Deep Linear Networks

Session 3 : Characterizing the Minimizer

Rathindra Nath Karmakar

References

- **Deep Linear Networks for Matrix Completion - An Infinite Depth Limit**, Nadav Cohen, Govind Menon, and Zsolt Veraszto (2023)
- **Implicit Regularization in Deep Learning May Not Be Explainable by Norms**, Noam Razin, Nadav Cohen (2020)
- **Implicit Regularization in Matrix Factorization**, Suriya Gunasekar, Blake Woodworth, Behnam Neyshabur, Srinadh Bhojanapalli, Nathan Srebro (2017)
- **The geometry of the deep linear network**, Govind Menon (2024)

Recap: Key Questions for Gradient Flow

For the gradient flow dynamics on a loss surface $\mathcal{L}(\mathbf{W})$:

$$\frac{d}{dt}\mathbf{W}(t) = -\nabla_{\mathbf{W}}\mathcal{L}(\mathbf{W}(t))$$

We want to understand:

- Convergence guarantees? (*Yes, for balanced cases*)
- Convergence rate? (*Can be accelerated by depth*)
- **Characterization of the minimizer reached?**
- Effect of noise and discretization?

Today, we address the third question: among all possible solutions, which one does gradient descent choose?

The Central Question of Implicit Regularization

In overparameterized settings, there are many (often infinite) parameter settings that achieve zero training loss. Gradient descent finds one of them. Why that specific one?

What objective is gradient descent implicitly minimizing?

We will investigate three competing hypotheses:

1. **Norm Minimization**
2. **Rank Minimization**
3. **Volume Maximization**

General Problem Formulation

- The problem is to recover a matrix $W \in \mathbb{R}^{d\times d'}$ from a set of observed entries $\{b_{i,j}\}$ at locations $(i,j) \in \Omega$.
- The goal is to find a matrix W that minimizes the squared error loss:

$$\ell(W) = \frac{1}{2}\sum_{(i,j)\in\Omega}(W_{i,j} - b_{i,j})^2$$

- Instead of optimizing over W directly, the matrix is parameterized as the product of L factor matrices:

$$W = W_L W_{L-1}\cdots W_1$$

- The factor matrices $\{W_l\}_{l=1}^{L}$ are then trained to minimize the loss.
- The optimization dynamics are studied under gradient flow:

$$\frac{d}{dt}W_l(t) = -\nabla_{W_l}\ell(W(t))$$

- This process starts from a random initialization close to zero that satisfies the "balancedness" condition: $W_{l+1}^T(0)W_{l+1}(0) = W_l^T(0)W_l(0)$.

Hypothesis 1: Norm Minimization

This is the classical explanation, rooted in linear regression where gradient descent with zero initialization famously converges to the minimum ℓ_2-norm solution. The hope is that this generalizes to deep learning, perhaps with a different norm (e.g., nuclear norm for matrices).

Conjecture 1 (Gunasekar et al., 2017). *For matrix completion, gradient descent with small initialization converges to the solution with the minimum **nuclear norm**.*

A Counterexample: Norms vs. Rank

Razin & Cohen (2020) constructed a simple 2x2 matrix completion problem to test this hypothesis. Given observations $w_{12} = 1, w_{21} = 1, w_{22} = 0$, the set of solutions is:

$$S = \left\{W_x = \begin{pmatrix} x & 1 \\ 1 & 0 \end{pmatrix} : x \in \mathbb{R}\right\}$$

This setup creates a direct conflict: Minimizing any norm (e.g., Frobenius $\|W_x\|_F = \sqrt{x^2+2}$) requires x to be bounded (minimum at $x = 0$).

Theorem: Norm Minimization is False

The paper proves that for this problem, gradient flow on a deep matrix factorization ($L \geq 2$) drives the unobserved entry to ∞.

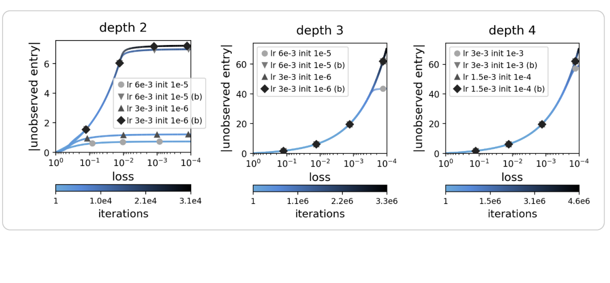
Theorem 1 (Razin & Cohen, 2020). *For the matrix completion problem above, with random near-zero balanced initialization and depth $L \geq 2$, if $\det(W(0)) > 0$ (a 50% chance), then as the loss $\ell(t) \to 0$:*

1. *For **any** norm or quasi-norm $\|\cdot\|$, the norm of the solution diverges: $\|W(t)\| \to \infty$.*
2. *The effective rank of the solution converges to its minimum possible value: $\mathrm{erank}(W(t)) \to 1$.*

This definitively shows that implicit regularization in this setting cannot be explained by the minimization of any norm.

Experimental Verification for Hypothesis 1

The theory predicts that as the loss decreases, the unobserved entry (w_{11}) should grow. The experiments (Fig 1 from Razin & Cohen, 2020) confirm this: as loss decreases (moving right to left on the x-axis), the absolute value of the unobserved entry increases. Since all norms must grow with this entry, this validates the theorem and refutes the norm minimization hypothesis.



Hypothesis 2: Rank Minimization

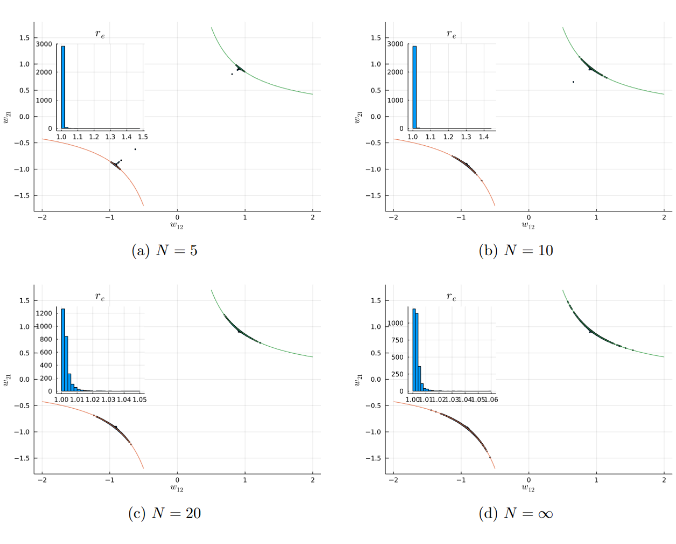
The failure of the norm hypothesis suggests a new candidate: perhaps the implicit bias is towards minimizing **rank** (or its continuous surrogate, **effective rank**).

Hypothesis 2. *Gradient descent converges to the solution with the minimum (effective) rank.*

This is consistent with the previous experiment and many empirical observations. But is it the full story?

Experiment: Is Rank/Effective Rank Sufficient?

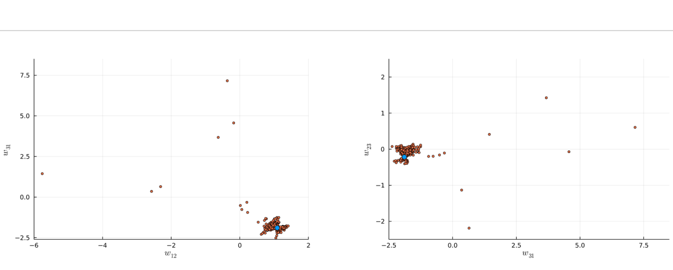
Consider a 2x2 diagonal matrix completion task where all rank-1 solutions lie on a hyperbola. The training outcomes (Fig 4 from Cohen, Menon, Veraszto, 2023) cluster tightly around the corners of the hyperbola, indicating a strong preference for these solutions.



(a) $N = 5$ (b) $N = 10$ (c) $N = 20$ (d) $N = \infty$

- All points on the red/green curves are rank-1 minimizers. A simple **rank minimization** principle fails to explain the clustering.
- The embedded histograms show the **effective rank** is also tightly clustered around 1.0 for all outcomes. Effective rank also fails to explain the preference for the corners.

In another experiment with a 3x3 diagonal matrix completion task, the majority of 500 outputs cluster near one particular rank-two minimizer out of many possibilities.



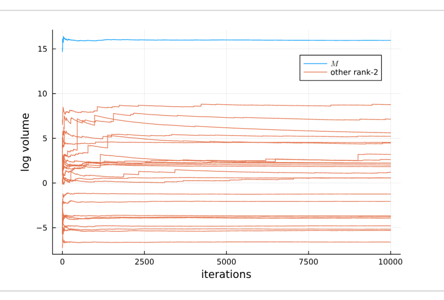
Hypothesis 3: Bias Towards High State Space Volume

The failure of the rank hypothesis suggests a more refined geometric principle is at play. The volume element tells us how to measure the "size" of a small region of matrices. Regions with higher volume are "bigger" targets for the optimization dynamics.

Hypothesis 3. *Gradient-based methods are biased towards solutions that can be represented in "more ways" in the parameter space. This size is measured by the volume element induced by the Riemannian metric g^N.*

Volume as the Predictor

Plotting the logarithm of the state space volume on the same plane of solutions (Fig 5 from Cohen, Menon, Veraszto, 2023), we see the volume (brighter colors) is highest precisely at the corners of the hyperbola where the training outcomes clustered. This suggests that the dynamics are not just minimizing rank, but are being attracted to the regions of maximal volume within the set of low-rank solutions.



Quantifying Volume on the Solution Space

The volume of a small region of end-to-end matrices W is given by $\sqrt{\det g^N dW}$.

Vandermonde Determinant: $\mathrm{van}(\Lambda) = \prod_{1\leq i<j\leq d}(\lambda_j - \lambda_i)$

Theorem 1.1 (Cohen, Menon, Veraszto, 2023). The volume element on the manifold of end-to-end matrices $(\mathcal{M}_d, g^N)$ is given in terms of the singular values Σ by:

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

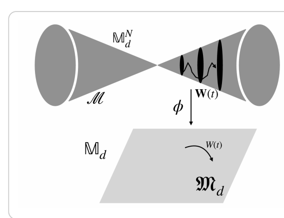
This formula shows that the volume density **diverges** as any singular value approaches zero (i.e., as the matrix W approaches a lower rank). This divergence indicates a strong geometric bias towards low-rank matrices during the training process.

The divergence comes from the term $\det(\Sigma^2)^{\frac{1-N}{2N}} = \prod_{i=1}^{d}\sigma_i^{\frac{1}{N}-1}$. For depth $N > 1$, the exponent is negative, so as any $\sigma_i \to 0$, the term $\sigma_i^{\frac{1}{N}-1} \to \infty$.

Connecting Geometries: Riemannian Submersion

How does the "downstairs" geometry on the space of end-to-end matrices relate to the simple "upstairs" Euclidean geometry on the space of factors?

The connection is a **Riemannian submersion**. This means the projection map $\phi : \mathcal{M}_r \to M_r$ (from the balanced manifold of factors to the space of rank-r matrices) is a geometry-preserving map.



Theorem 13 (Menon & Yu, 2023). For each rank r, the metric g^N on M_r is obtained from the map $\phi : \mathcal{M}_r \to M_r$ by Riemannian submersion.

Quantifying Volume of Representations

The "upstairs" space of weights $\mathbf{W} = (W_N, \ldots, W_1)$ contains many configurations that map to the same end-to-end matrix W. This set is the fiber, or group orbit, O_W. Its volume quantifies the degree of overparameterization.

Theorem 10 (Menon & Yu, 2023). The volume of the group orbit O_W corresponding to an end-to-end matrix W with singular values Σ is:

$$\mathrm{vol}(O_W) = c_d^{N-1}\frac{\mathrm{van}(\Sigma^2)}{\mathrm{van}(\Sigma^{2/N})}$$

This provides an "entropic" interpretation: solutions with a larger orbit volume are more numerous and thus more likely to be found. This volume also diverges as singular values approach zero.

Proof Sketch: Deriving the Volume Form

The foundation of the proof is the standard definition of a volume form in Riemannian geometry, $d\mathrm{vol}_g = \sqrt{\det(g_{ij})}\, dW$. The analysis is restricted to $W \in GL(d, \mathbb{R})$, the space of full-rank (invertible) matrices.

1. Determinant of the Metric Tensor

The determinant of g^N is computed by exploiting its spectral properties. The metric tensor can be diagonalized, and its determinant is invariant under this transformation.

- $g^N = (V\otimes U)D^N(\Sigma)(V\otimes U)^T$, where $(V\otimes U)$ is orthogonal.
- Thus, $\det g^N = \det D^N(\Sigma)$.
- The determinant of the diagonal matrix $D^N(\Sigma)$ is the product of its diagonal entries, which are the reciprocals of the eigenvalues, λ_{il}^N, of a related linear operator $A_{N,W}$.

$$\det g^N = \prod_{i,l=1}^{d}\frac{1}{\lambda_{il}^N}$$

2. Eigenvalues of $A_{N,W}$

The eigenvalues λ_{il}^N are given by Lemma 2.4. For a network of finite depth N and a matrix $W = U\Sigma V^T$:

$$\lambda_{il}^N = \frac{1}{N}\sum_{j=1}^{N}(\sigma_i^2)^{\frac{N-j}{N}}(\sigma_l^2)^{\frac{j-1}{N}}$$

3. The Jacobian of the SVD Map

To express the volume form consistently, we change from matrix entry coordinates (dW) to SVD coordinates ($d\Sigma, dU, dV$). This requires a Jacobian determinant from Lemma 2.7:

$$dW = \mathrm{van}(\Sigma^2)\, d\Sigma \wedge dU \wedge dV$$

where $\mathrm{van}(\Sigma^2) = \prod_{1\leq i<j\leq d}(\sigma_i^2 - \sigma_j^2)$.

4. Final Form of the Volume Element

Assembling the pieces, we start with the definition and substitute the expressions for the determinant and the Jacobian:

1. Start with $d\mathrm{vol}_{g^N} = \sqrt{\det g^N}\, dW$.
2. Substitute the determinant from step 1 and the Jacobian from step 3:

$$d\mathrm{vol}_{g^N} = \left(\prod_{i,l=1}^{d}\frac{1}{\lambda_{il}^N}\right)^{1/2}\mathrm{van}(\Sigma^2)\, d\Sigma\, dU\, dV$$

3. After performing the product over the eigenvalues λ_{il}^N and simplifying, we arrive at the final result from Theorem 1.1.

Summary of Findings

- The classical hypothesis that implicit regularization is equivalent to norm minimization is incorrect. There are natural problems where gradient descent drives all norms to infinity.
- A more robust heuristic is **rank minimization**, which correctly predicts the behavior in the counterexample.
- However, rank alone is insufficient to explain why specific low-rank solutions are preferred over others.
- The most fundamental explanation appears to be a bias towards regions of **maximal state space volume**, a concept made precise by the Riemannian geometry of the DLN. The volume is largest near low-rank solutions, and can distinguish between different solutions of the same rank.