, *Theorems on regularity and singularity of energy minimizing maps*, Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 1996. Based on lecture notes by Norbert Hungerbühler.

Division of Applied Mathematics, Brown University, 182 George St., Providence, RI 02912.

Email address: govind_menon@brown.edu