SIAM J. Applied Dynamical Systems
Vol. 22, No. 4, pp. 3208–3232

Deep Linear Networks for Matrix Completion—an Infinite Depth Limit*

Nadav Cohen†, Govind Menon‡, and Zsolt Veraszto‡

Abstract. The deep linear network (DLN) is a model for implicit regularization in gradient based optimization of overparametrized learning architectures. Training the DLN corresponds to a Riemannian gradient flow, where the Riemannian metric is defined by the architecture of the network and the loss function is defined by the learning task. We extend this geometric framework, obtaining explicit expressions for the volume form, including the case when the network has infinite depth. We investigate the link between the Riemannian geometry and the training asymptotics for matrix completion with rigorous analysis and numerics. We propose that under small initialization, implicit regularization is a result of bias towards high state space volume.

Key words. generalizability, implicit regularization, Riemannian gradient flow, deep linear network, matrix completion

MSC codes. 68T07, 58D17, 37N40

DOI. 10.1137/22M1530653

1. Introduction.

1.1. The deep linear network.

Deep learning has proven its general applicability in several fields of applied science in the past decade [15]. While neural networks are structurally simple function approximators (e.g., [14]), several questions around them remain unanswered. Two fundamental problems are to obtain first principles explanations for generalizability and implicit regularization. Generalizability refers to a network's performance on new, previously unseen data. Implicit regularization is the feature of deep networks to avoid overfitting despite overparametrization. In the context of classical regression problems, overfitting is mitigated by explicit regularization. Deep learning architectures are observed not to overfit despite the lack of explicit regularization. This is referred to as implicit regularization.

A simple model in which the effect of overparametrization in deep networks can be studied is the deep linear network (DLN) [1, 2]. The DLN is simple enough to serve as a minimal model for neural networks. It may also be applied directly to optimization problems, such as matrix completion [7], autoencoders [3], and multitask training. The most common approach

*Received by the editors October 25, 2022; accepted for publication (in revised form) by K. Josic July 25, 2023; published electronically November 28, 2023.
https://doi.org/10.1137/22M1530653
Funding: The work of the first author was partially supported by the Israel Science Foundation (grant 1780/21) and the Tel Aviv University Center for AI and Data Science (TAD). The work of the second and third authors was partially supported by NSF grant DMS-2107205.
†Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel (cohennadav@tauex.tau.ac.il).
‡Division of Applied Mathematics, Brown University, Providence, RI 02912 USA (govind_menon@brown.edu, zsolt_veraszto@brown.edu).

to the training process is gradient based minimization of a loss function, a process known as empirical risk minimization.

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

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

$$\mathbf{W} = (W_N, W_{N-1}, \dots, W_1). \tag{1.1}$$

The number of weight matrices, N, is referred to as the *depth* of the network. We have assumed that all matrices have the same size for simplicity. In fact, all that is required is that the matrix sizes be such that the product W in (1.2) is well defined. Properties of the DLN in such generality have been studied in [2, 3]. We build on these results in our work but restrict ourselves to $W_i \in \mathbb{M}_d$ for ease of computation. Some aspects of our analysis extend to a manifold of fixed-rank matrices as in [3] (see section 3.4.3).

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

$$W = \pi(\mathbf{W}) := \prod_{i=N}^{1} W_i. \tag{1.2}$$

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

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

$$\frac{d}{dt} W_j = -\partial_{W_j} E(W), \quad 1 \le j \le N, \tag{1.3}$$

is used as our training model. Equation (1.3) is an idealization for the manner in which training is implemented in practice. Since W is defined through (1.2), we also find that

$$\dot{W}_j = -W_{j+1}^T \dots W_N^T \, \partial_W E(W) \, W_1^T \dots W_{j-1}^T, \quad j = 1, \dots, N. \tag{1.4}$$

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

Our interest lies in the long-time behavior of the observable $W(t)$. The limit $\lim_{t\to\infty} W(t)$ is called the *training outcome.* Since the dynamics is governed by a gradient flow, this limit exists when the loss function $E(W)$ is bounded below and has compact sublevel sets. However, natural loss functions, such as those for certain matrix completion problems, do *not* have compact sublevel sets. In fact, the minima of these loss functions are submanifolds of $\mathbb{M}_d$. Thus, the prediction of training outcomes is a subtle problem. The existence of $\lim_{t\to\infty} W(t)$

when each of the W_i has full rank was established using Lojasiewicz's theorem [3, Thm. 10]. In order to use this convergence theorem as a foundation for the analysis in this paper, we restrict ourselves to $W_i \in GL(d)$ for most of the analysis. The space $GL(d)$ is the set of bijective linear transformations of $\mathbb{R}^d$, which is isomorphic to the set of $d \times d$ invertible matrices. Despite this assumption, training outcomes for many matrix completion problems have low rank. The repeated appearance of low-rank matrices as training outcomes has stimulated the use of several heuristics for the prediction of training outcomes (see the discussion in [20]). The goal of this paper is to shed light on this question using the Riemannian geometry of the DLN and numerical simulations.

Let us now review the Riemannian geometry underlying the Euclidean gradient flow defined in (1.3). An important concept (see [1, Def. 1]) is the notion of *balancedness*. We say that the weight matrices W_j and W_{j+1} are G_j-balanced if

$$G_j := W_{j+1}^T W_{j+1} - W_j W_j^T. \tag{1.5}$$

For fixed G_j, $1 \le j \le N$, (1.5) defines an algebraic variety $\mathbb{M}_d^N$ that is stratified by the rank. When $G_j = 0$, for each value of the rank r, the gradient flow (1.3) leaves the manifolds $\mathcal{M}_r$ of rank-r matrices solving (1.5) invariant [3, Cor. 6]. These are called the *balanced manifolds*. It is immediate from (1.5) that the singular values of the weight matrices $\{W_j\}_{j=1}^N$ are equal on the balanced manifolds.

In order to use existing convergence theory, we restrict our attention to full-rank matrices, and we denote $\mathcal{M}_d$ by $\mathcal{M}$ in what follows. This balanced manifold allows us to separate the dynamics into a flow "upstairs" in $\mathbb{M}_d^N$, described by (1.3), and a flow "downstairs" for the end-to-end matrix $W(t)$ in $GL(d)$. The flow downstairs is a Riemannian gradient flow with a metric computed in [3] that may be described as follows. We define the linear map $\mathcal{A}_{N,W} : T_W GL(d) \simeq \mathbb{M}_d \to \mathbb{M}_d$:

$$\mathcal{A}_{N,W}(Z) := \frac{1}{N} \sum_{j=1}^{N} \left(WW^T\right)^{\frac{N-j}{N}} Z \left(W^T W\right)^{\frac{j-1}{N}}. \tag{1.6}$$

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

$$\dot{W} = -\mathrm{grad}_{g^N} E(W), \tag{1.7}$$

under the metric

$$g^N(Z_1, Z_2) = \mathrm{tr}\left(\mathcal{A}_{N,W}^{-1}\left(Z_1\right)^T Z_2\right), \tag{1.8}$$

where $Z_1, Z_2 \in T_W GL(d)$. This structure allows us to extend the DLN geometry to the infinite depth limit, with the linear operator

$$\mathcal{A}_{\infty,W}(Z) = \lim_{N\to\infty} \mathcal{A}_{N,W}(Z) = \int_0^1 \left(WW^T\right)^{(1-\tau)} Z \left(W^T W\right)^{\tau} d\tau, \tag{1.9}$$

replacing $\mathcal{A}_N$ in (1.8) to define a limiting metric g^∞. When N is finite, the dynamical system (1.7) corresponds to a flow upstairs in $\mathbb{M}_d^N$. In the limit of infinite depth, (1.7) continues to hold, even though there is no longer a well-defined flow upstairs.

The existence of the infinite depth metric, in particular its resemblance to the Bogoliubov inner product in quantum statistical mechanics, was noted in [3, Remark 8]. We develop some properties of this metric below, emphasizing explicit formulas in singular value decomposition (SVD) coordinates in section 1.3.1. These formulas also show that the metric is at least C^1 in the open subset of $\mathbb{M}_d$ of matrices with distinct singular values. These formulas are better seen as first steps towards deeper exploration. We still lack an understanding of the curvature and geodesics of these metric though some partial results have been obtained in [21]. The appearance of such elegant Riemannian structures in the DLN is surprising at first sight. However, it is helpful to note that fundamental interior point methods for conic programs also have a surprising gradient structure (see [4, 5, 12]). In recent work [18], one of the authors and Yu have studied these metrics, seeking precise comparisons between gradient flows underlying conic programs and deep learning.

The above structure applies to an arbitrary loss function. In practice, the loss function is determined by the learning task and the ease of computation. We focus on matrix completion in this paper, choosing the loss function $E(W)$ to be the quadratic distance from a fixed matrix Φ termed the *optimization objective* [9]. Using $\circ$ for the elementwise (Hadamard) product, the family of loss functions we study takes the form

$$E_{\mathcal{B}}(W) = \frac{1}{2}\|\mathcal{B} \circ (\Phi - W)\|_2^2. \tag{1.10}$$

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

$$E_I(W) = \frac{1}{2}\|\mathrm{diag}\,(\Phi - W)\|_2^2. \tag{1.11}$$

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

The nature of the loss function is determined by the matrix $\mathcal{B}$. For example, the energy defined by (1.11) has a submanifold of global minima. Indeed, given a diagonal matrix Φ, $E_I(W)$ vanishes when W is chosen to be any matrix whose diagonal entries are Φ. Moreover, since the matrix can be completed to any rank from 1 to d, some of these minima are noninvertible matrices. We consider numerical examples with several such $\mathcal{B}$.

1.3. Statement of results.

1.3.1. The metric and volume forms. The SVD of W is denoted $W = U\Sigma V^T$, where $U, V \in O(d)$ and Σ denotes the diagonal matrix of singular values. We write $\Sigma_{ii} = \sigma_i$ and order the singular values in decreasing order: $\sigma_i \geq \sigma_j$ if $i < j$. The metric g^N may be expressed in a simple manner using the SVD. We find (see (2.17)) that

$$g^N = (V \otimes U) D^N(\Sigma)(V \otimes U)^T. \tag{1.12}$$

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

$$D^N_{il} = \frac{N}{\sum_{j=1}^N (\sigma_i^2)^{N-j/N}(\sigma_l^2)^{j/N}}, \quad 1 \leq i, l \leq d. \tag{1.13}$$

Here the subscript il denotes a double index on the diagonal elements of D^N. The expression is unambiguous due to the symmetry of (1.13) in i and l. The expression (1.12) holds in the limit $N = \infty$, with

$$D^\infty_{il} = \frac{2\log(\sigma_i/\sigma_l)}{\sigma_i^2 - \sigma_l^2}, \quad i \neq l, \quad D_{ii} = \frac{1}{\sigma_i^2}. \tag{1.14}$$

These expressions for the metric are used to compute the associated volume forms in $GL(d)$. Let $\mathrm{van}(\Lambda)$ denote the Vandermonde determinant of a diagonal matrix Λ, let $d\Sigma$ denote Lebesgue measure on $\mathbb{R}^N$, and let dU and dV denote Haar measure on $O(d)$.

Theorem 1.1. *The volume form of g^N is given by*

$$\sqrt{\det g^N}\, dW = N^{\frac{d(d-1)}{2}} \det(\Sigma^2)^{\frac{1-N}{2N}} \mathrm{van}\left(\Sigma^{2/N}\right) d\Sigma dU dV. \tag{1.15}$$

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

$$\sqrt{\det g^\infty}\, dW = \frac{\mathrm{van}\left(\log \Sigma^2\right)}{\sqrt{\det\left(\Sigma^2\right)}}\, d\Sigma dU dV. \tag{1.16}$$

The volume forms allow us to quantify the importance of regions of high volume that correspond to empirical observations of training outcomes. Notice that the volume density blows up when passing to the limit $\sigma_i \to 0$, showing a clear relationship between low rank and high volume.

The reader unfamiliar with Riemannian geometry should note that all our work reduces to explicit calculations in SVD coordinates. The symmetries of the metric implicit in (1.14) allow us to calculate several Jacobian determinants explicitly. The main subtlety in using SVD coordinates is that naive calculations must be restricted to the open set of $\mathbb{M}_d$ where W has distinct singular values, and the branches of the SVD must be resolved on the lower-dimensional varieties corresponding to repeated eigenvalues. Formulas, such as those in Theorem 1.1, are established under the assumption that the singular values are distinct, and then seen to hold in the limit of repeated singular values using continuity.

Such calculations follow the spirit of random matrix theory [17]. For example, Theorem 1.1 suggests interesting asymptotics for the DLN in the limits $N \to \infty$ and $d \to \infty$. We do not consider this question in this paper, but see [10] for a similar investigation.

1.3.2. Normal hyperbolicity. The spectral decomposition of g^∞ given in (1.12) (with $N = \infty$) and (1.14) has important consequences for the dynamics given by (1.7). For different choices of $\mathcal{B}$, let $\mathcal{N}_\mathcal{B}$ denote the set of global minima of the energy function (1.10),

$$\mathcal{N}_\mathcal{B} = \{W : W \in \mathbb{M}_d,\, \mathcal{B} \circ (\Phi - W) = 0\}. \tag{1.17}$$

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

$$K = d^2 - \sum_{i,j=1}^{d} \mathcal{B}_{ij}. \tag{1.18}$$

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

Theorem 1.2. *Any K-dimensional compact submanifold of $\mathcal{N}_{\mathcal{B}}$ is normally hyperbolic under the dynamical system (1.7) for $N = \infty$.*

The statement is a consequence of Lemma 3.4. The proof is presented in section 3.2.

1.3.3. Are the training outcomes low-rank matrices? The origin of implicit regularization was proposed to arise from (quasi-)norms in [19]. This idea is motivated by classical regression, where overfitting effects are often mitigated by adding explicit regularization terms to the optimization problem. Previous work in this direction tried to explicitly construct the regularizers in the form of (quasi-)norms. This explanation was challenged in [20], where the training outcomes were observed to be low-rank matrices, even though matrix norms blew up. We develop this idea in section 3. On one hand, we view the bias towards low-rank matrices as an entropic effect determined by the volume forms above (the volume forms diverge as the singular values approach zero). On the other hand, we construct numerical examples where using low rank as a criterion does not accurately predict the training outcome under small initial conditions. The numerical examples show that within the set of low-rank minimizers, the actually observed ones are also the ones with the largest concentration of volume around them.

1.3.4. Numerical simulations for finite and infinite depth. Section 3 details numerical simulations of system (1.7) for $N \leq \infty$. As (1.6) is very costly computationally for large N and (1.9) can only be explicitly evaluated in terms of SVD coordinates, (1.7) is inefficient to simulate directly. To keep track of the evolution of SVD coordinates, we derive their dynamics directly under the flow of (1.7).

In the first few examples, we make use of energy function (1.11). In section 3.3.1, we demonstrate that for large N and diagonal matrix completion, bias towards low rank should be viewed as bias towards *minimal* rank. This example also motivates the study of smaller matrices, where the matrix size and the minimal rank of minimizers are comparable. The example in section 3.3.2 is on 2×2 matrices; yet it illustrates the effect of depth, demonstrating the validity of the infinite depth limit.

The examples in section 3.4.1 show simulation outcomes for different choices of energy functions in the family (1.10). For these examples we chose $\mathcal{B}$ such that $E_{\mathcal{B}}$ has a finite number of rank-deficient minimizers. We see that not all of the rank-deficient minimizers are actually observed as training outcomes, contradicting the idea that low rank accurately predicts simulation outcomes. Instead, we find that the observed minimizers are the minimum-rank minimizers with maximal volume.

Section 3.4.3 details an example where the entire state space has minimal rank. In this case, it is impossible to predict simulation outcomes in terms of bias towards low rank. We demonstrate that high state space volume remains predictive, and the simulation outcomes are clustered in the region of state space where the volume is maximal.

1.4. Organization of the paper. The rest of this paper is organized as follows. In section 2, we review the Riemannian geometry of the DLN and establish equations (1.12)–(1.13). We then extend this geometry to the infinite depth limit and prove Theorem 1.1. In section 3 we use numerical experiments to show the correspondence between state space

volume and implicit regularization. We also present several comparisons between finite depth networks and the infinite depth limit. Our conclusions are summarized in section 4.

2. The Riemannian geometry of DLN. This section contains several results on the state space geometry for the DLN. We first review the balanced manifold as well as the Riemannian submersion that determines the Riemannian metric for the DLN, basing our discussion on past work [2, 3]. This is followed by computations in SVD coordinates that provide matrix representations of the metric, including the infinite depth limit $N \to \infty$ (equations (1.12)–(1.13)). Finally, we prove Theorem 1.1 on the volume forms.

2.1. Riemannian submersion of the balanced manifold. The geometry on $GL(d)$ is defined by a Riemannian submersion of the balanced manifold, introduced in [3] and illustrated in Figure 1. This observation effectively reduces the dynamics of the DLN to a space of dimension d^2, independent of the depth N.

Recall that $\mathbf{W} = (W_N, W_{N-1}, \dots, W_1) \in \mathbb{M}_d^N$ (see (1.1)).

Definition 2.1 (balanced manifold). *The balanced manifold of full-rank matrices is*

$$\mathcal{M} := \{\mathbf{W} | W_j \in GL(d), \text{ and } G_j = 0 \text{ for } j \in 1 \dots N-1\}. \tag{2.1}$$

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

Lemma 2.2 (invariant manifold). *The balanced manifold is invariant under the gradient flow* (1.3).

Proof. The invariance of G_j (not just $G_j = 0$) along trajectories can be checked by direct computation of the derivative along trajectories of (1.3):

$$\dot{G}_j = \dot{W}_{j+1}^T W_{j+1} + W_{j+1}^T \dot{W}_{j+1} - \dot{W}_j W_j^T - W_j \dot{W}_j^T = 0. \tag{2.2}$$

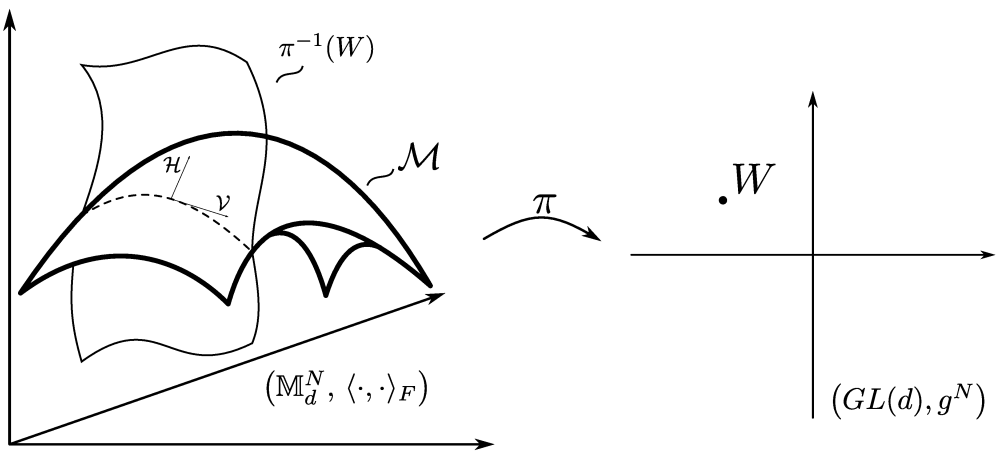


Figure 1. *The left side is the optimization space and shows the balanced manifold as an immersed manifold. The right side represents the space of trained networks equipped with a Riemannian metric given by the Riemannian submersion of the balanced manifold.*

We need to show that (2.1) implicitly defines a manifold. For that we need to check whether its differential is surjective as a linear map from $\mathbb{M}_d \times \mathbb{M}_d$ to the space of $d \times d$ symmetric matrices. In other words, we need to show that for $W_j, W_{j+1} \in GL(d)$ and arbitrary symmetric S the equation

$$\dot{W}_{j+1}^T W_{j+1} + W_{j+1}^T \dot{W}_{j+1} - \dot{W}_j W_j^T - W_j \dot{W}_j^T = S \tag{2.3}$$

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

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

Lemma 2.3 (symmetries). *For $Q \in O(d)$, let $L_i(Q)$ denote the linear map*

$$L_i(Q)(\mathbf{W}) = \left(W_N, W_{N-1}, \dots, W_{i+1}Q, Q^T W_i, \dots, W_1\right). \tag{2.4}$$

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

Proof. Both statements are obtained through direct computations. In order to see that $\pi\left(L_i\left(Q\right)\left(\mathbf{W}\right)\right) = \pi\left(\mathbf{W}\right)$, we compute

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

since $QQ^T = I$.

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

$$W_{i+1} Q Q^T W_{i+1}^T = W_{i+1} W_{i+1}^T = W_{i+2}^T W_{i+2}, \tag{2.5}$$

$$Q^T W_{i+1}^T W_{i+1} Q = Q^T W_i W_i^T Q, \tag{2.6}$$

and

$$W_i^T Q Q^T W_i Q = W_i^T W_i = W_{i-1} W_{i-1}^T. \tag{2.7}$$

∎

2.2. The operator $\mathcal{A}_{N,W}$ and the metric g^N. The metric g^N was defined in [3] using the linear operator $\mathcal{A}_{N,W} : T_W GL(d) \to \mathbb{M}_d$ (see (1.6)–(1.8) and [3, Def. 3]). We find it convenient to represent $\mathcal{A}_N$ and g^N in SVD coordinates since this yields explicit formulas for the metric and volume form. In all that follows we write the SVD of a matrix W as

$$W = U \Sigma V^T, \quad \Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_d), \quad \sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_d. \tag{2.8}$$

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

Lemma 2.4 (spectral decomposition of $\mathcal{A}_{N,W}$). *For finite N, the d^2 number of eigenvalues of $\mathcal{A}_{N,W}$ are*

$$\lambda_{il}^N = \frac{1}{N}\sum_{j=1}^{N} \left(\sigma_i^2\right)^{\frac{N-j}{N}} \left(\sigma_l^2\right)^{\frac{j-1}{N}} = \frac{\left(\sigma_i^2\right)^{\frac{N-1}{N}}}{N} \frac{1-\frac{\sigma_l^2}{\sigma_i^2}}{1-\left(\frac{\sigma_l^2}{\sigma_i^2}\right)^{\frac{1}{N}}}, \quad i,l \in 1,\dots,d. \tag{2.9}$$

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

$$\lambda_{il}^{\infty} = \frac{\sigma_i^2 - \sigma_l^2}{2\log\left(\sigma_i/\sigma_l\right)}, \quad i,l \in 1,\dots,d. \tag{2.10}$$

The corresponding eigenvectors are independent of N and are given by

$$T_{il} = UE_{il}V^T, \quad 1 \le i,l \le d. \tag{2.11}$$

Proof. Let

$$Y = \mathcal{A}_N(X), \tag{2.12}$$

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

$$\tilde{Y} = \frac{1}{N}\sum_{j=1}^{N} \left(\Sigma^2\right)^{\frac{N-j}{N}} \tilde{X} \left(\Sigma^2\right)^{\frac{j-1}{N}}. \tag{2.13}$$

Notice that we have obtained a diagonal form in coordinates:

$$\tilde{y}_{il} = \tilde{x}_{il}\frac{1}{N}\sum_{j=1}^{N} \left(\sigma_i^2\right)^{\frac{N-j}{N}} \left(\sigma_l^2\right)^{\frac{j-1}{N}}, \tag{2.14}$$

which proves (2.9).

The proof of (2.10) is similar. Fix $i \neq l$ and take the limit $N \to \infty$ in (2.14) to obtain

$$\lim_{N\to\infty} \lambda_{il}^N = \lim_{N\to\infty} \frac{1}{N}\sum_{j=1}^{N} \left(\sigma_i^2\right)^{\frac{N-j}{N}} \left(\sigma_l^2\right)^{\frac{j-1}{N}} = \int_0^1 \sigma_i^{2(1-t)}\sigma_l^{2t}dt = \frac{\sigma_i^2 - \sigma_l^2}{\log\left(\sigma_i^2/\sigma_l^2\right)}. \tag{2.15}$$

Finally, when $i = l$

$$\lim_{N\to\infty} \lambda_{il}^N = \lim_{N\to\infty} \left(\sigma_i^2\right)^{\frac{N-1}{N}} = \sigma_i^2. \tag{2.16}$$

■

As a direct consequence of Lemma 2.4, we obtain an explicit matrix representation for the DLN metric (1.8). Let the vectorization of the matrix coordinate one-forms be denoted dw^α, $\alpha = 1,\dots,d^2$. Then the metric $g^N = g_{\alpha\beta}^N dw^\alpha \otimes dw^\beta$, $N \le \infty$, at a point $W = U\Sigma V^T$ is given by

$$g^N = (V \otimes U)\, D^N(\Sigma)\, (V \otimes U)^T, \tag{2.17}$$

where D^N is the diagonal operator whose nonzero entries are

$$\frac{1}{\lambda^N_{il}}, \quad 1 \le i,l \le d. \tag{2.18}$$

The above calculations prove (1.12)–(1.13).

The following inequalities are used to establish normal hyperbolicity of an invariant manifold in Theorem 1.2 below.

Lemma 2.5 (inequalities on the spectrum of $\mathcal{A}_{\infty,W}$). *Let $1 \le i < j \le k < l \le d$; then $\lambda^\infty_{jj} \le \lambda^\infty_{ij} \le \lambda^\infty_{ii}$ and $\lambda^\infty_{ij} \le \lambda^\infty_{kl}$.*

Proof. $\lambda^\infty_{jj} = \sigma_j^2 \le \sigma_i^2 = \lambda^\infty_{ii}$ since $\sigma_j \le \sigma_i$.

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

$$f'(y) \le \frac{f(y) - f(x)}{y - x} \le f'(x). \tag{2.19}$$

Thus

$$\frac{d}{dx}\log(x)|_{\sigma_j^2} = \frac{1}{\sigma_j^2} \ge \frac{\log\sigma_j^2 - \log\sigma_i^2}{\sigma_j^2 - \sigma_i^2} \ge \frac{d}{dx}\log(x)|_{\sigma_i^2} = \frac{1}{\sigma_i^2}, \tag{2.20}$$

where the middle term is

$$\frac{\log\sigma_j^2 - \log\sigma_i^2}{\sigma_j^2 - \sigma_i^2} = \frac{1}{\lambda^\infty_{ij}}. \tag{2.21}$$

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

2.3. Volume forms and the proof of Theorem 1.1. Vandermonde determinants appear often in random matrix theory as the Jacobians for diagonalization (see, for example, [8, sect. 5.3], [17, sect. 2.2], and Lemma 2.7 below). Theorem 1.1 reflects an analogous feature of the DLN geometry. The proof is an easy consequence of Lemma 2.4.

Proof of Theorem 1.1. We compute the determinants of the matrices g^N and g^∞ as a product of the reciprocal of the eigenvalues of $\mathcal{A}_{N,W}$ given in Lemma 2.4. For g^∞ we find

$$\begin{aligned}\sqrt{\det g^\infty}dW &= \sqrt{\frac{1}{\sigma_1^2\dots\sigma_d^2}\prod_{i\ne j}\frac{\log\left(\sigma_i^2\right) - \log\left(\sigma_j^2\right)}{\sigma_i^2 - \sigma_j^2}}dW \\ &= \frac{1}{\sqrt{\det\left(\Sigma^2\right)}}\prod_{i<j}\frac{\log\left(\sigma_i^2\right) - \log\left(\sigma_j^2\right)}{\sigma_i^2 - \sigma_j^2}dW \\ &= \frac{\operatorname{van}\left(\log\Sigma^2\right)}{\sqrt{\det\left(\Sigma^2\right)\operatorname{van}\left(\Sigma^2\right)}}dW.\end{aligned} \tag{2.22}$$

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

$$\begin{aligned}\lambda_{il}^N &= \frac{1}{N}\sum_{j=1}^{N}\left(\sigma_i^2\right)^{\frac{N-j}{N}}\left(\sigma_l^2\right)^{\frac{j-1}{N}} = \frac{\left(\sigma_i^2\right)^{\frac{N-1}{N}}}{N}\left(1+\left(\frac{\sigma_l^2}{\sigma_i^2}\right)^{\frac{1}{N}}+\cdots+\left(\frac{\sigma_l^2}{\sigma_i^2}\right)^{\frac{N-1}{N}}\right)\\ &= \frac{\left(\sigma_i^2\right)^{\frac{N-1}{N}}}{N}\frac{1-\frac{\sigma_l^2}{\sigma_i^2}}{1-\left(\frac{\sigma_l^2}{\sigma_i^2}\right)^{\frac{1}{N}}}.\end{aligned} \tag{2.23}$$

In the case $i=l$ we obtain instead

$$\lambda_{il}^N = \left(\sigma_i^2\right)^{\frac{N-1}{N}}. \tag{2.24}$$

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

$$\begin{aligned}\sqrt{\det g^N}dW &= \sqrt{\prod_{i=1}^{d}\frac{1}{\left(\sigma_i^2\right)^{\frac{N-1}{N}}}\prod_{i\neq l}N\left(\sigma_i^2\right)^{\frac{1-N}{N}}\frac{1-\left(\frac{\sigma_l^2}{\sigma_i^2}\right)^{\frac{1}{N}}}{1-\frac{\sigma_l^2}{\sigma_i^2}}}dW\\ &= \frac{N^{\frac{d(d-1)}{2}}\det(\Sigma^2)^{\frac{1-N}{2N}}}{V(\Sigma^2)}V\left(\Sigma^{\frac{2}{N}}\right)dW.\end{aligned} \tag{2.25}$$

Finally, we use Lemma 2.7 below to express dW in SVD coordinates, completing the proof of Theorem 1.1. ∎

Remark 2.6 (volume form in new coordinates). Notice that using the coordinates $\Lambda=\log(\Sigma)$, we have $\frac{d\lambda_i}{d\sigma_i}=\frac{1}{\sigma_i}$ and thus

$$\sqrt{\det g^\infty}dW = 2^{\frac{d(d-1)}{2}}\mathrm{van}\Lambda\, d\Lambda dU\, dV. \tag{2.26}$$

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

2.4. The Jacobian of SVD. The following lemma is included for completeness since we were unable to find a convenient reference.

Lemma 2.7. *The Jacobian determinant of the SVD map:* $W\mapsto(U,\Sigma,V)\in O(d)\times\mathbb{R}^d\times O(d)$ *is given by the Vandermonde determinant* $\mathrm{van}(\Sigma^2)$.

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

$$\dot W = \dot U\Sigma V^T + U\dot\Sigma V^T + U\Sigma\dot V^T. \tag{2.27}$$

Since U,V are orthogonal and Σ is diagonal,

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

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

$$\dot{W} = U(A^U\Sigma + D + \Sigma A^V)V^T. \tag{2.28}$$

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

$$\mathcal{L}: TO(d)\times TO(d)\times T\mathbb{R}^d \to \mathbb{R}^{d\times d}, \quad (A^U, A^V, D)\mapsto A^U\Sigma + D + \Sigma A^V. \tag{2.29}$$

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

$$\sum_j a^U_{ij}\delta_{jk}\sigma_j + \delta_{ij}\sigma_i a^V_{jk}. \tag{2.30}$$

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

$$\begin{bmatrix} D_1 & D_2 \\ -D_2 & -D_1 \end{bmatrix}\begin{bmatrix} \operatorname{vec}(A^U) \\ \operatorname{vec}(A^V) \end{bmatrix}, \tag{2.31}$$

where D_1 is

$$D_1 = \operatorname{diag}(\underbrace{\sigma_1,\dots,\sigma_1}_{d-1},\underbrace{\sigma_2,\dots,\sigma_2}_{d-2},\dots,\underbrace{\sigma_{d-1}}_{1}), \tag{2.32}$$

and D_2 is

$$D_2 = \operatorname{diag}(\underbrace{\sigma_2,\dots,\sigma_d}_{d-1},\underbrace{\sigma_3,\dots,\sigma_d}_{d-2},\dots,\underbrace{\sigma_{d-1},\sigma_d}_{2},\underbrace{\sigma_d}_{1}). \tag{2.33}$$

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

$$|\det(D_2^2 - D_1^2)| = |\det\operatorname{diag}(\sigma_i^2 - \sigma_j^2,\, i>j)| = \prod_{i<j}(\sigma_i^2 - \sigma_j^2) = \operatorname{van}(\Sigma^2). \tag{2.34}$$

■

3. Dynamics of matrix completion. This section is devoted to the gradient dynamics of the DLN for matrix completion. We consider quadratic energy functions as in (1.10) and (1.11). As $\mathcal{B}$ varies, the energy $E_{\mathcal{B}}$ may have a unique minimum or a submanifold of minima. By (1.7) the dynamics of the DLN is determined by an interplay between the Riemannian geometry and the nature of $E_{\mathcal{B}}(W)$.

This section primarily focuses on describing numerical experiments. However, in order to generate efficient numerical simulations, it is necessary to first express the dynamics in SVD

coordinates. These expressions are summarized in Theorem 3.2. We also prove Theorem 1.2 on normal hyperbolicity of submanifolds of equilibria to shed light on the attraction to the balanced manifold.

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

3.1. Numerical integration of (1.7) using SVD. Numerically integrating (1.7) gets very costly with increasing depth, due to the large number of matrix powers needed. Alternatively, one can use the factorized formula (2.17) to compute the Riemannian gradient through the dual metric:

$$\dot{w} = -g^{N*}(w)\partial E(w). \tag{3.1}$$

This formulation, however, requires the SVD of W at every time step. Instead, we compute the SVD of the initial condition and directly evolve the singular coordinates using smooth SVD. All numerical results in this paper are obtained using the fourth order, fixed time step Runge–Kutta method. This formulation is stated in Theorem 3.2, building on the well-known result Lemma 3.1, which we review for completeness.

Lemma 3.1 (smooth SVD). *Given a smooth curve $W(t) : (t_1, t_2) \to GL(d)$, $W(t)$ having distinct singular values for all $t \in (t_1, t_2)$, a smooth SVD $W(t) = U(t)\Sigma(t)V(t)^T$ exists satisfying the following system of differential equations:*

$$\dot{\sigma}_i = u_i^T \dot{W} v_i, \tag{3.2}$$

$$\dot{u}_i = \sum_{j \neq i} \frac{1}{\sigma_i^2 - \sigma_j^2} \langle (\dot{W}W^T + W\dot{W}^T)u_i, u_j \rangle u_j, \tag{3.3}$$

$$\dot{v}_i = \sum_{j \neq i} \frac{1}{\sigma_i^2 - \sigma_j^2} \langle (\dot{W}^T W + W^T \dot{W})v_i, v_j \rangle v_j. \tag{3.4}$$

Proof. Under our assumptions these formulas can be verified by direct computation. For a more careful treatment for dealing with repeated singular values, see [6]. ∎

Using Lemma 3.1 and (3.1), we can write down the evolution equations for the singular coordinates $U(t)\Sigma(t)V^T(t) = W(t)$ directly. For $N < \infty$, this result is equivalent to statements in [2] (see Theorem 3 and Lemma 2). Let $\mathfrak{s}(M) := M - M^T$, $\alpha = 1 - 1/N$, and let $\circ$ denote the Hadamard (elementwise) product. For $N \leq \infty$ introduce the matrix

$$L_N^{il} = \begin{cases} \frac{\lambda_{il}^N}{\sigma_i^2 - \sigma_l^2} & \text{for } i \neq l, \\ 0 & \text{otherwise.} \end{cases} \tag{3.5}$$

Theorem 3.2. *Under the assumptions of Lemma 3.1, and differentiable loss function E, the SVD of the end-to-end matrix evolves according to the differential equations*

$$\dot{U} = U\mathfrak{s}\left(\left(L_N(\Sigma)\Sigma\right)\circ\left(U^T\partial_W E\, V\right)\right), \tag{3.6}$$

$$\dot{\Sigma} = -\Sigma^{2\alpha}\mathrm{diag}\left(U^T\partial_W E\, V\right), \tag{3.7}$$

$$\dot{V} = V\mathfrak{s}\left(\left(\Sigma L_N(\Sigma)\right)\circ\left(U^T\partial_W E\, V\right)\right). \tag{3.8}$$

Proof. For $N<\infty$, see Theorem 3 and Lemma 2 in [2]. Under infinite depth, direct computation using (3.1) and Lemma 3.1 verifies the result. ∎

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

Definition 3.3 (effective rank). *Let $W\in GL(d)$ with singular values σ_i and let s_i denote the normalized singular values*

$$s_i = \frac{\sigma_i}{\sum_j \sigma_j}. \tag{3.9}$$

Then the effective rank of W is

$$r_e(W) = \exp\left(-\sum_i s_i \log(s_i)\right). \tag{3.10}$$

3.2. Attraction rates and the proof of Theorem 1.2. We derive bounds on the normal attraction rates for the submanifold of equilibria $\mathcal{N}_{\mathcal{B}}$. Given a choice of $\mathcal{B}$, let $\mathcal{I}$ denote the vectorized index set of observed elements, that is,

$$\mathrm{vec}(\mathcal{B})_i = \begin{cases} 1 & \text{if } i\in\mathcal{I}, \\ 0 & \text{otherwise.}\end{cases} \tag{3.11}$$

In the following lemma, we compute the linearization of (1.7) in the normal direction of $\mathcal{B}_{\mathcal{N}}$. We use coordinates $W=\mathcal{B}\circ\Phi+X+Y$, where $\mathcal{B}\circ X=X$ and $\mathcal{B}\circ Y=0$. Lowercase x,y, and w denote the vectorization of these coordinates. The dual metric tensor (the matrix of which is represented by the inverse of g^N) is denoted g^{N*}. Let $W_0\in\mathcal{N}_{\mathcal{B}}$ be a fixed point.

Lemma 3.4. *At W_0, the linearization of (1.7) in the normal direction x is given by*

$$\dot{x} = -Ax, \tag{3.12}$$

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

$$A = \left[g^{N*}\big|_{W_0}\right]_{i,j\in\mathcal{I}}. \tag{3.13}$$

Proof. Using the general form of the Riemannian gradient flow (3.1) and taking a derivative

$$\frac{d}{d\epsilon} g^{N*}(\epsilon x, y)\partial E_I((\epsilon x, y)) = \left(\partial_x g^{N*}\right)\partial E_I(w_0) + g^{N*}(w_0)\partial^2 E(w_0)x, \tag{3.14}$$

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

$$\partial^2 E_i(W_0)_{ij} = \begin{cases} 1 & \text{if } i = j, i \in \mathcal{I}, \\ 0 & \text{otherwise,} \end{cases} \tag{3.15}$$

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

Proof of Theorem 1.2. We show a nonzero lower bound on the attraction rates. Let $\{\alpha_i\}_1^d$, $\alpha_i \le \alpha_j$ if $j \le i$ denote the eigenvalues of A. Then by the Cauchy interlacing theorem and Lemma 2.5,

$$\sigma_d^2 \le \alpha_d \le \alpha_{d-1} \le \cdots \le \alpha_1, \tag{3.16}$$

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

Notice that Lemma 2.5 and the Cauchy interlacing theorem completely characterizes the order of λ_{ij}^{∞} and the characteristic exponents of the $d = 2$ case:

$$\sigma_2^2 \le \alpha_2 \le \frac{\sigma_1^2 - \sigma_2^2}{\log \sigma_1^2 - \log \sigma_2^2} \le \alpha_1 \le \sigma_1^2. \tag{3.17}$$

3.3. Diagonal matrix completion.

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

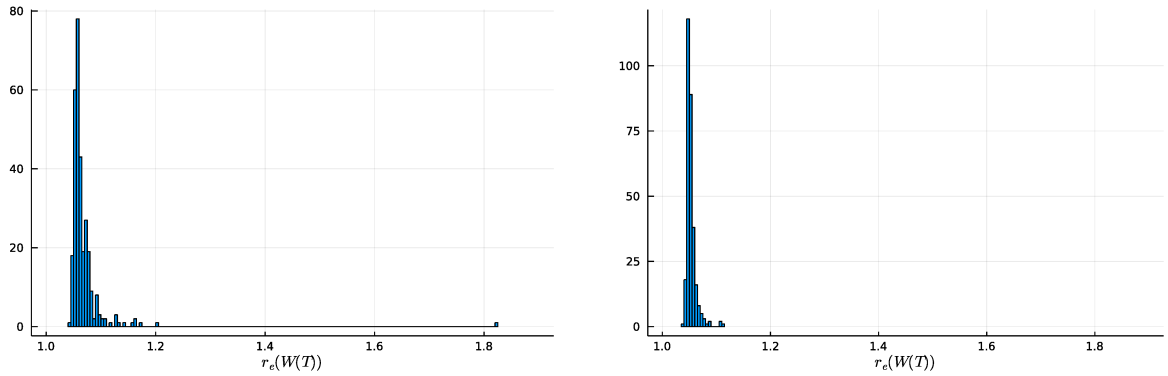
3.3.1. Example: $d = 20$.

We start with simulations of a larger, $d = 20$, example. Figures 2 and 3 show histograms of effective rank of optimization outcomes. Note that the rank and thus the effective rank here can be as large as 20; the size of the matrices and therefore bias towards low rank could mean convergence to a matrix of any effective rank smaller than 20.

Notice that even the shallowest case ($N = 3$) shows strong bias towards low-rank completions. In case of $N = 10$ and $N = \infty$, the obtained histograms are nearly identical, showing strong bias towards minimal, rank-one outcomes. This suggests that building intuition around the dynamics under sufficient depth is possible using small ($d = 2$ or $d = 3$) examples, in which case the rank-deficient cases (1.3) are easier to characterize.

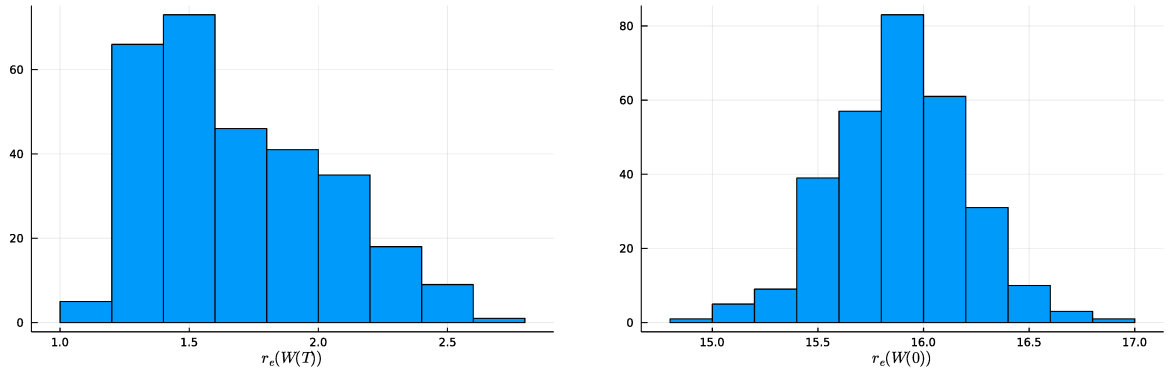
Approximately 300 optimization outcomes are included for each N. Figure 3(b) shows the distribution of effective rank upon initialization. The small random initial conditions are drawn from a Wigner ensemble; specifically, they consist of matrices of independent normal entries of mean zero and standard deviation 0.001, which distribution is denoted Wigner$(0, 0.001)$.

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



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

Figure 2. *Empirical distributions of effective rank in* 20×20 *matrix completion simulations.*



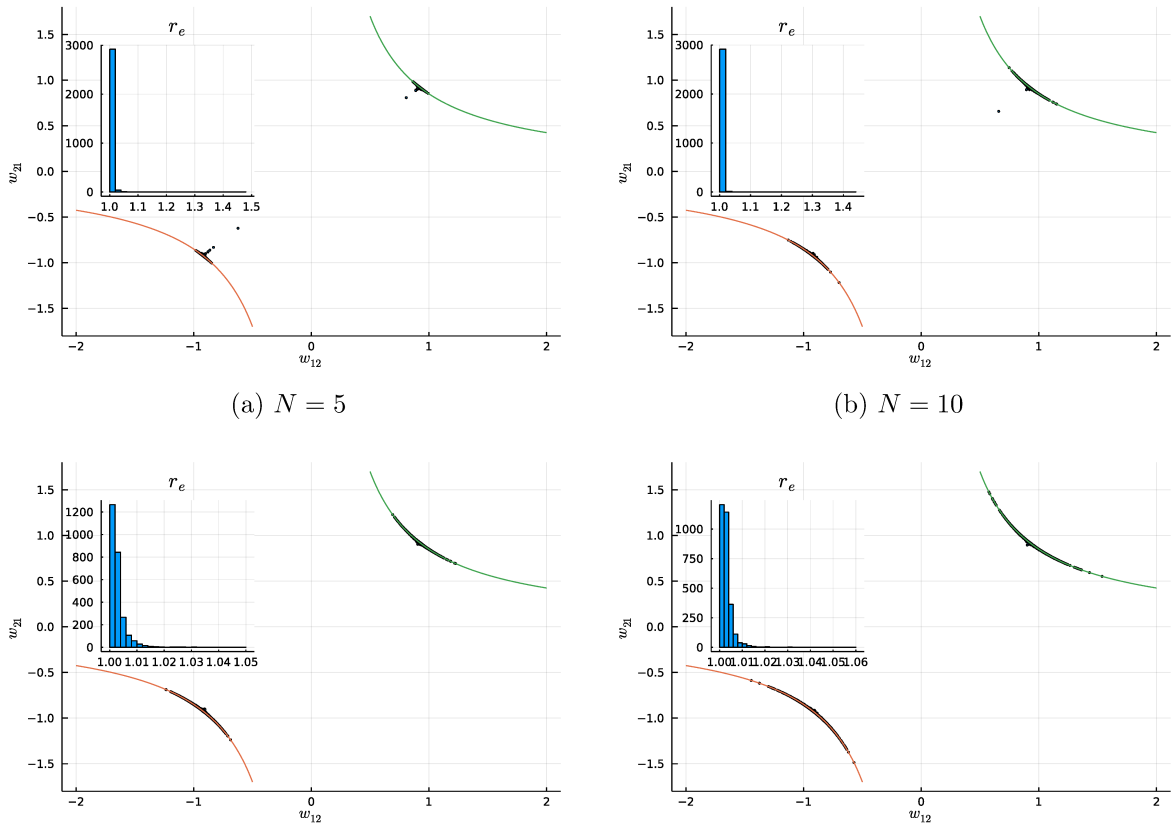
(a) $N = 3$ (b) Initialization

Figure 3. (a) *Effective rank distributions of shallow matrix completion simulations.* (b) *Sample distribution of effective rank at initialization for all above examples.*

3.3.2. Example: $d = 2$. In the following example for matrix completion, we illustrate how an increase in depth influences training outcome and at what depth do the results start to closely resemble our infinite depth limit. For easier visualization and building on the intuition developed in the previous section, we keep the width at the minimum $d = 2$. The results shown in Figure 4 are obtained by approximately 3000 runs for $N = 5, 10, 20$, and infinity each. Each of these figures shows the outcome of these batch simulations.

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

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



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

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

Figure 4. *Outcomes of batch simulations under $N = 5, 10, 20,$ and ∞, respectively. The black dots indicate training outcomes, while the red and green hyperbola lobes visualize rank-one minimizers. Sample distributions of effective rank are also included in the embedded graphs.*

Notice that here any matrix with fixed diagonal elements

$$\begin{bmatrix} \Phi_1 & w_{12} \\ w_{21} & \Phi_2 \end{bmatrix} \tag{3.18}$$

is a global minimizer. Batch simulations allow us to see whether there is a concentration of training outcomes on the w_{12}, w_{21} plane of possible global minimizers. Figure 4 shows the results of these simulations, and Figure 5 shows the clear correspondence to high phase space volume. The hyperbolas in the left panel correspond to the rank-one singularities on the manifold defined by

$$\frac{\Phi_1}{w_{21}} = \frac{w_{12}}{\Phi_2}. \tag{3.19}$$

At these singularities phase space volume blows up for any depth, and the probability of landing in this high volume region is expected to be high. The logarithmic volume can be interpreted as an entropic quantity [16], suggesting the use of statistical mechanical tools in our future analysis.

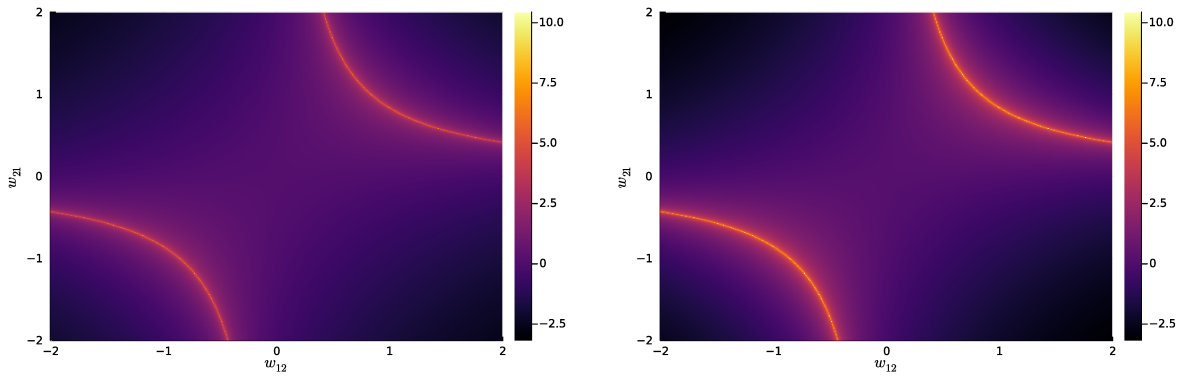


Figure 5. *Heatmap of the logarithmic volume on the plane of global minimizers for $N = 5$ and $N = \infty$, respectively. Note that the volume itself does not depend on the loss function, only the plane where the section was taken.*

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

Based on the example in section 3.3.1 and the results in Figure 4, we conclude that the infinite depth limit shows very similar behavior qualitatively to finite, sufficiently large depth models. Thus it is worth studying and a predictive theory of implicit regularization must be consistent with the infinite depth model.

Comparing Figures 4 and 6, we see that there is no clustering of simulation outcomes in high effective rank regions. However, effective rank does not take the effect of depth into consideration. As seen in Figure 5, state space volume shows higher concentration at higher depth, just as simulation outcomes in higher depth show more concentration in these regions. While we are not ready to make this relation quantitative, these results suggest that state space volume is a good candidate to rely on for a quantitative theory of implicit regularization.

We also see that the accumulation of training outcomes is strong near the corners of the hyperbola of rank-one completions. If high volume is a predictive quantity for training outcomes, we expect a subtle decrease of volume along the hyperbola starting from the corners. We characterize the blowup rates along the hyperbola by computing the normal perturbation of singular values and volume density along the hyperbola.

To simplify calculations, we fix the diagonal elements to 1, getting the one-parameter family of rank-one completions

$$W = \begin{bmatrix} 1 & \gamma \\ \frac{1}{\gamma} & 1 \end{bmatrix}. \tag{3.20}$$

To simplify some of the formulas, we assume $\gamma > 0$, restricting the analysis on the positive lobe of the hyperbola. The SVD n $W(\gamma) = U(\gamma)\Sigma(\gamma)V^T(\gamma)$ can be explicitly computed:

$$U = \frac{1}{\sqrt{1+\gamma^2}} \begin{bmatrix} \gamma & -1 \\ 1 & \gamma \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \gamma + \gamma^{-1} & 0 \\ 0 & 0 \end{bmatrix}, \quad V = \frac{1}{\sqrt{1+\gamma^2}} \begin{bmatrix} 1 & \gamma \\ \gamma & -1 \end{bmatrix}. \tag{3.21}$$

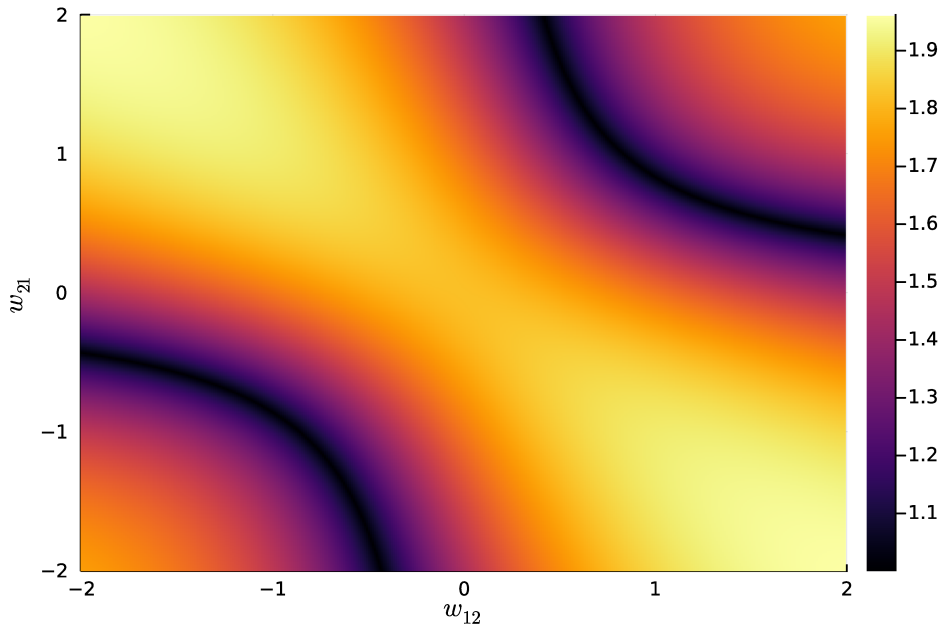


Figure 6. *Effective rank of global minimizers.*

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

$$\dot{W} = \frac{\eta}{\sqrt{1+\gamma^{-4}}} \begin{bmatrix} 0 & \gamma^{-2} \\ 1 & 0 \end{bmatrix}. \tag{3.22}$$

Using Lemma 3.1, the perturbed singular values can be computed:

$$\Sigma(\gamma,\eta) = \begin{bmatrix} \gamma+\gamma^{-1}+\frac{2\eta}{(1+\gamma^2)\sqrt{1+\gamma^{-4}}} & 0 \\ 0 & \eta\frac{\sqrt{\gamma^4+1}}{\gamma^2+1} \end{bmatrix} + \mathcal{O}(\eta^2). \tag{3.23}$$

And consequently, up to leading order, the determinant is

$$\det(\Sigma(\gamma,\eta)) = \eta\frac{\sqrt{\gamma^4+1}}{\gamma} + \mathcal{O}(\eta^2). \tag{3.24}$$

Plugging this back into the volume density function from (2.22),

$$\begin{aligned} \frac{\operatorname{van}(\log\Sigma^2)}{\det(\Sigma)\operatorname{van}(\Sigma^2)} &= \frac{2\gamma\left(\log\left(\gamma+\gamma^{-1}+\frac{2\eta}{(\gamma^2+1)\sqrt{1+\gamma^{-4}}}\right)-\log\left(\eta\frac{\sqrt{\gamma^4+1}}{\gamma^2+1}\right)\right)}{\left(\eta\sqrt{\gamma^4+1}+\mathcal{O}(\eta^2)\right)\left(\left(\gamma+\gamma^{-1}+\frac{2\eta}{(1+\gamma^2)\sqrt{1+\gamma^{-4}}}\right)^2-\left(\eta\frac{\sqrt{\gamma^4+1}}{\gamma^2+1}\right)^2\right)} \\ &= \mathcal{O}\left(\frac{|\log\eta|}{\eta}\right) \end{aligned} \tag{3.25}$$

as $\eta \to 0$. We gained quantitative understanding of how the singular values perturb. Specifically, we learn that the leading order coefficient of the smaller singular value, σ_2, is $\frac{\sqrt{\gamma^4+1}}{\gamma^2+1}$. This function has a minimum at $\gamma = 1$, the corner of the hyperbola. This shows a quantitative, but not qualitative, decrease of volume along the hyperbolas. The quickest blowup of the volume density is at the corners of the hyperbola.

3.4. Other configurations.

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

$$\mathcal{T} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \tag{3.26}$$

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

$$\begin{bmatrix} \Phi_{11} & \Phi_{12} \\ w_{21} & \Phi_{11} \end{bmatrix} \tag{3.27}$$

for arbitrary w_{21}, while imposing rank deficiency gives one solution, $w_{21} = \Phi_{11}\Phi_{22}/\Phi_{12}$. All simulation outcomes of 1000 random initial conditions drawn from Wigner$(0, 0.001)$ showed convergence to this minimizer. As an illustration, five sample trajectories are shown in Figure 7.

3.4.2. A 3×3 example. We continue with a more complicated $N = \infty$, $d = 3$ example. Computing the minimal rank to which a partially known matrix can be completed is in general nontrivial. For a detailed discussion of the computational complexity of matrix completion, see [11]. We covered cases before where either the small matrix size or the symmetry of the configuration in the observed elements (diagonal matrix completion) makes this problem trivial.

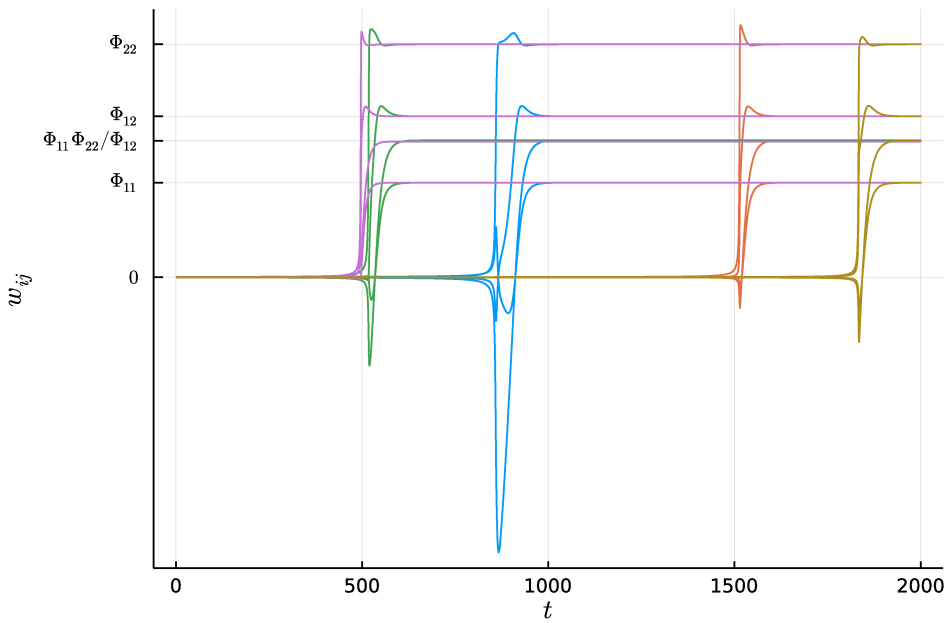


Figure 7. *Five sample trajectories of the matrix elements under training are shown. All of them converge to the rank-deficient minimizer.*

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

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

$$
\begin{aligned}
M(w_{31}, w_{12}, w_{23}) &= \begin{bmatrix} \Phi_{11} & w_{12} & \Phi_{13} \\ \Phi_{21} & \Phi_{22} & w_{23} \\ w_{31} & \Phi_{32} & \Phi_{33} \end{bmatrix} \\
&= \begin{bmatrix} \Phi_{11} \\ \Phi_{21} \\ w_{31} \end{bmatrix} \begin{bmatrix} 1 & \frac{w_{12}}{\Phi_{11}} & \frac{\Phi_{13}}{\Phi_{11}} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \Phi_{22} - \frac{\Phi_{21} w_{12}}{\Phi_{11}} & w_{23} - \frac{\Phi_{13}\Phi_{21}}{\Phi_{11}} \\ 0 & \Phi_{32} - \frac{w_{31} w_{12}}{\Phi_{11}} & \Phi_{33} - \frac{w_{31}\Phi_{13}}{\Phi_{11}} \end{bmatrix}.
\end{aligned} \tag{3.28}
$$

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

$$
0 = \Phi_{33} - \frac{\Phi_{13}}{\Phi_{11}} w_{31} - \frac{\left(\Phi_{32} - \frac{w_{31} w_{12}}{\Phi_{11}}\right)\left(w_{23} - \frac{\Phi_{13}\Phi_{21}}{\Phi_{11}}\right)}{\Phi_{22} - \frac{\Phi_{21} w_{12}}{\Phi_{11}}}. \tag{3.29}
$$

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

$$
w_{12}^1 = \frac{\Phi_{32}\Phi_{13}}{\Phi_{33}},\ w_{23}^1 = \frac{\Phi_{13}\Phi_{21}}{\Phi_{11}},\ w_{31}^1 = \frac{\Phi_{33}\Phi_{11}}{\Phi_{13}}. \tag{3.30}
$$

Let M denote the zero energy matrix with coordinates (3.30).

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

$$
\Phi = \begin{bmatrix} -1.55795 & \cdot & 1.58397 \\ 0.212869 & 0.0337805 & \cdot \\ \cdot & 1.32488 & 1.92653 \end{bmatrix}. \tag{3.31}
$$

Figure 8 shows the outcome of strong clustering in a region near M. We interpret this as a shortcoming of the rank based predictions, as the entire two-parameter family of minimizers defined by (3.29) has rank two. We attempt to explain the observation using phase space volume. We approximate phase space volume within the zero energy plane spanned by coordinates (w_{12}, w_{31}, w_{23}) upon simple Monte Carlo integration. We integrate the volume form (1.16) in a small cube around some rank-two minimizers.

Since these points are singular, volume form (1.16) blows up. As a result of this, the volume in a domain containing these points can be infinity. The volume, however, is locally finite, so working with domains bounded away from singular points allows us to make quantitative comparison between regions of the state space. Given a rank-two minimizer, we can study the concentration of volume around it by simple Monte Carlo integration of the volume density (2.22).

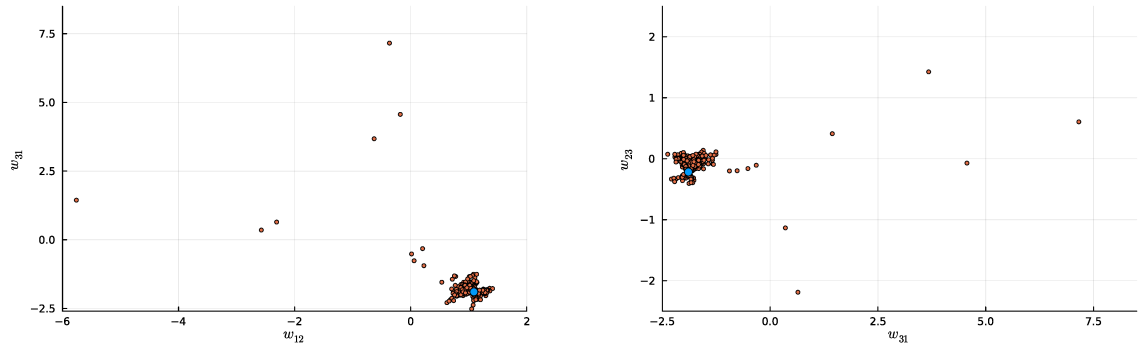


Figure 8. *Clustering of training outcomes obtained in the setup of* (3.28). *The majority of the* 500 *outputs cluster near one particular rank-two minimizer indicated by the blue dot. Random initial conditions were drawn from Wigner*$(0, 0.02)$.

Define the set of 3×3 matrices with smallest singular value strictly larger than h,

$$GL_3^h = \{X \in GL(3) : \sigma_3(X) > h\}, \tag{3.32}$$

and the three-dimensional H-cube around a point X,

$$\mathcal{C}_H(X) = \{W \in \mathbb{R}^{d\times d} : \|x_{31} - w_{31}\|_1 < H/2, \|x_{12} - w_{12}\|_1 < H/2, \|x_{23} - w_{23}\|_1 < H/2\}. \tag{3.33}$$

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

To carry out these integrations, random rank-two minimizers with undetermined elements (w_{12}, w_{31}, w_{23}) solving (3.29) are also required. We obtain these matrices by sampling w_{12} and w_{31} from a centered normal distribution of standard deviation 10. Parameters $H = 0.001$ and $h = 0.00001$ were chosen.

The outcome of simulations conducted on M and 24 other randomly selected rank-two minimizers is shown in Figure 9. M is the point of the highest volume vicinity, the difference from other equal-rank minimizers spanning several orders of magnitude. This is in alignment with the clustering of training outcomes.

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

3.4.3. Example: Minimal rank state space. Our last example is a pathological case showing an explicit shortcoming of the rank-hypothesis of implicit regularization. Let $\mathcal{R}^1 = \{X \in \mathbb{M}_2 : \text{rank}(X) = 1\}$. It is shown in Chapter 1 in [3] that the DLN geometry for $N < \infty$ can be defined on the manifold of fixed-rank matrices using (1.8) for the metric.

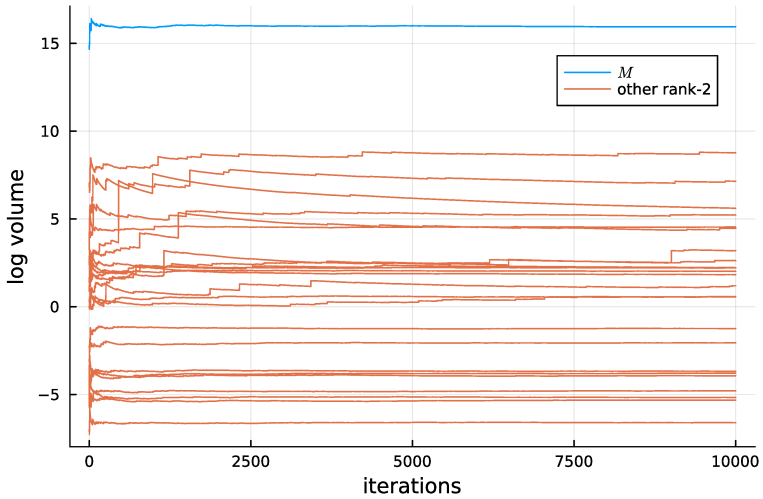


Figure 9. *Monte Carlo integration of volume around rank-two minimizers.*

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

$$\sqrt{G_1^N(W)} = \frac{N}{\sigma^{3\frac{N-1}{N}}}, \tag{3.34}$$

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

For numerical simulations, take the setup of section 3.3.2, save for setting $N = 20$ and the change in the generation of random initial conditions. We generate samples $u\sigma v^T = W_0' \sim \text{Wigner}(0, 0.001)$, then use

$$W_0 = u \begin{bmatrix} \sigma_1 & 0 \\ 0 & 0 \end{bmatrix} v^T \tag{3.35}$$

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

$$\begin{bmatrix} \Phi_1 & w_{12} \\ \frac{\Phi_2\Phi_1}{w_{12}} & \Phi_2 \end{bmatrix} = \begin{bmatrix} 0.58724 & w_{12} \\ \frac{0.8497}{w_{12}} & 1.447 \end{bmatrix}. \tag{3.36}$$

Volume form (3.34) shows that highest volume is obtained at smallest σ, but notice that in this case the volume form does not blow up. Nonetheless, we can predict highest concentration of training outcomes at the minimal singular value completions. Simple computation verifies that these are the same as the symmetric completions,

$$\begin{bmatrix} 0.58724 & \pm 0.921811 \\ \pm 0.921811 & 1.447 \end{bmatrix}, \tag{3.37}$$

which both admit $\sigma_{\min} = 2.03424$. Thus, based on phase space volume, matrices with singular value at $\sigma_{\min}$ and slightly above are expected to have high representation among training

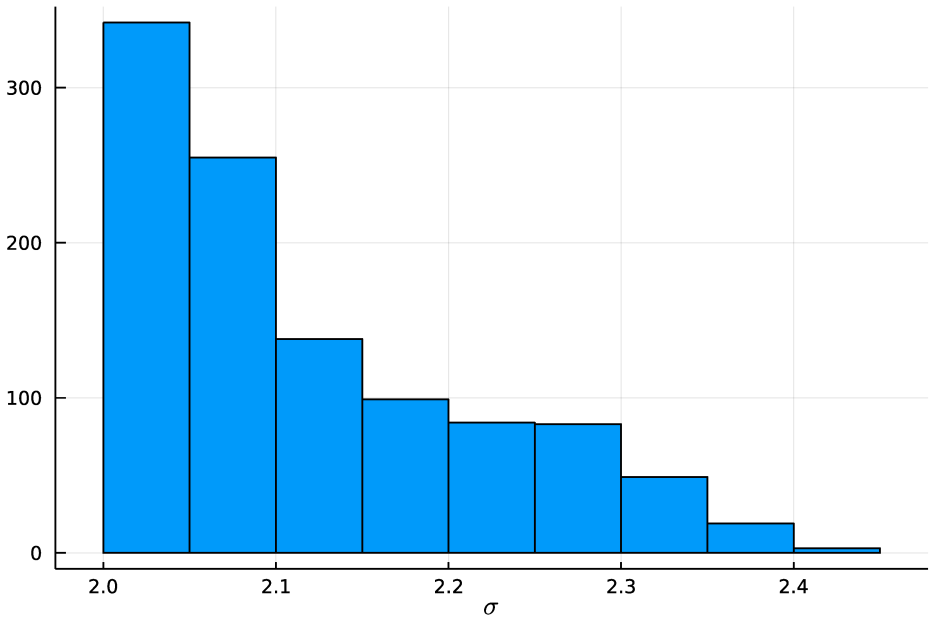


Figure 10. *Sample distribution of* 1072 *simulation outcomes of rank-one, diagonal matrix completion.*

outcomes. Figure 10 shows the outcome of 1072 simulations using random initial conditions verifying the predictions.

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

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

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

Looking forward, we would like to continue our study on the DLN geometry. The simplicity of the formulas presented in this paper suggests that further intrinsic quantities could be expressed using explicit formulas. The derivation of higher order geometric quantities could give rise to new results on the dynamics. The precise formulation of stochastic training dynamics requires the notion of Brownian motion on a Riemannian manifold. The stochastic differential equation describing intrinsic Brownian motion relies on computations of the Levi–Civita connection and curvatures [13, Chapter 3].

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

REFERENCES

[1] S. Arora, N. Cohen, N. Golowich, and W. Hu, *A convergence analysis of gradient descent for deep linear neural networks*, in 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, 2019, OpenReview.net, https://openreview.net/forum?id=SkMQg3C5K7.

[2] S. Arora, N. Cohen, W. Hu, and Y. Luo, *Implicit regularization in deep matrix factorization*, in Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, Vancouver, 2019, pp. 7411–7422, https://proceedings.neurips.cc/paper/2019/hash/c0c783b5fc0d7d808f1d14a6e9c8280d-Abstract.html.

[3] B. Bah, H. Rauhut, U. Terstiege, and M. Westdickenberg, *Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers*, Inf. Inference, 11 (2022), pp. 307–353, https://doi.org/10.1093/imaiai/iaaa039.

[4] D. A. Bayer and J. C. Lagarias, *The nonlinear geometry of linear programming.* I. *Affine and projective scaling trajectories*, Trans. Amer. Math. Soc., 314 (1989), pp. 499–526.

[5] D. A. Bayer and J. C. Lagarias, *The nonlinear geometry of linear programming.* II. *Legendre transform coordinates and central trajectories*, Trans. Amer. Math. Soc., 314 (1989), pp. 527–581.

[6] A. Bunse-Gerstner, R. Byers, V. Mehrmann, and N. K. Nichols, *Numerical computation of an analytic singular value decomposition of a matrix valued function*, Numer. Math., 60 (1991), pp. 1–39, https://doi.org/10.1007/BF01385712.

[7] E. J. Candès and B. Recht, *Exact matrix completion via convex optimization*, Found. Comput. Math., 9 (2009), pp. 717–772, https://doi.org/10.1007/s10208-009-9045-5.

[8] P. A. Deift, *Orthogonal Polynomials and Random Matrices: A Riemann-Hilbert Approach*, Courant Lect. Notes Math. 3, New York University, Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence, RI, 1999.

[9] 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, Long Beach, CA, 2017, pp. 6151–6159, https://proceedings.neurips.cc/paper/2017/hash/58191d2a914c6dae66371c9dcdc91b41-Abstract.html.

[10] B. Hanin and M. Nica, *Products of many large random matrices and gradients in deep neural networks*, Comm. Math. Phys., 376 (2020), pp. 287–322, https://doi.org/10.1007/s00220-019-03624-z.

[11] M. Hardt, R. Meka, P. Raghavendra, and B. Weitz, *Computational limits for matrix completion*, in Proceedings of the 27th Conference on Learning Theory (COLT), JMLR.org, 2014, pp. 703–725.

[12] R. Hildebrand, *Canonical barriers on convex cones*, Math. Oper. Res., 39 (2014), pp. 841–850, https://doi.org/10.1287/moor.2013.0640.

[13] E. P. Hsu, *Stochastic Analysis on Manifolds*, Grad. Stud. Math. 38, American Mathematical Society, Providence, RI, 2002, https://doi.org/10.1090/gsm/038.

[14] A. Lapedes and R. Farber, *How neural nets work*, in Evolution, Learning and Cognition, World Scientific Publishing, Teaneck, NJ, 1988, pp. 331–346.

[15] Y. LeCun, Y. Bengio, and G. E. Hinton, *Deep learning*, Nature, 521 (2015), pp. 436–444, https://doi.org/10.1038/nature14539.

[16] G. Menon, *Gibbs Measures for Semidefinite Programming*, manuscript, Brown University, 2020.

[17] G. Menon and T. Trogdon, *Random Matrix Theory and Numerical Linear Algebra*, manuscript, Brown University, 2020.

[18] G. Menon and T. Yu, *The Riemannian Langevin Equation and Conic Programs*, preprint, https://arxiv.org/abs/2302.11653, 2023.

[19] B. Neyshabur, R. Tomioka, and N. Srebro, *In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning*, preprint, https://arxiv.org/abs/1412.6614, 2014.

[20] N. Razin and N. Cohen, *Implicit regularization in deep learning may not be explainable by norms*, in NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020, pp. 21174–21187.

[21] Z. Veraszto, *The Deep Linear Network – Dynamics, Riemannian Geometry and Overparametrization*, Ph.D. thesis, Brown University, 2023.