Lecture 8 : High dimensional Statistics

In this lecture, we will present some phenomenon that appear in high dimensional regimes. In modern Machine learning, one often has to cope with two asymptotics: high number of data and high dimension of the data. That leads to sometimes counter intuitive behavior, that are classically regrouped under the term "curse of dimensionality". To be able to rigorously set theoretical results in this kind of regime, two probabilistic theories have been intensively developed in the last 50 years: the concentration of the measure theory and the random matrix theory. We will provide in this lecture first elementary results of these theories and deduce insights on simple Statistical learning procedure like Kernel clustering and Ridge regression.

1 Concentration of the measure tools

1.1 Concentration in high dimension

The concentration of the measure theory is only relevant in high dimension where as Talagrand noted in [3]: "A random variable that depends (in a "smooth" way) on the influence of many independent variable (but not too much on any of them) is essentially constant". In this sentence the whole question is to know what "smooth" means. The smoothness of a mapping can be first described as a Lipschitz property that we describe below.

DEF **Definition 14.** *Given a mapping $f : \mathbb{R}^p \to \mathbb{R}^q$ and a parameter $\lambda > 0$, we say that f is λ-Lipschitz iif:*

$$\forall x, y \in \mathbb{R}^p : \qquad \|f(x) - f(y)\| \leq \lambda \|x - y\|.$$

The simplest idea of concentration of the measure is then, given a high-dimensional random vector $X \in \mathbb{R}^p$, to understand the concentration of X through the behavior the real random variables $f(X)$ for $f : \mathbb{R}^p \to \mathbb{R}$, 1-Lipschitz. For X Gaussian, one can for instance set the following result.

RESULT **Theorem 8.40** (Concentration of Gaussian vector[2])**.** *Given a deterministic vector $\mu \in \mathbb{R}^p$ and a Gaussian vector $Z \sim \mathcal{N}(\mu, I_p)$, for any 1-Lipschitz mapping $f : \mathbb{R}^p \to \mathbb{R}$:*

$$\mathbb{P}\left(|f(Z) - \mathbb{E}[f(Z)]| \geq t\right) \leq 2e^{-t^2/2} \tag{8.8}$$

Notably, the result of the theorem does not let appear the dimension p. The proof that led to the constants 2 is quite elaborate, we will thus refer the interested reader to the book of Ledoux [2]. Let us apply this result on simple mappings f.

Example 8.41. *Let us consider $Z \sim \mathcal{N}(0, I_p)$ and apply Theorem 8.40 to the mappings:*

1. *$f : z = (z_1, \ldots, z_p) \mapsto \frac{1}{\sqrt{p}} \sum_{i=1}^p z_i = \frac{\mathbb{1}^T z}{\sqrt{p}}$, introducing the deterministic vector $\mathbb{1} = (1, \ldots, 1) \in \mathbb{R}^p$. We can bound for any $x, y \in \mathbb{R}^p$:*

$$|f(x) - f(y)| = \frac{1}{\sqrt{p}} \left|\mathbb{1}^T (x - y)\right| \leq \frac{\|x - y\| \|\mathbb{1}\|}{\sqrt{p}} \leq \|x - y\|,$$

thanks to Cauchy-Shwarz inequality. Therefore, one can bound:

$$\mathbb{P}\left(\left|\frac{1}{p} \sum_{i=1}^p Z_i\right| \geq t\right) = \mathbb{P}\left(|f(Z) - \mathbb{E}[f(Z)]| \geq \sqrt{p}t\right) \leq 2e^{-pt^2/2} \underset{p \to \infty}{\longrightarrow} 0,$$

it implies the law of large numbers for Gaussian vectors. The same result would have happened for projections on any deterministic vector $u \in \mathbb{R}^p$, that is why we say that Gaussian vectors concentrate around any equator $\{z \in \mathbb{R}^p : u^T z = 0\}$.

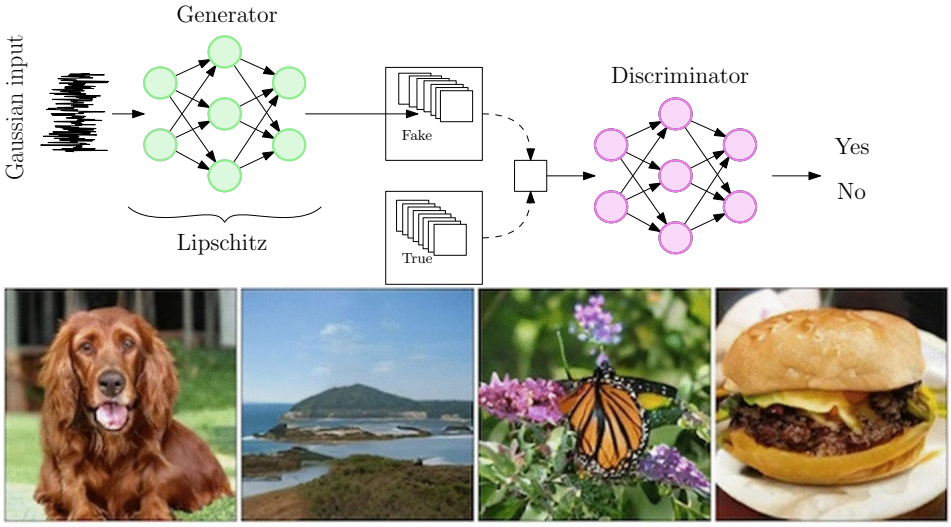

Figure 8.7: **(Top)** Schematic view of a GAN: the generator provides images that are concentrated by construction, Discriminator assert if the image is fake or real. **(Down)** Example of fake images created with GANs.

2. $f : z \mapsto \|z\|$.

$$\mathbb{P}\left(|\|Z\| - \mathbb{E}[\|Z\|]| \geq t\right) \leq 2e^{-t^2/2}$$

And recalling that $\mathbb{E}[\|Z\|]$ is of order $\sqrt{p}$, one can then understand the behavior represented on Figure 6.3: a Gaussian vector $Z \sim \mathcal{N}(\mu, I_p)$ of high dimension concentrates around the sphere $\sqrt{p}\mathbb{S}^{p-1}$.

The Gaussian example allows us to set similar concentration result for a wide range of random vectors since the class of "concentrated vectors" is stable through Lipschitz mappings.

RESULT
PROOF

Corollary 8.42. *Given a random vector $X : \Omega \to \mathbb{R}^p$, if there exists a random vector $Z \sim \mathcal{N}(0, I_q)$ and a λ-Lipschitz transformation $\Phi : \mathbb{R}^q \to \mathbb{R}^p$, for a parameter $\lambda > 0$, then for any 1-Lipschitz real-valued mapping $f : \mathbb{R}^p \to \mathbb{R}$:*

$$\mathbb{P}\left(|f(X) - \mathbb{E}[f(X)]| \geq t\right) \leq 2e^{-(t/\lambda)^2/2}. \tag{8.9}$$

Proof. Given a 1-Lipschitz mapping $f : \mathbb{R}^p \to \mathbb{R}$, let us bound:

$$\begin{aligned}\mathbb{P}\left(|f(X) - \mathbb{E}[f(X)]| \geq t\right) &\leq \mathbb{P}\left(|f(\Phi(Z)) - \mathbb{E}[f(\Phi(Z))]| \geq t\right) \\ &\leq \mathbb{P}\left(\left|\frac{1}{\lambda}f(\Phi(Z)) - \mathbb{E}\left[\frac{1}{\lambda}f(\Phi(Z))\right]\right| \geq \frac{t}{\lambda}\right) \leq 2e^{-(t/\lambda)^2/2},\end{aligned}$$

since the mapping $\frac{1}{\lambda}f \circ \Phi$ is 1-Lipschitz. □

Although the variations are bounded, Lipschitz transformation of a Gaussian vector $\Phi(Z)$ can allow complex dependencies between the entries.

This last corollary becomes particularly inspiring when considering the example of GAN images that are typically constructed as the image of a Gaussian vector through Lipschitz mappings. Therefore, by construction, their observations $f(X)$ follow the same concentration inequality as in (8.9) (the whole question is to compute the Lipschitz parameter of the generator, λ). Looking at the examples of fake images produced with GANs, one can understand why the concentration result given by Corollary 8.42 can be taken as a standard hypothesis in theoretical machine learning.

A question that naturally raises is then *"Can we express the concentration of non Lipschitz functionals?"* The answer is yes, there are many ways, however they are outside the scope of this course. In next section we present some key lemmas to deal with concentration inequalities on random variables.

1.2 Concentration of real observations

Theorem 8.40 and Corollary 8.42 allows to construct a wide range of concentration inequalities on random variables. We see in this section what conclusion can be made from those concentration inequalities. Let

us start with two fundamental inequalities of Probability theory that allows to go from concentration inequalities to bounds on moments and vice-versa:

RESULT **Lemma 8.43.** *Given a positive continuous random variable $Z : \Omega \to \mathbb{R}_+$, and an increasing mapping $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\mathbb{E}[\phi(Z)] \leq \infty$:*

$$\mathbb{P}(Z \geq t) \leq \frac{\mathbb{E}[\phi(Z)]}{\phi(t)} \qquad \text{and} \qquad \mathbb{E}[\phi(Z)] \leq \int_0^\infty \mathbb{P}(\phi(Z) \geq t)dt$$

PROOF *Proof.* The first result is the classical Markov inequality, the second result is proven with Fubini theorem (to swap integral signs) and using the identity $\phi(z) = \int_{\mathbb{R}_+} \mathbb{1}_{[0,\phi(z)]}(t)dt =:$

$$\mathbb{E}[\phi(Z)] = \int_{\mathbb{R}_+} \phi(z)d\mathbb{P}(z) = \int_{\mathbb{R}_+}\int_{\mathbb{R}_+} \mathbb{1}_{[0,\phi(z)]}(t)d\mathbb{P}(z)dt = \int_{\mathbb{R}_+}\int_{\mathbb{R}_+} \mathbb{1}_{[t,+\infty]}(\phi(z))d\mathbb{P}(z)dt = \int_{\mathbb{R}_+} \mathbb{P}(\phi(Z) \geq t)dt,$$

since, for all $z, t \in \mathbb{R}_+$:

$$\mathbb{1}_{[0,\phi(z)]}(t) = 1 \qquad \Longleftrightarrow \qquad \mathbb{1}_{[t,+\infty]}(\phi(z)) = 1.$$

□

With this Lemma, one can then bound all the moments of an exponentially concentrated[4] random variables.

RESULT **Lemma 8.44.** *Given a random variable $Z : \Omega \to \mathbb{R}$ such that $\mathbb{P}(|Z| \geq t) \leq 2e^{-(t/\eta)^q}$ for $q > 0$, then for any $r > 0$, there exists a constant $C > 0$ independent with η such that:*

$$\mathbb{E}[|Z|^r] \leq C\eta^r$$

Here we just track the dependence on η, this that is the only term that will contain some dimensional components as we will see in some examples (we have already seen it in Example 8.41, Item 1.)

Proof. From Lemma 8.43, one can deduce (with the change of variables $t^{1/r} \to u$ and $u/\eta \to v$):

$$\begin{aligned}\mathbb{E}[|Z|^r] &= \int_{\mathbb{R}_+} \mathbb{P}(|Z|^r \geq t)dt = \int_{\mathbb{R}_+} \mathbb{P}(|Z| \geq t^{\frac{1}{r}})dt = \int_{\mathbb{R}_+} ru^{r-1}\mathbb{P}(|Z| \geq u)dt \\ &\leq 2\int_{\mathbb{R}_+} ru^{r-1}e^{-(u/\eta)^q}dt = 2r\eta^r \int_{\mathbb{R}_+} v^{r-1}e^{-v^q}dt \leq C\eta^r.\end{aligned}$$

□

Proposition 8.45. *Given two random variables $X, Y : \Omega \to \mathbb{R}$ and four parameters a, b, σ, θ such that:*

$$\mathbb{P}(|X - a| \geq t) \leq 2e^{-(t/\sigma)^2} \qquad \text{and} \qquad \mathbb{P}(|Y - b| \geq t) \leq 2e^{-(t/\theta)^2}$$

one can express the concentration of the product XY around ab followingly:

$$\mathbb{P}(|XY - ab| \geq t) \leq 4e^{-(t/3\max(\theta|a|,\sigma|b|))^2} + 4e^{-t/3\max(\sigma^2,\theta^2)}$$

Proof. The proof relies on the implication, true for any $x, y, z, t \geq 0$:

$$x + y + z \geq t \quad \Rightarrow \quad x \geq \frac{t}{3} \quad \text{or} \quad y \geq \frac{t}{3} \quad \text{or} \quad z \geq \frac{t}{3}. \qquad \text{and} \qquad xy \geq t \quad \Rightarrow \quad x \geq \sqrt{t} \quad \text{or} \quad y \geq \sqrt{t}.$$

Then, the algebraic identity $xy - ab = (x - a)(y - b) + a(y - b) + b(x - a)$ leads to:

$$\begin{aligned}\mathbb{P}\left(|XY - ab| \geq t\right) &\leq \mathbb{P}\left(|X - a|\,|Y - b| + |X - a||b| + |Y - b||a| \geq t\right) \\ &\leq \mathbb{P}\left(|X - a| \geq \sqrt{\frac{t}{3}}\right) + \mathbb{P}\left(|Y - b| \geq \sqrt{\frac{t}{3}}\right) \\ &\quad + \mathbb{P}\left(|X - a|\,|b| \geq \frac{t}{3}\right) + \mathbb{P}\left(|Y - b|\,|a| \geq \frac{t}{3}\right).\end{aligned}$$

□

[4] An exponentially concentrated random variable is a random variable that follows a concentration inequality similar to the one on $f(X)$ in (8.8) with an exponential dependence on t on the right hand term.

If $a = \mathbb{E}[X]$ and $b = \mathbb{E}[Y]$ then Proposition 8.45 sets that XY concentrates around $\mathbb{E}[X]\mathbb{E}[Y]$ one can then wonder how expresses the concentration around $\mathbb{E}[XY]$. That question is solved by the two next lemmas.

Lemma 8.46. *Given a random variable $X : \Omega \to \mathbb{R}$ and a deterministic scalar $a \in \mathbb{R}$, such that:*

$$\mathbb{P}\left(|X - a| \geq t\right) \leq 2e^{-(t/\eta)^2},$$

if there exists a scalar $b \in \mathbb{R}$ and a constant $C > 0$, independent of η such that: $|a - b| \leq C\eta$ then:

$$\mathbb{P}\left(|X - b| \geq t\right) \leq C' e^{-(t/2\eta)^2},$$

for some constant C' independent with η.

One can employ this lemma in the case $b = \mathbb{E}[X]$ since one knows from Lemma 8.44 that there exists a constant $C > 0$ independent of η such that:

$$|a - \mathbb{E}[X]| \leq |\mathbb{E}[a - X]| \leq \mathbb{E}\left[|X - a|\right] \leq C\eta. \tag{8.10}$$

Proof. Let us start with the bound

$$|X - \mathbb{E}[X]| \leq |X - a| + |a - \mathbb{E}[X]|$$

and:

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq t\right) \leq \mathbb{P}\left(|X - a| \geq t - |a - \mathbb{E}[X]|\right) \leq 2\exp\left(-\left(\frac{t - C\eta}{\eta}\right)^2\right).$$

One can then note that if $t > 2C\eta$, then: $\frac{t}{2} = t - \frac{t}{2} \leq t - C\eta$ and therefore (since $u \mapsto e^{-(u/\eta)^2}$ is decreasing):

$$2\exp\left(-\left(\frac{t - C\eta}{\eta}\right)^2\right) \leq 2e^{-\left(\frac{t}{2\eta}\right)^2}.$$

if $t \leq 2C\eta$, one can still bound $\frac{t}{2} \leq C\eta$ and therefore:

$$\mathbb{P}\left(|X - \mathbb{E}[X]| \geq t\right) \leq 1 = \frac{e^{-\left(\frac{C\eta}{\eta}\right)^2}}{e^{-C^2}} \leq e^{C^2} e^{-\left(\frac{t}{2\eta}\right)^2}.$$

One then retrieve the second result of the lemma setting $C' \equiv \max\left(2, e^{C^2}\right)$. □

This last lemma allows us to prove a weak version the famous Hanson-Wright concentration inequality, very important in random matrix theory.

RESULT **Corollary 8.47** (Hanson-Wright)**.** *Given a deterministic symmetric positive matrix[5] $A \in \mathbb{R}^{p\times p}$ random vectors $Y : \Omega \to \mathbb{R}^p$ such that $X = \Phi(Z)$ with $Z \sim \mathcal{N}(0, I_q)$ and $\Phi : \mathbb{R}^q \to \mathbb{R}^p$, λ-Lipschitz:*

$$\mathbb{P}\left(|X^T AY - \mathbb{E}[X^T AY]| \geq t\right) \leq C\exp\left(-\frac{1}{2}\left(\frac{ct}{\lambda\|A\|\sqrt{p}\sigma}\right)^2\right) + C\exp\left(-\frac{1}{2}\left(\frac{ct}{\lambda^2\|A\|}\right)\right),$$

where $\sigma > 0$ is a parameter that satisfies $\sigma = \max(\sqrt{\frac{1}{p}\mathbb{E}[X^T X]}, \sqrt{\frac{1}{p}\mathbb{E}[Y^T Y]})$ and $C, c > 0$ are two constants independent of λ, σ.

Note that if $X \sim \mathcal{N}(0, I_p)$, $\sigma = \sqrt{\frac{1}{p}\mathbb{E}[X^T X]} = 1$.

Proof. One can express:

$$X^T AX = \|A^{\frac{1}{2}} X\|^2.$$

[5] Modifying a bit the concentration constants, the result is actually true for general matrices $A \in \mathbb{R}^{p\times p}$.

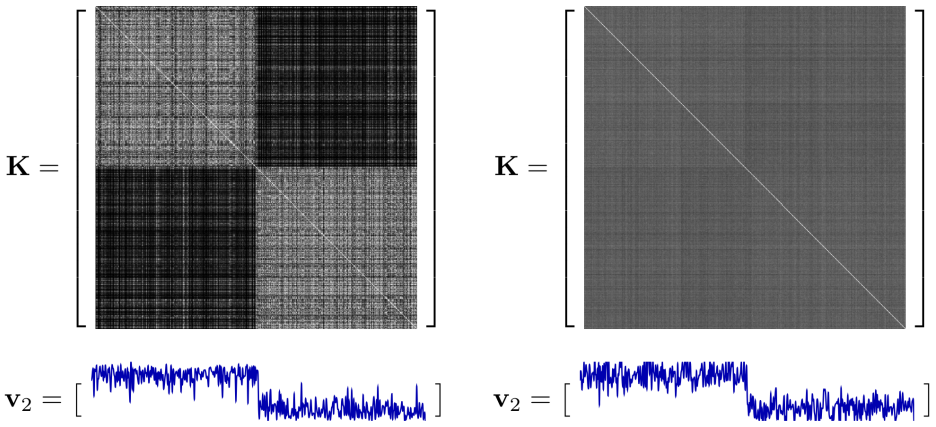

Figure 8.8: Gray scale view of heat Kernel matrices and the second top eigenvectors v_2 for small **(left, p = 5, n = 500)** and large **(right, p = 250, n = 500)** dimensional data. **(Down)**. The data random vectors belong to two balanced classes $C_1 : x \sim \mathcal{N}(\mu, Ip)$ and $C_2 : x \sim \mathcal{N}(-\mu, Ip)$ where $\mu = (1, 0, \dots, 0)$.

Knowing that:

$$\mathbb{P}\left(|\|A^{\frac{1}{2}}X\| - \mathbb{E}[\|A^{\frac{1}{2}}X\|]| \geq t\right) \leq 2\exp\left(-\frac{1}{2}\left(\frac{t}{\|A\|^{\frac{1}{2}}\lambda}\right)^2\right),$$

one can use Proposition 8.45 to set:

$$\mathbb{P}\left(|X^TAX - \mathbb{E}[\|A^{\frac{1}{2}}X\|^2]| \geq t\right) \leq 4\exp\left(-\frac{1}{2}\left(\frac{t}{3\lambda\|A^{\frac{1}{2}}\|\mathbb{E}[\|A^{\frac{1}{2}}X\|]}\right)^2\right) + 4\exp\left(-\frac{1}{2}\left(\frac{t}{3\lambda^2\|A\|}\right)\right). \tag{8.11}$$

Besides, thanks to Jensen inequality:

$$\mathbb{E}[\|A^{\frac{1}{2}}X\|] = \mathbb{E}[\sqrt{X^TAX}] \leq \sqrt{\mathbb{E}[X^TAX]} \leq \sqrt{p\|A\|}\sigma.$$

Finally, to replace "$\mathbb{E}[\|A^{\frac{1}{2}}X\|^2]$" with "$\mathbb{E}[X^TAX]$" in (8.11), one can employ Lemma 8.46 and (8.10). □

Remark 8.48. *One can obtain a similar concentration result for the concentration of X^TAY where $Y : \Omega \to \mathbb{R}^p$ is a supplementary random vector such that $(X, Y) = \Psi(Z)$ with $Z \sim \mathcal{N}(0, I_q)$ and $\Psi : \mathbb{R}^q \to \mathbb{R}^{2p}$, λ-Lipschitz. It just comes from the fact that:*

$$X^TAY = \frac{1}{4}\left((X+Y)^TA(X+Y) - (X-Y)A(X-Y)\right)$$

and that $X+Y$ and $X-Y$ are both 2λ-transformation of $Z \sim \mathcal{N}(0, I_q)$ under those hypotheses. One can then employ Corollary 8.47 to express the concentration of X^TAX.

Example 8.49 (Spectral clustering intuition)**.** *In spectral clustering the setting is generally unsupervised (no Y) and one wants to study the eigenvalues of the matrix:*

$$K \equiv (K(x_i, x_j))_{i,j\in[n]},$$

where $K : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ is a kernel matrix whose entry $K(x_i, x_j)$ should represent the similarity between x_i and x_j. If the kernel writes:

$$K(x_i, x_j) \equiv f\left(\frac{1}{p}\|x_i - x_j\|^2\right),$$

for a certain mapping $f : \mathbb{R}_+ \to \mathbb{R}_+$ (that would be $f : t \mapsto e^{-t^2}$ for the heat Kernel). The concentration of the measure theory then asserts that the similarity "$\|x_i - x_j\|^2$" will concentrates around its expectation, that is actually constant intra and interclass when the dimension increases. Considering two data x_i, x_j and denoting μ_i, μ_j their expectations, one can estimate:

$$\frac{1}{p}\|x_i - x_j\|^2 = \frac{1}{p}\|\mu_i - \mu_j\|^2 + \frac{2}{p}(\mu_i - \mu_j)^T(z_i - z_j) + \frac{1}{p}\|z_i\|^2 + \frac{1}{p}\|z_j\|^2 + \frac{2}{p}z_i^Tz_j \tag{8.12}$$

We know that:

$$\mathbb{E}[(\mu_i - \mu_j)^T(z_i - z_j)] = \mathbb{E}\left[\frac{2}{p} z_i^T z_j\right] = 0,$$

and we can assume that as for Gaussian vectors, $\forall i \in [n]$:

$$\mathbb{E}\left[\frac{1}{p}\|z_i\|^2\right] = 1.$$

Then Corollary 8.47 allows us to set that:

- $\mathbb{P}\left(\frac{1}{p}\left|(\mu_i - \mu_j)^T(z_i - z_j)\right| \geq t\right) \leq 2\exp\left(-\left(\frac{pt}{\|\mu_i - \mu_j\|}\right)^2\right)$
- $\forall i, j \in [n]$, $i \neq j$: $\mathbb{P}\left(\left|\frac{2}{p} z_i^T z_j\right| \geq t\right) \leq 4\exp\left(-\frac{1}{2}\left(\frac{\sqrt{p}t}{6}\right)^2\right) + 4\exp\left(-\frac{1}{2}\left(\frac{pt}{6}\right)\right)$,
- $\forall i \in [n] : \mathbb{P}\left(\left|\frac{1}{p} z_i^T z_i - 1\right| \geq t\right) \leq 4\exp\left(-\frac{1}{2}\left(\frac{\sqrt{p}t}{3}\right)^2\right) + 4\exp\left(-\frac{1}{2}\left(\frac{pt}{3}\right)\right)$,

In difficult settings of binary classification where μ_i either equals $+\mu$ or $-\mu$ and $\|\mu\| \leq_{p\to\infty} O(1)$, $\frac{1}{p}\|\mu_i - \mu_j\| \leq O(\frac{1}{p})$ and (8.12) can be approximated (thanks to concentration of the measures inferences):

$$\frac{1}{p}\|x_i - x_j\|^2 \approx \frac{1}{p}\|z_i\|^2 + \frac{1}{p}\|z_j\|^2 + \frac{2}{p} z_i^T z_j + O\left(\frac{1}{p}\right).$$

Therefore, the the intuition of low dimensional spectral clustering does not work anymore in high dimension as pictured on Figure 8.8, all the entries of K look the same iner and intraclass. One needs to resort to random matrix Theory to get precise insights on the performances of such methods.

2 Random matrix Theory

In linear algebra, the set of matrices is generally denoted $\mathcal{M}_{p,n} \equiv \mathbb{R}^{p\times n}$, we will therefore let appear those two notations in the sequel. The set of square matrices of size p is denoted $\mathcal{M}_p \equiv \mathbb{R}^{p\times p}$. The euclidean norm on $\mathbb{R}^p$ is denoted $\|\cdot\|$ ($\|x\| \equiv \sqrt{\sum_{i=1}^p x_i^2}$), then the Hilbert-Schmidt norm (or Frobenus norm) is denoted $\|\cdot\|_{\mathrm{F}}$ ($\forall M \in \mathcal{M}_{p,n}$: $\|M\|_{\mathrm{F}} = \sqrt{\mathrm{Tr}(MM^*)} = \sup_{\|A\|_{\mathrm{F}}\leq 1} |\mathrm{Tr}(AM)|$), the spectral norm is denoted $\|\cdot\|$ ($\|M\| = \sup_{\|x\|=1}\|Mx\|$).

2.1 Random matrix theory in machine learning

Random matrices naturally arise in machine learning since one often has to consider data matrices $X = (x_1, \ldots, x_n) \in \mathbb{R}^{p\times n}$ (one could imagine the $x_1, \ldots, x_n$ as being independent samples of a distribution of images, time series, features, or any other measurements). This matrix, as it contains all the information about the available data will appear explicitly or implicitly in the computations when one tries to prove theoretical guarantees to a given method.

Example 8.50 (Ridge regression)**.** *Given a training data set $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ with $(x_1, \ldots, x_n) \in \mathbb{R}^{p\times n}$ and $Y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, the Ridge regression problem is a linear method that looks for a parameter $\beta \in \mathbb{R}^p$ such that the projection $f_\beta(x_i) \equiv \beta^T x_i$ is close to y_i for all i. The loss one wants to minimize is the MSE:*

$$l(f_\beta(x), y) = \|\beta^T x - y\|^2,$$

but in Ridge regression, one adds a regularizing term $\gamma\|\beta\|^2$, with $\gamma > 0$ to avoid over-fitting. Then the Ridge-regression problem expresses:

$$\min_{\beta\in\mathbb{R}^p} L_\beta(X, Y).$$

where:

$$L_\beta(X, Y) = \frac{1}{n}\sum_{i=1}^n (x_i^T\beta - y_i)^2 + \gamma\|\beta\|^2 = \frac{1}{n}\|X^T\beta - Y\|^2 + \gamma\|\beta\|^2.$$

The solution can be computed cancelling the gradient:

$$\frac{\partial L_\beta(X,Y)}{\partial \beta} = 0 \quad \Longleftrightarrow \quad \frac{1}{n}XX^T\beta - \frac{1}{n}XY + \gamma\beta = 0 \quad \Longleftrightarrow \quad \left(\gamma I_p + \frac{1}{n}XX^T\right)\beta = \frac{1}{n}XY.$$

Therefore, denoting $Q \equiv \left(\gamma I_p + \frac{1}{n}XX^T\right)^{-1}$*, one obtains:*

$$\beta = \frac{1}{n}QXY,$$

we see appearing the so-called "sample covariance matrix" $\frac{1}{n}XX^T$ *that we will study later. Then the Mean square error (the MSE) on the training data set write:*

$$MSE = \left\|X^T\beta - Y\right\|^2 = \left\|\frac{1}{n}X^TQXY - Y\right\|^2$$

Let us then note that (it is trivial for square invertible matrices, it can be extended to rectangular matrices):

$$\frac{1}{n}X^TQX^T = \frac{1}{n}X\left(\frac{1}{n}XX^T + \gamma I_p\right)^{-1}X^T = \frac{1}{n}X^TX\left(\frac{1}{n}X^TX + \gamma I_p\right)^{-1} = I_p - \gamma\check{Q}, \tag{8.13}$$

where we introduced $\check{Q} = \left(\frac{1}{n}X^TX + \gamma I_p\right)^{-1}$*, the so-called "co-resolvent". One can then set:*

$$MSE_{tr} = \left\|X^T\beta - Y\right\|^2 = \gamma^2\left\|\check{Q}Y\right\|^2 = \gamma^2 Y^T\check{Q}^2Y. \tag{8.14}$$

We see here that the performances express with the resolvant $\check{Q}$*. One could be tempted to approximate:*

$$\check{Q} \approx \left(\gamma I_p + \frac{1}{n}\mathbb{E}[X^TX]\right)^{-1},$$

but that is wrong for big values of p*. To have an exact estimation of this matrix, one needs to resort to random matrix result whose the resolvent, and the co-resolvent are central object. We see here that already with a very simple object, random matrix theory seems indispensable.*

2.2 Assumptions and first properties of the resolvent

In all this section, we will consider a random matrix $X = (x_1, \ldots, x_n) : \Omega \to \mathbb{R}^{p\times n}$. To simplify the calculus, we will assume that all the columns $x_i : \Omega \to \mathbb{R}^p$, $i = 1, \ldots, n$ are identically distributed and denote:

$$\mu \equiv \mathbb{E}[x_1] = \cdots = \mathbb{E}[x_n] \qquad \text{and} \qquad \Sigma \equiv \mathbb{E}[x_1^Tx_1] = \cdots = \mathbb{E}[x_n^Tx_n]$$

and assume the following assumptions:

(A1) $x_1, \ldots, x_n$ are independent

(A2) $X = \phi(Z)$ with $Z \sim \mathcal{N}(0, I_q)$ for a given $q \in \mathbb{N}$ and $\Phi : \mathbb{R}^q \to \mathbb{R}^p$ 1-Lipschitz.

(A3) $\|\mu\| \leq 1$

To stay as simple as possible, given a parameter $\gamma > 0$, we will just study the resolvent[6]:

$$Q = Q(\gamma) \equiv (\gamma I_p + \frac{1}{n}XX^T)^{-1}$$

which is a random matrix as $\frac{1}{n}XX^T$.

Remark 8.51. *A famous result of complex analysis (that we will not detail here) ensures that if one is able to estimate* $Q(\gamma)$ *for all* $\gamma \in \mathbb{C}$*, with* $\Im(\gamma) > 0$*, then one is able to estimate the spectrum of* $\frac{1}{n}XX^T$ *(which is a random object like* $\frac{1}{n}XX^T$ *and concentrates around a certain distribution). This was the first motivation of random matrices, but we saw in* (8.14) *that the resolvent can actually appear in statistical learning independently of spectral inferences. We will thus study directly and strictly* Q *in the following, that will though still provide a flavor of random matrix calculus.*

[6]Looking at example 8.50, one might be more likely to work with $\check{Q}$, but the study of $\check{Q}$ is slightly more elaborate and anyway, the two resolvent satisfy the relation (8.13), that allows to go from one to the other easily.

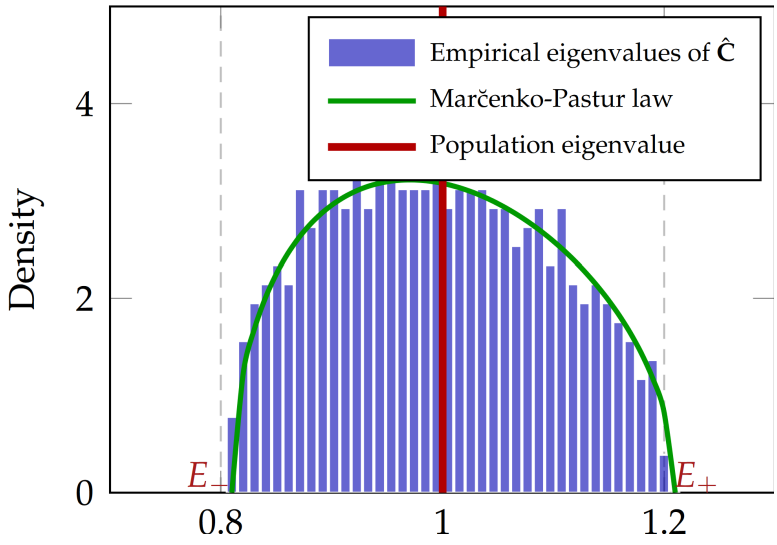

Figure 8.9: Histogram of the eigenvalues of the empirical covariance matrix $\frac{1}{n}XX^T$, Marcenko-Pastur distribution that approximates the spectral distribution $\mu = \frac{1}{p}\sum_{\gamma \in (\frac{1}{n}XX^T} \delta_\gamma$ and eigenvalues of the population covariance matrix $\mathbb{E}[\frac{1}{n}XX^T] = I_p$: δ_1. $X = (x_1, \dots, x_n) \in \mathbb{R}^{p\times n}$ with $x_1, \dots, x_n \in \mathbb{R}^p$ independent and satisfy $\forall i \in [n] : x_i \sim \mathcal{N}(0, I_p)$, $p = 500$, $n = 50000$.

Since we are working in the quasi-asymptotic regime where p and n are big, we allow ourselves to introduce two general notations "C, c" or "$O(1)$" to denote the constants that are independent with p, n. These universal constants C, c could be any positive real value that should stay the same in any setting with a different p and n. We then add our last assumption that will allow to simplify the result:

(A4) $p \leq O(n)$

All the following results will be valid under the assumptions $(A1-4)$, for simplicity, we will not mention those assumptions in the sequel.

Let us start with some important lemmas to control $\|\Sigma\|$, $\|Q\|$ and the concentration of Q.

Let us first show some important properties

RESULT

Lemma 8.52. *There exists a constant C, independent with p, n such that:*

$$\|\Sigma\| \leq C$$

PROOF

Proof. One knows that for any deterministic vector $u \in \mathbb{R}^p$ and any $i \in [n]$:

$$\mathbb{P}\left(|u^T(x_i - \mu_i)| \geq t\right) \leq 2e^{-\frac{t^2}{2}}.$$

which implies from Lemma 8.44 that:

$$\mathbb{E}\left[u^T(x_i - \mu_i)(x_i - \mu_i)^T u\right] = \mathbb{E}\left[|u^T(x_i - \mu_i)|^2\right] \leq C.$$

Noting that $\mathbb{E}\left[u^T(x_i - \mu_i)(x_i - \mu_i)^T u\right] = \mathbb{E}\left[u^T x_i x_i^T u\right] - (u^T\mu_i)^2 = u^T(\Sigma - \mu^T\mu)u$, one can finally bound:

$$\|\Sigma\| = \sup_{\|u\|\leq 1} u^T\Sigma u \leq \sup_{\|u\|\leq 1} u^T(\Sigma - \mu\mu^T)u + (u^T\mu)^2 \leq C,$$

for some constant $C > 0$ independent with p, n, since we know from assumption **(A3)** that $\|\mu\| \leq C$. □

Lemma 8.53. $\|Q\| \leq \frac{1}{\gamma}$ *and* $\frac{1}{n}\|XQ\| \leq \sqrt{\frac{2}{n\gamma}}$ *(recall that $\gamma > 0$).*

Proof. The proof relies on the fact that Q (and Q^{-1}) are positive symmetric matrices. One can then bound with the traditional order relation[7] on the set of positive symmetric matrices:

$$Q^{-1} = \gamma I_p + \frac{1}{p}XX^T \geq \gamma I_p.$$

[7] $A \leq B \Leftrightarrow \forall x \in \mathbb{R}^p : x^T(B - A)x \geq 0.$

That implies:

$$Q \leq (\gamma I_p)^{-1} = \frac{1}{\gamma} I_p.$$

The second bound come from the identity:

$$Q\frac{1}{n}XX^T == \left(\gamma I_p + \frac{1}{n}XX^T\right)^{-1}\frac{1}{n}XX^T = I_p - \gamma Q \tag{8.15}$$

Noting that for any $A \in \mathbb{R}^{p\times p}$:

$$\|A\| = \sup_{\|u\|,\|v\|\leq 1} u^T A v \leq \sup_{\|u\|,\|v\|\leq 1} \sqrt{u^T A A^T u v^T v} = \sqrt{\|AA^T\|},$$

thanks to the Cauchy-Shwarz inequality, one can bound thanks to (8.15):

$$\left\|\frac{1}{n}QX\right\| \leq \frac{1}{\sqrt{n}}\sqrt{\left\|\frac{1}{n}QXX^TQ\right\|} \leq \frac{1}{\sqrt{n}}\sqrt{\|Q-\gamma Q^2\|} \leq \frac{1}{\sqrt{n}}\sqrt{\|Q\| + \gamma\|Q^2\|} \leq \sqrt{\frac{2}{n\gamma}} \tag{8.16}$$

□

RESULT **Proposition 8.54.** *For any mapping $f : \mathbb{R}^{p\times p} \to \mathbb{R}$, 1-Lipschitz for the Frobenus norm on $\mathbb{R}^{p\times p}$, there*
PROOF *exist some constants $c > 0$, independent of p, n such that:*

$$\mathbb{P}\left(|f(Q) - \mathbb{E}[f(Q)]| \geq t\right) \leq 2e^{-cnt^2}.$$

Proof. Introducing the mapping $\Phi : \mathcal{M}_{p,n} \to \mathcal{M}_p$ defined as:

$$\Phi(M) = \left(\gamma I_p + \frac{MM^T}{n}\right)^{-1},$$

it is sufficient to show that Φ is $O(1/\sqrt{n})$-Lipschitz (for the Hilbert-Schmidt norm). For any $M \in \mathcal{M}_{n,p}$ and any $H \in \mathcal{M}_{p,n}$, we can bound:

$$\left\|d\Phi|_M \cdot H\right\|_{\mathrm{F}} = \left\|\Phi(M)\frac{1}{n}(MH^T + HM^T)\Phi(M)\right\|_{\mathrm{F}} \leq 2\sqrt{\frac{2}{n\gamma}}\|H\|_{\mathrm{F}}$$

thanks to Lemma 8.53. □

Now that we know that Q is concentrated, the next step is to find a deterministic matrix $\tilde{Q} \in \mathbb{R}^{p\times p}$ such that:

$$\|\mathbb{E}[Q] - \tilde{Q}\|_{\mathrm{F}} \leq O\left(\frac{1}{\sqrt{n}}\right).$$

Such a matrix is commonly called in random matrix literature a "deterministic equivalent" of Q.

2.3 Deterministic equivalent of the resolvent

To choose a deterministic equivalent equal to $(\gamma I_p + \Sigma)^{-1}$ (where $\Sigma = \frac{1}{n}\mathbb{E}[XX^T] = \mathbb{E}[x_i x_i^T]$ for all $i \in [n]$) would be too naive and indeed does not work. An efficient approach is to look for a deterministic equivalent of Q depending on a deterministic parameter $\Delta \in \mathbb{R}$ that we will provide later and having the form:

$$\tilde{Q}^\Delta = \left(\gamma I_p + \frac{\Sigma}{\Delta}\right)^{-1} \tag{8.17}$$

One can then express the difference with the expectation $\mathbb{E}[Q]$ followingly:

$$\mathbb{E}[Q] - \tilde{Q}^\Delta = \mathbb{E}\left[Q\left(\frac{1}{n}XX^T - \Sigma^\Delta\right)\tilde{Q}^\Delta\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[Q\left(x_i x_i^T - \frac{\Sigma}{\Delta}\right)\tilde{Q}^\Delta\right].$$

To pursue the estimation of the expectation, one needs to control the dependence between Q and x_i. For that purpose, one uses classically the Schur identities.

Lemma 8.55. $Q = Q_{-i} + \frac{1}{n}\frac{Q_{-i}x_i x_i^T Q_{-i}}{1-\frac{1}{n}x_i^T Q_{-i}x_i}$ *and* $Qx_i = \frac{Q_{-i}x_i}{1-\frac{1}{n}x_i^T Q_{-i}x_i}$ *where* $Q_{-i} = (\gamma I_p + \frac{1}{n}X_{-i}X_{-i}^T)^{-1}$ *and* $X_{-i} = (x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_n) \in \mathcal{M}_{p,n}$

Proof. The so called "resolvent identity":

$$A^{-1}(B-A)B^{-1} = A^{-1} - B^{-1} \tag{8.18}$$

applied to $A = Q^{-1}$ and $B = Q_{-i}^{-1}$ gives:

$$Q - Q_{-i} = Q\left(\frac{1}{n}X_{-i}X_{-i}^T - \frac{1}{n}XX^T\right)Q_{-i} = -\frac{1}{n}Qx_{-i}x_{-i}^T Q_{-i}. \tag{8.19}$$

Therefore multiplying on the right with x_{-i}, one gets:

$$Qx_i = Q_{-i}x_i - \frac{1}{n}Qx_{-i}x_{-i}^T Q_{-i}x_i,$$

thus:

$$Qx_i = \frac{Q_{-i}x_i}{1+\frac{1}{n}x_{-i}^T Q_{-i}x_i},$$

and one deduces easily the first result of the lemma from (8.19). □

Let us then pursue the computations with the new notation:

$$\forall i \in [n]: \qquad \Lambda_i \equiv 1 + \frac{1}{n}x_i^T Q_{-i}x_i$$

In particular the results of Lemma 8.55 rewrite:

$$Qx_i = \frac{Q_{-i}x_i}{\Lambda_i} \qquad \text{and} \qquad Q = Q_{-i} + \frac{Q_{-i}x_i x_i^T Q_{-i}}{\Lambda_i}. \tag{8.20}$$

Recalling that $Q - Q_{-i} = \frac{1}{n}Qx_i x_i^T Q_{-i}$, it is then possible to express:

$$\mathbb{E}[Q] - \tilde{Q}^{\Delta} = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[Q_{-i}\left(\frac{x_i x_i^T}{\Lambda_i} - \frac{\Sigma_i}{\Delta}\right)\tilde{Q}^{\Delta}\right] + \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[(Q_{-i}-Q)\frac{\Sigma_i}{\Delta}\tilde{Q}^{\Delta}\right] \tag{8.21}$$

From this decomposition, one is enticed into choosing in a first step for Δ the value:

$$\hat{\Lambda} \equiv \mathbb{E}[\Lambda_i] \in \mathbb{R}.$$

Since $x_1, \ldots, x_n$ are identically distributed, $\Lambda_1, \ldots, \Lambda_n$ are also identically distributed (but not independent!)

One can then bound for any $u \in \mathbb{R}^p$:

$$u^T\left(\mathbb{E}[Q] - \tilde{Q}^{\hat{\Lambda}}\right)u = \frac{1}{n}\sum_{i=1}^n \varepsilon_i + \frac{1}{n}\sum_{i=1}^n \delta_i \tag{8.22}$$

with:

- $\varepsilon_i = \mathbb{E}\left[u^T Q_{-i}x_i x_i^T \tilde{Q}^{\hat{\Lambda}}u\left(\frac{1}{\Lambda_i} - \frac{1}{\hat{\Lambda}}\right)\right]$
- $\delta_i = \mathbb{E}\left[u^T(Q_{-i}-Q)\frac{\Sigma_i}{\hat{\Lambda}}\tilde{Q}^{\hat{\Lambda}}u\right]$

To bound those two quantities, one will use the following lemmas:

Lemma 8.56. $\mathbb{P}\left(\left|u^T Q_{-i}x_i\right| \geq t\right) \leq Ce^{-ct^2}$ *and* $\mathbb{P}\left(\left|u\tilde{Q}^{\hat{\Lambda}}x_i\right| \geq t\right) \leq Ce^{-ct^2}$, *for some constants* $C, c > 0$ *independent with* n, p.

Proof. Let us only show the concentration of $u^T Q_{-i} x_i$ since it is more difficult than the concentration of $u^T \tilde{Q}^{\hat{\Lambda}} x_i$. Noting that $\mathbb{E}[u^T Q_{-i} x_i | X_{-i}] = u^T Q_{-i} \mu$ and $\mathbb{E}[u^T Q_{-i} x_i] = \mathbb{E}[u^T Q_{-i} \mu]$, one can bound with the formula given by Lemma 6.23:

$$\begin{aligned}
&\mathbb{P}\left(\left|u^T Q_{-i} x_i - \mathbb{E}[u^T Q_{-i} x_i]\right| \geq t\right) \\
&\quad \leq \mathbb{E}\left[\mathbb{P}\left(\left|u^T Q_{-i} x_i - \mathbb{E}[u^T Q_{-i} x_i | X_{-i}]\right| \geq t | X_{-i}\right)\right] + \mathbb{P}\left(\left|u^T Q_{-i} \mu - \mathbb{E}[u^T Q_{-i} \mu]\right| \geq t\right) \\
&\quad \leq \mathbb{E}\left[C' e^{-c(t/\|Q_{-i}\|)^2}\right] + C' e^{-c(t/\sqrt{n})^2} \quad \leq \quad C e^{-c' t^2}
\end{aligned}$$

thanks to the hypothesis on the concentration of x_i, the bound on $\|Q_{-i}\|$ given in Lemma 8.53 and Proposition 8.54. One can besides bound thanks to Jensen inequality:

$$\mathbb{E}[u^T Q_{-i} x_i] \leq \sqrt{\mathbb{E}[u^T Q_{-i} x_i x_i^T Q_{-i} u]} \leq \sqrt{\mathbb{E}[u^T Q_{-i} \Sigma_i Q_{-i} u]} \leq O(1).$$

One can then apply Lemma 8.46 with $a = 0$ and $b = \mathbb{E}[u^T Q_{-i} x_i]$ to conclude. □

Lemma 8.57. $\mathbb{P}\left(|\Lambda_i - \mathbb{E}[\Lambda_i]| \geq t\right) \leq C e^{-c(t/\sqrt{n})^2} + C e^{-ct/n}$, *for some constants* $C, c > 0$ *independent with* n.

Proof. Recalling that $\Lambda_i = 1 + \frac{1}{n} x_i Q_{-i} x_i$ and noting that $\mathbb{E}[\Lambda_i | X_i] = 1 + \frac{1}{n}\mathrm{Tr}(\Sigma Q_{-i})$, let us simply bound as in the proof of Lemma 8.56:

$$\begin{aligned}
&\mathbb{P}\left(|\Lambda_i - \mathbb{E}[\Lambda_i]| \geq t\right) \\
&\quad \leq \mathbb{E}\left[\mathbb{P}\left(|\Lambda_i - \mathbb{E}[\Lambda_i | X_{-i}]| \geq t | X_{-i}\right)\right] + \mathbb{P}\left(|\mathbb{E}[\Lambda_i | X_{-i}] - \mathbb{E}[\Lambda_i]| \geq t\right) \\
&\quad \leq \mathbb{E}\left[\mathbb{P}\left(\left|\frac{1}{n} x_i Q_{-i} x_i - \mathbb{E}\left[\frac{1}{n} x_i Q_{-i} x_i | X_{-i}\right]\right| \geq t | X_{-i}\right)\right] + \mathbb{P}\left(\left|\frac{1}{n}\mathrm{Tr}(\Sigma Q_{-i}) - \mathbb{E}\left[\frac{1}{n}\mathrm{Tr}(\Sigma Q_{-i})\right]\right| \geq t\right) \\
&\quad \leq \mathbb{E}\left[C e^{-c(t/\sqrt{p}\|Q_{-i}\|)^2} + C e^{-ct/n\|Q_{-i}\|}\right] + C e^{-c(t\sqrt{p}/n^{1/3})^2} \\
&\quad \leq C e^{-c'(t/\sqrt{n})^2} + C e^{-c' t/n}
\end{aligned}$$

thanks to Corollary 8.47 to bound the concentration of $\frac{1}{n} x_i Q_{-i} x_i$ and Proposition 8.54 to bound the concentration of $\frac{1}{n}\mathrm{Tr}(\Sigma Q_{-i})$ (the mapping $M \mapsto \frac{1}{n}\mathrm{Tr}(\Sigma M)$ is $\frac{\|\Sigma\|_F}{n}$-Lipschitz for the Frobenus norm and $\|\Sigma\|_F \leq O(\sqrt{p})$. □

Lemma 8.58. $|\Lambda| = |1 + \frac{1}{n} x_i^T Q_{-i} x_i| \geq 1$.

Lemma 8.59. $\|\mathbb{E}[Q_{-i}] - \mathbb{E}[Q]\| \leq O\left(\frac{1}{n}\right)$.

Proof. We know from (8.20) that $Q_{-i} - Q = \frac{1}{n\Lambda_i} Q_{-i} x_i x_i^T Q_{-i}$. One can then use Lemma 8.58 to bound for any deterministic vector $u \in \mathbb{R}^p$:

$$|u^T(\mathbb{E}[Q_{-i}] - \mathbb{E}[Q])u| \leq \frac{1}{n}\left|\mathbb{E}\left[u^T Q_{-i} x_i x_i^T Q_{-i} u\right]\right| = \frac{1}{n}\left|\mathbb{E}\left[u^T Q_{-i} \Sigma_i Q_{-i} u\right]\right| \leq O\left(\frac{1}{n}\right).$$

One can then directly conclude since for symmetric matrices like $Q - Q_{-i}$:

$$\|\mathbb{E}[Q_{-i}] - \mathbb{E}[Q]\| = \sup_{u \in \mathbb{R}^p} |u^T \mathbb{E}[Q - Q_{-i}] u| \leq O\left(\frac{1}{n}\right).$$

□

One now has all the elements to prove:

Proposition 8.60. $\|\mathbb{E}[Q] - \tilde{Q}^{\hat{\Lambda}}\| \leq O\left(\frac{1}{\sqrt{n}}\right)$.

Proof. Let us simply bound ε_i and δ_i appearing in (8.22). One can show directly from Lemma 8.59 that:

$$|\delta_i| \leq \|\mathbb{E}[Q] - \mathbb{E}[Q_{-i}]\| \, \|\Sigma\| \|\tilde{Q}^{\hat{\Lambda}}\| \leq O\left(\frac{1}{n}\right).$$

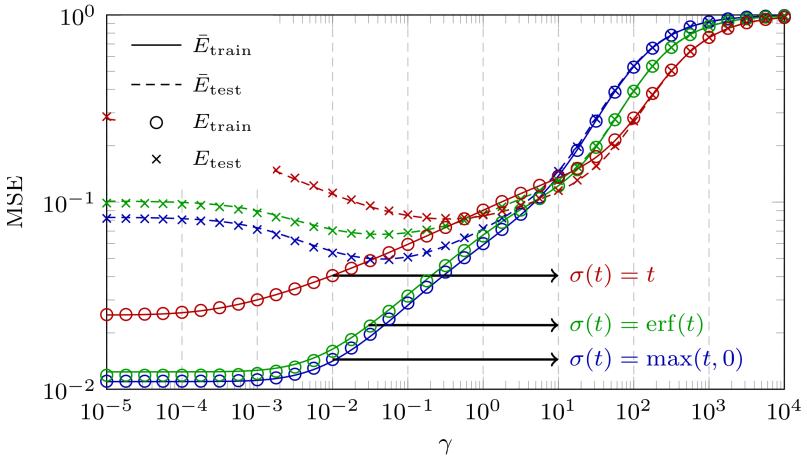

Figure 8.10: Test MSE and Train MSE of the Ridge regression as a function of γ for $X = \sigma(WZ)$ for some Gaussian matrix $W \in \mathcal{M}_{p,q}$ and $Z \in \mathcal{M}_{q,n}$ containing in columns MNIST digits ($q = 784$ is the number of pixels) for different activation functions σ. Figure taken from [1]

Then, one can use Hölder inequality[8] to set:

$$|\varepsilon_i| \leq \left(\mathbb{E}[|u^T Q_{-i} x_i|^3]\right)^{\frac{1}{3}} \left(\mathbb{E}[|x_i^T \tilde{Q}^{\hat{\Lambda}} u|^3]\right)^{\frac{1}{3}} \left(\mathbb{E}[|\Lambda_i - \hat{\Lambda}|^3]\right)^{\frac{1}{3}} \leq O(1) \cdot O(1) \cdot O\left(\frac{1}{\sqrt{n}}\right),$$

thanks to Lemma 8.44. □

The issue with last proposition is that $\tilde{Q}^{\hat{\Lambda}}$ still depends on $\hat{\Lambda} = 1 + \mathbb{E}[\frac{1}{n} x_i^T Q_{-i} x_i]$ that we do not know how to compute. This problem is solved thanks to the following heuristic (that can be rigorously justified but that lies outside the scope of this course):

$$\hat{\Lambda} = 1 + \frac{1}{n}\text{Tr}(\Sigma \mathbb{E}[Q_{-i}]) \approx 1 + \frac{1}{n}\text{Tr}(\Sigma \tilde{Q}^{\hat{\Lambda}}),$$

then introducing $\tilde{\Lambda} \in \mathbb{R}$ as the only solution to:

$$\Lambda = 1 + \frac{1}{n}\text{Tr}(\Sigma \tilde{Q}^{\Lambda}), \quad \Lambda \in \mathbb{R},$$

one can show that:

$$|\hat{\Lambda} - \tilde{\Lambda}| \leq O(1/\sqrt{n}) \tag{8.23}$$

The resolvent identity (8.18) provides the bound:

$$\left\|\tilde{Q}^{\hat{\Lambda}} - \tilde{Q}^{\tilde{\Lambda}}\right\| = \left|\frac{1}{\hat{\Lambda}} - \frac{1}{\tilde{\Lambda}}\right| \left\|\tilde{Q}^{\hat{\Lambda}} \Sigma \tilde{Q}^{\tilde{\Lambda}}\right\| \leq |\tilde{\Lambda} - \hat{\Lambda}| \frac{\|\tilde{Q}^{\hat{\Lambda}}\| \|\Sigma\| \|\tilde{Q}^{\tilde{\Lambda}}\|}{|\tilde{\Lambda}\hat{\Lambda}|} \leq O\left(\frac{1}{\sqrt{n}}\right).$$

Thanks to Lemmas 8.52, 8.53 and 8.58 and (8.23). Consequently Proposition 8.60 finally gives us:

$$\|\mathbb{E}[Q] - \tilde{Q}^{\tilde{\Lambda}}\| \leq \|\mathbb{E}[Q] - \tilde{Q}^{\hat{\Lambda}}\| + \|\tilde{Q}^{\hat{\Lambda}} - \tilde{Q}^{\tilde{\Lambda}}\| \leq O\left(\frac{1}{\sqrt{n}}\right).$$

Unlike $\tilde{Q}^{\hat{\Lambda}}$, $\tilde{Q}^{\tilde{\Lambda}}$ can be numerically computed if one knows Σ, therefore that gives us a precise estimate of $\mathbb{E}[Q]$ and subsequently of the MSE of the Ridge regression given in Exercise 8.50. This deterministic equivalent $\tilde{Q}^{\tilde{\Lambda}}$ then intervenes in a precise estimate of the Mean square error of the Ridge regression. The prediction then strictly depends on the population covariance matrix of the data. As depicted on Figure 8.10, such predictions are extremely accurate.

[8] Hölder inequality sets that for any random variables $X, Y, Z : \Omega \to \mathbb{R}$:

$$|\mathbb{E}[XYZ]| \leq \left(\mathbb{E}[|X|^3]\right)^{\frac{1}{3}} \left(\mathbb{E}[|Y|^3]\right)^{\frac{1}{3}} \left(\mathbb{E}[|Z|^3]\right)^{\frac{1}{3}}$$

Bibliography

[1] Romain Couillet and Zhenyu Liao. *Random matrix methods for machine learning*. Cambridge University Press, 2022.

[2] Michel Ledoux. The concentration of measure phenomenon. ed. by peter landweber et al. vol. 89. *Mathematical Surveys and Monographs. Providence, Rhode Island: American Mathematical Society*, page 181, 2005.

[3] Michel Talagrand. A new look at independence. *The Annals of probability*, pages 1–34, 1996.