STAT 946 Fall 2025 Scribes: Jonathan Borenstein
September 29

Lecture 8: Marchenko-Pastur Law

Setup

- Data Matrix $X \in \mathbb{R}^{n \times d}$, $X_{ij} \overset{iid}{\sim} N(0,1)$ where $n =$ number of data points and $d =$ data dimension
- $\lambda_i := \lambda_i(\frac{1}{n} X^T X)$, i^{th} largest eigenvalue
- Empirical spectral distribution (ESD) $\hat{\rho} := \frac{1}{d} \sum_{i=1}^{d} \delta_{\lambda_i}$

Stieltjes Transform

Definition 1. *Stieltjes Transform*

$$\hat{\varphi}(z) = \int \frac{1}{x - z} d\hat{\rho}(x) = \frac{1}{d} \sum_{i=1}^{d} \frac{1}{\lambda_i - z}$$

where $z \in \mathbb{C} \setminus \operatorname{supp}(\hat{\rho})$.

Then, as $n, d \to \infty$ and $\frac{d}{n} \to \gamma > 0$, $\hat{\varphi} \to \varphi$ satisfying

$$\frac{1}{\varphi(z)} + z = \frac{\gamma}{1 + \varphi(z)} \quad \text{(Self-consistency equation)}$$

and $\hat{\rho} \to \rho$ where

$$\rho(x) = \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{2\pi\gamma x}$$

with $x \in [\lambda_-, \lambda_+]$ and $\lambda_\pm = (1 \pm \sqrt{\gamma})^2$. Note that $\lambda_- = 0 \iff \gamma = 1$ and when $\gamma > 1$, $d - n$ of $\lambda_i = 0$, contributing point mass $(1 - \frac{1}{\gamma})\delta_0$ in the limit.

Stochastic Calculus

- $\{B_t\}_{t\geq 0}$ Brownian Motion (B.M.)
- $f : \mathbb{R} \to \mathbb{R}$

Definition 2. *Itô integral*

$$\int_0^T \underbrace{f(B_t)}_{\textit{left hand rule}} \, dB_t = \lim_{\Delta \to 0} \sum_{k=0}^{N-1} f(B_{t_k}) \Delta B_{t_k}$$

where $\Delta = \frac{T}{N}$, $t_k = k\Delta$, *and* $\Delta B_{t_k} = B_{t_{k+1}} - B_{t_k}$, *noting* $\Delta B_{t_k} \overset{iid}{\sim} N(0, \frac{T}{N}) \overset{d}{=} \frac{1}{\sqrt{N}} Z_k$ *gives us "CLT scaling".*

Stochastic Differential Equations (SDEs)

Note: more accurately, stochastic *integral* equations.

Definition 3. *Stochastic differential equation*

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t \implies X_t - X_0 = \underbrace{\int_0^t b(X_s)ds}_{\textit{Riemann}} + \underbrace{\int_0^t \sigma(X_s)dB_t}_{\textit{Itô}}$$

for $b, \sigma : \mathbb{R} \to \mathbb{R}$.

Heuristically, $dB_t \cdot dB_t = dt$, since

$$\sum_{k=0}^{N-1} f(B_{t_k}) \Delta B_{t_k}^2 \overset{d}{=} \sum_{k=0}^{N-1} f(B_{t_k}) \underbrace{\frac{Z_k^2}{N}}_{\text{"LLN scaling"}} \overset{N\to\infty}{\longrightarrow} \int_0^T f(B_t)dt.$$

Note: This can be made more precise, but this suffices for our purpose.

Itô's Lemma (chain rule)

- In ODE

$$\dot{X}_t = b(X_t)$$
$$\partial_t f(X_t) = f'(X_t)\dot{X}_t$$

- In SDE

$$dX_t = b(X_t)dt + \sigma(X_t)dB_t$$
$$df(X_t) = f'(X_t)b(X_t)dt + f'(X_t)\sigma(X_t)dB_t + \frac{1}{2}f''(X_t)\sigma(X_t)^2dt$$

Multivariate Form

- $f : \mathbb{R}^d \to \mathbb{R}$
- $b : \mathbb{R}^d \to \mathbb{R}^d$
- $\sigma = I_d$ (for simplicity)
- $\Delta = \sum_{i=1}^d \partial_{x_i}^2$ (Laplacian)

$$df(X_t) = \left\langle \nabla f(X_t), b(X_t) \right\rangle dt + \left\langle \nabla f(X_t), dB_t \right\rangle + \frac{1}{2}\Delta f(X_t)dt$$

Covariance Matrix

First, consider $\hat{\Sigma} = \frac{1}{n}X^TX$ with $X_{ij} \overset{iid}{\sim} N(0,1)$. We generalize this to the time dependent case by taking $\hat{\Sigma}(t) = \frac{1}{n}X(t)^TX(t)$ with $X_{ij} \overset{iid}{\sim}$ B.M.. Let $\lambda_i : \underbrace{\mathbb{R}^{d\times d}}_{SPD(d)} \to \mathbb{R}$.

We have

$$\text{Itô} \implies d \underbrace{\lambda_i(\hat{\Sigma}(t))}_{\text{“particle” } \lambda_i(t)} = \frac{2}{\sqrt{n}}\sqrt{\lambda_i(t)}dB_i(t) + \left(\frac{n-d}{n} + \underbrace{\frac{1}{n}\sum_{j=1, j\neq i}^{d} \frac{2\lambda_i(t)}{\lambda_i(t) - \lambda_j(t)}}_{\text{“Interaction”}} \right) dt.$$

Recall, $\hat{\rho}_t = \frac{1}{d}\sum_{i=1}^d \delta_{\lambda_i(t)}$ and $\hat{\varphi}_t(z) = \int \frac{1}{x-z}d\hat{\rho}_t(x) = \frac{1}{d}\sum_{i=1}^d \frac{1}{\lambda_i(t)-z}$. When $d, n \to \infty$ with $\frac{d}{n} \to \gamma$, we have

$$d\lambda_i(t) = 0dB_t + \left[(1-\gamma) + \gamma \int \frac{2\lambda_i(t)}{\lambda_i(t) - y} d\rho_t(y) \right] dt.$$

Now, back in the finite d, n case, we have (d notation is a bit confusing here)

$$d\hat{\varphi}_t(z) = \frac{1}{d}\sum_{i=1}^{d} d\left(\frac{1}{\lambda_i(t)-z}\right) = \frac{1}{d}\sum_{i=1}^{d}\frac{-1}{(\lambda_i(t)-z)^2}\left[(1-\gamma)+\frac{\gamma}{d}\sum_{j=1,j\neq i}^{d}\frac{2\lambda_i(t)}{\lambda_i(t)-\lambda_j(t)}\right]dt.$$

When $d, n \to \infty$ with $\frac{d}{n} \to \gamma$, we have

$$d\varphi_t(z) = \int \frac{-1}{(x-z)^2}\left[(1-\gamma)+\gamma\int\frac{2x}{x-y}d\rho_t(y)\right]d\rho_t(x)dt$$

With some algebraic steps, we get

$$\partial_t\varphi_t(z) = -\partial_z\left[(1-\gamma)\varphi_t(z)+\gamma z\varphi_t(z)^2\right],$$

which is a first order nonlinear PDE and may be solved by the method of characteristics to obtain

$$\varphi_t(z) = \frac{z}{z+(\gamma-1)t+\sqrt{(z-(\gamma+1)t)^2-4\gamma t^2}}$$

and we may note that, at $t = 1$, this solves

$$\frac{1}{\varphi(z)}+z = \frac{\gamma}{1+\varphi(z)}.$$

Inversion Formula

Using the Stieltjes inversion formula gives us

$$\rho(x) = \lim_{b\to 0_+}\operatorname{Im}\frac{\varphi_1(x+ib)}{\pi} = \frac{\sqrt{(\lambda_+-x)(x-\lambda_-)}}{2\pi\gamma x}$$

with $\lambda_\pm = (1\pm\sqrt{\gamma})^2$

Propagation of Chaos

We may observe that

$$\hat{\rho}(t) = \frac{1}{d}\sum_{i=1}^{d}\delta_{\lambda_i(t)} \xrightarrow{v} \rho_t \xleftarrow{v} \mathscr{L}(\lambda_i(t))$$

and

$$\mathscr{L}((\lambda_1,\ldots,\lambda_k)(t)) \xrightarrow{v} \rho_t^{\otimes k} \quad \text{(Independent!)}$$

Asymptotic Equivalence

Definition 4. *Asymptotic equivalence*
We say sequences $\{a_n\}$ and $\{b_n\}$ are asymptotically equivalent, denoted $a_n \sim b_n$, if $\frac{a_n}{b_n} \to 1$ as $n \to \infty$.

Proposition 1. *In the context of ridge regularization, with regularization strength $\lambda > 0$, we have*

$$\operatorname{Tr}\left(\hat{\Sigma}(\hat{\Sigma} + \lambda I)^{-1}\right) \sim \operatorname{Tr}\left(\Sigma(\Sigma + \frac{1}{\varphi(-\lambda)})^{-1}\right),$$

where $\hat{\Sigma} = \frac{1}{n} X^T X$ and $\Sigma = I$.

This will be relevant for our later discussion of double descent.