STAT 946 - Deep Learning Theory

Kevin Zhao

Session 2 - September 8th, 2025

1 Review from last class - Motivations

1.1 Modern deep architectures and datasets are very large

This naturally leads us to analyze algorithms in **asymptotic regimes**, where problem size grows and limit theorems simplify the mathematics. Some open problems includes:

- **Comparing methods:** Comparing optimization procedures/architectures at scale.
- **Why DL works:** Explaining optimization behavior and generalization observed in large models.

1.2 Asymptotics often simplify the math

A simple example of this is the central limit theorem where number of random variables (or sources) $\to \infty$ but the contribution of each variable $\to 0$:

$$\frac{\sum_{i=1}^{n}(X_i - \mu)}{\sigma\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1).$$

2 Scaling Examples in Probability Theory

Here we present 3 major examples of scaling limits in probability theory that is useful in modern deep learning.

2.1 Random Walk

Random walk is a discrete object in probability theory and it has the following setup. Let $(X_i)_{i\geq 1}$ be i.i.d. real random variables with

$$\mathbb{E}[X_i] = 0, \qquad 0 < \sigma^2 = \mathrm{Var}(X_i) < \infty \quad \forall i.$$

For each n, define the scaled partial sums

$$Y_k^{(n)} := \frac{1}{\sqrt{n}} \sum_{i=1}^{k} X_i, \qquad k = 0, 1, \ldots, n.$$

2.1.1 Continuous time interpolation of RW

Define the piecewise-constant process

$$Z_t^{(n)} := Y_{\lfloor nt \rfloor}^{(n)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} X_i,$$

Now, observe that for a fixed t,

$$Z_t^{(n)} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, t),$$

by the central limit theorem. Let's now introduce brownian motion for further building a process limit on top of the point wise limit.

Definition 1 (Brownian Motion). *A process $B : [0, T] \to \mathbb{R}$ is a Brownian motion if:*

1. $B_0 = 0$ *almost surely;*
2. *for $0 \leq s \leq t$, $B_t - B_s \sim \mathcal{N}(0, t - s)$ and increments over disjoint intervals are independent;*
3. *the sample paths are almost surely continuous.*

The following theorem from Donsker enables the analysis of discrete time processes (e.g. Stochastic Gradient Descent) via the Brownian model.

Theorem 1 (Donsker). *Recall the definition of $Z_t^{(n)}$ from before,*

$$Z_t^{(n)} := Y_{\lfloor nt \rfloor}^{(n)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} X_i, , \quad t \in [0, T].$$

Define the space

$$\mathcal{X} = (C([0, 1]), || \cdot ||_\infty)$$

and the function space

$$\mathcal{X}^* = \{\psi : \mathcal{X} \to \mathbb{R} \text{ continuous and bounded}\}.$$

Then

$$\big(Z_t^{(n)}\big)_{t\in[0,T]} \xrightarrow[n\to\infty]{d} \big(B_t\big)_{t\in[0,T]} \quad \text{in } \mathcal{X}.$$

Equivalently, for all $\psi \in \mathcal{X}^*$,

$$\mathbb{E}\left[\psi\big(Z^{(n)}\big)\right] \longrightarrow \mathbb{E}\left[\psi(B)\right].$$

2.2 Random Matrix Theory

Let's now shift the focus to Random Matrix Theory, which has the following setup. Let

$$X = \big[x_{ij}\big]_{i,j=1}^{n} \in \mathbb{R}^{n\times n}, \qquad X = X^{\top},$$

with $\{x_{ij} : 1 \le i \le j \le n\}$ independent, $\mathbb{E}[x_{ij}] = 0$ and $\mathbb{E}[x_{ij}^2] = 1$. Consider the scaling $\frac{1}{\sqrt{n}}X$ and denote its ordered eigenvalues by

$$\lambda_1^{(n)} \le \cdots \le \lambda_n^{(n)}, \qquad \lambda_i^{(n)} = \lambda_i\Big(\tfrac{1}{\sqrt{n}}X\Big).$$

Define the probability measure

$$\rho^{(n)}(dx) \;=\; \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i^{(n)}}(dx),$$

so that for any test function f,

$$\int_{\mathbb{R}} f(x)\,\delta_{x_0}(dx) = f(x_0), \qquad \text{and} \qquad \int_{\mathbb{R}} f(x)\,\rho^{(n)}(dx) \;=\; \frac{1}{n}\sum_{i=1}^{n} f\big(\lambda_i^{(n)}\big).$$

Now, the following theorem describes the limiting distribution of the eigenvalue distribution

Theorem 2 (Wigner semicircle law)**.** *Following the above notations,*

$$\rho^{(n)} \;\to\; \rho_{\mathrm{sc}} \quad \textit{weakly}$$

where ρ_{sc} has density

$$\rho_{\mathrm{sc}}(x) = \frac{\sqrt{(4-x^2)_+}}{2\pi}, \qquad \textit{support on } [-2,2].$$

Equivalently, for every bounded continuous $\psi \in C_b(\mathbb{R})$,

$$\int \psi(x)\,\rho^{(n)}(dx) \;\longrightarrow\; \int \psi(x)\,\rho_{\mathrm{sc}}(x)\,dx \quad \textit{a.s.}$$

*The same convergence also holds for $\psi \in C_0(\mathbb{R})$ i.e. continuous and vanishing at ∞. This is often called **Vague** convergence but in the compact case, it is the same as weakly convergence.*

Note that as $n \to \infty$, the eigenvalue distribution of the random matrix consisting of i.i.d random variables with mean 0 and variance 1 tends to a deterministic shape of a semi-circle with the following properties:

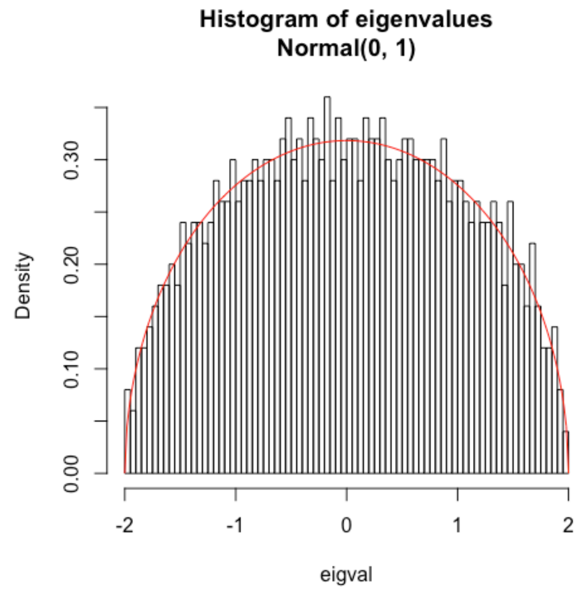

Figure 1: Empirical spectral distribution approaching the semicircle law. See these notes for details.

- The bulk of eigenvalues lies in the interval $[-2, 2]$; the fraction outside this interval converges to 0 as $n \to \infty$.
- The extreme eigenvalues concentrate at the edges: $\lambda_{\max} \to 2$ and $\lambda_{\min} \to -2$ almost surely.

In an example of a matrix consisting of $\mathcal{N}(0, 1)$, its limiting eigenvalue behaviours can be characterized by the diagram 1:

2.3 Mean-filed particle systems

Consider particles $X_i(t) \in \mathbb{R}$ for $i = 1, \ldots, n$ that follows the ODE evolution:

$$\dot{X}_i(t) = \frac{1}{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{1}{X_i(t) - X_j(t)}$$

This describes a kind of repulsion between the particles. Recall that the *empirical measure* is defined as

$$\rho_t^{(n)}(dx) = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i(t)}(dx), \qquad \int f(x)\, \rho_t^{(n)}(dx) = \frac{1}{n} \sum_{i=1}^{n} f\big(X_i(t)\big)$$

for a given time t. Therefore, using the same trick as the random matrix theory case, we obtain the following approximation using a integral:

$$\dot{X}_i(t) \approx \int_{\mathbb{R}} \frac{1}{X_i(t) - y}\, \rho_t^{(n)}(dy),$$

where the error between the two terms comes from the self-term $i = j$.

By considering the initial conditions $X_i(0) \overset{\text{i.i.d.}}{\sim} \rho_0$ for all i, we have $\rho_t^{(n)} \to \rho_t$ and therefore

$$(X_1(t), \dots, X_k(t)) \xrightarrow{law} (\rho_t, \dots, \rho_t) \qquad \text{for each k}$$

which is often being referred as the propagation of chaos where each particle is asymptotically independent. Now by taking the mean-field limit on the ODE as $n \to \infty$, we obtain the deterministic differential equation that characterize the path of the particles $X(t)$ in an unifying way:

$$\begin{cases} \dfrac{d}{dt} X(t) = \displaystyle\int \frac{1}{X(t) - y}\, \rho_t(dy), \\ X(0) \sim \rho_0(dx), \end{cases}$$

where the law ρ_t follows a transport equation of the form:

$$\partial_t \rho_t(x) = -div_x \Big(\rho_t(x) \int \frac{1}{x - y} \rho_t(dy) \Big).$$

In deep learning, mean particle systems are closely related to mean-field neural networks: neurons/parameters play the role of particles, and their empirical distribution evolves by a transport PDE.