Deep Linear Networks

Session 4: Effect of Noise and Discretization

Rathindra Nath Karmakar

References

- **Motion by mean curvature and Dyson Brownian motion**, C.-P. Huang, D. Inauen, and G. Menon (2023)
- **An entropy formula for the deep linear network**, G. Menon and T. Yu (2024)
- **The geometry of the deep linear network**, G. 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? (*Bias towards max volume*)
- **Effect of noise and discretization?**

Today, we address the final question. What happens when we add stochastic noise to the dynamics?

Stochasticity in Deep Learning

Real-world training is not a clean gradient descent. Noise arises from many sources:

- **Stochastic Gradient Descent (SGD):** Gradients are computed on mini-batches of data, which is a noisy estimate of the true gradient.
- **Discretization Error:** Using a finite step size $\eta > 0$ introduces error.
- **Round-off Errors:** Floating point arithmetic introduces small, random perturbations.

How do these stochastic effects interact with the Riemannian geometry of the problem? One tool is the **Riemannian Langevin Equation (RLE)**. It can be used to model **noise due to round-off errors**

The Central Idea

Noise in the "upstairs" parameter space (along redundant directions) induces a deterministic, curvature-driven drift in the "downstairs" solution space.

Warm-Up: Dyson Brownian Motion

To build intuition, we study a famous model from random matrix theory.

- **Downstairs Dynamics (Eigenvalues):** A system of interacting particles (eigenvalues) $x_1 < \cdots < x_d$ evolves according to:

$$dx_i = \sum_{j\neq i}\frac{1}{x_i - x_j}dt + \sqrt{\frac{2}{\beta}}dW_i$$

  This is **Dyson Brownian Motion**, combining repulsion and random noise.

Prerequisites for Theorem 15

- Let Her_d be the space of $d \times d$ Hermitian matrices.
- Given a vector of eigenvalues $x \in \mathbb{R}^d$, the **isospectral orbit** O_x is the set of all Hermitian matrices with those eigenvalues:

$$O_x = \{M \in \mathrm{Her}_d \,|\, M = UXU^*, U \in U_d\}$$

  where $X = \mathrm{diag}(x)$ and U_d is the unitary group.
- We can model noise on the full matrix space ("upstairs") using a standard Wiener process H_t on Her_d. Let P_M and $P_M^{\perp}$ be projections onto the tangent and normal spaces of the orbit O_x at point M.

Theorem 15: Noise Upstairs, Curvature Downstairs

Consider the "upstairs" SDE for a matrix M_t evolving on the isospectral orbit O_{x_t} subject to isotropic noise:

$$dM_t = P_{M_t}dH_t + \sqrt{\frac{2}{\beta}}P_{M_t}^{\perp}dH_t$$

Theorem 15 (Huang, Inauen, Menon, informal).

(a) *The eigenvalues of the matrix M_t have the same law as the solution x_t to Dyson Brownian Motion.*

(b) *In the zero-noise limit ($\beta \to \infty$), noise purely tangential to the orbit ($P_{M_t}dH_t$) induces a deterministic drift normal to the orbit, equal to motion by (minus one half) **mean curvature**.*

This is a profound result: purely random fluctuations in the "gauge" directions (upstairs) create a deterministic, geometry-driven motion (a drift) in the observable space (downstairs).

General Geometric Framework

General Principles 1: Submersion with Group Action

We formalize the "upstairs/downstairs" picture.

- Let $(\mathcal{M}, g)$ be a reference Riemannian manifold .
- Let G be a Lie group that acts on $\mathcal{M}$ by isometries .
- The "downstairs" space is the quotient space $M = \mathcal{M}/G$, equipped with a metric h via **Riemannian submersion**.
- The map $\phi : \mathcal{M} \to M$ is the projection. The inverse image $\phi^{-1}(x) = O_x$ is the group orbit over x.

General Principles 2: The "Upstairs" RLE

We define stochastic gradient descent via the RLE of a lifted loss function $L = E \circ \phi$.

Consider the "upstairs" dynamics for $m \in \mathcal{M}$ given by the SDE:

$$dm^{\beta,\kappa} = -\mathrm{grad}_g L(m)dt + P_m d\mathbf{M}^{\beta} + \sqrt{\kappa}P_m^{\perp}d\mathbf{M}^{\beta}$$

- The first term is the standard gradient of the lifted loss.
- P_m and $P_m^{\perp}$ are projections onto directions tangential and normal to the group orbit $O_{\phi(m)}$.
- $\mathbf{M}^{\beta}$ is Brownian motion on $\mathcal{M}$. κ modulates the anisotropy of the noise.

General Principles 3: The "Downstairs" SDE

The "upstairs" stochastic flow projects to a stochastic flow downstairs.

The flow of m_t projects to the RLE for the **free energy** downstairs:

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

where the free energy is defined as:

$$F_{\beta}(x) = L(x) - \frac{1}{\beta}S(x) \quad \text{and} \quad S(x) = \log \mathrm{vol}(O_x)$$

Noise in the redundant "gauge" directions upstairs manifests as an entropic term that modifies the energy landscape downstairs. In the limit $\kappa \to 0$, the upstairs noise is purely tangential, and the downstairs flow becomes deterministic: $\dot{x} = -\mathrm{grad}_h F_{\beta}(x)$.

Application to the Deep Linear Network

RLE for DLN: Upstairs Dynamics

Applying the general principle to the DLN, the "upstairs" RLE on the balanced manifold $\mathcal{M}$ is:

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

This is standard gradient descent on the lifted loss function $L = E \circ \phi$, plus a noise term $d\mathbf{M}^{\beta,\kappa}$ that represents Brownian motion on the balanced manifold (with the Frobenius metric).

RLE for DLN: Downstairs Dynamics

The law of the end-to-end matrix $W_t = \phi(\mathbf{W}_t)$ is then given by the "downstairs" RLE:

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

- The drift term is the Riemannian gradient of the free energy $F_{\beta}(W) = E(W) - \frac{1}{\beta}S(W)$.
- The noise term $dX^{\beta/\kappa}$ is Brownian motion on the downstairs manifold (M_d, g^N).

The Free Energy Gradient in the DLN

The gradient of the free energy, which drives the system's evolution, can be computed explicitly. It balances the drive to minimize loss with an opposing entropic force.

$$\mathrm{grad}_{g^N}F_{\beta}(W) = \underbrace{A_{N,W}(E'(W))}_{\text{Loss Term}} - \underbrace{\frac{1}{\beta}\mathrm{grad}_{g^N}S(W)}_{\text{Entropic/Curvature Term}}$$

The second term is the entropic force arising from the geometry of overparameterization. For the DLN, it takes the specific form:

$$\frac{1}{\beta}\mathrm{grad}_{g^N}S(W) = \frac{1}{\beta}Q_N\Sigma'Q_0^T$$

- $W = Q_N\Sigma Q_0^T$: This is the Singular Value Decomposition (SVD) of the end-to-end matrix. Q_N and Q_0 are orthogonal matrices, and Σ is a diagonal matrix containing the singular values σ_k.
- Σ': A diagonal matrix whose entries are functions of the singular values σ_k and network depth N. It acts as a repulsive force between the singular values, pushing the system away from low-rank solutions. The paper notes this "physical effect" is strictly due to the geometry of the DLN.

The Explicit Downstairs SDE

The Menon & Yu paper provides an explicit formula for the Brownian motion $dX^{\beta/\kappa}$ on (M_d, g^N) .

Theorem 18 (Menon & Yu, 2023). The solution X_t^{β} to the following Itô SDE with initial condition $X_0 = W_0$ is Brownian motion on (M_d, g^N):

$$dX_t^{\beta} = \sqrt{\frac{2}{\beta}}\begin{pmatrix}\sqrt{N}\lambda_1^{N-1}dB_{11}^{1,1} & \cdots \\ \vdots & \ddots\end{pmatrix} + \frac{1}{\beta}Q_N\Sigma''Q_0^T dt$$

Here, $\Lambda = \Sigma^{1/N}$, dB is a matrix of standard Wiener processes, and Σ'' is a diagonal matrix of drift terms (related to mean curvature) that arises from the Itô correction.

Proof Sketch for Theorem 15 (2x2 Case)

The Goal

Let's prove Theorem 15 for the simple case of $d = 2$. We want to show that the eigenvalues x_1, x_2 of the matrix M_t evolving by:

$$dM_t = P_{M_t}dH_t + \sqrt{\frac{2}{\beta}}P_{M_t}^{\perp}dH_t$$

follow the Dyson Brownian Motion equations:

$$dx_1 = \frac{1}{x_1 - x_2}dt + \sqrt{\frac{2}{\beta}}dW_1$$

$$dx_2 = \frac{1}{x_2 - x_1}dt + \sqrt{\frac{2}{\beta}}dW_2$$

Step 1: Simplification

The dynamics are invariant under unitary transformations ($M \to UMU^*$). This allows us to analyze the process at a point where the matrix M_t is diagonal, without loss of generality.

- Let's fix a time t and assume M_t is the diagonal matrix $X = \mathrm{diag}(x_1, x_2)$.
- **Normal Space ($T_X^{\perp}O_x$):** The normal space to the orbit at X consists of all 2x2 Hermitian matrices that commute with X. These are the diagonal matrices.

$$T_X^{\perp}O_x = \left\{\begin{pmatrix} a & 0 \\ 0 & b\end{pmatrix}, a, b \in \mathbb{R}\right\}$$

- **Tangent Space (T_XO_x):** The tangent space is the orthogonal complement. It consists of all 2x2 Hermitian matrices with zeros on the diagonal.

$$T_XO_x = \left\{\begin{pmatrix} 0 & z \\ \bar{z} & 0\end{pmatrix}, z \in \mathbb{C}\right\}$$

Step 2: Decomposing the Noise

We express the "upstairs" noise dH_t using an orthonormal basis for 2x2 Hermitian matrices that respects our tangent/normal split.

Normal Basis: $E_a = \begin{pmatrix} 1 & 0 \\ 0 & 0\end{pmatrix}$, $E_b = \begin{pmatrix} 0 & 0 \\ 0 & 1\end{pmatrix}$

Tangent Basis: $E_c = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 & 1 \\ 1 & 0\end{pmatrix}$, $E_d = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 & i \\ -i & 0\end{pmatrix}$

The matrix noise term dH_t can be written with four independent Wiener processes W_a, W_b, W_c, W_d:

$$dH_t = E_a dW_a + E_b dW_b + E_c dW_c + E_d dW_d$$

Projecting this noise onto the tangent (P_X) and normal ($P_X^{\perp}$) spaces at X gives our SDE for dM_t:

$$dM_t = \underbrace{(E_c dW_c + E_d dW_d)}_{\text{Tangent Part}} + \sqrt{\frac{2}{\beta}}\underbrace{(E_a dW_a + E_b dW_b)}_{\text{Normal Part}}$$

Step 3: Itô's Formula for Eigenvalues

To find the dynamics of an eigenvalue, say $x_1(M_t)$, we use Itô's formula. This requires the first and second derivatives of the eigenvalue with respect to changes in the matrix.

For a function $f(M_t)$, Itô's formula is: $df = \mathrm{D}f(dM_t) + \frac{1}{2}\mathrm{D}^2 f(dM_t, dM_t)$

- **First Derivative (captures noise):** The change in an eigenvalue is most sensitive to the corresponding diagonal entry.

$$\mathrm{D}x_1(A) = A_{11}$$

- **Second Derivative (captures drift):** The second-order change depends on interactions between off-diagonal elements. For a matrix A:

$$\mathrm{D}^2x_1(A, A) = 2\frac{|A_{12}|^2}{x_1 - x_2}$$

Now we just need to plug our dM_t into these formulas.

Step 4: The Final Calculation

Let's compute the terms for dx_1.

- **Noise Term:** We apply the first derivative to dM_t. Only the normal part contributes to the diagonal.

$$\mathrm{D}x_1(dM_t) = (dM_t)_{11} = \left(\sqrt{\frac{2}{\beta}}(E_a dW_a + E_b dW_b)\right)_{11} = \sqrt{\frac{2}{\beta}}dW_a$$

  This gives the Brownian motion part of the equation. Let's call W_a our new W_1.
- **Drift Term:** We apply the second derivative. Only the tangent part has off-diagonal entries. The quadratic variation of dM_t in the off-diagonal is $[E_c, E_c]dt + [E_d, E_d]dt$.

$$\frac{1}{2}\mathrm{D}^2x_1(dM_t, dM_t) = \frac{1}{2}\left(\underbrace{\mathrm{D}^2x_1(E_c, E_c)dt}_{\text{from } dW_c^2} + \underbrace{\mathrm{D}^2x_1(E_d, E_d)dt}_{\text{from } dW_d^2}\right)$$

$$= \frac{1}{2}\left(2\frac{|(E_c)_{12}|^2}{x_1 - x_2} + 2\frac{|(E_d)_{12}|^2}{x_1 - x_2}\right)dt = \frac{1}{x_1 - x_2}\left(\left|\frac{1}{\sqrt{2}}\right|^2 + \left|\frac{i}{\sqrt{2}}\right|^2\right)dt = \frac{1}{x_1 - x_2}dt$$

Combining these gives: $dx_1 = \frac{1}{x_1 - x_2}dt + \sqrt{\frac{2}{\beta}}dW_1$. The proof for x_2 is identical.

Summary of Findings

- Noise in training can be modeled rigorously using the **Riemannian Langevin Equation**.
- There is a deep connection, via Riemannian submersion, between stochastic dynamics in the "upstairs" parameter space and the resulting dynamics in the "downstairs" solution space.
- Noise that is tangential to the fibers (group orbits) upstairs induces a deterministic drift downstairs related to the **mean curvature** of the fibers. This drift is equivalent to the gradient of the **Boltzmann entropy** (volume).
- For the DLN, we have explicit formulas for this stochastic process, which corresponds to gradient descent of a **free energy** functional, combining the original loss with an entropic term that favors high-volume, low-rank solutions.