Lectures on High Dimensional Probability

Contents

1 Ridge estimators **5**
- 1.1 Setting . . . 5
- 1.2 Localization using deterministic arguments . . . 7
- 1.3 Bounding the linear process . . . 9
- 1.4 Bounding quadratic processes . . . 11
  - 1.4.1 Conclusion of the proof . . . 12
- 1.5 Problem . . . 13

2 Community detection **15**
- 2.1 Graphs . . . 15
- 2.2 Two-classes stochastic block model (SBM) . . . 15
- 2.3 Spectral strategies . . . 16
  - 2.3.1 General remarks . . . 16
  - 2.3.2 Preliminaries . . . 17
- 2.4 The spectral clustering algorithm . . . 18
- 2.5 Guedon-Vershynin spectral algorithm . . . 19
  - 2.5.1 Critical discussion of the SCA bounds . . . 19
  - 2.5.2 Definition of the new algorithm . . . 19
  - 2.5.3 Bounding the risk . . . 21
- 2.6 Proof of Grothendieck's inequality . . . 25
- 2.7 Problem . . . 27

3 Concentration of measure **29**
- 3.1 Motivating example . . . 29
- 3.2 Sub-Gaussian random variables . . . 30
- 3.3 Hoeffding's inequality . . . 33
- 3.4 Bounded difference inequality . . . 34
- 3.5 Gaussian concentration inequality . . . 36
- 3.6 Sub-Gamma random variables . . . 37
- 3.7 Link with sub-exponential random variables . . . 39
- 3.8 Bernstein's inequality . . . 42
- 3.9 Problem . . . 44

4 Deviation inequalities for random matrices **45**
4.1 Calculus on matrices . . . 45
4.2 Deviation bounds for sums of independent random matrices . 46
4.2.1 Extension of the tensorization argument . . . 46
4.2.2 Matrix deviation inequalities . . . 48
4.3 Applications of Matrix Hoeffding's inequality . . . 51
4.3.1 Matrix Khintchine's inequality . . . 51
4.3.2 Application to Matrix completion . . . 52
4.4 Applications of Matrix Bernstein's inequality . . . 55
4.5 Decoupling and quadratic forms . . . 56
4.5.1 Decoupling . . . 57
4.5.2 Concentration of Gaussian Chaos . . . 58
4.5.3 Hanson-Wright's inequality . . . 59
4.5.4 Problem . . . 60

5 PAC-Bayesian bounds **63**
5.1 Setting . . . 63
5.2 Basic tools . . . 65
5.3 Sub-Gaussian vectors . . . 66
5.3.1 Choice of μ and ρ's . . . 66
5.3.2 Bounding the second moment . . . 67
5.3.3 Conclusion . . . 67
5.4 Sub-exponential random vectors . . . 68
5.4.1 Priors . . . 68
5.4.2 Conclusion . . . 69
5.5 Extension to random matrices . . . 70
5.5.1 Probabilistic assumption . . . 70
5.5.2 Global strategy . . . 71
5.5.3 Priors and quantities of interest . . . 72
5.5.4 Optimization . . . 73
5.5.5 Application to quadratic processes . . . 74

6 Upper bounds on random processes **77**
6.1 Dudley's inequality . . . 78
6.2 VC dimension . . . 80
6.2.1 Examples . . . 80
6.2.2 Pajor's lemma . . . 81
6.3 Covering numbers and VC dimension . . . 83
6.3.1 Bounding covering numbers by VC dimension . . . 83
6.3.2 Application to ERM for classification . . . 86
6.4 Generic chaining bound . . . 87
6.5 Application to linear SVM estimators . . . 89

7 Gaussian Processes **93**
7.1 Setting . . . 93
7.2 Examples . . . 93
7.2.1 Canonical Gaussian process on $\mathbb{R}^n$ . . . 93
7.2.2 Canonical Gaussian vector on Hilbert spaces . . . 94
7.3 Bounding suprema . . . 94
7.4 The generic chaining bound . . . 96
7.4.1 Hierarchical clustering . . . 96
7.4.2 How do we choose Γ_n? . . . 96
7.5 The majorizing measure theorem . . . 98
7.5.1 Another look at $\gamma_2(\Gamma)$ . . . 98
7.5.2 Gaussian Calculus . . . 98
7.5.3 Slepian's and Sudakov-Fernique results . . . 100
7.5.4 Talagrand's recursive bound . . . 103
7.5.5 Proof of the Majorizing measure theorem . . . 103

Chapter 1

Ridge estimators

We start these lectures with a particular learning problem in high dimension that we solve using generic arguments. The proof can easily be adapted to more involved problems, see in particular the problem at the end of the chapter, and motivates the development of probabilistic tools in the following chapters.

1.1 Setting

Consider the high dimensional Gaussian linear regression setting where we observe i.i.d. couples (x_i, y_i), $i \in \{1, \dots, n\}$, distributed as (x, y) and such that

$$y = \langle \theta^*, x \rangle + \sigma \xi \ .$$

In all the chapter, we assume that $x, \theta^* \in \mathbb{R}^d = \Theta$, $y, \sigma, \xi \in \mathbb{R}$, with d possibly much larger than n. The parameters θ^* and σ are fixed, while x, y, ξ are random and satisfy $x \sim \mathrm{N}(0, \Sigma)$, $\xi \sim \mathrm{N}(0, 1)$ are independent. We will also write $X \sim \mathrm{N}(0, \mathbf{I})$ a standard Gaussian vector in $\mathbb{R}^d$ such that $x = \Sigma^{1/2} X$.

Given $\lambda > 0$, the ridge estimator is defined as

$$\hat{\theta}_\lambda \in \mathrm{argmin}_{\theta \in \Theta} \, P_n \ell_\theta + \lambda \|\theta\|_2^2 \ ,$$

where the loss is the square loss $\ell_\theta(x, y) = (y - \langle x, \theta \rangle)^2$ and the empirical means are denoted $P_n g = n^{-1} \sum_{i=1}^n g(x_i, y_i)$ for any function g taking values in a convex set $\mathcal{C}$.

The goal of our analysis is to bound the excess risk of $\hat{\theta}_\lambda$, $P(\ell_{\hat{\theta}_\lambda} - \ell_{\theta^*})$, where $Pg = \mathbb{E}_{x,y}[g(x, y)]$, so, if g is measurable with respect to $\mathcal{D}_n = \{(x_i, y_i), i = 1, \dots, n\}$, $Pg = \mathbb{E}[g(x, y) | \mathcal{D}_n]$. To state the result, we introduce the effective dimension

$$\mathcal{D}_\lambda = \mathrm{Tr}(\Sigma(\Sigma + \lambda \mathbf{I})^{-1}) = \sum_{i=1}^d \frac{\lambda_i}{\lambda_i + \lambda} \ ,$$

where $\lambda_1 \geqslant \ldots \geqslant \lambda_d \geqslant 0$ are the eigenvalues of Σ. Note that the effective dimension always satisfies $\mathcal{D}_\lambda \leqslant d$, but the inequality can be pessimistic: It also holds that $\mathcal{D}_\lambda \leqslant \mathrm{Tr}(\Sigma)/\lambda$ for any $\lambda > 0$, so $\mathcal{D}_\lambda$ can be finite even if $d = \infty$ provided that $\mathrm{Tr}(\Sigma) = \sum_{i=1}^{+\infty} \lambda_i < \infty$. We will also repeatedly use the notation $\|u\|_{\mathbf{A}}$, defined for any vector u and positive semi-definite matrix $\mathbf{A}$, by $\|u\|_{\mathbf{A}} = \sqrt{u^T \mathbf{A} u} = \|\mathbf{A}^{1/2} u\|_2 = \sqrt{\langle u, \mathbf{A} u \rangle}$. When $\mathbf{A}$ is non singular, Cauchy-Schwarz inequality shows that

$$\langle u, v \rangle = \left\langle \mathbf{A}^{1/2} u, \mathbf{A}^{-1/2} v \right\rangle \leqslant \|u\|_{\mathbf{A}} \|v\|_{\mathbf{A}^{-1}} \ .$$

We prove the following result.

Theorem 1. *Assume that $x \sim \mathrm{N}(0, \Sigma)$ and $\xi \sim \mathrm{N}(0, 1)$ are independent, then, if $\mathcal{D}_\lambda \leqslant n/100$, the Ridge regressor $\hat{\theta}_\lambda$ satisfies, for any $z \in (0, n/100)$, with probability $1 - 2\exp(-z)$,*

$$P(\ell_{\hat{\theta}_\lambda} - \ell_{\theta^*}) \leqslant C \left(\sigma^2 \frac{\mathcal{D}_\lambda + z}{n} + \lambda^2 \|\theta^*\|^2_{\Sigma^{-1}} \right) \ .$$

One can make a few remarks to appreciate this result:

1. The complexity term can be bounded from above, for any $k \in \{1, \ldots, d\}$, by

$$\mathcal{D}_\lambda \leqslant k + \frac{1}{\lambda} \sum_{i=k+1}^{d} \lambda_i \ .$$

For example, if the spectrum satisfies $\lambda_i \asymp i^{-\alpha}$ for some $\alpha > 1$, we see that this bound is optimized for $k \asymp \lambda^{-1/\alpha}$ and gives $\mathcal{D}_\lambda \lesssim \lambda^{-1/\alpha}$

2. The regularization term can be written using the decomposition of θ^* onto the orthonormal basis of eigenvectors of Σ: writing u_i these vectors, we write $\theta^* = \sum_{i=1}^d \theta_i^* u_i$ and

$$\|\theta^*\|^2_{\Sigma^{-1}} = \sum_{i=1}^{d} \frac{(\theta_i^*)^2}{\lambda_i} \ .$$

As expected, this term is smaller if θ^* is well represented on the directions of Σ associated with its largest eigenvalues: In the extreme case where

$$\theta^* = \theta_1^* u_1, \qquad \|\theta^*\|^2_{\Sigma^{-1}} \asymp \frac{\|\theta^*\|_2^2}{\lambda_1} \ ,$$

while if

$$\theta^* = \theta_d^* u_d, \qquad \|\theta^*\|^2_{\Sigma^{-1}} \asymp \frac{\|\theta^*\|_2^2}{\lambda_d} \ .$$

3. Combining these two bounds, we get an upper bound on the Ridge estimator equal to
$$\frac{\lambda^{-1/\alpha}}{n} + \lambda^2 \|\theta^*\|^2_{\Sigma^{-1}} \ ,$$
where $\alpha > 1$ depends on the spectrum of Σ, it is minimal for $\lambda = (1/\|\theta^*\|^2_{\Sigma^{-1}} n)^{\alpha/(2\alpha+1)}$ and gives the rate $\|\theta^*\|^{2/(2\alpha+1)}_{\Sigma^{-1}} n^{-2\alpha/(2\alpha+1)}$.

In the remaining of the chapter, we prove the main theorem. The proof is decomposed to be easily generalized to other frameworks.

1.2 Localization using deterministic arguments

The purpose of this section is to show that it is sufficient to bound localized *linear* and *quadratic* processes to bound the excess risk of $\hat{\theta}_\lambda$. To proceed with this step, we do not use any probabilistic argument.

Step 1: If there exists an ellipsoïd $\mathcal{C}$ centered in θ^* such that any $\theta \notin \mathcal{C}$ satisfies
$$P_n \ell_\theta + \lambda \|\theta\|_2^2 > P_n \ell_{\theta^*} + \lambda \|\theta^*\|_2^2 \ ,$$
then $\hat{\theta}_\lambda \in \mathcal{C}$.

This elementary remark follows directly from the definition of $\hat{\theta}_\lambda$. Indeed,
$$P_n \ell_{\hat{\theta}_\lambda} + \lambda \|\hat{\theta}_\lambda\|_2^2 \leqslant P_n \ell_{\theta^*} + \lambda \|\theta^*\|_2^2 \ .$$
Therefore, by definition of $\mathcal{C}$, $\hat{\theta}_\lambda$ cannot belong to the complementary of $\mathcal{C}$.

Remark 2. *Step 1 would be true for any form of the set $\mathcal{C}$. The focus on ellipsoids will become clear later in Step 3.*

Step 2: If all θ in the frontier $\partial\mathcal{C}$ satisfy
$$\varphi(\theta) = P_n \ell_\theta + \lambda \|\theta\|_2^2 - (P_n \ell_{\theta^*} + \lambda \|\theta^*\|_2^2) \geqslant 0 \ , \tag{1.1}$$
then $\hat{\theta}_\lambda \in \mathcal{C}$.

This second elementary remark follows from the convexity of square loss. As φ is strictly convex and non negative on $\partial\mathcal{C}$. Therefore, if $\theta \notin \mathcal{C}$, there exists $\alpha \in (0,1)$ such that $\bar{\theta} = \alpha\theta + (1-\alpha)\theta^* \in [\theta^*, \theta] \cap \partial\mathcal{C}$. Thus, by strict convexity,
$$0 \leqslant \varphi(\bar{\theta}) < \alpha\varphi(\theta) + (1-\alpha)\varphi(\theta^*) \ ,$$
that is
$$0 < \alpha(P_n \ell_\theta + \lambda \|\theta\|_2^2 - (P_n \ell_{\theta^*} + \lambda \|\theta^*\|_2^2)) \ .$$

As this is true for any $\theta \notin \mathcal{C}$, the conclusion follows from Step 1.

Discussion: Step 2 would be true for any convex set $\mathcal{C}$. In the following, we focus on the function φ defined in Eq (1.1) and our goal now is to choose $\mathcal{C}$ such that $\varphi(\theta) \geqslant 0$ on $\partial\mathcal{C}$. In particular, it will become clear why $\mathcal{C}$ is chosen as an ellipsoid, and which ellipsoid we should choose.

Step 3: The following decomposition of φ holds: Let $\Sigma_n = P_n(xx^T)$ denote the matrix of empirical second moments of the design x,

$$\varphi(\theta) = \|\theta - \theta^*\|^2_{\Sigma_n + \lambda \mathbf{I}} + 2\sigma \left\langle P_n(\xi X), \Sigma^{1/2}(\theta - \theta^*) \right\rangle + 2\lambda \langle \theta - \theta^*, \theta^* \rangle \quad . \tag{1.2}$$

The key to prove this bound is the following simple quadratic/multiplier decomposition of the square loss:

$$\begin{aligned} \ell_\theta(x,y) - \ell_{\theta^*}(x,y) &= (\langle x, \theta - \theta^* \rangle + \sigma\xi)^2 - (\sigma\xi)^2 \\ &= 2\sigma \langle \xi x, \theta - \theta^* \rangle + \langle x, \theta - \theta^* \rangle^2 \quad , \end{aligned}$$

Writing $x = \Sigma^{1/2} X$, we get

$$\ell_\theta(x,y) - \ell_{\theta^*}(x,y) = 2\sigma \left\langle \xi X, \Sigma^{1/2}(\theta - \theta^*) \right\rangle + \left\langle (xx^T)(\theta - \theta^*), \theta - \theta^* \right\rangle \quad .$$

Thus

$$P_n(\ell_\theta - \ell_{\theta^*}) = 2\sigma \left\langle P_n(\xi X), \Sigma^{1/2}(\theta - \theta^*) \right\rangle + \langle \Sigma_n(\theta - \theta^*), \theta - \theta^* \rangle \quad .$$

Together with the bound

$$\lambda(\|\theta\|_2^2 - \|\theta^*\|_2^2) = \langle \lambda \mathbf{I}(\theta - \theta^*), \theta - \theta^* \rangle + 2\lambda \langle \theta^*, \theta - \theta^* \rangle \quad ,$$

we get

$$\begin{aligned} \varphi(\theta) = 2\sigma &\left\langle P_n(\xi X), \Sigma^{1/2}(\theta - \theta^*) \right\rangle + \langle (\Sigma_n + \lambda \mathbf{I})(\theta - \theta^*), \theta - \theta^* \rangle \\ &+ 2\lambda \langle \theta^*, \theta - \theta^* \rangle \quad . \end{aligned}$$

This is equivalent to the desired conclusion.

Discussion: The decomposition given in Step 3 suggests to take for $\mathcal{C}$ the ellipsoid

$$\mathcal{C} = \{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|^2_{\Sigma + \lambda \mathbf{I}} \leqslant r^2\} \quad , \tag{1.3}$$

where r is an hyperparameter that remains to be calibrated. Indeed, this choice makes the positive term $\|\theta - \theta^*\|^2_{\Sigma_n + \lambda \mathbf{I}}$ in the decomposition of $\varphi(\theta)$ as large as possible, provided that the random matrix $\Sigma_n + \lambda \mathbf{I}$ is close to its expected value $\Sigma + \lambda \mathbf{I}$.

Then, to conclude, we have two remaining tasks:

1. Bound the linear process $\left\langle P_n(\xi X), \Sigma^{1/2}(\theta - \theta^*) \right\rangle$ for all $\theta \in \mathcal{C}$.
2. Check that the quadratic process $\|\theta - \theta^*\|^2_{\Sigma_n + \lambda \mathbf{I}}$ behaves as its expected value $\|\theta - \theta^*\|^2_{\Sigma + \lambda \mathbf{I}}$ uniformly over the frontier $\partial\mathcal{C}$.

1.3 Bounding the linear process

In this section, we focus on obtaining upper bounds on the supremum of the linear process

$$\sup_{\theta\in\mathcal{C}} \left\langle P_n(\xi X), \Sigma^{1/2}(\theta-\theta^*) \right\rangle \ .$$

Let us reparametrize the problem to simplify notations and write

$$T = \{t = \Sigma^{1/2}(\theta-\theta^*)/r,\ \theta\in\mathcal{C}\} \ ,$$

so we have

$$\sup_{\theta\in\mathcal{C}} \left\langle P_n(\xi X), \Sigma^{1/2}(\theta-\theta^*) \right\rangle = r \sup_{t\in T} \langle P_n(\xi X), t\rangle \ .$$

We decompose this control into elementary steps.

Control for a single t in the unit sphere. The purpose of this paragraph is to bound the Laplace transform of $\langle P_n(\xi X), t\rangle$, for any t in the unit sphere. Fix $i\in\{1,\dots,n\}$ and $t\in\mathbb{R}^d$ such that $\|t\|_2 = 1$, we start by computing the Laplace transform of $\langle \xi_i X_i, t\rangle$. For any $s\in\mathbb{R}$, we have $\langle X, t\rangle \sim \mathrm{N}(0,1)$ is independent of ξ_i, so

$$\mathbb{E}[\exp(s\xi_i \langle X_i, t\rangle)|\xi_i] = \exp\left(\frac{s^2\xi_i^2}{2}\right) \ .$$

We deduce that, for any $|s| < 1/\sqrt{2}$,

$$\begin{aligned}
\mathbb{E}[\exp(s\xi_i \langle X_i, t\rangle)] &= \int \exp\left(\frac{-y^2(1-s^2)}{2}\right)\frac{\mathrm{d}y}{\sqrt{2\pi}} \\
&= \frac{1}{\sqrt{1-s^2}} \\
&= \left(1+\frac{s^2}{1-s^2}\right)^{1/2} \\
&\leqslant \exp\left(s^2\right) \ .
\end{aligned}$$

By independence, we deduce that, for any $|s| < n/\sqrt{2}$,

$$\mathbb{E}[\exp(s\langle P_n(\xi X), t\rangle)] = \prod_{i=1}^n \mathbb{E}\left[\exp\left(\frac{s}{n}\xi_i \langle X_i, t\rangle\right)\right] \leqslant \exp\left(\frac{s^2}{n}\right) \ .$$

As we shall see in Chapter 3, it is possible to derive directly from this upper bound on the Laplace transform a concentration bound for $\langle P_n(\xi X), t\rangle$. We do not pursue this path here and use a more involved consequence of this bound on the Laplace transform given in Chapter 7, that directly allows to derive from a uniform deviation bound on $\sup_{t\in T}\langle P_n(\xi X), t\rangle$.

From a single t to uniform bounds over ellipsoids In Chapter 7 at the end of the lecture notes, we develop the PAC-Bayesian method which allows to deduce uniform upper bounds over processes from bounds on the Laplace transform for a single t. The conclusion of the previous paragraph is that the vector $P_n(\xi X)$ satisfies the sub-Gamma assumption (5.4) with $b = \sqrt{2}/n$ and $K = 1/\sqrt{n}$.

To use the results of this section, we also have to write the ellipsoid T properly. We have

$$\begin{aligned} T &= \{t = \Sigma^{1/2}(\theta - \theta^*)/r : \ \|\theta - \theta^*\|^2_{\Sigma+\lambda \mathbf{I}} \leqslant r^2\} \\ &= \{t = \Sigma^{1/2} u : \ \|u\|^2_{\Sigma+\lambda \mathbf{I}} \leqslant 1\} \\ &= \{t : \|(\Sigma + \lambda \mathbf{I})^{1/2} \Sigma^{-1/2} t\|_2^2 \leqslant 1\} \ . \end{aligned}$$

The conclusion of this is that

$$T = \{t : \|\Gamma^{-1/2} t\|_2 \leqslant 1\}, \qquad \text{with} \qquad \Gamma = \Sigma(\Sigma + \lambda \mathbf{I})^{-1} \ .$$

To express the result, we define, for any matrix A, its operator norm $\|A\|$ and effective rank $r(A) = \mathrm{Tr}(A)/\|A\|$. By the Pac-Bayesian bound (5.6), w.p.l.t. $1 - \exp(-z)$,

$$\sup_{t \in T} \langle P_n(\xi X), t \rangle \leqslant \sqrt{\|\Gamma\|} \left(4 \sqrt{\frac{r(\Gamma)}{n}} + \sqrt{1+z} \left(\frac{4}{\sqrt{n}} \vee \frac{3\sqrt{2r(\Gamma)}}{n} \right) + \frac{3\sqrt{2}(1+z)}{n} \right) \ .$$

We can rearrange this result saying that

$$\|\Gamma\| r(\Gamma) = \mathcal{D}_\lambda, \qquad \|\Gamma\| \leqslant 1 \ ,$$

we deduce that

$$\sup_{t \in T} \langle P_n(\xi X), t \rangle \leqslant 4 \sqrt{\frac{\mathcal{D}_\lambda}{n}} + \sqrt{1+z} \left(\frac{4}{\sqrt{n}} \vee \frac{3\sqrt{2\mathcal{D}_\lambda}}{n} \right) + \frac{3\sqrt{2}(1+z)}{n} \ .$$

Finally, we can use the assumptions stated in the theorem $\mathcal{D}_\lambda < n/100$ and $z \leqslant n/100$ to say that

$$\frac{3\sqrt{2\mathcal{D}_\lambda}}{n} \leqslant \frac{3}{2\sqrt{n}}, \qquad \frac{3\sqrt{2}(1+z)}{n} \leqslant \frac{3\sqrt{1+z}}{\sqrt{50n}} \ ,$$

we conclude that, w.p.l.t. $1 - \exp(-z)$

$$\sup_{t \in T} \langle P_n(\xi X), t \rangle \leqslant 4 \sqrt{\frac{\mathcal{D}_\lambda}{n}} + 5 \sqrt{\frac{1+z}{n}} \ . \tag{1.4}$$

1.4 Bounding quadratic processes

In this section, we are interested in the quadratic process. We write, for any $\theta \in \partial\mathcal{C}$,

$$\begin{aligned}\|\theta - \theta^*\|^2_{\bar{\Sigma}_n + \lambda \mathbf{I}} &= (P_n - P)[\langle x, \theta - \theta^* \rangle^2] + r^2 \\ &\geqslant r^2 \left(1 - \sup_{t,t' \in T} t^T \big(P_n X X^T - \mathbf{I} \big) t' \right) \ , \end{aligned} \tag{1.5}$$

where the ellipsoid T was defined in the previous section. We proceed in two steps as in the previous section to bound this term.

Step 1: Bounding the Laplace transform. Let us first fix u and v in the unit sphere and compute the Laplace transform of $u^T(X_i X_i^T - \mathbf{I})v$. We write $v = \alpha u + z$, with $\alpha = \langle u, v \rangle$, $z = v - \langle u, v \rangle \, u \perp u$ and, for any $s \in \mathbb{R}$,

$$\begin{aligned}\mathbb{E}[\exp(s u^T (X_i X_i^T - \mathbf{I})v) | \, \langle X_i, u \rangle] &= \mathbb{E}[\exp(s(\langle X_i, u \rangle \, \langle X_i, v \rangle - \langle u, v \rangle)) | \, \langle X_i, u \rangle] \\ &= \exp(s\alpha(\langle X_i, u \rangle^2 - 1)) \mathbb{E}[\exp(s \, \langle X_i, u \rangle \, \langle X_i, z \rangle \, | \, \langle X_i, u \rangle] \\ &= \exp\left(\left(s\alpha + \frac{s^2 \|z\|^2}{2} \right) \langle X_i, u \rangle^2 - s\alpha \right) \ .\end{aligned}$$

We deduce that, for any s such that $|2s\alpha + s^2\|z\|^2| < 1/2$,

$$\mathbb{E}[\exp(s \left\langle X_i X_i^T - \Sigma, u v^T \right\rangle_F)] = \frac{\exp(-s\alpha)}{\sqrt{1 - 2s\alpha - s^2\|z\|^2}} \leqslant \exp(s^2) \ .$$

As $\|z\|^2 = 1 - \alpha^2$, we have, for any $|s| < 1/6$,

$$2s\alpha + s^2\|z\|^2 \leqslant |s|(1 + 2\alpha - \alpha^2) < 1/2 \ .$$

By independence, we obtain that, for any $|s| < n/6$,

$$\begin{aligned}\mathbb{E}[\exp(s \left\langle P_n(XX^T) - \Sigma, u v^T \right\rangle_F)] &= \prod_{i=1}^n \mathbb{E}\left[\exp\left(\frac{s}{n} \left\langle X_i X_i^T - \Sigma, u v^T \right\rangle_F \right) \right] \\ &\leqslant \exp\left(\frac{s^2}{n} \right) \ .\end{aligned}$$

Step 2: Uniform bounds using the PAC-Bayesian approach. As for the linear process, we can now move from simple bounds to uniform bounds using the Pac Bayesian approach of Chapter 7. The conclusion of the previous paragraph is that the matrix $P_n(XX^T) - \mathbf{I}$ satisfies Assumption (5.9) with $b = 6/n$ and $K = 1/\sqrt{n}$. We will use inequality (5.10) with $\mathcal{E}_U = \mathcal{E}_V = T$, the ellipsoid defined for the linear process:

$$T = \{t : \|\Gamma^{-1/2} t\|_2 \leqslant 1\}, \qquad \Gamma = \Sigma(\Sigma + \lambda \mathbf{I})^{-1} \ .$$

The conclusion is that, for all $z > 0$, with probability $1 - \exp(-z)$,

$$\sup_{t,t' \in T} t^T (P_n(XX^T) - \mathbf{I}) t' \leqslant \|\Gamma\| \left(\sqrt{\frac{4(r(\Gamma) + 1 + z)}{n}} \vee \frac{24(r(\Gamma) + 1 + z)}{n} \right) .$$

Using as in the previous section that $\|\Gamma\| r(\Gamma) = \mathcal{D}_\lambda$ and $\|\Gamma\| \leqslant 1$, we can rearrange the terms to get

$$\sup_{t,t' \in T} t^T (P_n(XX^T) - \mathbf{I}) t' \leqslant \left(\sqrt{\frac{4(\mathcal{D}_\lambda + 1 + z)}{n}} \vee \frac{24(\mathcal{D}_\lambda + 1 + z)}{n} \right) .$$

Now, we use the bounds $\mathcal{D}_\lambda < n/100$, $z < n/100$ to say that, with probability $1 - \exp(-n/100)$,

$$\sup_{t,t' \in T} t^T (P_n(XX^T) - \mathbf{I}) t' \leqslant \frac{1}{2} .$$

Plugging this bound into the basic lower bound (1.5) gives the final lower bound on the quadratic process: with probability $1 - \exp(-n/100)$, for any $\theta \in \partial\mathcal{C}$,

$$\|\theta - \theta^*\|^2_{\Sigma_n + \lambda \mathbf{I}} \geqslant \frac{r^2}{2} , \tag{1.6}$$

1.4.1 Conclusion of the proof

To conclude the proof, we go back to (1.2). By (1.6), with probability $1 - \exp(-n/100)$, for any $\theta \in \partial\mathcal{C}$,

$$\|\theta - \theta^*\|^2_{\Sigma_n + \lambda \mathbf{I}} \geqslant \frac{r^2}{2} ,$$

Besides, by (1.4), with probability $1 - \exp(-z)$, for any $\theta \in \partial\mathcal{C}$,

$$2\sigma \left\langle P_n(\xi X), \Sigma^{1/2}(\theta - \theta^*) \right\rangle \geqslant -\sigma r \left(8\sqrt{\frac{\mathcal{D}_\lambda}{n}} + 10\sqrt{\frac{1+z}{n}} \right) .$$

Finally, we always have, by Cauchy-Schwarz inequality, for any $\theta \in \partial\mathcal{C}$,

$$2\lambda \langle \theta - \theta^*, \theta^* \rangle \geqslant -2\lambda r \|\theta^*\|_{\Sigma^{-1}} .$$

Together, these three informations show that, with probability $1 - \exp(-z) - \exp(-n/100)$, for any $\theta \in \partial\mathcal{C}$,

$$\varphi(\theta) \geqslant \frac{r}{2} \left(r - \sigma \left(16\sqrt{\frac{\mathcal{D}_\lambda}{n}} + 20\sqrt{\frac{1+z}{n}} \right) - 4\lambda \|\theta^*\|_{\Sigma^{-1}} \right) .$$

This last lower bound is clearly $\geqslant 0$ if

$$r^* = \sigma \left(16\sqrt{\frac{\mathcal{D}_\lambda}{n}} + 20\sqrt{\frac{1+z}{n}} \right) + 4\lambda \|\theta^*\|_{\Sigma^{-1}} .$$

It follows therefore from Step 2 of Section 1.2 that, with probability $1 - \exp(-z) - \exp(-n/100)$, $\hat{\theta}_\lambda \in \mathcal{C}$ for $r = r^*$, that is, by definition of this set, we have proved that

$$\|\hat{\theta}_\lambda - \theta^*\|_{\Sigma+\lambda\mathbf{I}} \leqslant r^* \ .$$

The theorem is proved as we have

$$P(\ell_{\hat{\theta}_\lambda} - \ell_{\theta^*}) = \|\hat{\theta}_\lambda - \theta^*\|_\Sigma^2 \leqslant \|\hat{\theta}_\lambda - \theta^*\|_{\Sigma+\lambda\mathbf{I}}^2 \ .$$

1.5 Problem

In this problem, we try to replicate the analysis we developed for linear regression into the slightly more challenging problem of logistic regression. We present typical steps that can be followed to analyse other M-estimators. Recall that logistic regression produces a linear classifier in a high dimension setting where we observe n couples (x_i, y_i) independent and identically distributed, where the couples $(x, y) \in \mathcal{X} \times \mathcal{Y}$, with $\mathcal{X} = \mathbb{R}^d$ (with possibly $d \geqslant n$) and $\mathcal{Y} = \{-1, 1\}$ and where, for each $\theta \in \Theta = \mathbb{R}^d$, we define the loss

$$\ell_\theta(x, y) = \varphi(-y \langle \theta, x \rangle), \qquad \varphi(u) = \log(1 + \exp(x)) \ .$$

We analyse the estimator

$$\hat{\theta}_\lambda \in \operatorname{argmin}_{\theta\in\Theta}\{P_n\ell_\theta + \lambda\|\theta\|_2^2\} \ .$$

1. The purpose of this question is to adapt the localization argument, the question is decomposed in two steps.

   (a) Show that, if there exists a neighborhood $\mathcal{V}(\theta^*)$ of θ^* such that, for all θ in the frontier $\partial\mathcal{V}(\theta^*)$,

   $$P_n\ell_\theta + \lambda\|\theta\|_2^2 \geqslant P_n\ell_{\theta^*} + \lambda\|\theta^*\|_2^2 \ .$$

   then $\hat{\theta}_\lambda \in \mathcal{V}(\theta^*)$.

   (b) Explain why we will consider as neighborhood $\mathcal{V}(\theta^*)$ an ellipsoid $\mathcal{E}(r)$ of the following form

   $$\mathcal{E}(r) = \{\theta \in \Theta : \langle H_{\theta^*}(\theta - \theta^*), \theta - \theta^* \rangle \leqslant r^2\} \ ,$$

   where the matrix $H_{\theta^*} = P[\varphi''(-y \langle x, \theta^* \rangle) x x^T] + 2\lambda\mathbf{I}$. We expect a heuristic argument here, the purpose of the following questions is to formalize arguments that can be used to actually prove that this heuristic is correct.

   In the following, we fix $\mathcal{V}(\theta^*) = \mathcal{E}(r)$ as in question 1.(b) and focus on proving the equation proposed in question 1.(a). We consider

   $$\varphi(\theta) = P_n(\ell_\theta - \ell_{\theta^*}) + \lambda(\|\theta\|^2 - \|\theta^*\|^2) \ .$$

2. Show that

$$\varphi(\theta) = \langle P_n \nabla_\theta \ell_{\theta^*}, \theta - \theta^* \rangle + \frac{1}{2} \langle H_n(\theta - \theta^*), \theta - \theta^* \rangle + 2\lambda \langle \theta - \theta^*, \theta^* \rangle \ ,$$

where $\nabla_\theta \ell_{\theta'}(x, y) = \varphi'(-y \langle x, \theta' \rangle)(-yx)$,

$$H_n = \int_0^1 P_n[\varphi''(-y \langle x, t\theta + (1-t)\theta^* \rangle) xx^T] \mathrm{d}t + 2\lambda \mathbf{I} \ .$$

3. Give a concentration inequality for the linear process $\langle P_n \nabla_\theta \ell_{\theta^*}, \theta - \theta^* \rangle$ using the PAC-Bayesian approach mentioned in Section 1.3.

4. Show that

$$\left| \int_0^1 \varphi''(-y \langle x, t\theta + (1-t)\theta^* \rangle) \mathrm{d}t - \varphi''(-y \langle x, \theta^* \rangle) \right| \leqslant \frac{1}{2} | \langle x, \theta - \theta^* \rangle | \ .$$

Deduce that

$$\langle H_n(\theta - \theta^*), \theta - \theta^* \rangle \geqslant \langle H_n^*(\theta - \theta^*), \theta - \theta^* \rangle - \frac{1}{2} \langle R_n(\theta - \theta^*), \theta - \theta^* \rangle \ ,$$

where

$$H_n^* = P_n[\varphi''(-y \langle x, \theta^* \rangle) xx^T] + \lambda \mathbf{I}, \qquad R_n = P_n[| \langle x, \theta - \theta^* \rangle | xx^T] \ .$$

We see here that the analysis of the quadratic process in general can be handled by finding lower bound on the quadratic process $\langle H_n^* t, t \rangle$ over ellipsoids, which can be done using the PAC-Bayesian approach presented in Section 1.4. Then we have to bound from above the process $\langle R_n(\theta - \theta^*), \theta - \theta^* \rangle = P_n[| \langle x, \theta - \theta^* \rangle |^3]$ which is a remainder term as it is at first order of order r^3 compared to the quadratic term $\langle H_{\theta^*}(\theta - \theta^*), \theta - \theta^* \rangle = r^2$ on the frontier $\partial \mathcal{E}(r)$. Note that the concentration of this term is not easily derived from classical tools as the Laplace transform $\mathbb{E}[\exp(s \langle x, \theta - \theta^* \rangle^3)]$ is not defined, for any $s \neq 0$.

Chapter 2

Community detection

Community detection is a basic problem of clustering in graphs, where we want to recover well connected nodes. In this chapter, we present spectral strategies to solve this problem in the toy statistical model of balanced two-classes SBM.

2.1 Graphs

Hereafter, a graph $G = (V, E)$ is a couple where $V = \{1, \dots, n\}$ is a finite set of *vertices* or *nodes* and $E \subset V \times V$ is a set of *edges*. All graphs here are undirected, meaning that E is a set of *pairs* $\{i, j\}$ or that, $(i, j) \in E$ iff $(j, i) \in E$. A graph is represented by its *adjacency matrix* A, which is the $n \times n$ matrix such that

$$A_{i,j} = \begin{cases} 1 & \text{if } \{i,j\} \in E \ , \\ 0 & \text{if } \{i,j\} \notin E \ . \end{cases}$$

The matrix A is symmetric as the graph is undirected. A graph $G = (V, E)$ is random if the set E of edges is random. For example, the Erdös-Renyi graph $G(n, p)$ is the random graph where

$$\forall i \leqslant j, \qquad A_{i,j} \text{ are i.d.d. Bernoulli random variables } \mathcal{B}(p) \ .$$

In these graphs, we are typically interested in the asymptotic behavior when $n \to \infty$ and $p \to 0$. For example, we may wonder, as $n \to \infty$, the smallest p such that the Erdös-Renyii graph has an infinite connex component.

2.2 Two-classes stochastic block model (SBM)

The balanced two classes SBM is an extension of Erdös-Renyi model, denoted by $G(N, p, q)$, where $0 < q < p < 1$. The set of vertices $V = \{1, \dots, N\}$, where N is odd ($N = 2n$) is divided into two communities

C_1^* and C_{-1}^* of equal size $|C_i^*| = n$ and the set of vertices is random: $\forall i \leqslant j$, $A_{i,j}$ are independent random variables with parameters

$$\begin{cases} p & \text{if } i,j \text{ belong to the same community } , \\ q & \text{if } i,j \text{ belong to different communities } . \end{cases}$$

The goal of community detection is to recover the communities from the observation of the adacency matrix A.

Community detection aims at discovering a partition of $\{1,\dots,2n\}$ or, equivalently, an element of the hypercube $\Theta = \{-1,1\}^{2n}$. Indeed, each partition $C_{-1} \cup C_1$ of $\{1,\dots,2n\}$ is encoded by the vector $\theta \in \{-1,1\}^{2n}$ such that $\theta_i = 1$ if $i \in C_1$ and $\theta_i = -1$ if $i \in C_{-1}$. We denote by $\theta^* \in \Theta$ the vector of the hypercube encoding the true partition of interest, i.e. the one with coefficients $\theta_i^* = 1$ if $i \in C_1^*$ and $\theta_i^* = -1$ if $i \in C_{-1}^*$.

To evaluate a community detection algorithm, we define, for each $\theta \in \Theta$, the proportion of indices where θ and θ^* disagree, that, is the *risk* of θ is defined by its Hamming distance to θ^* divided by $2n$:

$$\mathcal{R}(\theta) = \frac{1}{2n}\#\{i \in \{1,\dots,2n\} : \theta_i \neq \theta_i^*\} = \frac{1}{2n}\sum_{i=1}^{2n} \mathbf{1}_{\{\theta_i \neq \theta_i^*\}} \ .$$

2.3 Spectral strategies

2.3.1 General remarks

Spectral estimation strategies are based on the elementary remark that, up to renumbering of the nodes, one can assume that $C_{-1} = \{1,\dots,n\}$ and $C_1 = \{n+1,\dots,2n\}$. In this case, writing $\mathbf{J}_n = \mathbf{1}_n\mathbf{1}_n^T$ the $n \times n$ matrix filled with 1, it holds that the adjacency matrix A of the graph satisfies

$$\mathbb{E}[A] = \begin{bmatrix} p\mathbf{J}_n & q\mathbf{J}_n \\ q\mathbf{J}_n & p\mathbf{J}_n \end{bmatrix} \ .$$

This elementary remark shows that $\mathbb{E}[A]$ is rank 2, that its largest eigenvalue $\lambda_1 = n(p+q)$ is the average degree of the random graph with distribution $G(2n,p,q)$, it is associated to the (normalized) eigenvector $\mathbf{u}_1 = \mathbf{1}_{2n}/\sqrt{2n}$. The second eigenvalue of $\mathbb{E}[A]$ is $\lambda_2 = n(p-q)$. The most important remark is that this second eigenvalue is associated with the (normalized) eigenvector

$$\mathbf{u}_2 = \frac{1}{\sqrt{2n}}\begin{bmatrix} \mathbf{1}_n \\ -\mathbf{1}_n \end{bmatrix} \ .$$

In words, **the second eigenvector $\mathbf{u}_2$ of $\mathbb{E}[Z]$ classifies perfectly the communities**.

2.3.2 Preliminaries

Spectral strategies consist in building unobserved matrices R, based on $\mathbb{E}[A]$ in particular, admitting $\mathbf{u}_2$ as i-th eigenvector. Then, we estimate $\mathbf{u}_2$ as the i-th eigenvector of an estimator $\hat{R}$ of R (obtained by replacing $\mathbb{E}[A]$ by A in the definition of R for example). Then, we classify each i according to the sign of the i-th coordinate $\hat{u}_i$ of $\hat{u}$, that is, we build

$$\hat{\theta} = (\text{sign}(\hat{u}_i))_{i\in\{1,\dots,2n\}} \ . \tag{2.1}$$

In this case, assuming $\|\hat{u}\|_2 = 1$ and writing $v = \sqrt{2n}\mathbf{u}_2$, $\hat{v} = \sqrt{2n}\hat{u}$,

$$\|\hat{u} - \mathbf{u}_2\|_2^2 = \frac{1}{2n}\sum_{i=1}^{2n}(v_i - \hat{v}_i)^2 \ .$$

As $v_i \in \{-1, 1\}$ and $\text{Sign}(\hat{v}_i) = \text{Sign}(\hat{u}_i)$, for any i misclassified by our algorithm, we have

$$(v_i - \hat{v}_i)^2 \geqslant 1 \ .$$

Therefore,

$$\mathcal{R}(\hat{\theta}) = \frac{1}{2n}\sum_{i=1}^{2n}\mathbb{I}\{i \text{ misclassified}\} \leqslant \frac{1}{2n}\sum_{i=1}^{2n}(v_i - \hat{v}_i)^2 = \|\hat{u} - \mathbf{u}_2\|_2^2 \ .$$

Let $\mathbf{a}$ denote the angle between $\mathbf{u}_2$ and $\hat{u}$, we have

$$\|\hat{u} - \mathbf{u}_2\|_2^2 = 2(1 - \cos(\mathbf{a})) \leqslant 2(1 - \cos^2(\mathbf{a})) \leqslant 2\sin^2(\mathbf{a}) \ .$$

This last remark is useful since $\sin(\mathbf{a})$ can be bounded from above using Davis-Kahan theorem.

Theorem 3 (Davis-Kahan)**.** *Let $\mathbf{S}, \mathbf{T}$ denote two symmetric $n \times n$ matrices. Fix $i \in \{1, \dots, n\}$ and assume that the i-th eigenvalue of $\mathbf{S}$ is well separated from the rest of the spectrum*

$$\min_{j\neq i} |\lambda_i(\mathbf{S}) - \lambda_j(\mathbf{S})| = \delta > 0 \ .$$

Then the angle $\boldsymbol{a}$ between $\mathbf{u}_i(\mathbf{S})$ and $\mathbf{u}_i(\mathbf{T})$ satisfies

$$\sin(\boldsymbol{a}) \leqslant \frac{2}{\delta}\|\mathbf{S} - \mathbf{T}\| \ ,$$

where $\|\cdot\|$ denote the operator norm, that is the largest singular value.

We do not prove Davis-Kahan Theorem in these notes, a reference where it is done is given in Slack.

Therefore, as $\mathbf{u}_2$ is the i-th eigenvector of the matrix R, we can bound the risk of $\hat{\theta}$ by computing the spectral gap

$$\delta = \min_{j \neq i} |\lambda_i(R) - \lambda_j(R)| \ . \tag{2.2}$$

By Davis-Kahan theorem, we obtain then that the risk is bounded by

$$\mathcal{R}(\hat{\theta}) \leqslant \frac{8}{\delta^2} \|R - \hat{R}\|^2 \ . \tag{2.3}$$

The final task is thus to bound from above the operator norm of the random matrix $\|R - \hat{R}\|$. In the following, we present two strategies based on this principle.

2.4 The spectral clustering algorithm

The spectral clustering algorithm (SCA) is the algorithm given in (2.1) when the reference matrix $R = \mathbb{E}[A]$ and its estimator $\hat{R}$ is the adjacency matrix A. For clarity's sake, let us recall here this algorithm.

Algorithm 1: Spectral clustering for community detection. *Compute the second (normalized) eigenvector $\hat{u}$ of A and classify each i according to the sign of the ith coordinate of $\hat{u}$, $\hat{\theta}_i = sign(\hat{u}_i)$, that is $i \in C_1$ iff $\hat{\theta}_i = 1$ or classify $i \in C_{\hat{\theta}_i}$.*

According (2.3), the proportion of misclassification of this algorithm is bounded by

$$\mathcal{R}(\hat{\theta}) \leqslant \frac{8}{\delta^2} \|A - \mathbb{E}[A]\|^2 \ ,$$

where δ is the spectral gap of $\mathbb{E}[A]$ corresponding to its second eigenvalue, which is, according to the general remarks of Section 2.3.1:

$$\delta = \min(\lambda_1 - \lambda_2, \lambda_2) = n\mu, \qquad \mu = \min(p - q, 2q) \ .$$

To conclude the analysis of this algorithm, it remains to bound $\|A - \mathbb{E}[A]\|$. In Chapter 4 (more precisely, in Theorem 43), we will prove that, with probability $1 - \delta$, $\|A - \mathbb{E}[A]\| \leqslant C \max(\sqrt{np \log(n/\delta)}, \log(n/\delta))$, this shows that, for the SCA,

$$\mathcal{R}(\hat{\theta}) \leqslant C \max\left(\frac{p \log(n/\delta)}{n\mu^2}, \left(\frac{\log(n/\delta)}{n\mu} \right)^2 \right) \ .$$

Equivalently, we get that $\mathcal{R}(\hat{\theta}) \leqslant \epsilon$ if

$$n \frac{\mu^2}{p} \gtrsim \frac{\log(n/\delta)}{\epsilon} \ ,$$

as the second condition $n\mu \gtrsim \log(n/\delta)/\sqrt{\epsilon}$ is then automatically satisfied.

2.5 Guedon-Vershynin spectral algorithm

The material of this section is borrowed from the paper "Community detection in sparse networks via Grothendieck's inequality" by O. Guédon and R. Vershynin.

2.5.1 Critical discussion of the SCA bounds

There are two main limitations to the SCA algorithm. The first one is that the quality of the algorithm gets worse as q decreases, which is non-satisfying as $q = 0$ would mean that both communities are disjoints so recovering them should be easier. The reason is that the first two eigenvalues of A have to be different for SCA to succeed.

The second limitation is that the rate $\log(n)/n$ is slightly sub-optimal. The reason here is that we don't exploit the important information that the second eigenvector belongs to the hypercube.

The purpose of the new strategy is to bypass both issues.

2.5.2 Definition of the new algorithm

To build a new spectral algorithm, we have to build a new reference matrix R as a solution of a tractable problem that can easily be approximated, then, we estimate R by the solution $\hat{R}$ of an approximating problem.

Building the reference matrix R. We proceed in two steps to define R. In the first step, we build a preliminary matrix P by removing from $\mathbb{E}[A]$ the uninformative part regarding its main eigenvalue. This makes our informative eigenvector $\mathbf{u}_2$ the first eigenvector of this preliminary matrix P. This can easily be done. Indeed, write the eigendecomposition of $\mathbb{E}[A]$ as

$$\mathbb{E}[A] = \lambda_1 \frac{\mathbf{J}_{2n}}{2n} + \lambda_2 \mathbf{u}_2 \mathbf{u}_2^T \ ,$$

so $\mathbf{u}_2$ is also the largest eigenvector of the preliminary matrix

$$P = \mathbb{E}[A] - \lambda_1 \frac{\mathbf{J}_{2n}}{2n} = \frac{\lambda_2}{2n} \begin{bmatrix} \mathbf{J}_n & -\mathbf{J}_n \\ -\mathbf{J}_n & \mathbf{J}_n \end{bmatrix} \ .$$

The second problem is a bit more tricky. A first idea would be to simply say that

$$\sqrt{2n}\mathbf{u}_2 = \text{argmax}\{\mathbf{x}^T P \mathbf{x},\ \mathbf{x} \in \{-1, 1\}^{2n}\} \ .$$

Exploiting this idea would yield to a strategy that is statistically optimal (look at the problem at the end of the chapter), but that cannot be exploited in practice. Indeed, the max-cut problem defining $\sqrt{n}\mathbf{u}_2$ in the last

formula is NP-hard. Therefore, we propose here a convex relaxation of it (this construction based on a SDP relaxation is extremely classical). Write

$$\mathbf{x}^T P \mathbf{x} = \left\langle P, \mathbf{x}\mathbf{x}^T \right\rangle_F \ ,$$

where $\langle \cdot, \cdot \rangle_F$ denote the Frobenius inner product between matrices. To build a convex relaxation of

$$\max_{\mathbf{x} \in \{-1,1\}^{2n}} \left\langle P, \mathbf{x}\mathbf{x}^T \right\rangle_F \ ,$$

we just need to build a convex set that contains the matrices $\mathbf{x}\mathbf{x}^T$, with $\mathbf{x} \in \{-1,1\}^{2n}$. The idea of the algorithm is to choose

$$\mathcal{S}_+ = \{\mathbb{X} \succcurlyeq 0 : \mathbb{X}_{i,i} = 1\} \ .$$

Indeed, $\mathcal{S}_+$ is a convex set and all matrices $\mathbb{X} = \mathbf{x}\mathbf{x}^T$, with $\mathbf{x} \in \{-1,1\}^{2n}$ are symmetric positive semi-definite and satisfy $\mathbb{X}_{i,i} = 1$, hence $\mathcal{S}_+ \supset \{\mathbf{x}\mathbf{x}^T : \mathbf{x} \in \{-1,1\}^n\}$. The maximization of $\langle P, \mathbb{X} \rangle_F$ over $\mathcal{S}_+$ is easy: Define

$$R = \begin{bmatrix} \mathbf{J}_n & -\mathbf{J}_n \\ -\mathbf{J}_n & \mathbf{J}_n \end{bmatrix} = \frac{2n}{\lambda_2} P = 2n \mathbf{u}_2 \mathbf{u}_2^T \ .$$

First remark that

$$\langle \mathbb{X}, P \rangle_F = \frac{\lambda_2}{2n} \langle \mathbb{X}, R \rangle_F \ .$$

Then, by Cauchy-Schwarz inequality, for any $\mathbb{X} \in \mathcal{S}_+$,

$$\langle \mathbb{X}, R \rangle_F \leqslant \|\mathbb{X}\|_F \|R^*\|_F \ .$$

Then, as $\mathbb{X} \in \mathcal{S}_+$, there exists $\mathbf{X}$ such that $\mathbf{X}^T\mathbf{X}$, so $\mathbb{X}_{i,j} = \langle C_i(\mathbf{X}), C_j(\mathbf{X}) \rangle$, where $C_i(\mathbf{X})$ denote the i-th column of $\mathbf{X}$. The condition $\mathbb{X}_{i,i} = 1$ then means that $\|C_i(\mathbf{X})\|_2 = 1$, thus by Cauchy-Schwarz inequality, all $|\mathbb{X}_{i,j}| \leqslant 1$, therefore,

$$\|\mathbb{X}\|_F \leqslant 2n = \|R\|_F \ .$$

As R belongs to $\mathcal{S}_+$, we have thus, for any $\mathbb{X} \in \mathcal{S}_+$,

$$\langle \mathbb{X}, R \rangle_F \leqslant \|\mathbb{X}\|_F \|R\|_F \leqslant \|R\|_F^2 = \langle R, R \rangle_F \ .$$

By the equality condition in Cauchy-Schwarz inequality, we have thus

$$\text{argmax}_{\mathbb{X} \in \mathcal{S}_+} \langle P, \mathbb{X} \rangle = R = \begin{bmatrix} \mathbf{J}_n & -\mathbf{J}_n \\ -\mathbf{J}_n & \mathbf{J}_n \end{bmatrix} \ . \tag{2.4}$$

The matrix R defined above has clearly $\mathbf{u}_2$ as first eigenvector. We use the representation of R as solution of the convex problem $R = \text{argmax}_{\mathbb{X} \in \mathcal{S}_+} \langle P, \mathbb{X} \rangle$ to estimate it and classify finally according to the general spectral scheme described at the beginning of the chapter.

Step 2: Building $\hat{R}$. The construction of $\hat{R}$ is then a plugging strategy: we estimate P by an estimator $\hat{P}$ and then R by

$$\hat{R} \in \text{argmax}_{\mathbb{X} \in \mathcal{S}_+} \left\langle \hat{P}, \mathbb{X} \right\rangle \ . \tag{2.5}$$

To build an estimator $\hat{P}$ of P, we estimate $\mathbb{E}[A]$ by A and we need an estimator of $\lambda_1 = n(p+q)$. We have

$$\mathbb{E}[\sum_{1 \leqslant i \leqslant j \leqslant 2n} A_{i,j}] = \frac{1}{2}\mathbb{E}[\sum_{1 \leqslant i,j \leqslant 2n} A_{i,j}] + \frac{1}{2}\mathbb{E}[\sum_{i=1}^{2n} A_{i,i}] = \frac{n^2}{2}(p+q) + np \ .$$

This shows that

$$\hat{\lambda}_1 = \frac{2}{n} \sum_{1 \leqslant i \leqslant j \leqslant 2n} A_{i,j} \ ,$$

is an asymptotically unbiased estimator of λ_1. We deduce from this estimator our approximating matrix

$$\hat{P} = A - \frac{\hat{\lambda}_1}{2n} \mathbf{J}_{2n} \ . \tag{2.6}$$

2.5.3 Bounding the risk

Guedon and Vershynin's algorithm can now be formally defined:

Algorithm 2: (GVSA) GV Spectral algorithm. *Compute the first (normalized) eigenvector $\hat{u}$ of $\hat{R}$ defined in* (2.5) *and classify each i according to the sign of the ith coordinate of $\hat{u}$: $\hat{\theta}_i = sign(\hat{u}_i)$, that is $i \in C_1$ iff $\hat{\theta}_i = 1$ or classify $i \in C_{\hat{\theta}_i}$.*

The performance of this algorithm is controlled by our general bound (2.3). Here, the spectral gap δ is simply

$$\delta = \lambda_1(R) = 2n \ .$$

Therefore, it is sufficient to bound the operator norm of $\hat{R} - R$. As the operator norm $\|\hat{R} - R\|$ is the sup-norm of its spectrum, it is bounded from above by its ℓ_2 norm, which is $\|\hat{R} - R\|_F$. We deduce that

$$\mathcal{R}(\hat{\theta}) \leqslant \frac{2}{n^2}\|\hat{R} - R\|_F^2 = \frac{2}{n^2}(\|\hat{R}\|_F^2 + \|R\|_F^2 - 2\left\langle \hat{R}, R \right\rangle_F) \ .$$

We have $\|R\|_F^2 = 4n^2$ and $\|\hat{R}\|_F^2 \leqslant 4n^2$ as we have seen that its entries are all in $[-1, 1]$. We get that

$$\mathcal{R}(\hat{\theta}) \leqslant \frac{4}{n^2}(\|R\|_F^2 - \left\langle \hat{R}, R \right\rangle_F) \ . \tag{2.7}$$

Thus, our main task is to bound from bellow the inner product $\left\langle \hat{R}, R \right\rangle_F$.

Step 1: Basic remarks. To proceed, we first use repeatedly that

$$R = \begin{bmatrix} \mathbf{J}_n & -\mathbf{J}_n \\ -\mathbf{J}_n & \mathbf{J}_n \end{bmatrix} = \frac{2n}{\lambda_2} P \ .$$

Indeed, we get

$$\begin{aligned}
\left\langle \hat{R}, R \right\rangle_F &= \frac{2n}{\lambda_2} \left\langle \hat{R}, P \right\rangle_F \\
&= \frac{2n}{\lambda_2} \left(\left\langle \hat{R}, \hat{P} \right\rangle_F - \left\langle \hat{R}, \hat{P} - P \right\rangle_F \right) \\
&\geqslant \frac{2n}{\lambda_2} \left(\left\langle \hat{R}, \hat{P} \right\rangle_F - \sup_{\mathbb{X} \in \mathcal{S}_+} \left\langle \mathbb{X}, \hat{P} - P \right\rangle_F \right) \\
&\geqslant \frac{2n}{\lambda_2} \left(\left\langle R, \hat{P} \right\rangle_F - \sup_{\mathbb{X} \in \mathcal{S}_+} \left\langle \mathbb{X}, \hat{P} - P \right\rangle_F \right) \\
&\geqslant \frac{2n}{\lambda_2} \left(\langle R, P \rangle_F - \left\langle R, P - \hat{P} \right\rangle - \sup_{\mathbb{X} \in \mathcal{S}_+} \left\langle \mathbb{X}, \hat{P} - P \right\rangle_F \right) \\
&\geqslant \frac{2n}{\lambda_2} \left(\langle R, P \rangle_F - 2 \sup_{\mathbb{X} \in \mathcal{S}_+} | \left\langle \mathbb{X}, \hat{P} - P \right\rangle_F | \right) \\
&= \|R\|_F^2 - \frac{4n}{\lambda_2} \sup_{\mathbb{X} \in \mathcal{S}_+} | \left\langle \mathbb{X}, \hat{P} - P \right\rangle_F |
\end{aligned}$$

Plugging this bound into (2.7) yields

$$\mathcal{R}(\hat{\theta}) \leqslant \frac{16}{n\lambda_2} \sup_{\mathbb{X} \in \mathcal{S}_+} | \left\langle \mathbb{X}, \hat{P} - P \right\rangle_F | \ .$$

Step 2: Grothendieck's inequality. The second step of the proof is the following inequality due to Grothendieck:

Theorem 4. *For any matrix B such that*

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \{-1, 1\}^n : \qquad \sum_{1 \leqslant i,j \leqslant n} B_{i,j} x_i y_j \leqslant 1 \ ,$$

we have, for any vectors $\boldsymbol{u}, \boldsymbol{v}$ in $\mathbb{B}_2 = \{\boldsymbol{u} \in \mathbb{R}^n : \|\boldsymbol{u}\|_2 \leqslant 1\}$,

$$\sum_{1 \leqslant i,j \leqslant n} B_{i,j} \langle \boldsymbol{u}_i, \boldsymbol{v}_j \rangle \leqslant 2 \ .$$

Remark 5. *We prove a slightly sub-optimal version of this inequality at the end of the chapter, where the constant 2 is replaced by some absolute constant C.*

Let us now reformulate Grothendieck's inequality to make it more related to our problem. First, define the set $\mathbf{S} = \{\mathbf{X} \text{ sym} : \mathbf{X}^T\mathbf{X} \in \mathcal{S}_+\}$. Grothendieck inequality can be rewritten: For any matrix A,

$$\sup_{\mathbf{X},\mathbf{Y}\in\mathbf{S}} \left\langle A, \mathbf{X}^T\mathbf{Y}\right\rangle_F \leqslant 2 \sup_{x,y\in\{-1,1\}^n} \langle x, Ay\rangle \ .$$

The second remark is that, for any matrix A,

$$\sup_{\mathbb{X}\in\mathcal{S}_+} |\langle A, \mathbb{X}\rangle_F| = \sup_{\mathbf{X}\in\mathbf{S}} |\left\langle A, \mathbf{X}^T\mathbf{X}\right\rangle_F| \leqslant \sup_{\mathbf{X},\mathbf{Y}\in\mathbf{S}} \left\langle A, \mathbf{X}^T\mathbf{Y}\right\rangle_F \ ,$$

where the last inequality follows by taking $\mathbf{Y}^* = \epsilon\mathbf{X}^*$, with $\mathbf{X}^*$ a maximizer of $|\left\langle A, \mathbf{X}^T\mathbf{X}\right\rangle_F|$ and $\epsilon = \text{Sign}(\left\langle A, (\mathbf{X}^*)^T\mathbf{X}^*\right\rangle_F)$.

Together with the conclusion of Step 1, these remarks imply that

$$\mathcal{R}(\hat{\theta}) \leqslant \frac{32}{n\lambda_2} \sup_{x,y\in\{-1,1\}^n} \left\langle x, (\hat{P} - P)y\right\rangle \ . \tag{2.8}$$

Step 3: Probabilistic bounds. Let us now recall that

$$P = \mathbb{E}[A] - \frac{\lambda_1}{2n}\mathbf{J}_{2n}, \qquad \hat{P} = A - \frac{\hat{\lambda}_1}{2n}\mathbf{J}_{2n} \ ,$$

so

$$\begin{aligned}\sup_{x,y\in\{-1,1\}^{2n}} \left\langle x, (\hat{P} - P)y\right\rangle \leqslant & \sup_{x,y\in\{-1,1\}^{2n}} \langle x, (A - \mathbb{E}[A])y\rangle \\ & + \frac{\lambda_1 - \hat{\lambda}_1}{2n} \sup_{x,y\in\{-1,1\}^{2n}} \langle x, \mathbf{J}_{2n}y\rangle \ .\end{aligned}$$

We have

$$\sup_{x,y\in\{-1,1\}^{2n}} \langle x, \mathbf{J}_{2n}y\rangle = 4n^2 \ .$$

Besides

$$\hat{\lambda}_1 - \lambda_1 = \frac{2}{n} \sum_{1\leqslant i\leqslant j\leqslant 2n} (A_{i,j} - \mathbb{E}[A_{i,j}]) + 2p \ .$$

Bernstein's inequality shows that, for any $z > 0$, with probability $1 - 2\exp(-z^2)$,

$$\left| \sum_{1\leqslant i\leqslant j\leqslant 2n} (A_{i,j} - \mathbb{E}[A_{i,j}]) \right| \lesssim nz\sqrt{p} \vee z^2 \ .$$

thus, with the same probability,

$$\frac{|\lambda_1 - \hat{\lambda}_1|}{2n} \sup_{x,y\in\{-1,1\}^{2n}} \langle x, \mathbf{J}_{2n}y\rangle \lesssim np \vee nz\sqrt{p} \vee z^2 \ . \tag{2.9}$$

Now, for any x, y in the hypercube $\{-1,1\}^{2n}$, we have

$$\langle x, (A - \mathbb{E}[A])y\rangle = \sum_{1\leqslant i\leqslant j\leqslant 2n} (A_{i,j} - \mathbb{E}[A_{i,j}])(x_i y_j + x_j y_i) \ .$$

Bernstein's inequality shows that, for any $z > 0$, with probability $1 - 2\exp(-z^2)$,

$$\big|\, \langle x, (A - \mathbb{E}[A])y\rangle \,\big| \lesssim nz\sqrt{p} \vee z^2 \ .$$

By a union bound, we get that, with probability $1 - 2 * 4^{2n}\exp(-z^2)$,

$$\sup_{x,y\in\{-1,1\}^{2n}} \langle x, (A - \mathbb{E}[A])y\rangle \lesssim nz\sqrt{p} \vee z^2 \ .$$

Together with (2.9), this shows that, with probability $1-2*(4^{2n}+1)\exp(-z^2)$,

$$\sup_{x,y\in\{-1,1\}^{2n}} \left\langle x, (\hat{P} - P)y\right\rangle \lesssim np \vee nz\sqrt{p} \vee z^2 \ .$$

Plugging this bound into (2.8) finally shows that, with probability $1 - 2 * (4^{2n} + 1)\exp(-z^2)$,

$$\mathcal{R}(\hat{\theta}) \lesssim \frac{np \vee nz\sqrt{p} \vee z^2}{n^2(p-q)} \ .$$

To discuss this result, it is helpful to write $z = n\sqrt{p}z'$, so, for any $z' > 0$, we have, with probability $1 - 2 * (4^{2n} + 1)\exp(-n^2 p(z')^2)$,

$$\mathcal{R}(\hat{\theta}) \lesssim \frac{np \vee n^2 pz' \vee n^2 p(z')^2}{n^2(p-q)} \ . \tag{2.10}$$

This last upper bound is smaller than ϵ if

$$\frac{p-q}{p} \gtrsim \frac{1}{n\epsilon} \vee \frac{z'}{\epsilon} \vee \frac{(z')^2}{\epsilon} \ .$$

Besides, the probability is large iff

$$np(z')^2 \gtrsim 1 \ .$$

These conditions are compatible iff

$$\frac{p-q}{p} \gtrsim \frac{1}{\sqrt{np}\epsilon} \ .$$

Putting everything together, we have obtained that

Theorem 6. *Assume that*

$$\frac{p-q}{p} \geqslant \frac{C}{\sqrt{np}\epsilon} \ ,$$

then, with probability $1 - 2\exp(-n)$*, GVSA makes less than* ϵn *errors.*

Remark 7. *The condition in Theorem 6 cannot be fulfilled unless* $p \gtrsim (n\epsilon^2)^{-1}$ *and in this case, it is fulfilled when* $p - q$ *is sufficiently large compared to* p*. It is clear that this result improves upon the result proved for SCA as it holds*

1. *without lower bound on* q*, which is much more reasonable,*

2. *in sparse networks where* $p \asymp 1/n$*.*

On the other hand, the dependency of p *with respect to* ϵ *is much worse for this algorithm than for SCA.*

2.6 Proof of Grothendieck's inequality

Let $K(B)$ denote the smallest constant such that, for any vectors X_i, Y_i taking values in an Hilbert space H and such that $\|X_i\|, \|Y_j\| \leqslant 1$,

$$\sum_{1 \leqslant i,j \leqslant n} B_{i,j} \langle X_i, Y_j \rangle \leqslant K(B) \ .$$

Such constant always exists and is always smaller than $\sum_{1 \leqslant i,j \leqslant n} |B_{i,j}|$. Notice that, as one can restrict this to the space spanned by X_i (resp. Y_j), and as this space is isometric to $\mathbb{R}^n$, we have

$$K(B) = \sup_{\mathbf{u}_i, \mathbf{v}_j \in \mathbb{B}_2} \sum_{1 \leqslant i,j \leqslant n} B_{i,j} \langle \mathbf{u}_i, \mathbf{v}_j \rangle$$

Let $Z \sim \mathrm{N}(0, \mathbf{I})$ denote a standard Gaussian vector and, for any $\mathbf{u} \in \mathbb{B}_2$, let $X_{\mathbf{u}} = \langle Z, \mathbf{u} \rangle$. $X_{\mathbf{u}}$ is a centered Gaussian process with covariance

$$\Sigma_{\mathbf{u},\mathbf{v}} = \mathbb{E}[X_{\mathbf{u}} X_{\mathbf{v}}] = \mathbf{u}^T \mathbb{E}[ZZ^T]\mathbf{v} = \langle \mathbf{u}, \mathbf{v} \rangle \ \ ,$$

so

$$\sum_{1 \leqslant i,j \leqslant n} B_{i,j} \langle \mathbf{u}_i, \mathbf{v}_j \rangle = \mathbb{E}\Big[\sum_{1 \leqslant i,j \leqslant n} B_{i,j} X_{\mathbf{u}_i} X_{\mathbf{v}_j} \Big] \ .$$

Now let $R > 0$ and, for any $\mathbf{u} \in \mathbb{B}_2$, let

$$X^a_{\mathbf{u}} = X_{\mathbf{u}} \mathbf{1}_{\{|X_{\mathbf{u}}| \leqslant R\}}, \qquad X^b_{\mathbf{u}} = X_{\mathbf{u}} \mathbf{1}_{\{|X_{\mathbf{u}}| > R\}} \ .$$

We deduce that

$$\sum_{1 \leqslant i,j \leqslant n} B_{i,j} \langle \mathbf{u}_i, \mathbf{v}_j \rangle = \sum_{i=1}^{4} E_i \ \ ,$$

where

$$E_1 = \mathbb{E}\Big[\sum_{1\leqslant i,j\leqslant n} B_{i,j} X^a_{\mathbf{u}_i} X^a_{\mathbf{v}_j}\Big] \ ,$$
$$E_2 = \mathbb{E}\Big[\sum_{1\leqslant i,j\leqslant n} B_{i,j} X^a_{\mathbf{u}_i} X^b_{\mathbf{v}_j}\Big] \ ,$$
$$E_3 = \mathbb{E}\Big[\sum_{1\leqslant i,j\leqslant n} B_{i,j} X^b_{\mathbf{u}_i} X^a_{\mathbf{v}_j}\Big] \ ,$$
$$E_4 = \mathbb{E}\Big[\sum_{1\leqslant i,j\leqslant n} B_{i,j} X^b_{\mathbf{u}_i} X^b_{\mathbf{v}_j}\Big] \ .$$

As $|X^a_{\mathbf{u}_i}| \leqslant R$, we have, by assumption

$$\sum_{1\leqslant i,j\leqslant n} B_{i,j} X^a_{\mathbf{u}_i} X^a_{\mathbf{v}_j} \leqslant R^2 \ ,$$

so $E_1 \leqslant R^2$.

Besides, we have

$$\mathbb{E}[(X^a_{\mathbf{u}})^2] \leqslant R^2$$

$$\begin{aligned}\mathbb{E}[(X^b_{\mathbf{u}})^2] &= 2\int_R^{+\infty} x^2 \exp(-x^2/2)\frac{\mathrm{d}x}{\sqrt{2\pi}} \\ &= 2[-x\exp(-x^2/2)]_R^{+\infty} + \mathbb{P}(|\,\mathrm{N}(0,1)| > R) \\ &\leqslant (2R/\sqrt{2\pi}+1)\exp(-R^2/2) \ .\end{aligned}$$

Therefore, as all $X^a_{\mathbf{u}}$ and $X^b_{\mathbf{u}}$ belong to the Hilbert space

$$H = \{f : \mathbb{R}^n \to \mathbb{R} : \mathbb{E}[f^2(Z)] < \infty\} \ ,$$

endowed with the inner product $\langle f, g\rangle = \mathbb{E}[f(Z)g(Z)]$, by definition of $K(B)$, we have

$$E_2 \leqslant K(B) R\sqrt{2R/\sqrt{2\pi}+1}\exp(-R^2/4) \ ,$$
$$E_3 \leqslant K(B) R\sqrt{2R/\sqrt{2\pi}+1}\exp(-R^2/4) \ ,$$
$$E_4 \leqslant K(B)(2R/\sqrt{2\pi}+1)\exp(-R^2/2) \ .$$

We have obtained that

$$K(B) \leqslant R^2 + K(B)\psi(R) \ ,$$

where

$$\psi(R) \leqslant 2R\sqrt{2R/\sqrt{2\pi}+1}\exp(-R^2/4) + (2R/\sqrt{2\pi}+1)\exp(-R^2/2) \ .$$

Therefore, for any $R > 0$ such that $\psi(R) < 1$, we have

$$K(B) \leqslant \frac{R^2}{1-\psi(R)} \ .$$

2.7 Problem

The problem is decomposed into two essentially independent parts. In the first part, we propose to consider the reference matrix

$$R = \mathbb{E}[A] - \lambda_1 \mathbf{u}_1 \mathbf{u}_1^T \ .$$

1. Propose an estimator $\hat{R}$ of R and deduce a spectral clustering algorithm associated.
2. Compute the spectral gap δ of R.
3. Control the operator norm $\|\hat{R} - R\|$.
4. Deduce an upper bound on the frequency of misclassified nodes of the algorithm in question 1: $\mathcal{R}(\hat{\theta})$. Discuss the result.

In the second part of this problem, we define $P = \mathbb{E}[A] - \lambda_1 \mathbf{u}_1 \mathbf{u}_1^T$ and the reference matrix

$$\theta^* \in \mathrm{argmax}_{\theta \in \{-1,1\}^{2n}} \theta^T P \theta \ .$$

5. Prove that $\theta^* = \mathbf{u}_2$. We estimate thus P by $\hat{P}$ given in question 1 and θ^* by
$$\hat{\theta} \in \mathrm{argmax}_{\theta \in \{-1,1\}^{2n}} \theta^T \hat{P} \theta \ .$$
6. Prove that
$$\left\langle \theta^*, \hat{\theta} \right\rangle^2 = C \left\langle P, \hat{\theta}\hat{\theta}^T \right\rangle_F \ ,$$
for some constant C that should be computed.
7. Prove that
$$\left\langle \theta^*, \hat{\theta} \right\rangle^2 \geqslant 4n^2 - 2C \sup_{\theta \in \{-1,1\}^{2n}} |\left\langle \hat{P} - P, \theta\theta^T \right\rangle| \ .$$
8. Deduce an upper bound on $\|\theta^* - \hat{\theta}\|^2$ and then on $\mathcal{R}(\hat{\theta})$. Discuss the result.

Chapter 3

Concentration of measure

In the first lectures, we saw that several problems in machine learning reduce to the problem of building deterministic bounds on suprema of sums of independent random variables. These kind of deviation inequalities can be deduced from concentration inequalities, which show that regular functions of independent random variables do not deviate much from their expectation. Then, we have to bound expected suprema. This chapter presents basic tools to prove concentration inequalities for a random variable X around its expectation, that is upper bounds on the probabilities $\mathbb{P}(|X - \mathbb{E}[X]| > x)$ for any $x > 0$. We focus here on non necessarily bounded random variables such as sub-Gaussian and sub-Gamma random variables, for which these tails are exponentially small. We conclude the chapter with the application of the general result to the particular case where X is a sum of possibly unbounded independent random variables. It turns out that deriving asymptotically optimal bounds in this case is still quite open.

In all the chapter, C denotes a numerical constant whose value may change from line to line.

3.1 Motivating example

Suppose we want to analyse the SVM predictor (see also the problem in Chapter 1):

$$\hat{\theta}_\lambda \in \operatorname{argmin}_{\theta \in \Theta} P_n \ell_\theta + \lambda \|\theta\|_2^2 \ ,$$

where $x \in \mathbb{R}^p = \Theta$, $y \in \{-1, 1\}$ and $\ell_\theta(x, y) = \varphi(-y \langle x, \theta \rangle)$, with φ a 1-Lipschitz function, say $\varphi(x) = \log(1 + \exp(x))$.

Following the agenda presented in the first lecture, it is not hard to get convinced that this analysis requires in particular an upper bound on the linear process

$$\sup_{\theta \in \mathcal{C}} \langle P_n \nabla_\theta \ell_{\theta^*}, \theta - \theta^* \rangle \ ,$$

over an ellipsoid $\mathcal{C}$. Under a simple assumption such as $x \sim \mathrm{N}(0, \Sigma)$, the concentration of this process is not so easy to prove as $P_n \nabla_\theta \ell_{\theta^*} = P_n[\varphi'(-y \langle x, \theta^* \rangle)(-yx)] = n^{-1} \sum_{i=1}^n \varphi'(-y_i \langle x_i, \theta^* \rangle)(-y_i x_i)$, so:

1. This concentration cannot be reduced to the one of a Lipschitz function of a Gaussian vector, so the GCI does not apply.

2. The random variables $Z_\theta = \langle \varphi'(-y_i \langle x_i, \theta^* \rangle)(-y_i x_i), \theta - \theta^* \rangle$ are not bounded.

The purpose of the following section is to introduce the proper setting to prove concentration of $P_n Z_\theta$. Deviation inequalities for $\sup_{\theta \in \mathcal{C}} P_n[Z_\theta] - \mathbb{E}[Z_\theta]$ will be provided in Lectures 6 and 5.

3.2 Sub-Gaussian random variables

The variable Z_θ is actually a sub-Gaussian random variable. The following result provides several equivalent ways to define sub-Gaussian random variables. The first of these characterizations is that these random variables are those concentrating around 0 at least as fast as Gaussian random variables.

Theorem 8. *Let X denote a random variable. Then, the following properties are equivalent.*

(i) For any $x > 0$, $\mathbb{P}(|X| > x) \leqslant 2 \exp(-x^2/K_1^2)$.

(ii) For any integer $p \geqslant 1$, $\|X\|_p := \mathbb{E}[|X|^p]^{1/p} \leqslant K_2 \sqrt{p}$.

(iii) For any $|s| < 1/K_3$, $\mathbb{E}[\exp(s^2 X^2)] \leqslant \exp(s^2 K_3^2)$.

(iv) $\mathbb{E}[\exp(X^2/K_4^2)] \leqslant 2$.

If moreover, X is centered, then all properties are also equivalent to

(v) For any $s \in \mathbb{R}$, $\mathbb{E}[\exp(sX)] \leqslant \exp(s^2 K_5^2)$.

If one of these properties holds, then we say that X is sub-Gaussian and the different K_i differ by at most a multiplicative numerical constant. The smallest constant K_4 such that (iv) holds is called the sub-Gaussian norm of X and it is denoted by $\|X\|_{\psi_2}$.

Theorem 8 is the most important result of this chapter. We shall mostly use the implication $(v) \implies (i)$ in the following. In this case, the proof based on Chernoff's bound shows that if (v) is satisfied with K_5, then (i) holds with $K_1 = 2K_5$.

As an exercise, show that $\mathbb{E}[X] = 0$ if (v) holds.

Proof. We prove a chain of implications.

(i) $\Longrightarrow$ (ii) Assume that $K_1 = 1$. Then, for any $p \geqslant 1$, we have

$$\begin{aligned}\mathbb{E}[|X|^p] &= \int_0^{+\infty} \mathbb{P}(|X|^p > t)\mathrm{d}t \quad \text{since } |X|^p \geqslant 0 \text{ a.s.} \\ &= p\int_0^{+\infty} \mathbb{P}(|X| > u)u^{p-1}\mathrm{d}u \quad \text{posing } t = u^p \\ &\leqslant p\int_0^{+\infty} (u^2)^{p/2-1}\exp(-u^2)2u\mathrm{d}u \\ &= p\Gamma(p/2) \quad \text{posing } u^2 = v \ .\end{aligned}$$

Then, we use the following classical Stirling's approximation, valid for any $x > 0$,

$$\left(\frac{x}{e}\right)^x \leqslant \Gamma(x+1) \leqslant x^x \ . \tag{3.1}$$

It implies directly

$$\|X\|_p \leqslant p^{1/p}\sqrt{\frac{p}{2}} \leqslant 2\sqrt{p} \ .$$

Now let $K_1 > 0$, then $\mathbb{P}(|X/K_1| > x/K_1) \leqslant \exp(-(x/K_1)^2)$, so $|X/K_1|$ satisfies (i) with $K_1 = 1$, thus $\|X/K_1\|_p \leqslant 2\sqrt{p}$ and therefore $\|X\|_p \leqslant 2K_1\sqrt{p}$.

(ii) $\Longrightarrow$ (iii) We assume that $K_2 = 1$. We use a Taylor expansion to say that

$$\mathbb{E}[\exp(s^2X^2)] \leqslant \sum_{k=0}^{+\infty} \frac{s^{2k}k^k}{\Gamma(k+1)} \leqslant \sum_{k=0}^{+\infty} s^{2k}e^k \ .$$

The last inequality follows from (3.1). The last upper bound is finite if $|s| < 1/\sqrt{e}$ and then

$$\mathbb{E}[\exp(s^2X^2)] \leqslant \frac{1}{1-es^2} \ .$$

Moreover, for any $|s| < 1/\sqrt{2e}$, it follows that

$$\mathbb{E}[\exp(s^2X^2)] \leqslant 1 + \frac{es^2}{1-es^2} \leqslant 1 + 2es^2 \leqslant \exp(2es^2) \ .$$

Let now $K_2 > 0$, then $\|X/K_2\|_p \leqslant \sqrt{p}$ for any $p \geqslant 1$ so for any $|s| < 1/\sqrt{2e}$

$$\mathbb{E}[\exp((s/K_2)^2X^2)] \leqslant \exp(2eK_2^2(s/K_2)^2) \ ,$$

that, is for any $|s| < 1/\sqrt{2e}K_2$, $\mathbb{E}[\exp(s^2X^2)] \leqslant \exp(s^2 2eK_2^2)$.

(iii) $\Longrightarrow$ (iv) is trivial.

(iv) $\Longrightarrow$ (i) Let $x > 0$, we have

$$\mathbb{P}(|X| > x) = \mathbb{P}(\exp(X^2/K_4^2) > \exp(x^2/K_4^2)) \leqslant 2\exp(-x^2/K_4^2) \ ,$$

where the last inequality follows from Markov's inequality.

(iii) $\Longrightarrow$ (v) Let $s \in \mathbb{R}$. If $|s| < 1/K_3$, we use the inequality $\exp(x) \leqslant x + \exp(x^2)$ to say that

$$\mathbb{E}[\exp(sX)] \leqslant \mathbb{E}[\exp(s^2X^2)] \leqslant \exp(s^2K_3^2) \ .$$

If $|s| \geqslant 1/K_3$, we use that $sx \leqslant \frac{1}{2}(s^2K_3^2 + x^2/K_3^2)$ to obtain

$$\mathbb{E}[\exp(sX)] \leqslant \exp(s^2K_3^2/2)\mathbb{E}[\exp(X^2/2(K_3^2))] \leqslant \exp(s^2K_3^2) \ .$$

(v) $\Longrightarrow$ (i) We use Chernoff's bound: Let $x > 0$, then we have

$$\begin{aligned}\mathbb{P}(X > x) &= \inf_{s>0} \mathbb{P}(\exp(sX) > \exp(sx)) \\ &\leqslant \inf_{s>0} \exp(-sx + \log \mathbb{E}[\exp(sX)]) \\ &= \exp(-\sup_{s>0}\{sx - \log \mathbb{E}[\exp(sX)]\}) \ . \end{aligned} \tag{3.2}$$

When (v) holds, we have therefore

$$\mathbb{P}(X > x) \leqslant \exp(-\sup_{s>0}\{sx - s^2K_5^2\}) = \exp(-x^2/4K_5^2) \ .$$

Likewise, we have

$$\mathbb{P}(X < -x) = \mathbb{P}(-X > x) \leqslant \exp(-x^2/4K_5^2) \ .$$

And finally

$$\mathbb{P}(|X| > x) \leqslant 2\exp(-x^2/4K_5^2) \ .$$

□

Let us now gather some interesting remarks that will be useful in the following.

1. To prove that Z_θ is sub-Gaussian, we can use for example, the characterization by the moments, indeed, as $\langle x, \theta - \theta^* \rangle \sim \|\theta - \theta^*\|_\Sigma \, \mathrm{N}(0,1)$ and the standard Gaussian distribution is sub-Gaussian with $\|\mathrm{N}(0,1)\|_{\psi_2} = \kappa = \sqrt{2\log 2}$,

$$\mathbb{E}[\exp(Z_\theta^2/K^2)] \leqslant \mathbb{E}[\exp(\langle x, \theta - \theta^* \rangle / K^2)] \leqslant 2 \ ,$$

if $K \geqslant \kappa\|\theta - \theta^*\|_\Sigma$. Actually, modifying slightly the proof, we obtain a much more important result which is that $\|Z_\theta - Z_{\theta'}\|_{\psi_2} \leqslant \kappa\|\theta - \theta'\|_\Sigma$. This property says that Z_θ has sub-Gaussian increments, a property that will allow to bound the deviation of the supremum using chaining arguments in Lecture 5.

2. A second remark that can be useful is that $\|X\|_{\psi_2}$ defines indeed a norm. (Check in particular that the triangle inequality follows from convexity of $x \mapsto \exp(x^2)$).

3. A third remark is that, if $\|X\|_{\psi_2} \leqslant K$, then

$$\|X - \mathbb{E}[X]\|_{\psi_2} \leqslant \|X\|_{\psi_2} + \|\mathbb{E}[X]\|_{\psi_2} \leqslant 2\|X\|_{\psi_2} \leqslant 2K \ .$$

The first inequality follows from the fact that $\|\cdot\|_{\psi_2}$ is a norm while the second one comes from Characterization (iv) of sub-Gaussian random variables and Jensen's inequality. A consequence here is that

$$\begin{aligned} \|Z_\theta - \mathbb{E}[Z_\theta]\|_{\psi_2} &\leqslant 2\kappa\|\theta - \theta^*\|_\Sigma \ , \\ \|Z_\theta - Z_{\theta'} - \mathbb{E}[Z_\theta - Z_{\theta'}]\|_{\psi_2} &\leqslant 2\kappa\|\theta - \theta'\|_\Sigma \ . \end{aligned} \tag{3.3}$$

This centering step is key to derive from the sub-Gaussianity of independent random variables the sub-Gaussianity of the sum. Indeed, this property is immediate using characterization (v) of sub-Gaussian random variables. It is known as the general Hoeffding's inequality (see Theorem 9 in Section 3.3).

4. Bounded random variables are sub-Gaussian. Indeed, if $X \in [a, b]$ almost surely, then $\|X - \mathbb{E}[X]\|_{\psi_2} \leqslant (b-a)/\log 2$. This follows directly from the upper bound $\exp(-\log 2(X - \mathbb{E}[X])^2/(b-a)^2) \leqslant 2$ a.s., which yields that characterisation (iv) of sub-Gaussian random variables in Theorem 8 holds. A sharper analysis shows that (v) holds in this case with $K_5 = (b-a)/\sqrt{8}$.

3.3 Hoeffding's inequality

Theorem 9 (General Hoeffding's inequality)**.** *If $X_1, \ldots, X_n$ are independent centered and sub-Gaussian, then $\|\sum_{i=1}^n X_i\|_{\psi_2} \leqslant C\sqrt{\sum_{i=1}^n \|X_i\|_{\psi_2}^2}$.*

Proof. By independence, for any $s \in \mathbb{R}$,

$$\mathbb{E}[\exp(s\sum_{i=1}^n X_i)] = \prod_{i=1}^n \mathbb{E}[\exp(sX_i)] \ .$$

The result then follows directly from the characterisation (v) of sub-Gaussian random variables in Theorem 8. In particular, we have therefore that the constant K_5 for the sum satisfies $K_5^2 = \sum_{i=1}^n K_{5,i}^2$, where $K_{5,i}$ is the constant K_5 for the variable X_i and it follows that

$$\forall x > 0, \qquad \mathbb{P}(|\sum_{i=1}^n X_i| > x) \leqslant \exp\left(-\frac{x^2}{4\sum_{i=1}^n K_{5,i}^2}\right) \ .$$

□

Theorem 9 (with sharp constants) and the fact that bounded random variables are sub-Gaussian yield the classical Hoeffding's inequality.

Theorem 10 (Hoeffding's inequality)**.** *If $X_1, \dots, X_n$ are independent random variables such that $X_i \in [a_i, b_i]$ a.s., then*

$$\forall z > 0, \qquad \mathbb{P}(|\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])| > z) \leqslant \exp\left(-\frac{2z^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) .$$

To conclude this section, we present the concentration of the process $(P_n - P)Z_\theta$ that appear in the analysis of SVM. We have

$$\begin{aligned}
\|(P_n - P)(Z_\theta)\|_{\psi_2} &= \frac{1}{n}\|\sum_{i=1}^{n}(Z_\theta(x_i, y_i) - \mathbb{E}[Z_\theta(x, y)])\|_{\psi_2} \\
&\leqslant \frac{C}{n}\sqrt{\sum_{i=1}^{n}\|(Z_\theta(x_i, y_i) - \mathbb{E}[Z_\theta(x, y)])\|_{\psi_2}^2} \quad \text{Hoeffding} \\
&\leqslant \frac{C}{n}\sqrt{\sum_{i=1}^{n}\|\theta - \theta^*\|_\Sigma^2} \quad \text{by (3.3)} \\
&= \frac{C\|\theta - \theta^*\|_\Sigma}{\sqrt{n}} .
\end{aligned}$$

The same arguments can be used to show also that

$$\|(P_n - P)(Z_\theta - Z_{\theta'})\|_{\psi_2} \leqslant \frac{C\|\theta - \theta'\|_\Sigma}{\sqrt{n}} ,$$

therefore, that, for any $z > 0$,

$$\mathbb{P}((P_n - P)(Z_\theta - Z_{\theta'}) > z) \leqslant \exp\left(-\frac{nz^2}{C\|\theta - \theta'\|_\Sigma^2}\right) .$$

The following sections gather the proofs of the bounded difference inequality and the GCI, which provide two non-trivial examples of sub-Gaussian random variables.

3.4 Bounded difference inequality

The bounded difference inequality is an extension of Hoeffding's inequality that allows to bound suprema of bounded empirical processes.

Let $\mathbf{c} \in \mathbb{R}^n$ and $\mathcal{X}_1, \dots, \mathcal{X}_n$ denote measurable spaces. The set $\mathbb{BD}(\mathbf{c})$ is the set of functions $f : \prod_{i=1}^n \mathcal{X}_i \to \mathbb{R}$ such that

$$\forall \mathbf{x}, \mathbf{y} \in \prod_{i=1}^{n}\mathcal{X}_i, \qquad |f(\mathbf{x}) - f(\mathbf{y})| \leqslant \sum_{i=1}^{n} c_i \mathbf{1}_{\{x_i \neq y_i\}} .$$

Theorem 11 (Bounded difference inequality)**.** *Let $X_1, \dots, X_n$ denote independent random variables such that $X_i \in \mathcal{X}_i$ and let $f \in \mathbb{BD}(\boldsymbol{c})$. Then $\|f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)]\|_{\psi_2} \leqslant C\|\boldsymbol{c}\|_2$.*

Remark 12. *The proof shows that we have tight constants*

$$\forall s \in \mathbb{R}, \qquad \mathbb{E}[\exp(s(f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)]))] \leqslant \exp\left(\frac{s^2\|\boldsymbol{c}\|_2^2}{8}\right) .$$

Remark 13. *The result can be extended to functions with sub-Gaussian increments (see the problem at the end of the chapter).*

Proof. Let $s \in \mathbb{R}$ and $\mathcal{F}_i$ denote the sigma-algebra generated by $X_1, \dots, X_i$. Denoting by $\mathcal{F}_0 = \{\Omega, \emptyset\}$, we have

$$f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)] = \sum_{i=1}^n \Delta_i,$$

where

$$\Delta_i = \mathbb{E}[f(X_1, \dots, X_n)|\mathcal{F}_i] - \mathbb{E}[f(X_1, \dots, X_n)|\mathcal{F}_{i-1}] \ .$$

Let

$$B_i^+ = \sup_{x_i \in \mathcal{X}_i} \mathbb{E}[f(X_1, \dots, x_i, \dots, X_n)|\mathcal{F}_{i-1}] \ ,$$
$$B_i^- = \inf_{x_i \in \mathcal{X}_i} \mathbb{E}[f(X_1, \dots, x_i, \dots, X_n)|\mathcal{F}_{i-1}] \ .$$

We have $B_i^- \leqslant \mathbb{E}[f(X_1, \dots, X_n)|\mathcal{F}_i] \leqslant B_i^+$, B_i^+ and B_i^- are $\mathcal{F}_{i-1}$ measurable and, as $f \in \mathbb{BD}(\mathbf{c})$, $B_i^+ - B_i^- \leqslant c_i$. Therefore, by Hoeffding's lemma

$$\mathbb{E}[\exp(s\Delta_i)|\mathcal{F}_{i-1}] \leqslant \exp\left(\frac{s^2c_i^2}{8}\right) .$$

Proceeding recursively, it follows that

$$\mathbb{E}[\exp(s\sum_{i=1}^n \Delta_i)] \leqslant \exp\left(\frac{s^2\|\mathbf{c}\|_2^2}{8}\right) .$$

□

Let $X_1, \dots, X_n$ denote independent random variables such that $X_i \in [a_i, b_i]$ a.s.. Let

$$f : \prod_{i=1}^n [a_i, b_i] \to \mathbb{R}, \qquad f(x_1, \dots, x_n) = \sum_{i=1}^n x_i \ .$$

Let $\mathbf{x}, \mathbf{y} \in \prod_{i=1}^n [a_i, b_i]$ tels que $x_i = y_i$ for any $i \neq i_0$. Then

$$f(\mathbf{x}) - f(\mathbf{y}) = x_{i_0} - y_{i_0} \leqslant b_{i_0} - a_{i_0} \ .$$

Therefore, by the bounded difference inequality

$$\forall s \in \mathbb{R}, \qquad \mathbb{E}[\exp(s(\sum_{i=1}^{n} X_i - \mathbb{E}[X_i]))] \leqslant \exp\left(\frac{s^2 \sum_{i=1}^{n}(b_i - a_i)^2}{8}\right) .$$

The bounded difference therefore extends Hoeffding's inequality.

It also applies to the following more general situation that is of interest in learning theory. Let T denote a countable set and let $\{(X_{i,t})_{t\in T}, i = 1, \dots, n\}$ denote independent processes indexed by T. We assume that

$$\forall t \in T, \forall i \in \{1, \dots, n\}, \qquad X_{i,t} \in [a_i, b_i], \text{ a.s. } .$$

Let $\mathcal{X}_i = [a_i, b_i]^T$ endowed with the cylinder sigma-algebra and let

$$f : \prod_{i=1}^{n} \mathcal{X}_i \to \mathbb{R}, \qquad f((x_{1,t})_{t\in T}, \dots, (x_{n,t})_{t\in T}) = \sup_{t\in T}\left\{ \sum_{i=1}^{n} x_{i,t} - \mathbb{E}[\sum_{i=1}^{n} X_{i,t}]\right\} .$$

Using that $\sup_{t\in T} a_t - \sup_{t\in T} b_t \leqslant \sup_{t\in T}(a_t - b_t)$, we get that, for any $\mathbf{x}, \mathbf{y} \in \prod_{i=1}^{n} \mathcal{X}_i$ such that $(x_{i,t})_{t\in T} = (y_{i,t})_{t\in T}$ for any $i \neq i_0$,

$$f(\mathbf{x}) - f(\mathbf{y}) \leqslant \sup_{t\in T}\{x_{i_0,t} - y_{i_0,t}\} \leqslant b_{i_0} - a_{i_0} .$$

Therefore, by the bounded difference inequality, the random variable $Z = \sup_{t\in T} \sum_{i=1}^{n} X_{i,t} - \mathbb{E}[X_{i,t}]$ satisfies

$$\forall s \in \mathbb{R}, \qquad \mathbb{E}[\exp(s(Z - \mathbb{E}[Z]))] \leqslant \exp\left(\frac{s^2 \sum_{i=1}^{n}(b_i - a_i)^2}{8}\right) .$$

The bounded difference inequality shows that the supremum of empirical processes $Z = \sup_{t\in T} \sum_{i=1}^{n} X_{i,t} - \mathbb{E}[X_{i,t}]$ concentrates as well as if $T = \{t_0\}$ is reduced to a singleton!

See also the problem at the end of the section for suprema of sums of sub-Gaussian random variables.

3.5 Gaussian concentration inequality

Theorem 14. *Let $X \sim \mathrm{N}(0, \boldsymbol{I})$ denote a standard Gaussian vector on $\mathbb{R}^d$ and let $f : \mathbb{R}^d \to \mathbb{R}$ denote a function such that*

$$\forall x, y \in \mathbb{R}^d, \qquad |f(x) - f(y)| \leqslant L\|x - y\|_2 .$$

Then $\|f(X) - \mathbb{E}[f(X)]\|_{\psi_2} \leqslant CL$.

Remark 15. *We provide an elementary proof of the result with sub-optimal constants. Using Herbst's argument together with log-Sobolev's inequality for Gaussian distribution, one can prove (see for example P. Massart's lectures)*

$$\mathbb{E}[\exp(s(f(X) - \mathbb{E}[f(X)]))] \leqslant \exp\left(\frac{s^2 L^2}{2}\right) .$$

This bound is sharp as can be seen when $d = 1$ and $f(x) = x$.

Proof. Let $s \in \mathbb{R}$, we want to bound from above

$$\mathbb{E}[\exp(s(f(X) - \mathbb{E}[f(X)]))] \ .$$

Let $X' \sim \mathrm{N}(0, \mathbf{I})$ be independent from X. By Jensen's inequality

$$\begin{aligned}\mathbb{E}[\exp(s(f(X) - \mathbb{E}[f(X)]))] &= \mathbb{E}[\exp(s(f(X) - \mathbb{E}[f(X')|X]))] \\ &= \mathbb{E}[\exp(\mathbb{E}[s(f(X) - f(X'))|X])] \\ &\leqslant \mathbb{E}[\exp(s(f(X) - f(X')))] \ .\end{aligned}$$

Now, for any $\theta \in [0, \pi/2]$, let

$$U(\theta) = \sin(\theta)X + \cos(\theta)X', \qquad V(\theta) = \partial_\theta U(\theta) = \cos(\theta)X - \sin(\theta)X' \ .$$

We have

$$\begin{gathered}U(\pi/2) = X, \qquad U(0) = X', \qquad U(\theta) \sim \mathrm{N}(0, \mathbf{I}), \qquad V(\theta) \sim \mathrm{N}(0, \mathbf{I}) \ , \\ (U(\theta), V(\theta))^T \text{ centered Gaussian vector}, \qquad \mathbb{E}[U(\theta)V(\theta)] = 0 \ .\end{gathered}$$

It follows that $U(\theta)$ and $V(\theta)$ are independent and, by the fundamental theorem of analysis,

$$f(X) - f(X') = \int_0^{\pi/2} \langle \nabla f(U(\theta)), V(\theta) \rangle \, \mathrm{d}\theta \ .$$

Conditioning on $U(\theta)$, $\langle \nabla f(U(\theta)), V(\theta) \rangle \sim \mathrm{N}(0, \|\nabla f(U(\theta))\|_2^2)$. By Jensen's inequality

$$\begin{aligned}\mathbb{E}[\exp(s(f(X) - f(X')))] &= \mathbb{E}[\exp(\int_0^{\pi/2} s \, \langle \nabla f(U(\theta)), V(\theta) \rangle \, \mathrm{d}\theta)] \\ &\leqslant \frac{2}{\pi} \int_0^{\pi/2} \mathbb{E}[\exp((\pi s/2) \, \langle \nabla f(U(\theta)), V(\theta) \rangle)] \mathrm{d}\theta\end{aligned}$$

It remains to recall that, if $X \sim \mathrm{N}(0, \sigma^2)$, $\mathbb{E}[\exp(sX)] = \exp(s^2\sigma^2/2)$ to conclude that

$$\mathbb{E}[\exp(s(f(X) - f(X')))] \leqslant \exp(s^2\pi^2L^2/8) \ .$$

□

3.6 Sub-Gamma random variables

Sub-Gaussian concentration inequalities are easy to use but usually do not provide the correct order of magnitude of deviations. A simple situation where this phenomenon occurs is as follows: Assume $X_1, \ldots, X_n$ are i.i.d.

with Bernoulli distribution $\mathcal{B}(p)$. Then Hoeffding's inequality states that, for any $x > 0$,

$$\mathbb{P}\left(\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - p)\right| > x\right) \leqslant 2\exp\big(-Cx^2\big) \ .$$

On the other hand, the central limit theorem states that $n^{-1/2}\sum_{i=1}^{n}(X_i - p) \implies \mathrm{N}(0, p(1-p))$, so by the Gaussian concentration inequality,

$$\lim_{n\to\infty}\mathbb{P}\left(\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - p)\right| > x\right) \leqslant 2\exp\left(-\frac{Cx^2}{p(1-p)}\right) \ .$$

The question that we investigate in this section is whether one can obtain deviation from the mean for random variables with probability $1 - 2\exp(-Cx^2/\mathrm{Var}(X))$, at least for small values of x.

Definition 16. *A random variable X is (b, σ^2)-sub-Gamma if*

$$\forall |s| < 1/b, \qquad \mathbb{E}[\exp(sX)] \leqslant \exp(s^2\sigma^2) \ .$$

A sufficient condition for sub-Gamma random variables can be expressed in terms of moments.

Proposition 17. *If X is centered and there exists b, σ^2 such that, for any $k \geqslant 2$,*

$$\mathbb{E}[|X|^k] \leqslant k! b^{k-2}\sigma^2 \ ,$$

then, for any $\epsilon > 0$, X is $(2b, 2\sigma^2)$-sub-Gamma.

Proof. Write, for any $s < 1/b$,

$$\mathbb{E}[\exp(sX)] \leqslant 1 + \sum_{k\geqslant 2}\frac{s^k\mathbb{E}[|X|^k]}{k!} \leqslant 1 + s^2\sigma^2\sum_{k\geqslant 0} b^k s^k \leqslant 1 + 2s^2\sigma^2 \leqslant \exp(2s^2\sigma^2) \ .$$

□

If X is bounded, the following result shows that X is sub-Gamma with parameter $\sigma^2 \asymp \mathrm{Var}(X)$.

Proposition 18. *Assume that X is centered and $|X| \leqslant b$ a.s., then, for any $\epsilon > 0$, X is $(b(1+\epsilon)/3\epsilon, \mathit{Var}(X)(1+\epsilon)/2)$ sub-Gamma.*

Remark 19. *Proposition 18 is a refinement of the previous proposition that would have shown directly that X is $(2b, 2\,\mathit{Var}(X))$ sub-Gamma.*

Proof. The proof goes along the same argument. Write $\sigma^2 = \mathrm{Var}(X)$. We have

$$\mathbb{E}[\exp(sX)] = 1 + \frac{s^2}{2} \sum_{k \geqslant 2} \frac{s^{k-2} \mathbb{E}[X^k]}{k!/2} \ .$$

Using that

$$\forall k \geqslant 2, \qquad \mathbb{E}[X^k] \leqslant \sigma^2 b^{k-2}, \qquad k! \geqslant 2 * 3^{k-2} \ ,$$

we get

$$\mathbb{E}[\exp(sX)] \leqslant 1 + \frac{s^2\sigma^2}{2} \frac{1}{1 - bs/3} \ .$$

So if $s < 3\epsilon/[b(1+\epsilon)]$, then

$$\mathbb{E}[\exp(sX)] \leqslant 1 + \frac{s^2\sigma^2(1+\epsilon)}{2} \leqslant \exp\left(\frac{s^2\sigma^2(1+\epsilon)}{2} \right) \ .$$

□

Finally, the following proposition establishes that the sum of independent sub-Gamma random variables is also sub-Gamma.

Proposition 20. *Assume that $X_1, \ldots, X_n$ are independent and for any $i \in \{1, \ldots, n\}$, X_i is (b_i, σ_i^2) sub-Gamma. Then $\sum_{i=1}^n X_i$ is $(\bar{b}, \bar{\sigma}^2)$ sub-Gamma, with*

$$\bar{b} = \max_{i \in \{1,\ldots,n\}} b_i, \qquad \bar{\sigma}^2 = \sum_{i=1}^n \sigma_i^2 \ .$$

Proof. Just write that, by independence, for any s for which it makes sense (in particular, for any $|s| < 1/\bar{b}$ therefore)

$$\mathbb{E}[\exp(s \sum_{i=1}^n X_i)] = \prod_{i=1}^n \mathbb{E}[\exp(sX_i)] \ .$$

□

3.7 Link with sub-exponential random variables

The first result provides various equivalent characterization of sub-exponential random variables.

Theorem 21. *Let X be a random variable. The following statement are equivalent.*

(i) For any $x > 0$, $\mathbb{P}(|X| > x) \leqslant 2\exp(-x/K_1)$.

(ii) *For any* $p \geqslant 1$, $(\mathbb{E}[|X|^p])^{1/p} \leqslant K_2 p$.

(iii) *For any* $|s| < 1/K_3$, $\mathbb{E}[\exp(s|X|)] \leqslant \exp(sK_3)$.

(iv) $\mathbb{E}[\exp(|X|/K_4)] \leqslant 2$.

If one of these properties holds, all of them do and the different K_i *differ by at most a multiplicative constant.*

Besides, the smallest K_4 *such that (iv) holds is called the sub-exponential norm of* X *and is denoted by* $\|X\|_{\psi_1}$.

Proof. **(i)** $\implies$ **(ii):** Assume that $K_1 = 1$ and write

$$\begin{aligned}\mathbb{E}[|X|^p] &= \int_0^{+\infty} \mathbb{P}(|X| > u^{1/p})\mathrm{d}u \\ &= p\int_0^{+\infty} \mathbb{P}(|X| > v)v^{p-1}\mathrm{d}v \\ &\leqslant 2p\int_0^{+\infty} \exp(-v)v^{p-1}\mathrm{d}v = 2p\Gamma(p) \ .\end{aligned}$$

By Stirling's estimate $\Gamma(p) \leqslant (p-1)^{p-1}$ this yields $\mathbb{E}[|X|^p] \leqslant 2p^p$. The result for general K_1 follows by applying the case $K_1 = 1$ to X/K_1.

(ii) $\implies$ **(iii):** Assume that $K_2 = 1$ and write

$$\mathbb{E}[\exp(s|X|)] = \sum_{k\geqslant 0} \frac{s^k\mathbb{E}[|X|^k]}{k!} \leqslant \sum_{k\geqslant 0} \frac{s^k k^k}{k!} \ .$$

By Stirling's estimate $k! \geqslant (k/e)^k$, we get that the series is convergent if $|s| < 1/2e$ and

$$\mathbb{E}[\exp(s|X|)] \leqslant \frac{1}{1-se} = 1 + \frac{se}{1-se} \leqslant 1 + 2se \leqslant \exp(2es) \ .$$

This proves the result for $K_2 = 1$. The result for general K_2 follows by applying the result for $K_2 = 1$ to X/K_2.

(iii) $\implies$ **(iv)** is straightforward.

(iv) $\implies$ **(i):** By Markov's inequality

$$\begin{aligned}\mathbb{P}(|X| > x) &= \mathbb{P}(\exp(|X|/K_4) > \exp(x/K_4)) \\ &\leqslant \mathbb{E}[\exp(|X|/K_4)]\exp(-x/K_4) \\ &\leqslant 2\exp(-x/K_4) \ .\end{aligned}$$

□

It is clear from (ii) for example that any sub-Gaussian random variable is sub-exponential and $\|X\|_{\psi_1} \leqslant C\|X\|_{\psi_2}$. Besides, point (iv) allows to show the important fact that, if X is sub-Gaussian, X^2 is sub-exponential and $\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2$. Finally, the basic remark $2ab = \inf_{\epsilon\in(0,1)} \epsilon a^2 + \epsilon^{-1}b^2$ allows to show the extension of this important fact that, for any X, Y, $\|XY\|_{\psi_1} \leqslant \|X\|_{\psi_2}\|Y\|_{\psi_2}$

The link with sub-Gamma random variables is given in the following result.

Theorem 22. *If X is centered with $\mathbb{E}[X^2] = \sigma^2$ and $\|X\|_{\psi_1} \leqslant K$, then*

(i) X is (CK, CK^2) sub-Gamma.

(ii) X is $(CK\log(K/\sigma), C\sigma^2)$ sub-Gamma.

Remark 23. *The same proof shows that, if X is centered with $\mathbb{E}[X^2] = \sigma^2$ and $\|X\|_{\psi_2} \leqslant K$, then X is $(CK\sqrt{\log(K/\sigma)}, C\sigma^2)$ sub-Gamma. These results complement Proposition 18 for possibly unbounded sub-Gamma (or sub-Gaussian) random variables.*

Proof. For the first part of the proof, we have by Characterization (ii) of sub-exponential random variables, for all $k \geqslant 2$,

$$\mathbb{E}[|X|^k] \leqslant (CKk)^k \leqslant k!(CeK)^k \ ,$$

where the second inequality follows by Stirling's estimate $k! \geqslant (k/e)^k$. The conclusion then follows from Proposition 17.

The second part uses the same ingredient but it's slightly more tricky. Let $k \geqslant 2$ be an integer and $x \geqslant 1$. We have, by Hölder's inequality,

$$\mathbb{E}[|X|^k] = \mathbb{E}[|X|^{2-1/x}|X|^{k-2+1/x}] \leqslant \sigma^{2-1/x}\mathbb{E}[|X|^{2x(k-2)+2}]^{1/2x} \ .$$

Now by Characterisation (ii) of sub-exponential random variables, we have

$$\mathbb{E}[|X|^{2x(k-2)+2}] \leqslant (CK(2x(k-2)+2))^{2x(k-2)+2}$$

Therefore,

$$\begin{aligned}\mathbb{E}[|X|^k] &\leqslant (CK)^{k-2+1/x}\sigma^{2-1/x}(2x(k-2)+2)^{k-2+1/x} \\ &\leqslant C\sigma^2(CK)^{k-2}(k-1)!\left(\frac{K}{\sigma}\right)^{1/x} x^{k-2}\end{aligned}$$

As this is true for any $x \geqslant 1$, one can choose $x = \log(K/\sigma)$ and the result follows by Proposition 17. □

We conclude this section with the following straightforward but useful corollary.

Proposition 24. *If X, X' are sub-Gaussian, then the random variable $XX' - \mathbb{E}[XX']$ is $(C\|X\|_{\psi_2}\|X'\|_{\psi_2}, C\|X\|_{\psi_2}^2\|X'\|_{\psi_2}^2)$-sub-Gamma.*

3.8 Bernstein's inequality

Bernstein's inequality is a concentration result for sub-Gamma random variables.

Theorem 25 (Bernstein's inequality)**.** *Assume that X is (b, σ^2) sub-Gamma. Then, for any $x > 0$,*

$$\mathbb{P}(|X| > x) \leqslant 2 \exp\left(- \min\left(\frac{x^2}{2\sigma^2}, \frac{x}{b} \right) \right) \ .$$

Proof. We fix $x > 0$ and use Chernoff's bound. We have

$$\mathbb{P}(X > x) \leqslant \exp\left(- \sup_{s>0} sx - \log \mathbb{E}[\exp(sX)] \right) \ .$$

For any $s < 1/b$, we have

$$\log \mathbb{E}[\exp(sX)] \leqslant s^2\sigma^2 \ ,$$

so

$$\sup_{s>0} sx - \log \mathbb{E}[\exp(sX)] \leqslant \sup_{s \in [0,1/b]} sx - s^2\sigma^2 = \begin{cases} x/b & \text{if } x > 2\sigma^2/b \ , \\ x^2/2\sigma^2 & \text{if } x \leqslant 2\sigma^2/b \end{cases}$$

This proves that, if X is (b, σ^2) sub-Gamma. Then, for any $x > 0$,

$$\mathbb{P}(X > x) \leqslant \exp\left(- \min\left(\frac{x^2}{2\sigma^2}, \frac{x}{b} \right) \right) \ .$$

By definition, if X is (b, σ^2) sub-Gamma, $-X$ also so the result follows by applying the previous bound to $-X$ and conclude with a union bound. □

Together with Proposition 20, Bernstein's inequality yields the following useful corollary.

Theorem 26. *Let $X_1, \dots, X_n$ denote independent random variables such that each X_i is (b_i, σ_i^2) sub-Gamma. Then*

$$\forall x > 0, \qquad \mathbb{P}(|\sum_{i=1}^n X_i| > x) \leqslant 2 \exp\left(-\min\left(\frac{x^2}{2\sum_{i=1}^n \sigma_i^2}, \frac{x}{\max_{i \in \{1,\dots,n\}} b_i} \right) \right) \ .$$

Theorem 26 together with the characterization of sub-Gamma random variables stated in Proposition 17 boils down to the proof of the classical Bernstein's inequality.

Theorem 27 (Bernstein's inequality)**.** *If $X_1, \dots, X_n$ are independent random variables such that, for any $i \in \{1, \dots, n\}$ and $k \geqslant 2$,*

$$\mathbb{E}[|X_i|^k] \leqslant k! \sigma_i^2 K_i^{k-2} \ ,$$

then, for any $z > 0$,

$$\mathbb{P}\bigg(\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) > z \bigg) \leqslant \exp\bigg(- c \min\bigg(\frac{x^2}{\sigma^2}, \frac{x}{K} \bigg) \bigg) \ ,$$

with $\sigma^2 = \sum_{i=1}^n \sigma_i^2$, $K = \max_{i \in \{1,\dots,n\}} K_i$.

A corollary of this result is obtained by combining Theorem 26 with Theorem 22. It yields the following corollary.

Corollary 28. *Assume that $X_1, \dots, X_n$ are independent and sub-Exponential, then, for any $z > 0$,*

$$\mathbb{P}\bigg(\bigg| \sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \bigg| > z \bigg) \leqslant 2 \exp\bigg(- c \min\bigg(\frac{t^2}{\sum_{i=1}^n \|X_i\|_{\psi_1}^2}, \frac{t}{\max_{i \in \{1,\dots,n\}} \|X_i\|_{\psi_1}} \bigg) \bigg) \ .$$

A less classical consequence can be obtained by putting together Theorem 25 and Theorem 22. Indeed, if $X_1, \dots, X_n$ are centered independent random variables such that $\|X_i\|_{\psi_1} \leqslant K_i$ and $\mathbb{E}[X_i^2] = \sigma_i^2$, then

$$\forall x > 0, \qquad \mathbb{P}\bigg(\bigg| \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \bigg| > x \bigg) \leqslant 2 \exp\bigg(- c \min\bigg(\frac{x^2}{\sigma^2}, \frac{\sqrt{n} x}{\bar{K}} \bigg) \bigg) \ ,$$

where

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n \sigma_i^2, \qquad \bar{K} = \max_{i=1,\dots,n} K_i \log(K_i/\sigma_i) \ .$$

This completes the first part of the program we planned in this chapter by showing that sums of sub-exponential random variables deviates as predicted by the central limit theorem for any

$$x \leqslant \frac{\sqrt{n} \sigma^2}{\bar{K}} \ .$$

We conclude with a simple corollary that we use repeatedly in the following.

Corollary 29. *Let $(X_1, X_1'), \dots, (X_n, X_n')$ denote independent couples of sub-Gaussian random variables, let*

$$K^2 = \sum_{i=1}^n \|X_i\|_{\psi_2}^2 \|X_i'\|_{\psi_2}^2, \qquad b = \max_{i \in \{1,\dots,n\}} \|X_i\|_{\psi_2} \|X_i'\|_{\psi_2} \ .$$

Then, for any $z > 0$, we have

$$\mathbb{P}\bigg(\sum_{i=1}^n X_i X_i' - \mathbb{E}[X_i X_i'] > z \bigg) \leqslant \exp\bigg(- C \min\bigg(\frac{z^2}{K^2}, \frac{z}{b} \bigg) \bigg) \ .$$

3.9 Problem

The purpose of this problem is to prove an extension of the bounded difference inequality for suprema of sums of sub-Gaussian random variables

$$\sup_{t\in T}\sum_{i=1}^{n} X_{i,t} \ ,$$

where $(X_{i,t})_{t\in T}$ are independent random variables such that

$$K = \| \sup_{i\in\{1,\dots,n\}} \sup_{t\in T} X_{i,t}\|_{\psi_2} < \infty \ .$$

Before we move to this application, we consider as in the bounded difference inequality a function

$$\gamma : \begin{cases} \mathcal{X}_1 \times \ldots \times \mathcal{X}_n \to \mathbb{R} \\ (x_1,\dots,x_n) \mapsto \gamma(x_1,\dots,x_n) \ . \end{cases}$$

We denote by $\mathcal{D}_n = \{X_1,\dots,X_n\}$ a set of independent random variables taking values respectively in $\mathcal{X}_1,\dots,\mathcal{X}_n$. We let $\mathcal{F}_i$ denote the sigma-algebra generated by $X_1,\dots,X_i$ and denote by

$$K_i \geqslant \| \sup_{x_1,\dots,x_n} \gamma(x_1,\dots,x_{i-1},X_i,x_{i+1},\dots,x_n) - \mathbb{E}[\gamma(x_1,\dots,x_{i-1},X_i,x_{i+1},\dots,x_n)]\|_{\psi_2} \ .$$

1. Prove that

$$\gamma(X_1,\dots,X_n) - \mathbb{E}[\gamma(X_1,\dots,X_n)] = \sum_{i=1}^{n} \Delta_i \ ,$$

where $\Delta_i = \mathbb{E}[\gamma(X_1,\dots,X_n) - \mathbb{E}[\gamma(X_1,\dots,X_n)|\mathcal{D}_n \setminus X_i]|\mathcal{F}_i]$

2. Prove that, for any $s \in \mathbb{R}$,

$$\mathbb{E}\left[\left(s\{\gamma(X_1,\dots,X_n) - \mathbb{E}[\gamma(X_1,\dots,X_n)|\mathcal{D}_n\setminus X_i]\}|\mathcal{D}_n\setminus X_i\right)\right] \leqslant \exp\left(C^2 s^2 K_i^2\right) \ .$$

3. Prove that, for any $z > 0$,

$$\mathbb{P}\big(\gamma(X_1,\dots,X_n) - \mathbb{E}[\gamma(X_1,\dots,X_n)] > z\big) \leqslant \exp\left(-\frac{C^2 z^2}{\sum_{i=1}^n K_i^2}\right) \ .$$

4. Let T denote a finite space, let $X_1,\dots,X_n$ denote independent random variables such that

$$X_i = (X_{i,t})_{t\in T}, \qquad \mathbb{E}[X_{i,t} = 0], \qquad \| \sup_{t\in T} X_{i,t}\|_{\psi_2} \leqslant K_i < \infty \ .$$

Prove that, for any $z > 0$,

$$\mathbb{P}\big(\sup_{t\in T} X_{i,t} - \mathbb{E}[\sup_{t\in T} X_{i,t}] > z\big) \leqslant \exp\left(-\frac{C^2 z^2}{\sum_{i=1}^n K_i^2}\right) \ .$$

Chapter 4

Deviation inequalities for random matrices

This chapter extends the tools we saw in the previous lecture to random matrices. We start with basic notions of matrix calculus that will be useful for these extensions. Then, we prove matrix Hoeffding's inequality, matrix Bernstein's inequality and Hanson-Wright's inequality.

4.1 Calculus on matrices

For any symmetric $n \times n$ matrix X, we write its eigenvalues decomposition

$$X = \sum_{i=1}^{n} \lambda_i u_i u_i^T \ ,$$

where the spectrum is ordered so that $\lambda_1 \geqslant \ldots \geqslant \lambda_n$ and $u_1, \ldots, u_n$ an orthonormal basis of eigenvectors of X. We also denote by $\|X\|$ the operator norm of any matrix, that is its largest singular value (the sup-norm of its spectrum if X is symmetric).

Definition 30. *Let $X = \sum_{i=1}^{n} \lambda_i u_i u_i^T$, Y denote two $n \times n$ symmetric matrices and let $f : \mathbb{R} \to \mathbb{R}$ denote a function. We define*

1. *$X \succcurlyeq Y$ if $X - Y$ is positive semi-definite.*
2. *$f(X) = \sum_{i=1}^{n} f(\lambda_i) u_i u_i^T$.*

Remark 31. *The definition of $f(X)$ extends the one when f is a polynomial.*

As an exercise, check the following properties that will be useful in the remaining of this chapter.

Proposition 32. *Let X, X' and Y, Y' denote two $n \times n$ symmetric matrices and let f and g denote two $\mathbb{R} \to \mathbb{R}$ functions.*

1. *If* $X \succcurlyeq Y$, $Tr(X) \geqslant Tr(Y)$.
2. *If* $X \preccurlyeq X'$ *and* $Y \preccurlyeq Y'$, *then* $X + Y \preccurlyeq X' + Y'$.
3. $\|X\| \leqslant t$ *is equivalent to* $-t\boldsymbol{I} \preccurlyeq X \preccurlyeq t\boldsymbol{I}$.
4. *If* $f(x) \leqslant g(x)$ *for any* $|x| \leqslant K$ *and* $\|X\| \leqslant K$, *then* $f(X) \preccurlyeq g(X)$.
5. *If* $X \succcurlyeq Y$, $XY = YX$ *and* f *is non decreasing,* $f(X) \succcurlyeq f(Y)$.
6. *If* $X \succcurlyeq Y$ *and* f *is non decreasing,* $Tr(f(X)) \geqslant Tr(f(Y))$.
7. *If* $0 \preccurlyeq X \preccurlyeq Y$, $\log(X) \preccurlyeq \log(Y)$.

4.2 Deviation bounds for sums of independent random matrices

In this section, we establish the extension of Hoeffding and Bernstein's inequalities for sums of independent random matrices. The main difficulty in this extension is that the set of matrices is not commutative so the basic inequality

$$\exp(X + Y) = \exp(X)\exp(Y) \ ,$$

does not hold anymore. As this inequality yields the tensorization property

$$\mathbb{E}\Big[\exp\Big(s\sum_{i=1}^{N} X_i\Big)\Big] = \exp\Big(\sum_{i=1}^{N} \log \mathbb{E}[\exp(sX_i)]\Big) \ ,$$

we need to work on the extension of this argument. The section therefore starts with an extension of this argument that is then applied to prove deviation inequalities for random matrices.

4.2.1 Extension of the tensorization argument

In this section, we establish the following result. Let $\lambda_1(A)$ denote the largest eigenvalue of A.

Lemma 33 (Tensorization for random matrices)**.** *Let* $X_1, \ldots, X_N$ *denote independent* $n \times n$ *symmetric random matrices, let* $s > 0$

$$\mathbb{E}\Big[\exp\Big(s\lambda_1\Big(\sum_{i=1}^{N} X_i\Big)\Big)\Big] \leqslant Tr\Big(\exp\Big(\sum_{i=1}^{N} \log \mathbb{E}\big[\exp\big(sX_i\big)\big]\Big)\Big) \ .$$

Proof. The proof is divided into two lemmas. The first one is completely elementary.

Lemma 34. *Let X denote a symmetric $n \times n$ matrix, we have, for any $s > 0$,*

$$\exp\big(s\lambda_1(X)\big) \leqslant Tr(\exp(sX)) \ .$$

Proof of Lemma 34. We first use that $x \mapsto \exp(sx)$ is non decreasing to say that

$$\exp\big(s\lambda_1(X)\big) = \lambda_1\big(\exp\big(sX\big)\big) \ .$$

By Point 4.,

$$\exp\big(sX\big) \succcurlyeq 0 \ .$$

Therefore,

$$\lambda_1\big(\exp\big(sX\big)\big) \leqslant \mathrm{Tr}\,\big(\exp\big(sX\big)\big) \ .$$

□

By Lemma 34, we have thus

$$\exp\left(s\lambda_1\Big(\sum_{i=1}^N X_i\Big)\right) \leqslant \mathrm{Tr}\left(\exp\left(s\sum_{i=1}^N X_i\right)\right) \ . \tag{4.1}$$

The next step of the proof is the following result known as Lieb's inequality.

Theorem 35 (Lieb's inequality for random matrices)**.** *For any $n \times n$ symmetric matrices H deterministic and X random, we have*

$$\mathbb{E}[Tr(\exp(H+X))] \leqslant Tr(\exp(H + \log(\mathbb{E}[\exp(X)]))) \ .$$

We do not prove Lieb's inequality here, an article providing a proof is given in Slack. A consequence of this result is the following lemma.

Lemma 36. *For any independent $n \times n$ symmetric random matrices $X_1, \ldots, X_N$, we have*

$$\mathbb{E}\bigg[\ Tr\left(\exp\left(s\sum_{i=1}^N X_i\right)\right)\bigg] \leqslant Tr\left(\exp\left(\sum_{i=1}^N \log\big(\mathbb{E}[\exp(sX_i)]\big)\right)\right) \ .$$

Proof of Lemma 36. We use recursively Lieb's inequality conditionally on $\mathcal{F}_i = Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_N$, with $Z_j = sX_j$ if $j \leqslant i$ and $Z_j = \log \mathbb{E}[\exp(sX_j)]$ if $j > i$ to say that

$$\begin{aligned}
&\mathbb{E}\bigg[\ \mathrm{Tr}\left(\exp\Big(\sum_{j=1, j\neq i}^N Z_j + Z_i\Big)\right)|\mathcal{F}_i\bigg] \\
&\qquad\qquad \leqslant \mathrm{Tr}\left(\exp\Big(\sum_{j=1, j\neq i}^N Z_j + \log \mathbb{E}[\exp(Z_i)|\mathcal{F}_i]\Big)\right) \\
&\qquad\qquad = \mathrm{Tr}\left(\exp\Big(\sum_{j=1, j\neq i}^N Z_j + \log \mathbb{E}[\exp(Z_i)]\Big)\right) \ ,
\end{aligned}$$

where the last inequality holds by independence of the X_i. □

Together with (4.1), Lemma 36 proves that

$$\mathbb{E}\Big[\exp\Big(s\lambda_1\Big(\sum_{i=1}^N X_i\Big)\Big)\Big] \leqslant \mathrm{Tr}\,\Big(\exp\Big(\sum_{i=1}^N \log\mathbb{E}\big[\exp\big(sX_i\big)\big]\Big)\Big) \ .$$

□

4.2.2 Matrix deviation inequalities

In this section, we extend Hoeffding and Bernstein's inequality to sums of independent random matrices.

Theorem 37 (Matrix Hoeffding's inequality)**.** *Let $A_1, \dots, A_N$ denote $n \times n$ symmetric deterministic matrices and let $\epsilon_1, \dots, \epsilon_N$ denote independent Rademacher random variables. Then, for any $z > 0$,*

$$\mathbb{P}\Big(\Big\|\sum_{i=1}^N \epsilon_i A_i\Big\| > z\Big) \leqslant 2n\exp\Big(-\frac{cz^2}{\|\sum_{i=1}^N A_i^2\|}\Big) \ .$$

Proof. First, we have that $\|A\| = \max(\lambda_1(A), -\lambda_n(A))$, where $\lambda_n(A)$ is the smallest eigenvalue of A. We focus on $\lambda_1(A)$ and let the reader check that the argument generalizes to $\|A\|$ up to a union bound. The proof uses Chernoff's bound

$$\forall s > 0, \qquad \mathbb{P}\Big(\lambda_1\Big(\sum_{i=1}^N \epsilon_i A_i\Big) > z\Big) \leqslant \exp(-sz)\mathbb{E}\Big[\exp\Big(s\lambda_1\Big(\sum_{i=1}^N \epsilon_i A_i\Big)\Big)\Big] \ . \tag{4.2}$$

Then, by tensorization for random matrices

$$\mathbb{E}\Big[\exp\Big(s\lambda_1\Big(\sum_{i=1}^N \epsilon_i A_i\Big)\Big)\Big] \leqslant \mathrm{Tr}\,\Big(\exp\Big(\sum_{i=1}^N \log\mathbb{E}\big[\exp\big(s\epsilon_i A_i\big)\big]\Big)\Big) \ . \tag{4.3}$$

Now, to bound the Laplace transform of a single random matrix X, we use Point 4 in Proposition 32. First, we use that, for any real number x,

$$\mathbb{E}[\exp(\epsilon x)] = \frac{1}{2}(\exp(x) + \exp(-x)) = \sum_{k\geqslant 0}\frac{x^{2k}}{(2k)!} \leqslant \sum_{k\geqslant 0}\frac{x^{2k}}{2^k k!} = \exp\Big(\frac{x^2}{2}\Big) \ .$$

We deduce from this bound and Points 4 and 7 in Proposition 32 that

$$\log\mathbb{E}\big[\exp\big(s\epsilon_i A_i\big)\big] \preccurlyeq \frac{s^2}{2}A_i^2 \ .$$

By Point 2 in Proposition 32, we get

$$\sum_{i=1}^N \log\mathbb{E}\big[\exp\big(s\epsilon_i A_i\big)\big] \leqslant \frac{s^2}{2}\sum_{i=1}^N A_i^2 \ .$$

By Point 6 in Proposition 32, we get

$$\mathrm{Tr}\left(\exp\left(\sum_{i=1}^{N}\log\mathbb{E}\big[\exp\left(s\epsilon_i A_i\right)\big]\right)\right) \leqslant \mathrm{Tr}\left(\exp\left(\frac{s^2}{2}\sum_{i=1}^{N}A_i^2\right)\right) \ .$$

Plugging this into (4.3), we get

$$\begin{aligned}\mathbb{E}\Big[\exp\Big(s\lambda_1\Big(\sum_{i=1}^{N}\epsilon_i A_i\Big)\Big)\Big] &\leqslant \mathrm{Tr}\left(\exp\left(\frac{s^2}{2}\sum_{i=1}^{N}A_i^2\right)\right)\\ &\leqslant n\exp\left(\frac{s^2}{2}\lambda_1\Big(\sum_{i=1}^{N}A_i^2\Big)\right) \ .\end{aligned}$$

Therefore,

$$\mathbb{E}\Big[\exp\Big(s\lambda_1\Big(\sum_{i=1}^{N}\epsilon_i A_i\Big)\Big)\Big] \leqslant n\exp\left(\frac{s^2}{2}\Big\|\sum_{i=1}^{N}A_i^2\Big\|\right) \ . \tag{4.4}$$

Plugging this into (4.2), we finally get

$$\forall s>0, \qquad \mathbb{P}\Big(\lambda_1\Big(\sum_{i=1}^{N}\epsilon_i A_i\Big)>z\Big) \leqslant n\exp\left(-sz+\frac{s^2}{2}\Big\|\sum_{i=1}^{N}A_i^2\Big\|\right) \ .$$

Optimizing over $s>0$ proves the concentration of $\lambda_1\big(\sum_{i=1}^{N}\epsilon_i A_i\big)$. The same argument would show the concentration of $-\lambda_n\big(\sum_{i=1}^{N}\epsilon_i A_i\big)$ and a union bound concludes the proof of the theorem. □

Let us now turn to Bernstein's inequality.

Theorem 38 (Matrix Bernstein's inequality)**.** *Let $X_1,\dots,X_N$ denote independent, centered, $n\times n$, symmetric random matrices such that $\|X_i\|\leqslant K$ a.s.. Then, for any $z>0$,*

$$\mathbb{P}\Big(\Big\|\sum_{i=1}^{N}X_i\Big\|>z\Big) \leqslant 2n\exp\left(-c\min\left(\frac{t^2}{\sigma^2},\frac{t}{K}\right)\right) \ ,$$

with $\sigma^2=\big\|\sum_{i=1}^{N}\mathbb{E}[X_i^2]\big\|$.

Proof. We focus as in the previous proof on the concentration of $\lambda_1\big(\sum_{i=1}^{N}X_i\big)$. The proof starts with Chernoff's bound

$$\forall s>0, \qquad \mathbb{P}\Big(\lambda_1\Big(\sum_{i=1}^{N}X_i\Big)>z\Big) \leqslant \exp(-sz)\mathbb{E}\Big[\exp\Big(s\lambda_1\Big(\sum_{i=1}^{N}X_i\Big)\Big)\Big] \ .$$

Then, by tensorization's bound for random matrices

$$\mathbb{E}\Big[\exp\Big(s\lambda_1\Big(\sum_{i=1}^N X_i\Big)\Big)\Big] \leqslant \mathrm{Tr}\,\Big(\exp\Big(\sum_{i=1}^N \log \mathbb{E}\big[\exp\big(sX_i\big)\big]\Big)\Big) \ .$$

Then, we use that, for any $|x| \leqslant K$, and $|s| \leqslant 1/K$,

$$\begin{aligned}\exp(sx) &= 1 + sx + \sum_{k\geqslant 2} \frac{(sx)^k}{k!} \\ &\leqslant 1 + sx + \frac{s^2x^2}{2(1-|sK|/3)} \leqslant 1 + sx + s^2x^2 \ .\end{aligned}$$

By Point 4 in Proposition 32, this implies

$$\exp(sX_i) \preccurlyeq \mathbf{I} + sX_i + s^2X_i^2 \ .$$

Taking expectation on both sides, and using one more time Point 4 in Proposition 32 with the inequality $1 + x \leqslant \exp(x)$ valid for any $x \in \mathbb{R}$, yields

$$\mathbb{E}[\exp(sX_i)] \preccurlyeq \mathbf{I} + s^2\mathbb{E}[X_i^2] \preccurlyeq \exp\big(s^2\mathbb{E}[X_i^2]\big) \ .$$

By Point 7 in Proposition 32, this implies

$$\log(\mathbb{E}[\exp(sX_i)]) \preccurlyeq s^2\mathbb{E}[X_i^2] \ .$$

By Point 2 in Proposition 32, we deduce

$$\sum_{i=1}^N \log(\mathbb{E}[\exp(sX_i)]) \preccurlyeq s^2 \sum_{i=1}^N \mathbb{E}[X_i^2] \ .$$

By Point 6 in Proposition 32, we get

$$\begin{aligned}\mathrm{Tr}\,\Big(\exp\Big(\sum_{i=1}^N \log\big(\mathbb{E}[\exp(sX_i)]\big)\Big)\Big) &\leqslant \mathrm{Tr}\,\Big(\exp\Big(s^2\sum_{i=1}^N \mathbb{E}[X_i^2]\Big)\Big) \\ &\leqslant n\exp\big(s^2\sigma^2\big) \ .\end{aligned}$$

We conclude that, for any $|s| \leqslant K$,

$$\mathbb{P}\Big(\lambda_1\Big(\sum_{i=1}^N X_i\Big) > z\Big) \leqslant n\exp\big(-sz + s^2\sigma^2\big) \ .$$

Optimizing over $|s| \leqslant K$ shows the deviation inequality for $\lambda_1\big(\sum_{i=1}^N X_i\big)$. The same argument shows the deviation of $-\lambda_n\big(\sum_{i=1}^N X_i\big)$ (check this part!) and the conclusion follows from a union bound. □

4.3 Applications of Matrix Hoeffding's inequality

4.3.1 Matrix Khintchine's inequality

A consequence of Matrix Hoeffding's inequality is the matrix version of Khintchine's inequality.

Theorem 39 (Matrix Khintchine's inequality)**.** *Let $A_1, \dots, A_N$ denote $n \times n$ symmetric deterministic matrices and let $\epsilon_1, \dots, \epsilon_N$ denote independent Rademacher random variables. Then*

$$\mathbb{E}\left[\left\| \sum_{i=1}^N \epsilon_i A_i \right\|\right] \leqslant C\sqrt{1+\log n}\left\| \sum_{i=1}^N A_i^2 \right\|^{1/2} .$$

Proof. The proof follows by integration of Hoeffding's inequality and is left as an exercise. □

Khintchine's inequality admits the following corollary for non necessarily symmetric matrices.

Corollary 40 (General Matrix Khintchine's inequality)**.** *Let $A_1, \dots, A_N$ denote $n \times p$ deterministic matrices and let $\epsilon_1, \dots, \epsilon_N$ denote independent Rademacher random variables. Then,*

$$\mathbb{E}\left[\left\| \sum_{i=1}^N \epsilon_i A_i \right\|\right] \lesssim \sqrt{\log(p+n)\max\left(\sup_{x_1 \in \mathbb{B}_p} \sum_{i=1}^N \|A_i^T x_1\|_2^2, \sup_{x_2 \in \mathbb{B}_n} \sum_{i=1}^N \|A_i x_2\|_2^2 \right)} .$$

Proof. The proof relies on the following trick. For any $n \times p$ matrix A, we denote by

$$S(A) = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} .$$

Then, S is a linear map such that, for any $n \times p$ matrix A, $S(A)$ is a symmetric $(n+p) \times (n+p)$ matrix such that

$$S(A)^2 = \begin{bmatrix} AA^T & 0 \\ 0 & A^T A \end{bmatrix}, \qquad \|S(A)\| = \|A\| . \tag{4.5}$$

We have therefore,

$$\begin{aligned}\mathbb{E}\Big[\Big\|\sum_{i=1}^N \epsilon_i A_i\Big\|\Big] &= \mathbb{E}\Big[\Big\|S\Big(\sum_{i=1}^N \epsilon_i A_i\Big)\Big\|\Big] \quad \text{by (4.5)} \\ &= \mathbb{E}\Big[\Big\|\sum_{i=1}^N \epsilon_i S\big(A_i\big)\Big\|\Big] \\ &\leqslant C\sqrt{\log(p+n)}\Big\|\sum_{i=1}^N S(A_i)^2\Big\|^{1/2} \quad \text{Khintchine .}\end{aligned}$$

Finally, for any $x = (x_1^T, x_2^T)^T$ in $\mathbb{R}^n \times \mathbb{R}^p \setminus \{0\}$,

$$\begin{aligned}x^T \sum_{i=1}^N S(A_i)^2 x &= \sum_{i=1}^N \|A_i^T x_1\|_2^2 + \sum_{i=1}^N \|A_i x_2\|_2^2 \\ &= \|x\|_2^2\Big(\lambda \sum_{i=1}^N \Big\|A_i^T \frac{x_1}{\|x_1\|_2}\Big\|_2^2 + (1-\lambda)\sum_{i=1}^N \Big\|A_i \frac{x_2}{\|x_2\|_2}\Big\|_2^2\Big) \ ,\end{aligned}$$

with $\lambda = \|x_1\|_2^2/\|x\|_2^2$, thus $1-\lambda = \|x_2\|_2^2/\|x\|_2^2$, using the convention $0/0 = 0$. □

4.3.2 Application to Matrix completion

A classical application of Matrix Khintchine's inequality is the problem of Matrix completion. Suppose we want to recover a $n \times p$ matrix X based the observation of the matrix Y with entries $y_{i,j} = \delta_{i,j} x_{i,j}$, where $\delta_{i,j}$ are i.i.d. Bernoulli $B(q)$ random variables, where $0 < q \leqslant 1/2$. The key assumption for this task to be feasible is to assume that X has low rank r. The main idea behind the algorithm for matrix completion is to approximate

$$\hat{X} \in \operatorname{argmin}_{r(Z)\leqslant r} \|q^{-1}Y - Z\| \ .$$

$\hat{X}$ is not an estimator unless we know r and the optimization problem is computationally hard due to the constraints $r(Z) \leqslant r$ so convex relaxations are necessary to actually approximate $\hat{X}$. Yet, to understand the relevance of this strategy, we bound here the performance of $\hat{X}$.

The key remark is that, by construction

$$\|\hat{X} - X\| \leqslant \|\hat{X} - q^{-1}Y\| + \|q^{-1}Y - X\| \leqslant \frac{2}{q}\|Y - qX\| \ .$$

Moreover,

$$Y - qX = ((\delta_{i,j} - q)x_{i,j})_{1\leqslant i\leqslant n, 1\leqslant j\leqslant p} \ .$$

As $Y - qX$ is centered, we can apply the following symmetrization lemma.

Lemma 41 (Symmetrization)**.** *Let $(X_{1,t})_{t\in T},\dots,(X_{n,t})_{t\in T}$ denote independent processes indexed by a separable space T such that $\mathbb{E}[X_{i,t}]=0$ for any i and t. Let $\epsilon_1,\dots,\epsilon_n$ denote independent Rademacher random variables, then*

$$\mathbb{E}\left[\sup_{t\in T}\sum_{i=1}^n X_{i,t}\right]\leqslant 2\mathbb{E}\left[\sup_{t\in T}\sum_{i=1}^n \epsilon_i X_{i,t}\right] .$$

Proof. We introduce $(X'_{1,t})_{t\in T},\dots,(X'_{n,t})_{t\in T}$ independent from the processes $\mathcal{D}=\{(X_{1,t})_{t\in T},\dots,(X_{n,t})_{t\in T}\}$ and with the same distribution, so

$$\sup_{t\in T}\sum_{i=1}^n X_{i,t}=\sup_{t\in T}\sum_{i=1}^n X_{i,t}-\mathbb{E}[X'_{i,t}|\mathcal{D}]\leqslant \mathbb{E}\left[\sup_{t\in T}\sum_{i=1}^n X_{i,t}-X'_{i,t}|\mathcal{D}\right] .$$

Taking expectation on both sides, we get

$$\mathbb{E}\left[\sup_{t\in T}\sum_{i=1}^n X_{i,t}\right]\leqslant \mathbb{E}\left[\sup_{t\in T}\sum_{i=1}^n X_{i,t}-X'_{i,t}\right] .$$

Now we use that the processes $(X_{1,t}-X'_{1,t})_{t\in T},\dots,(X_{n,t}-X'_{n,t})_{t\in T}$ and $(\epsilon_1(X_{1,t}-X'_{1,t}))_{t\in T},\dots,(\epsilon_n(X_{n,t}-X'_{n,t}))_{t\in T}$ have the same distribution to say that

$$\mathbb{E}\left[\sup_{t\in T}\sum_{i=1}^n X_{i,t}-X'_{i,t}\right]=\mathbb{E}\left[\sup_{t\in T}\sum_{i=1}^n \epsilon_i(X_{i,t}-X'_{i,t})\right] .$$

We conclude saying that

$$\mathbb{E}\left[\sup_{t\in T}\sum_{i=1}^n \epsilon_i(X_{i,t}-X'_{i,t})\right]\leqslant \mathbb{E}\left[\sup_{t\in T}\sum_{i=1}^n \epsilon_i X_{i,t}\right]+\mathbb{E}\left[\sup_{t\in T}\sum_{i=1}^n (-\epsilon_i)X_{i,t}\right] ,$$

and the fact that ϵ_i and $-\epsilon_i$ have the same distributions. □

A consequence of the symmetrization lemma for random matrices is that, if $A_1,\dots,A_N$ are independent random matrices such that $\mathbb{E}[A_i]=0$, then

$$\mathbb{E}\left[\left\|\sum_{i=1}^N A_i\right\|\right]\leqslant 2\mathbb{E}\left[\left\|\sum_{i=1}^N \epsilon_i A_i\right\|\right] .$$

For matrix completion, this yields

$$\mathbb{E}[\|Y-qX\|]\leqslant 2\mathbb{E}\left[\left\|\sum_{i=1}^n\sum_{j=1}^p \epsilon_{i,j}(\delta_{i,j}-p)x_{i,j}E_{i,j}\right\|\right] ,$$

where $E_{i,j}=e_i\mathbf{e}_j^T$ is the canonical basis of $n\times p$ matrices.

We can now apply Khintchine's inequality conditionally on all $\delta_{i,j}$ and for this, let $u \in \mathbb{B}_p$, we have

$$E_{i,j}u = u_j e_i \ ,$$

thus

$$\sum_{i=1}^{n}\sum_{j=1}^{p} \|\epsilon_{i,j}(\delta_{i,j}-q)x_{i,j}E_{i,j}u\|_2^2 = \sum_{i=1}^{n}\sum_{j=1}^{p}(\delta_{i,j}-q)^2 x_{i,j}^2 u_j^2$$
$$\leqslant \max_{j\in\{1,\dots,p\}} \sum_{i=1}^{n}(\delta_{i,j}-q)^2 x_{i,j}^2 \ .$$

Now, we introduce $C_j(X)$, the j-th column of X and we have, by Bernstein's inequality

$$\mathbb{E}[\max_{j\in\{1,\dots,p\}} \sum_{i=1}^{n}(\delta_{i,j}-q)^2 x_{i,j}^2] \lesssim q \max_{j\in\{1,\dots,p\}} \|C_j(X)\|_2^2 + \log p \|X\|_\infty^2 \ .$$

Proceeding similarly for the lines, we get finally

$$\mathbb{E}[\|\hat{X}-X\|]$$
$$\lesssim \sqrt{\log(n+p)}\max\left(\frac{\max_i \|L_i(X)\|_2}{\sqrt{q}}, \frac{\max_j \|C_j(X)\|_2}{\sqrt{q}}, \frac{\sqrt{\log(n+p)}\|X\|_\infty}{q}\right) \ .$$

Finally, using the fact that both X and $\hat{X}$ have rank r, we have

$$\|\hat{X}-X\|_F^2 \leqslant 2r\|\hat{X}-X\|^2 \ ,$$

thus

$$\mathbb{E}[\|\hat{X}-X\|_F^2]$$
$$\lesssim r\log(n+p)\max\left(\frac{\max_i \|L_i(X)\|_2^2}{q}, \frac{\max_j \|C_j(X)\|_2^2}{q}, \frac{\log(n+p)\|X\|_\infty^2}{q^2}\right) \ .$$

Using the rough bound $\|L_i(X)\|_2^2 \leqslant p\|X\|_\infty^2$ and $\|C_j(X)\|_2^2 \leqslant n\|X\|_\infty$, we deduce that, if $npq \geqslant r(n+p)\log(n+p)$, we have

$$\frac{1}{np}\mathbb{E}[\|\hat{X}-X\|_F^2] \leqslant C\frac{r(n+p)\log(n+p)}{npq}\|X\|_\infty^2 \ .$$

Informally, if the number of observations npq exceeds the number of parameters $r(n+p)$ by a logarithmic factor, matrix recovery is possible in the sense that the average error in the L^2-sense $\frac{1}{np}\mathbb{E}[\|\hat{X}-X\|_F^2]$ is smaller than the size of the original matrix $\|X\|_\infty^2$.

4.4 Applications of Matrix Bernstein's inequality

Theorem 42. *Let $\boldsymbol{M}$ denote a $n \times p$ random matrix with independent centered rows $\boldsymbol{M}_i^T$ such that $\|\boldsymbol{M}_i\|_2 \leqslant K$ a.s.. Then, for any $\epsilon \in (0,1)$ and $\eta \in (0,1)$, with probability larger than $1-\epsilon$,*

$$\|\boldsymbol{M}^T\boldsymbol{M} - \mathbb{E}[\boldsymbol{M}^T\boldsymbol{M}]\| \leqslant \eta\|\mathbb{E}[\boldsymbol{M}^T\boldsymbol{M}]\| + \frac{C}{\eta}K^2\log(p/\epsilon) \ .$$

Proof. We write $\mathbb{B}_2 = \{u \in \mathbb{R}^p : \|u\|_2 \leqslant 1\}$ and, for any $u \in \mathbb{B}_2$,

$$u^T\mathbf{M}^T\mathbf{M}u = \|(\langle \mathbf{M}_i, u\rangle)\|_2^2 = \sum_{i=1}^n \langle \mathbf{M}_i, u\rangle^2 = u^T(\sum_{i=1}^n \mathbf{M}_i\mathbf{M}_i^T)u \ .$$

This proves that $\mathbf{M}^T\mathbf{M} = \sum_{i=1}^n \mathbf{M}_i\mathbf{M}_i^T$ is a sum of independent symmetric random $p \times p$ matrices. Besides, these matrices satisfy, for any $u \in \mathbb{B}_2$, by Cauchy-Schwarz inequality,

$$u^T\mathbf{M}_i\mathbf{M}_i^T u = \langle \mathbf{M}_i, u\rangle^2 \leqslant \|\mathbf{M}_i\|_2^2 \leqslant K^2 \ .$$

The last inequality holds a.s. By the matrix Bernstein's inequality, it follows therefore that, for any $t > 0$,

$$\mathbb{P}\big(\big\|\mathbf{M}^T\mathbf{M} - \mathbb{E}[\mathbf{M}^T\mathbf{M}]\big\| > t\big) \leqslant 2p\exp\left(-c\min\left(\frac{t^2}{\sigma^2}, \frac{t}{K}\right)\right) \ ,$$

where

$$\begin{aligned}
\sigma^2 &= \left\| \sum_{i=1}^n \mathbb{E}[\mathbf{M}_i\mathbf{M}_i^T\mathbf{M}_i\mathbf{M}_i^T] \right\| \\
&= \left\| \sum_{i=1}^n \mathbb{E}[\|\mathbf{M}_i\|_2^2\mathbf{M}_i\mathbf{M}_i^T] \right\| \\
&\leqslant K^2 \left\| \sum_{i=1}^n \mathbb{E}[\mathbf{M}_i\mathbf{M}_i^T] \right\| \\
&= K^2\big\|\mathbb{E}[\mathbf{M}^T\mathbf{M}]\big\| \ .
\end{aligned}$$

□

For the second application, we go back to the community detection problem presented in Chapter 2. We presented the spectral clustering algorithm in this chapter to solve the problem and provided an error bound for the proportion of misclassified nodes under a control on the spectral norm of $\mathbf{A} - \mathbb{E}[\mathbf{A}]$, where $\mathbf{A}$ is the adjacency matrix of the observed random graph. Our purpose here is to provide a control on this spectral norm using matrix Bernstein's inequality.

Theorem 43. *Let $\boldsymbol{A}$ denote the adjacency matrix of the graph of a balanced 2-classes SBM with size $2n$ and parameters p and q. Then, for any $\epsilon \in (0,1)$, with probability at least $1-\epsilon$,*

$$\|\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]\| \leqslant C \log(n/\epsilon) \ .$$

Proof. We denote by $\mathbf{e}_{i,j}$ the canonical basis of the $(2n) \times (2n)$ matrices. We can write

$$\mathbf{A} = \sum_{i=1}^{2n} B_{i,i}\mathbf{e}_{i,i} + \sum_{1 \leqslant i < j \leqslant 2n} B_{i,j}(\mathbf{e}_{i,j} + \mathbf{e}_{j,i}) \ ,$$

where $B_{i,j}$ are independent Bernoulli variables with parameter p if i and j belong to the same community and q if they belong to different communities. Therefore, $\mathbf{A} - \mathbb{E}[\mathbf{A}]$ is a sum of independent, centered, random matrices with spectra bounded by 1. By the matrix Bernstein's inequality, it follows therefore that, for any $t > 0$,

$$\mathbb{P}(\|\mathbf{A} - \mathbb{E}[\mathbf{A}]\| > t) \leqslant 4n \exp(-c \min(t^2/\sigma^2, t)) \ ,$$

where

$$\begin{aligned}
\sigma^2 &= \left\| \sum_{i=1}^{n} p\mathbf{e}_{i,i} + \sum_{1 \leqslant i < j \leqslant 2n} \mathbb{E}[B_{i,j}](\mathbf{e}_{i,j} + \mathbf{e}_{j,i})^2 \right\| \\
&= \left\| \sum_{i=1}^{n} p\mathbf{e}_{i,i} + \frac{1}{2} \sum_{i=1}^{2n} \sum_{j=1, j \neq i}^{n} \mathbb{E}[B_{i,j}](\mathbf{e}_{i,i} + \mathbf{e}_{j,j}) \right\| \\
&= \left\| \frac{1}{2} \sum_{i,j=1}^{2n} \mathbb{E}[B_{i,j}](\mathbf{e}_{i,i} + \mathbf{e}_{j,j}) \right\| \\
&= \left\| \sum_{i=1}^{2n} \mathbf{e}_{i,i} \sum_{j=1}^{n} \mathbb{E}[B_{i,j}] \right\| \\
&= n(p+q) \ .
\end{aligned}$$

□

4.5 Decoupling and quadratic forms

In this section, we prove tools to show the concentration of the linear function

$$\langle A, \mathbb{X} \rangle \ ,$$

when $\mathbb{X}$ is a Gram Matrix, that is, a matrix whose entries are of the form $\langle X_i, X_j \rangle$ with independent vectors $X_1, \ldots, X_n$ taking values in a Hilbert

space $\mathcal{H}$. The final result is given in the problem concluding the chapter. Remark that these matrices cannot be written $\sum_{i=1}^N X_i$ with independent random matrices X_i, so this result is different from the ones considered in the previous sections. Let thus A denote a $n \times n$ symmetric matrix A with null diagonal and independent centered random vectors $X_1, \dots, X_n$ taking values in a Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$, we are interested in the concentration of

$$\langle A, \mathbb{X} \rangle = \sum_{1 \leqslant i \neq j \leqslant n} A_{i,j} \langle X_i, X_j \rangle \ .$$

The tricky part is that these random variables are not independent and to proceed, we are going to see an elegant and powerful argument called decoupling that allows to replace the variables X_j by independent copies X'_j.

4.5.1 Decoupling

Theorem 44 (Decoupling)**.** *For any convex function $F : \mathbb{R} \to \mathbb{R}$, we have*

$$\mathbb{E}\left[F\left(\sum_{1 \leqslant i \neq j \leqslant n} A_{i,j} \langle X_i, X_j \rangle\right)\right] \leqslant \mathbb{E}\left[F\left(4 \sum_{1 \leqslant i \neq j \leqslant n} A_{i,j} \langle X_i, X'_j \rangle\right)\right] \ ,$$

where $X'_1, \dots, X'_n$ are independent copies of $X_1, \dots, X_n$.

Proof. The proof is decomposed in several elementary steps.

Step 1: For every random variables Y and Z such that $\mathbb{E}[Z|Y] = 0$, we have, by Jensen's inequality

$$\mathbb{E}[F(Y)] = \mathbb{E}[F(Y + \mathbb{E}[Z|Y])] \leqslant \mathbb{E}[F(Y + Z)] \ .$$

Step 2: Let $B_1, \dots, B_n$ denote independent Bernoulli $\mathcal{B}(1/2)$ random variables, independent for $X_1, \dots, X_n$. Let $I = \{i \in \{1, \dots, n\} : B_i = 1\}$. We have

$$\begin{aligned}
\sum_{1 \leqslant i \neq j \leqslant n} A_{i,j} \langle X_i, X_j \rangle &= \sum_{1 \leqslant i \neq j \leqslant n} 4\mathbb{E}_B[B_i(1 - B_j)] A_{i,j} \langle X_i, X_j \rangle \\
&= 4\mathbb{E}_B\left[\sum_{1 \leqslant i \neq j \leqslant n} B_i(1 - B_j) A_{i,j} \langle X_i, X_j \rangle\right] \\
&= 4\mathbb{E}_I\left[\sum_{i \in I} \sum_{j \in I^c} A_{i,j} \langle X_i, X_j \rangle\right] \ .
\end{aligned}$$

Using Jensen's inequality once again, we get that

$$\begin{aligned}
\mathbb{E}\left[F\left(\sum_{1 \leqslant i \neq j \leqslant n} A_{i,j} \langle X_i, X_j \rangle\right)\right] &\leqslant \mathbb{E}_X \mathbb{E}_I\left[F\left(4 \sum_{i \in I} \sum_{j \in I^c} A_{i,j} \langle X_i, X_j \rangle\right)\right] \\
&= \mathbb{E}_I \mathbb{E}_X\left[F\left(4 \sum_{i \in I} \sum_{j \in I^c} A_{i,j} \langle X_i, X_j \rangle\right)\right] \ .
\end{aligned}$$

Step 3: Conditionally on I, the random variables

$$\sum_{i\in I}\sum_{j\in I^c} A_{i,j}\left\langle X_i, X_j\right\rangle, \qquad \sum_{i\in I}\sum_{j\in I^c} A_{i,j}\left\langle X_i, X_j'\right\rangle \ ,$$

have the same distribution, so

$$\mathbb{E}\left[F\Big(\sum_{1\leqslant i\neq j\leqslant n} A_{i,j}\left\langle X_i, X_j\right\rangle\Big)\right] \leqslant \mathbb{E}_I\mathbb{E}_X\left[F\Big(4\sum_{i\in I}\sum_{j\in I^c} A_{i,j}\left\langle X_i, X_j'\right\rangle\right] \ . \tag{4.6}$$

Now, for every $I, J \subset \{1,\dots,n\}$, we write

$$Q(I,J) = \sum_{i\in I}\sum_{j\in J} A_{i,j}\left\langle X_i, X_j'\right\rangle \ .$$

For any $I \subset T_n = \{1,\dots,n\}$

$$Q(T_n,T_n) = Q(I,I^c) + Q(I^c,I^c) + Q(T_n,I) \ ,$$

where

$$\mathbb{E}[Q(I^c,I^c)|Q(I,I^c)] = 0, \qquad \mathbb{E}[Q(T_n,I)|Q(I,I^c)+Q(I^c,I^c)] = 0 \ .$$

Therefore, using Step 1 twice, for any $I \subset T_n$,

$$\begin{aligned}\mathbb{E}\left[F\Big(4\sum_{i\in I}\sum_{j\in I^c} A_{i,j}\left\langle X_i, X_j'\right\rangle\right] &= \mathbb{E}[F(4Q(I,I^c))] \\ &\leqslant \mathbb{E}[F(4(Q(I,I^c)+Q(I^c,I^c)))] \\ &\leqslant \mathbb{E}[F(4Q(T_n,T_n))] \ .\end{aligned}$$

As this is true for any value of I, it is true by taking the expectation with respect to I. Plugging this result into (4.6) yields the result. □

4.5.2 Concentration of Gaussian Chaos

In this section, we prove two results in the Gaussian case. The first one is a basic application of Bernstein's inequality for sums of Gaussian random variables that will be used as a technical tool in the following.

Theorem 45. *Let $(g_1,g_1'),\dots,(g_n,g_n')$ denote n independent couples of Gaussian random variables and let $\boldsymbol{s} = (s_1,\dots,s_n)^T \in \mathbb{R}^n$. Then, $\sum_{i=1}^n s_i g_i g_i'$ is (b,σ^2)-sub-Gamma, where $b = C\|\boldsymbol{s}\|_\infty$ and $\sigma^2 = C\|\boldsymbol{s}\|_2^2$. In particular, for any $z>0$,*

$$\mathbb{P}\Big(\sum_{i=1}^n s_i g_i g_i' > z\Big) \leqslant \exp\left(-c\min\left(\frac{z^2}{\|\boldsymbol{s}\|_2^2}, \frac{z}{\|\boldsymbol{s}\|_\infty}\right)\right) \ .$$

Proof. As g_i and g'_i are sub-Gaussian, $s_i g_i g'_i$ is sub-Exponential with $\|s_i g_i g'_i\|_{\psi_1} = Cs_i^2$. The result thus follows from Corollary 28. □

We can now move to the proof of the concentration of Gaussian Chaos.

Theorem 46 (Concentration of Gaussian Chaos)**.** *Let A denote a $n \times n$ matrix and let g, g' denote two standard Gaussian vectors. Then, $g^T A g'$ is (b, σ^2) sub-Gamma, where $b = C\|A\|$ and $\sigma^2 = C\|A\|_F^2$. In particular, for any $z > 0$,*

$$\mathbb{P}\Big(g^T A g' > z \Big) \leqslant \exp\left(- c \min\left(\frac{z^2}{\|A\|_F^2}, \frac{z}{\|A\|} \right) \right) \ .$$

Proof. Let $A = \sum_{i=1}^n s_i u_i v_i^T$ denote the SVD of A, so

$$g^T A g' = \sum_{i=1}^n s_i \left\langle g, u_i \right\rangle \left\langle g', v_i \right\rangle \ .$$

The result then follows from the previous lemma, using independence of the couples of standard Gaussian random variables (g_i, g'_i), where $g_i = \langle g, u_i \rangle$ and $g'_i = \langle g', v_i \rangle$. □

4.5.3 Hanson-Wright's inequality

Hanson-Wright's inequality is an extension of the previous result to the case where X_i are not necessarily standard Gaussian but are sub-Gaussian random variables. Let $X = (X_1, \dots, X_n)^T$ denote a vector in $\mathbb{R}^n$ with independent sub-Gaussian entries and let A denote a $n \times n$ deterministic matrix with 0 entries in the diagonal. We are interested in the concentration of

$$X^T A X = \sum_{1 \leqslant i \neq j \leqslant n} a_{i,j} X_i X_j \ ,$$

The following theorem is known as Hanson-Wright's inequality.

Theorem 47. *Assume that $X_1, \dots, X_n$ are independent and that each X_i is K-sub-Gaussian. Then, for any $z > 0$,*

$$\mathbb{P}\Big(X^T A X > z \Big) \leqslant 2 \exp\left(- c \min\left(\frac{z^2}{K^4 \|A\|_F^2}, \frac{z}{K^2 \|A\|} \right) \right) \ .$$

Proof. We can assume by homogeneity that $K = 1$. The first step of the proof is to use decoupling to get, for any $s \in \mathbb{R}$,

$$\mathbb{E}\big[\exp\left(s X^T A X \right) \big] \leqslant \mathbb{E}\big[\exp\left(4 s X^T A X' \right) \big] \ .$$

The second step is to move from sub-Gamma random variables to Gaussian ones. Let G, G' denote independent standard Gaussian vectors in $\mathbb{R}^n$, independent from X, X'. We have, for any vector $u \in \mathbb{R}^n$,

$$\mathbb{E}[\exp(\langle u, X\rangle)] = \prod_{i=1}^{n} \mathbb{E}[\exp(u_i X_i)] \leqslant \prod_{i=1}^{n} \exp(C u_i^2) \leqslant \exp(C\|u\|_2^2) \ .$$

Thus, for any $s > 0$,

$$\begin{aligned}\mathbb{E}_X\big[\exp\big(sX^T AX'\big)\big] &= \mathbb{E}_X[\exp(\big\langle X, sAX'\big\rangle)] \\ &\leqslant \exp(Cs^2\|AX'\|_2^2)] \\ &= \mathbb{E}_G\big[\exp\big(CsG^T AX'\big)\big] \ .\end{aligned}$$

Reproducing the argument proves that

$$\mathbb{E}\big[\exp\big(4sX^T AX'\big)\big] \leqslant \mathbb{E}\big[\exp\big(CsG^T AG'\big)\big] \ .$$

The third and last step is to bound the Laplace transform of Gaussian Chaos. By Theorem 46, we have, for any $|s| < 1/C\|A\|$,

$$\mathbb{E}[\exp(sG^T AG')] \leqslant \exp(Cs^2\|A\|_F^2) \ .$$

In particular thus, for any $|s| < 1/(C\|A\|)$,

$$\mathbb{E}\big[\exp\big(sX^T AX\big)\big] \leqslant \exp(Cs^2\|A\|_F^2) \ .$$

This means that $X^T AX$ is $(C\|A\|, C\|A\|_F^2)$-sub-Gamma and the result follows therefore from Bernstein's inequality Theorem 25. □

4.5.4 Problem

The question we investigate in this problem is the following. Let $X_1, \ldots, X_n$ denote independent random vectors of $\mathbb{R}^d$ satisfying $\mathbb{E}[X_i] = 0$ and the following sub-Gaussian assumption: there exists a symmetric matrix Γ such that

$$\forall u \in \mathbb{R}^d, \qquad \mathbb{E}\big[\exp(\langle u, X_i\rangle)\big] \leqslant \exp(u^T \Gamma u) \ .$$

Let $A = (a_{i,j})$ denote a $n \times n$ deterministic matrix with 0 in the diagonal. We are interested in this problem in providing concentration bounds for

$$Z = \sum_{i \neq j = 1}^{n} a_{i,j} \langle X_i, X_j\rangle \ .$$

1. Denote by $X_1', \ldots, X_n'$ independent copies of $X_1, \ldots, X_n$. Prove that, for any sufficiently small $s \in \mathbb{R}$,

$$\mathbb{E}\big[\exp\big(sZ\big)\big] \leqslant \mathbb{E}\Big[\exp\Big(4s \sum_{1 \leqslant i \neq j \leqslant n} a_{i,j} \langle X_i, X_j'\rangle\Big)\Big] \ .$$

2. Let $g_1, \dots, g_n, g'_1, \dots, g'_n$ denote independent standard Gaussian vectors. Show that there exists a numerical constant C such that, for any sufficiently small $s \in \mathbb{R}$,

$$\mathbb{E}\Big[\exp\Big(s \sum_{1 \leqslant i \neq j \leqslant n} \langle X_i, X_j \rangle \Big) \Big] \leqslant \mathbb{E}\Big[\exp\Big(Cs \sum_{1 \leqslant i \neq j \leqslant n} a_{i,j} \langle \Gamma g_i, g'_j \rangle \Big) \Big] \ .$$

3. Prove that there exist independent standard Gaussian vectors $G_1, \dots, G_d$, $G'_1, \dots, G'_d$ in $\mathbb{R}^n$ and non negative real numbers $\lambda_1, \dots, \lambda_d$ such that

$$\sum_{1 \leqslant i \neq j \leqslant n} a_{i,j} \langle \Gamma g_i, g'_j \rangle = \sum_{k=1}^{d} \lambda_k G_k^T A G'_k \ .$$

4. Prove that, for any $z > 0$,

$$\mathbb{P}(Z > z) \leqslant \exp\Big(- C \min\Big(\frac{z^2}{\sigma^2}, \frac{z}{b} \Big) \Big) \ ,$$

where

$$\sigma^2 = \|\Gamma\|_F^2 \|A\|_F^2, \qquad b = \|\Gamma\| \|A\| \ .$$

Chapter 5

PAC-Bayesian bounds

In the first part of this chapter, we consider a random vector X such that $\mathbb{E}[X] = 0$. The vector X defines the linear process $X_t = \langle X, t\rangle$ that we can bound using chaining arguments as will be seen in Chapter 6, . This allows to obtain deviation bounds for suprema of X_t over some sets T. This chapter explains how to prove similar deviation inequalities using the PAC-Bayesian approach. We know that each X_t concentrates if $\|X_t\|_{\psi_\alpha} \leqslant K$, where $\alpha \in \{1, 2\}$ using the methods of Chapter 3. The Pac-Bayesian approach allows to go from the concentration of each X_t to deviation inequalities for $\sup_{t\in T} X_t$ when T is an ellipsoïd. Although less general than chaining methods, this approach yields tighter constants and is sometimes much easier to develop in examples.

In a second part of the chapter, we extend the PAC-Bayesian approach to prove deviation inequalities for the operator norm of random matrices.

5.1 Setting

In the first sections of the chapter, X is a random vector of $\mathbb{R}^d$ such that $\mathbb{E}[X] = 0$. Besides, we will use repeatedly the following definitions.

Definition 48. *The vector X is called K-sub-Gaussian if, for any t in the sphere $\mathbb{S} \subset \mathbb{R}^d$ and any $s \in \mathbb{R}$, $\mathbb{E}[\exp(s\langle X, t\rangle)] \leqslant \exp(s^2K^2)$. It is called (b, K)-sub-Gamma if, for any $t \in \mathbb{S}$ and any $s \leqslant 1/b$, $\mathbb{E}[\exp(s\langle X, t\rangle)] \leqslant \exp(s^2K^2)$.*

When X is sub-Gaussian (resp. sub-Gamma), the concentration of each $X_t = \langle X, t\rangle$ is derived from Hoeffding's (resp. Bernstein's) inequality. We want to make these deviations uniform over an ellipsoïd $T \subset \mathbb{R}^d$. For this, we use a Bayesian approach in the sense that we assume that the process X_t is evaluated at a random parameter θ. This parameter will always be assumed independent from the vector X and we will consider several possible distributions $(\rho_t)_{t\in T}$ for θ. With this approach, we replace the problem

of bounding the probability of an event $\{\forall t \in T, \ldots\}$ by the problem of bounding the supremum of expectations of functions over all $\rho \in (\rho_t)_{t\in T}$.

There exists a way to bound the expectation of a random variable w.r.t. several probability measures. Given a probability distribution μ on $\mathbb{R}^d$, we will prove in the following section (see Lemma 49) a variational formula which states that, for any function $g : \mathbb{R}^d \to \mathbb{R}$,

$$\exp\left(\sup_{\rho \preccurlyeq \mu} \mathbb{E}_\rho[g(\theta)] - K(\rho,\mu)\right) = \mathbb{E}_\mu[\exp(g(\theta))] \ , \tag{5.1}$$

where $K(\rho,\mu) = \int \log(\mathrm{d}\rho/\mathrm{d}\mu)\mathrm{d}\rho$ stands for the Küllback-Leibler divergence.

Consider now the function $g(\theta) = s\langle X,\theta\rangle - \log \mathbb{E}_X[\exp(s\langle X,\theta\rangle)]$ (here and in the following, for any function $f(X,\theta)$ we denote by $\mathbb{E}_X[f(X,\theta)]$ the expectation of f with respect to the distribution of X and by $\mathbb{E}_\rho[f(X,\theta)]$ the expectation w.r.t. $\theta \sim \rho$, assuming always that X and θ are independent). This function satisfies

$$\exp(g(\theta)) = \frac{\exp(s\langle X,\theta\rangle)}{\mathbb{E}_X[\exp(s\langle X,\theta\rangle)]} \ ,$$

so $\mathbb{E}_X[\exp(g(\theta))] = 1$ and therefore $\mathbb{E}_X[\mathbb{E}_\mu[\exp(g(\theta))]] = \mathbb{E}_\mu[\mathbb{E}_X[\exp(g(\theta))]] = 1$ and, for any $z > 0$,

$$\mathbb{P}_X(\mathbb{E}_\mu[\exp(g(\theta))] > \exp(z)) \leqslant \exp(-z) \ .$$

On the other hand, the variational formula implies that

$$\log \mathbb{E}_\mu[\exp(g(\theta))] = \sup_{\rho \preccurlyeq \mu} \mathbb{E}_\rho[g(\theta)] - K(\rho,\mu) \ ,$$

thus, for any $z > 0$,

$$\mathbb{P}_X(\sup_{\rho \preccurlyeq \mu} \mathbb{E}_\rho[g(\theta)] - K(\rho,\mu) > z) = \mathbb{P}_X(\mathbb{E}_\mu[\exp(g(\theta))] > \exp(z)) \leqslant \exp(-z) \ .$$

Replacing by the expression of $g(\theta)$, we get that

$$\mathbb{P}_X\left(\forall t \in T, s\mathbb{E}_{\rho_t}[\langle X,\theta\rangle] \leqslant \mathbb{E}_{\rho_t}[\log \mathbb{E}_X[\exp(s\langle X,\theta\rangle)]] + K(\rho_t,\mu) + z\right) \leqslant \exp(-z) \ .$$

We can readily see here why this formula can be interesting. For any distribution ρ_t centered at t, we have in particular $\mathbb{E}_{\rho_t}[\langle X,\theta\rangle] = \langle X, \mathbb{E}_{\rho_t}[\theta]\rangle = \langle X,t\rangle$. Besides, the assumption that the vector X is sub-Gaussian (resp. sub-Gamma) yields that

$$\log \mathbb{E}_X[\exp(s\langle X,\theta\rangle)] \leqslant K^2 s^2 \|\theta\|_2^2, \qquad \forall \theta \in \mathbb{R}^d, s \in \mathbb{R} \quad ,$$

(resp. $\forall \theta \in \mathbb{R}^d, s \in \mathbb{R} : |s|\|\theta\| \leqslant 1/b$).

Therefore, if ρ_t are centered at t, we get that, for any $z > 0$,

$$\mathbb{P}_X\left(\forall t \in T, \langle X,t\rangle \leqslant sK^2\mathbb{E}_{\rho_t}[\|\theta\|_2^2] + \frac{K(\rho_t,\mu)+z}{s}\right) \leqslant \exp(-z) \ .$$

This inequality is valid for any distribution μ and either

- in the sub-Gaussian case, for any probability distributions ρ_t and any $s \in \mathbb{R}$ or,
- in the sub-Gamma case, for any $r > 0$, for any probability distribution ρ_t supported in the ball $B(0, r)$ and all $|s| \leqslant 1/rb$.

In all cases, we see that we have to pick distributions ρ_t centered at t and such that $\mathbb{E}_{\rho_t}[\|\theta\|_2^2] = \|t\|_2^2 + \mathbb{E}_{\rho_t}[\|\theta - t\|_2^2]$ can easily be computed. Then, we can choose μ such that all Küllback divergences $K(\rho_t, \mu)$ can be upper bounded.

In the following, we first establish the basic tools mentioned in this section, then show that Gaussian priors can be used in the sub-Gaussian case and build truncated Gaussian for the exponential case.

5.2 Basic tools

For any probability measures ρ and μ on a measurable space Ω, we denote by $\rho \preccurlyeq \mu$ if ρ is absolutely continuous with respect to μ. When $\rho \preccurlyeq \mu$, we also denote the KL divergence by

$$K(\rho, \mu) = \int \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\mu}\right)\mathrm{d}\rho \ .$$

PAC-Bayesian approach builds on a variational formula for entropies that reads as follows:

Lemma 49 (Variational fomula)**.** *Let (Ω, μ) denote a probability space and let g denote a real valued function on Θ such that $\log \mathbb{E}_\mu[\exp(g(\theta))] < +\infty$. Then, for any $\rho \preccurlyeq \mu$, we have*

$$\log \mathbb{E}_\mu[\exp(g(\theta))] \geqslant \mathbb{E}_\rho[g(\theta)] - K(\rho, \mu) \ .$$

Besides, the probability distribution such that $\mathrm{d}\rho/\mathrm{d}\mu = \exp(g)/\mathbb{E}_\mu[\exp(g(\theta))]$ satisfies

$$\log \mathbb{E}_\mu[\exp(g(\theta))] = \mathbb{E}_\rho[g(\theta)] - K(\rho, \mu) \ .$$

Hence, we have

$$\log \mathbb{E}_\mu[\exp(g(\theta))] = \sup_{\rho:\rho\preccurlyeq\mu} \left\{\mathbb{E}_\rho[g(\theta)] - K(\rho, \mu)\right\} \ .$$

Proof. The proof is elementary: For the first item, write $f = \mathrm{d}\rho/\mathrm{d}\mu$ so $K(\rho, \mu) = \mathbb{E}_\rho[\log(f(\theta))]$. Then, remark that

$$\log \mathbb{E}_\mu[\exp(g(\theta))] \geqslant \log \int_{\theta: f(\theta)\neq 0} \frac{\exp(g(\theta))}{f(\theta)}\mathrm{d}\rho(\theta) = \log \mathbb{E}_\rho\left[\frac{\exp(g(\theta))}{f(\theta)}\mathbf{1}_{f(\theta)\neq 0}\right] \ .$$

Thus, by Jensen's inequality

$$\log \mathbb{E}_\mu[\exp(g(\theta))] \geqslant \mathbb{E}_\rho\big[g(\theta) - \log\big(f(\theta)\big)\big] = \mathbb{E}_\rho[g(\theta)] - K(\rho, \mu) \ .$$

For the second item, we compute directly

$$\log\left(\frac{\mathrm{d}\rho}{\mathrm{d}\mu}\right) = g(\theta) - \log(\mathbb{E}_\mu[\exp(g(\theta))]) \ ,$$

so

$$K(\rho, \mu) = \mathbb{E}_\rho[g(\theta)] - \log(\mathbb{E}_\mu[\exp(g(\theta))]) \ ,$$

which is equivalent to the result. □

5.3 Sub-Gaussian vectors

In this section, we show why the general PAC-Bayesian bound allows to derive the concentration of the linear process we met in Section 5.1. As precise constants can be obtained using the PAC-Bayesian approach, we specify here the constants in the sub-Gaussian assumption. Let X denote a random vector in $\mathbb{R}^d$ such that, for any $t \in \mathbb{R}^d$,

$$\mathbb{E}[\exp(\langle t, X\rangle)] \leqslant \exp\left(\frac{\|t\|_2^2}{2}\right) \ .$$

Let T denote the ellipsoid described by the positive definite matrix Γ by the formula

$$T = \Gamma^{1/2}\mathbb{B}_2 = \{t \in \mathbb{R}^d : \|\Gamma^{-1/2}t\|_2 = \|t\|_{\Gamma^{-1}} \leqslant 1\} \ .$$

The goal here is to obtain deviation bounds for $\sup_{t\in T} \langle X, t\rangle$.

Recall that, if X is a standard Gaussian vector, the Gaussian concentration inequality implies that, for any $z > 0$, w.p.a.l. $1 - 2\exp(-z)$,

$$\sup_{t\in T} \langle X, t\rangle \leqslant \sqrt{\mathrm{Tr}(\Gamma)} + \sqrt{2\|\Gamma\|z} \ . \tag{5.2}$$

5.3.1 Choice of μ and ρ's

In this section, we show that Gaussian distributions can be used as priors. Let Σ denote a non singular positive matrix and choose μ to be the Gaussian distribution $\mathrm{N}(0, \Sigma)$ and, for any $t \in T$, we let ρ_t denote the Gaussian distribution $\mathrm{N}(t, \Sigma)$ (centered at t with the same variance as μ). With these choices, it is indeed easy to compute the Küllback-Leibler divergence. We have

$$\begin{aligned}\log\left(\frac{\mathrm{d}\rho_t(x)}{\mathrm{d}\mu(x)}\right) &= \log\left(\exp\left(\frac{\|x\|_{\Sigma^{-1}}^2 - \|x-t\|_{\Sigma^{-1}}^2}{2}\right)\right) \\ &= \frac{1}{2}\big(2\langle x - t, t\rangle_{\Sigma^{-1}} + \|t\|_{\Sigma^{-1}}^2\big) \ .\end{aligned}$$

Taking the expectation w.r.t. $x \sim \rho_t$ in the expression yields directly

$$K(\rho_t, \mu) = \frac{1}{2}\|t\|^2_{\Sigma^{-1}} \ .$$

5.3.2 Bounding the second moment

For any $t \in T$,

$$\mathbb{E}_{\rho_t}[\|\theta\|_2^2] = \|t\|_2^2 + \mathbb{E}_{\rho_t}[\|\theta - t\|_2^2] = \|t\|_2^2 + \mathrm{Tr}(\Sigma) \ ,$$

where the last inequality directly follows from $\|\theta - t\|_2^2 = \mathrm{Tr}\left((\theta - t)(\theta - t)^T\right)$.

5.3.3 Conclusion

The Pac-Bayesian bound directly implies in this case that, for any $z > 0$, with probability larger than $1 - \exp(-z)$, we have, simultaneously for all $t \in T$,

$$s\,\langle X, t\rangle \leqslant \frac{s^2}{2}\big(\|t\|_2^2 + \mathrm{Tr}(\Sigma)\big) + \frac{1}{2}\|t\|^2_{\Sigma^{-1}} + z \ .$$

This result holds for any $s > 0$ and any non singular $\Sigma \succcurlyeq 0$.

Let us specify now $\Sigma = \Gamma/\beta$, where $\beta > 0$. We deduce that, for any $z > 0$, with probability larger than $1 - \exp(-z)$, we have, simultaneously for all $t \in T$,

$$\langle X, t\rangle \leqslant \frac{s}{2}\left(\|t\|_2^2 + \frac{1}{\beta}\,\mathrm{Tr}(\Gamma)\right) + \frac{\beta}{2s}\|t\|^2_{\Gamma^{-1}} + \frac{z}{s} \ .$$

Using the definition of T, we get

$$\langle X, t\rangle \leqslant \frac{s}{2}\left(\|t\|_2^2 + \frac{1}{\beta}\,\mathrm{Tr}(\Gamma)\right) + \frac{\beta}{2s} + \frac{z}{s} \ .$$

We let $s > 0$ free and choose $\beta = s\sqrt{\mathrm{Tr}(\Gamma)}$ to get

$$\langle X, t\rangle \leqslant \sqrt{\mathrm{Tr}(\Gamma)} + \frac{s}{2}\|t\|_2^2 + \frac{z}{s} \ .$$

Finally, we optimize the choice of $s = \sqrt{2z}/\|t\|_2$ to get

$$\forall z > 0, \qquad \mathbb{P}\big(\forall t \in T, \quad \langle X, t\rangle \leqslant \sqrt{\mathrm{Tr}(\Gamma)} + \|t\|_2\sqrt{2z}\big) \geqslant 1 - \exp(-z) \ . \quad (5.3)$$

This bound directly implies (5.2). An interesting feature is that we recover this bound from the general Pac-Bayesian bound *up to the constants*. To the best of our knowledge, this is the only approach showing this bound with tight constants for sub-Gaussian random vectors.

On the other hand, it does not seem easy to extend the result beyond ellipsoids to cover the case of a general subset $T \subset \mathbb{R}^d$.

5.4 Sub-exponential random vectors

In this section we consider vectors such that

$$\forall t \in \mathbb{S}, \qquad \forall |s| \leqslant 1/b, \qquad \mathbb{E}[\exp(s\,\langle t, X\rangle)] \leqslant \exp(s^2 K^2) \ . \tag{5.4}$$

For these vectors, the general approach developed in Section 5.1 shows that, for any $r > 0$, as long as ρ_t is centered at t and supported on a ball centered at t with radius r, for any $|s| < 1/b(r + \sup_{t\in T} \|t\|_2)$, for any $z > 0$,

$$\mathbb{P}_X\left(\forall t \in T, \langle X, t\rangle \leqslant CsK^2\mathbb{E}_{\rho_t}[\|\theta\|_2^2] + \frac{K(\rho_t, \mu) + z}{s}\right) \leqslant \exp(-z) \ .$$

As in the previous section, T denote the ellipsoid described by the positive definite matrix Γ by the formula

$$T = \Gamma^{1/2}\mathbb{B}_2 = \{t \in \mathbb{R}^d : \|\Gamma^{-1/2}t\|_2 = \|t\|_{\Gamma^{-1}} \leqslant 1\} \ .$$

5.4.1 Priors

We define for ρ_t the truncated Gaussian defined for any $t \in T$, a radius r and non-singular covariance matrix Σ to be specified later by the distribution with density

$$f_t(x) = \frac{1}{(\sqrt{2\pi})^d\sqrt{\det(\Sigma)}C_r}\exp\left(-\frac{1}{2}(x-t)^T\Sigma^{-1}(x-t)\right)\mathbf{1}_{\{\|x-t\|_2\leqslant r\}} \ ,$$

with normalizing constant $C_r = \mathbb{P}(\mathsf{N}(0,\Sigma) \in B(0,r))$. Remark first that, as the density f_t is symmetric around t, this distribution is centered at t as expected. As mentioned earlier, the key is to be able to bound the second moment of these measures and choose a prior measure μ for which the Küllback divergences $K(\rho_t, \mu)$ can be uniformly bounded.

Start with the second moment. As ρ_t is centered at t, we have

$$\mathbb{E}_{\rho_t}[\|\theta\|_2^2] = \|t\|_2^2 + \mathbb{E}_{\rho_0}[\|\theta\|_2^2] \ .$$

Then, denoting by φ the density of the Gaussian distribution $\mathsf{N}(0,\Sigma)$ and by $G \sim \mathrm{N}(0, \mathbf{I})$, we have

$$\mathbb{E}_{\rho_0}[\|\theta\|_2^2] = \frac{1}{C_r}\int_{B(0,r)} \|x\|_2^2\varphi(x)\mathrm{d}x \leqslant \frac{1}{C_r}\mathbb{E}[\|\Sigma^{1/2}G\|_2^2] = \frac{1}{C_r}\operatorname{Tr}(\Sigma) \ .$$

Moving to the Küllback divergence, we denote by $\mu = \mathrm{N}(0,\Sigma)$, so

$$\begin{aligned}\frac{\mathrm{d}\rho_t(x)}{\mathrm{d}\mu(x)} &= \frac{1}{C_r}\exp\left(\frac{1}{2}\big(\|x\|_{\Sigma^{-1}} - \|x-t\|_{\Sigma^{-1}}^2\big)\right)\mathbf{1}_{\{\|x-t\|_2\leqslant r\}} \\ &= \frac{1}{C_r}\exp\left(\langle x-t, t\rangle_{\Sigma^{-1}} + \frac{1}{2}\|t\|_{\Sigma^{-1}}^2\right)\mathbf{1}_{\{\|x-t\|_2\leqslant r\}} \ .\end{aligned}$$

Integrating the logarithm of this expression and using again that ρ_t is centered in t, we get

$$K(\rho_t, \mu) \leqslant C_r \left(\log\left(\frac{1}{C_r}\right) + \frac{1}{2}\|t\|^2_{\Sigma^{-1}} \right) \leqslant \frac{1}{e} + \frac{1}{2}\|t\|^2_{\Sigma^{-1}} \ .$$

5.4.2 Conclusion

We choose $\Sigma = \Gamma/\beta$ for a parameter β to be chosen later. By (5.3), $C_r \geqslant 1/2$ if $r = \sqrt{\mathrm{Tr}(\Gamma)/\beta} + \sqrt{\|\Gamma\| \log(1/2)/\beta} \leqslant 2\sqrt{\mathrm{Tr}(\Gamma)/\beta} = r_0$. We have thus, for $r = r_0$, $\beta > 0$ and $|s| \leqslant 1/b(r_0 + \sqrt{\|\Gamma\|})$, for any $z > 0$,

$$\mathbb{P}_X \left(\forall t \in T, \langle X, t\rangle \leqslant sK^2 \left(\|\Gamma\| + \frac{\mathrm{Tr}(\Gamma)}{\beta} \right) + \frac{\beta + 1 + z}{s} \right) \leqslant \exp(-z) \ .$$

We choose $s = [\alpha(\sqrt{\mathrm{Tr}(\Gamma)/\beta} + \sqrt{\|\Gamma\|})]^{-1}$, with $\alpha \geqslant b$, we have, with probability larger than $1 - \exp(-z)$,

$$\forall t \in T, \langle X, t\rangle \leqslant \left(\sqrt{\|\Gamma\|} + \sqrt{\frac{\mathrm{Tr}(\Gamma)}{\beta}} \right) \left(\frac{K^2}{\alpha} + \alpha(\beta + 1 + z) \right) \ . \qquad (5.5)$$

Define now the effective rank

$$r(\Gamma) = \frac{\mathrm{Tr}(\Gamma)}{\|\Gamma\|} \ .$$

We consider two possible situations: Either $r(\Gamma) + 1 + z \leqslant (K/b)^2$ or $r(\Gamma) + 1 + z > (K/b)^2$.

Let us start with the case where $r(\Gamma) + 1 + z \leqslant (K/b)^2$. We choose $\beta = r(\Gamma)$. We get, $\forall t \in T$,

$$\langle X, t\rangle \leqslant 2\sqrt{\|\Gamma\|} \left(\frac{K^2}{\alpha} + \alpha(r(\Gamma) + 1 + z) \right) \ .$$

There, we choose $\alpha = K/\sqrt{r(\Gamma) + 1 + z} \geqslant b$ and we have, with probability larger than $1 - \exp(-z)$,

$$\forall t \in T, \qquad \langle X, t\rangle \leqslant 4K\sqrt{\mathrm{Tr}(\Gamma) + \|\Gamma\|(1 + z)} \ .$$

Let us now move to the case where $r(\Gamma) + 1 + z > (K/b)^2$. We further divide this case into two disjoint situations: Either $1 + z \leqslant (K/b)^2$ or $1 + z > (K/b)^2$. Start with $1 + z \leqslant (K/b)^2$ so $b \leqslant K/\sqrt{1 + z}$. We choose $\beta = 1 + z$, $\alpha = K/\sqrt{1 + z}$ to get, $\forall t \in T$,

$$\langle X, t\rangle \leqslant 3K\sqrt{1 + z}\sqrt{\|\Gamma\| \left(1 + \frac{r(\Gamma)}{1 + z} \right)} = 3K\sqrt{\mathrm{Tr}(\Gamma) + \|\Gamma\|(1 + z)} \ .$$

Finally, consider the case $1+z > (K/b)^2$. Choose $\beta = 1+z$, $\alpha = b$ to get, for all $t \in T$,

$$\begin{aligned}\langle X, t\rangle &\leqslant \left(\sqrt{\|\Gamma\|} + \sqrt{\frac{\mathrm{Tr}(\Gamma)}{1+z}}\right)\left(\frac{K^2}{b} + 2b(1+z)\right) \\ &\leqslant 3b(1+z)\left(\sqrt{\|\Gamma\|} + \sqrt{\frac{\mathrm{Tr}(\Gamma)}{1+z}}\right) \\ &= 3b\big(\sqrt{\mathrm{Tr}(\Gamma)(1+z)} + \sqrt{\|\Gamma\|}(1+z)\big) \ .\end{aligned}$$

In all cases, we conclude that, with probability larger than $1-\exp(-z)$,

$$\sup_{t\in T} \langle X, t\rangle \leqslant \sqrt{\|\Gamma\|}\left(4K\sqrt{r(\Gamma)} + \sqrt{1+z}(4K \vee 3b\sqrt{r(\Gamma)}) + 3b(1+z)\right) \ . \quad (5.6)$$

5.5 Extension to random matrices

In this section, we are interested in the extension of the previous results to the case where $\mathbf{X}$ is a random $n \times p$ matrix such that $\mathbb{E}[\mathbf{X}] = 0$. Let $\mathcal{E}_U$ and $\mathcal{E}_V$ denote two ellipsoïds: given two symmetric and non singular matrices Γ_U and Γ_V, let

$$\mathcal{E}_U = \{u \in \mathbb{R}^n : \|\Gamma_U^{-1/2} u\|_2 \leqslant 1\}, \qquad \mathcal{E}_V = \{v \in \mathbb{R}^p : \|\Gamma_V^{-1/2} v\|_2 \leqslant 1\} \ .$$

We are interested in this section in providing upper bounds on

$$\sup_{u\in\mathcal{E}_U, v\in\mathcal{E}_V} u^T \mathbf{X} v = \sup_{u\in\mathcal{E}_U, v\in\mathcal{E}_V} \left\langle \mathbf{X}, vu^T \right\rangle_F \ ,$$

that hold with high probability.

5.5.1 Probabilistic assumption

We are particularly interested in the case where $\mathbf{X} = \sum_{i=1}^n (X_i X_i^T - \Sigma)$ for some independent random vectors $X_i \in \mathbb{R}^n$ such that $\mathbb{E}[X_i] = 0$, $\mathbb{E}[X_i X_i^T] = \Sigma$ and

$$\forall u \in \mathbb{R}^n, \qquad \mathbb{E}[\exp(\langle X_i, u\rangle)] \leqslant \exp(K^2\|u\|^2) \ .$$

In this case, for any $u \in \mathbb{S}$, the random variable $\langle u, X_i\rangle$ are K-sub-Gaussian, so $\langle u, X_i\rangle^2 - u^T\Sigma u$ are (CK^2, CK^4) sub-Gamma, so

$$\forall u \in \mathbb{R}^n : \|u\|_2 \leqslant 1/CK^2, \qquad \mathbb{E}\left[\exp\left(\left\langle \sum_{i=1}^n (X_i X_i^T - \Sigma), uu^T \right\rangle_F\right)\right] \leqslant \exp(CnK^2) \ .$$

To take this situation into account, we assume in the following that our matrix of interest $\mathbf{X}$ satisfies the following assumption: $\forall u, v : \|u\|_2 \vee \|v\|_2 \leqslant 1$ and $\forall s : |s| < 1/b$,

$$\mathbb{E}[\exp(s\left\langle \mathbf{X}, uv^T \right\rangle_F)] \leqslant \exp(s^2K^2) \ . \quad (5.7)$$

5.5.2 Global strategy

To bound the random variable of interest, we extend the approach seen for random vectors. Consider the function

$$g(U,V) = s\left\langle \mathbf{X}, UV^T \right\rangle_F - \log \mathbb{E}_{\mathbf{X}}[\exp(s\left\langle \mathbf{X}, UV^T \right\rangle_F)] \ ,$$

assuming always that $\mathbf{X}$, U and V are independent. This function satisfies $\mathbb{E}_{\mathbf{X}}[\exp(g(U,V))] = 1$ and therefore, for any $z > 0$,

$$\mathbb{P}_{\mathbf{X}}(\mathbb{E}_{\mu}[\exp(g(U,V))] > \exp(z)) \leqslant \exp(-z) \ .$$

By the variational formula it follows that, for any $z > 0$,

$$\mathbb{P}_{\mathbf{X}}(\sup_{\rho \preccurlyeq \mu} \mathbb{E}_{\rho}[g(U,V)] - K(\rho,\mu) > z) \leqslant \exp(-z) \ .$$

Replacing by the expression of $g(U,V)$, we get that, with $\mathbb{P}_{\mathbf{X}}$-probability larger than $1-\exp(-z)$, $\forall (u,v) \in \mathcal{E}_U \times \mathcal{E}_V$,

$$s\mathbb{E}_{\rho_u \otimes \rho'_v}[\left\langle \mathbf{X}, UV^T \right\rangle_F] \leqslant \mathbb{E}_{\rho_u \otimes \rho'_v}[\log \mathbb{E}_{\mathbf{X}}[\exp(s\left\langle \mathbf{X}, UV^T \right\rangle_F)]] + K(\rho_u \otimes \rho'_v, \mu) + z \ .$$

Now we have, by independence,

$$\mathbb{E}_{\rho_u \otimes \rho'_v}[\left\langle \mathbf{X}, UV^T \right\rangle_F] = \left\langle \mathbf{X}, \mathbb{E}_{\rho_u}[U]\mathbb{E}_{\rho'_v}[V]^T \right\rangle_F = \left\langle \mathbf{X}, uv^T \right\rangle_F \ ,$$

provided that ρ_u is centered at u and ρ'_v at v.

Furthermore, Assumption (5.7) ensures that, if ρ_u is supported in the Euclidean ball $\mathbb{B}(u, r_U)$ and ρ'_v in $\mathbb{B}(u, r_V)$, for any s such that

$$|s| \leqslant \frac{1}{(\|u\| + r_U)(\|v\| + r_V)} \ ,$$

we have $|s|\|U\|\|V\| \leqslant 1/b$ a.s. and thus

$$\log \mathbb{E}_{\mathbf{X}}[\exp(s\left\langle \mathbf{X}, UV^T \right\rangle_F)] \leqslant K^2 s^2 \|U\|^2 \|V\|^2 \ ,$$

so, by independence

$$\mathbb{E}_{\rho_u \otimes \rho'_v}[\log \mathbb{E}_{\mathbf{X}}[\exp(s\left\langle \mathbf{X}, UV^T \right\rangle_F)]] \leqslant K^2 s^2 \mathbb{E}_{\rho_u}[\|U\|^2]\mathbb{E}_{\rho'_v}[\|V\|^2] \ .$$

Finally, assuming μ is of the form $\mu = \mu_U \otimes \mu_V$, we have

$$\begin{aligned} K(\rho_u \otimes \rho'_v, \mu) &= \int \log\left(\frac{f_{\rho_u}(x) f_{\rho'_v}(y)}{f_{\mu_1}(x) f_{\mu_2}(y)}\right) f_{\rho_u}(x) f_{\rho'_v}(y) \mathrm{d}(x,y) \\ &= K(\rho_u, \mu_1) + K(\rho'_v, \mu_2) \ . \end{aligned}$$

To conclude this section, we recall that we have obtained that, for any $\rho = \rho_u \otimes \rho'_v$ and any $\mu = \mu_1 \otimes \mu_2$,

(i) if ρ_u is centered at u and ρ'_v at v,

(ii) if ρ_u is supported in the Euclidean ball $\mathbb{B}(u, r_U)$ and ρ'_v in $\mathbb{B}(u, r_V)$,

for any s such that

$$|s| \leqslant \frac{1}{(\|u\| + r_U)(\|v\| + r_V)} \ ,$$

with $\mathbb{P}_{\mathbf{X}}$-probability larger than $1 - \exp(-z)$, $\forall (u, v) \in \mathcal{E}_U \times \mathcal{E}_V$,

$$s \left\langle \mathbf{X}, uv^T \right\rangle_F \leqslant K^2 s^2 \mathbb{E}_{\rho_u}[\|U\|^2] \mathbb{E}_{\rho'_v}[\|V\|^2] + K(\rho_u, \mu_1) + K(\rho'_v, \mu_2) + z \ . \quad (5.8)$$

5.5.3 Priors and quantities of interest

We define for ρ_u the truncated Gaussian defined for any $u \in \mathcal{E}_U$, a radius r_U and covariance matrix Γ_U/β_U with β_U to be specified later by the distribution with density

$$f_u(x) = \frac{1}{(\sqrt{2\pi/\beta_U})^n \sqrt{\det(\Gamma_U)} C_U} \exp\left(-\frac{\beta_U}{2} (x-u)^T \Gamma_U^{-1} (x-u) \right) \mathbf{1}_{\{\|x-u\|_2 \leqslant r_U\}} \ ,$$

with normalizing constant $C_U = \mathbb{P}(\mathsf{N}(0, \Gamma_U/\beta_U) \in B(0, r_U))$. Hereafter, we fix $r_U = 2\sqrt{\mathrm{Tr}(\Gamma_U)/\beta_U}$ so $1/2 \leqslant C_U \leqslant 1$. This distribution is centered at u and supported in $\mathbb{B}(u, r_U)$.

We define μ_1 as the Gaussian distribution $\mathrm{N}(0, \Gamma_U/\beta_U)$. We define similarly ρ'_v as truncated Gaussian distributions and μ_2 as the Gaussian distribution $\mathrm{N}(0, \Gamma_V/\beta_V)$.

The computations of Section 5.4.1 show that, with these choices,

$$\begin{aligned}
\mathbb{E}_{\rho_u}[\|U\|^2] &\leqslant \frac{\mathrm{Tr}(\Gamma_U)}{\beta_U} + \|\Gamma_U\| \ , \\
\mathbb{E}_{\rho'_v}[\|V\|^2] &\leqslant \frac{\mathrm{Tr}(\Gamma_V)}{\beta_V} + \|\Gamma_V\| \ , \\
K(\rho_u, \mu_1) &\leqslant \frac{1}{e} + \frac{\beta_U}{2} \|t\|^2_{\Gamma_U^{-1}} \leqslant \frac{\beta_U + 1}{2} \ , \\
K(\rho'_u, \mu_2) &\leqslant \frac{\beta_V + 1}{2} \ .
\end{aligned}$$

To conclude this section, we get that, with our choices, for any s such that

$$|s| \leqslant \left[b \sqrt{\frac{\mathrm{Tr}(\Gamma_U)}{\beta_U} + \|\Gamma_U\|} \sqrt{\frac{\mathrm{Tr}(\Gamma_V)}{\beta_V} + \|\Gamma_V\|} \right]^{-1} \ ,$$

we have

$$\left\langle \mathbf{X}, uv^T \right\rangle_F \leqslant K^2 s \left(\frac{\mathrm{Tr}(\Gamma_U)}{\beta_U} + \|\Gamma_U\| \right) \left(\frac{\mathrm{Tr}(\Gamma_V)}{\beta_V} + \|\Gamma_V\| \right) + \frac{\beta_U + \beta_V}{2s} + \frac{1+z}{s} \ .$$

5.5.4 Optimization

Let $\alpha \geqslant b$ and

$$s = \left[\alpha\sqrt{\frac{\mathrm{Tr}(\Gamma_U)}{\beta_U} + \|\Gamma_U\|}\sqrt{\frac{\mathrm{Tr}(\Gamma_V)}{\beta_V} + \|\Gamma_V\|}\right]^{-1}$$

so

$$\left\langle \mathbf{X}, uv^T \right\rangle_F \leqslant \sqrt{\left(\frac{\mathrm{Tr}(\Gamma_U)}{\beta_U} + \|\Gamma_U\|\right)\left(\frac{\mathrm{Tr}(\Gamma_V)}{\beta_V} + \|\Gamma_V\|\right)}\left(\frac{K^2}{\alpha} + \alpha\left(\frac{\beta_U + \beta_V}{2} + 1 + z\right)\right) \ .$$

Let us introduce the following effective ranks

$$r(\Gamma_U) = \frac{\mathrm{Tr}(\Gamma_U)}{\|\Gamma_U\|}, \qquad r(\Gamma_V) = \frac{\mathrm{Tr}(\Gamma_V)}{\|\Gamma_V\|} \ .$$

so

$$\left\langle \mathbf{X}, uv^T \right\rangle_F \leqslant \sqrt{\|\Gamma_U\|\|\Gamma_V\|\left(\frac{r(\Gamma_U)}{\beta_U} + 1\right)\left(\frac{r(\Gamma_V)}{\beta_V} + 1\right)}\left(\frac{K^2}{\alpha} + \alpha\left(\frac{\beta_U + \beta_V}{2} + 1 + z\right)\right) \ .$$

Choose $\beta_U = r(\Gamma_U)$, $\beta_V = r(\Gamma_V)$ to get

$$\left\langle \mathbf{X}, uv^T \right\rangle_F \leqslant \sqrt{\|\Gamma_U\|\|\Gamma_V\|}\left(\frac{2K^2}{\alpha} + \alpha\left(r(\Gamma_U) + r(\Gamma_V) + 2(1+z)\right)\right) \ .$$

We now distinguish between two situations: Either $r(\Gamma_U) + r(\Gamma_V) + 2(1+z) \leqslant 2(K/b)^2$ or $r(\Gamma_U) + r(\Gamma_V) + 2(1+z) > 2(K/b)^2$.

We first consider the case where $r(\Gamma_U) + r(\Gamma_V) + 2(1+z) \leqslant 2(K/b)^2$, so

$$b \leqslant \frac{\sqrt{2}K}{\sqrt{r(\Gamma_U) + r(\Gamma_V) + 2(1+z)}} \ .$$

We choose thus

$$\alpha = \frac{\sqrt{2}K}{\sqrt{r(\Gamma_U) + r(\Gamma_V) + 2(1+z)}} \geqslant b \ .$$

so

$$\left\langle \mathbf{X}, uv^T \right\rangle_F \leqslant K\sqrt{2\|\Gamma_U\|\|\Gamma_V\|}\sqrt{r(\Gamma_U) + r(\Gamma_V) + 2(1+z)} \ .$$

We now consider the case where $r(\Gamma_U) + r(\Gamma_V) + 2(1+z) > 2(K/b)^2$ and choose $\alpha = b$ to get

$$\begin{aligned}\left\langle \mathbf{X}, uv^T \right\rangle_F &\leqslant \sqrt{\|\Gamma_U\|\|\Gamma_V\|}\left(\frac{2K^2}{b} + b\left(r(\Gamma_U) + r(\Gamma_V) + 2(1+z)\right)\right) \\ &\leqslant 2b\sqrt{\|\Gamma_U\|\|\Gamma_V\|}\left(r(\Gamma_U) + r(\Gamma_V) + 2(1+z)\right) \ .\end{aligned}$$

We conclude that, in all cases, we have, if $\mathbf{X}$ satisfies the following assumption: $\forall u, v : \|u\|_2 \vee \|v\|_2 \leqslant 1$ and $\forall s : |s| < 1/b$,

$$\mathbb{E}[\exp(s \left\langle \mathbf{X}, uv^T \right\rangle_F)] \leqslant \exp(s^2 K^2) \ , \tag{5.9}$$

then, for all $z > 0$, with probability $1 - \exp(-z)$,

$$\sup_{u \in \mathcal{E}_U, v \in \mathcal{E}_V} u^T \mathbf{X} v \leqslant \sqrt{\|\Gamma_U\| \|\Gamma_V\|} \left(K\mathcal{C}(U, V, z) \vee b\mathcal{C}(U, V, z)^2 \right) \ , \tag{5.10}$$

where $r(\Gamma_U) = \mathrm{Tr}(\Gamma_U)/\|\Gamma_U\|$, $r(\Gamma_V) = \mathrm{Tr}(\Gamma_V)/\|\Gamma_V\|$ and

$$\mathcal{C}(U, V, z) = \sqrt{2(r(\Gamma_U) + r(\Gamma_V)) + 4(1 + z)}$$
$$\mathcal{E}_U = \{u \in \mathbb{R}^n : \|\Gamma_U^{-1/2} u\|_2 \leqslant 1\}, \qquad \mathcal{E}_V = \{v \in \mathbb{R}^p : \|\Gamma_V^{-1/2} v\|_2 \leqslant 1\} \ .$$

5.5.5 Application to quadratic processes

In this section, we are interested in the following problem. Let $X_1, \ldots, X_n$ denote i.i.d. random vectors in $\mathbb{R}^d$ such that $\mathbb{E}[X_i X_i^T] = \Sigma$,

$$\forall i \in \{1, \ldots, n\}, \forall u \in \mathbb{R}^d, \qquad \| \langle u, X_i \rangle \|_{\psi_2} \leqslant \|u\|_2 \ .$$

Let Γ_U and Γ_V denote two non-singular positive semi-definite matrices and let

$$\mathcal{E}_U = \{x \in \mathbb{R}^d : \|\Gamma_U^{-1/2} x\| \leqslant 1\}, \qquad \mathcal{E}_V = \{x \in \mathbb{R}^d : \|\Gamma_V^{-1/2} x\| \leqslant 1\} \ .$$

We are interested in the random variable

$$Z = \sup_{u \in \mathcal{E}_U, v \in \mathcal{E}_V} \frac{1}{n} \sum_{i=1}^n \{\langle X_i, u \rangle \langle X_i, v \rangle - u^T \Sigma v\} \ .$$

We define the matrix

$$\mathbf{X} = \frac{1}{n} \sum_{i=1}^n \{X_i X_i^T - \Sigma\} \ ,$$

so $Z = \sup_{u \in \mathcal{E}_U, v \in \mathcal{E}_V} u^T \mathbf{X} v$. We have, for any $u, v \in \mathbb{R}^d$,

$$\begin{aligned} \| \langle X_i, u \rangle \langle X_i, v \rangle \|_{\psi_1} &\leqslant \frac{1}{2} \big(\alpha \| \langle X_i, u \rangle^2 \|_{\psi_1} + \frac{1}{\alpha} \| \langle X_i, v \rangle^2 \|_{\psi_1} \big) \\ &= \frac{1}{2} \big(\alpha \| \langle X_i, u \rangle \|_{\psi_2}^2 + \frac{1}{\alpha} \| \langle X_i, v \rangle \|_{\psi_2}^2 \big) \\ &\leqslant \|u\|_2 \|v\|_2 \ . \end{aligned}$$

It follows that

$$\|u^T (X_i X_i^T - \Sigma) v\|_{\psi_1} \leqslant 2 \|u\|_2 \|v\|_2 \ .$$

Thus by Theorem 22, there exists an absolute constant $C > 0$ such that, for any $|s| \leqslant 1/C\|u\|_2\|v\|_2$,

$$\mathbb{E}[\exp(su^T(X_iX_i^T - \Sigma)v)] \leqslant \exp(C^2s^2\|u\|_2^2\|v\|_2^2) \ .$$

By independence, it follows therefore that, for any $|s| \leqslant 1/C\|u\|_2\|v\|_2$,

$$\mathbb{E}[\exp(su^T(\sum_{i=1}^n\{X_iX_i^T - \Sigma\})v)] \leqslant \exp(nC^2s^2\|u\|_2^2\|v\|_2^2) \ .$$

This can be written, for any $|s| \leqslant 1/C\|u\|_2\|v\|_2$,

$$\mathbb{E}[\exp(nsu^T\mathbf{X}v)] \leqslant \exp(nC^2s^2\|u\|_2^2\|v\|_2^2) \ .$$

Fix now u and v such that $\|u\|_2 \vee \|v\|_2 \leqslant 1$. The following condition implies that, for any $|s| \leqslant n/C$,

$$\mathbb{E}[\exp(su^T\mathbf{X}v)] \leqslant \exp\left(\frac{C^2s^2}{n}\right) \ .$$

In other words, the matrix $\mathbf{X}$ satisfies condition (5.7) with $b = C/n$ and $K = C/\sqrt{n}$. It follows from the previous section that we have therefore, for all $z > 0$, with probability $1 - \exp(-z)$,

$$\sup_{u\in\mathcal{E}_U, v\in\mathcal{E}_V} u^T\mathbf{X}v \leqslant C\sqrt{\|\Gamma_U\|\|\Gamma_V\|}\left(\mathcal{C}(U,V,z) \vee \mathcal{C}(U,V,z)^2\right) \ ,$$

where

$$\mathcal{C}(U,V,z) = \sqrt{\frac{r(\Gamma_U) + r(\Gamma_V) + 1 + z}{n}} \ ,$$
$$\mathcal{E}_U = \{u \in \mathbb{R}^n : \|\Gamma_U^{-1/2}u\|_2 \leqslant 1\}, \qquad \mathcal{E}_V = \{v \in \mathbb{R}^p : \|\Gamma_V^{-1/2}v\|_2 \leqslant 1\} \ .$$

Chapter 6

Upper bounds on random processes

In this chapter, we provide various chaining bound used to upper bound suprema of processes with sub-Gaussian increments. Let $\{X_t, t \in T\}$ denotes a random process, that is a collection of random variables indexed by a separable set T. We assume that the process X_t is centered, i.e. that $\mathbb{E}[X_t] = 0$, and that it has sub-Gaussian increments.

Definition 50. *Assume that d is a distance on T. The process $\{X_t, t \in T\}$ is said to have sub-Gaussian increments with respect to d if there exists K such that, for any $s, t \in T$,*

$$\|X_s - X_t\|_{\psi_2} \leqslant K d(s,t) \ .$$

Remark 51. *The standard Gaussian process has sub-Gaussian increments with respect to the Euclidean distance on $\mathbb{R}^p$. This distance will play a particularly important role in this chapter.*

The purpose of this chapter is to give the main known methods to obtain upper bounds on the following extension of the Gaussian width of T:

$$\mathbb{E}[\sup_{t \in T} X_t] \ .$$

To avoid measurability issues, we focus on cases where T is separable so

$$\mathbb{E}[\sup_{t \in T} X_t] = \sup_{T_0 \subset T, |T_0| < \infty} \mathbb{E}[\sup_{t \in T_0} X_t] \ ,$$

and therefore, without loss of generality, we only consider cases where T is finite. We derive these bounds from *chaining arguments*. We start with Dudley's argument, which is a multi-scale refinement of the ϵ-net argument we have seen to obtain upper bounds on linear processes over the Euclidean ball. We show a nice application of Dudley's bound to suprema of boolean

functions where it can be used to bound the expected supremum using Vapnik Chervonenkis dimension of sets of Boolean functions. We conclude the chapter with the generic chaining bound to prove deviation inequalities for suprema of random processes with sub-Gaussian increments and discuss some applications to statistical learning theory and bounds on quadratic processes.

6.1 Dudley's inequality

Dudley's inequality is a multi-scale refinement of the ϵ-net argument. This argument involves the notion of covering number of a metric set T.

Definition 52. *Assume that T is precompact. The covering number $\mathcal{N}(T, d, \epsilon)$ is the smallest number of balls of radius ϵ necessary to cover T.*

The ϵ-net argument can be used to bound for example to bound the linear process $X_t = \langle X, t \rangle$ over the unit Euclidean ball $T = \mathbb{B}_2$. It is based on the decomposition

$$X_t = (X_t - X_{\pi(t)}) + X_{\pi(t)} \ ,$$

valid for any $t \in T$, where $\pi(t)$ is a point in an ϵ-net N_ϵ of T such that $d(t, \pi(t)) \leqslant \epsilon$. Using this decomposition, we obtained that, for any $\epsilon \in (0, 1)$,

$$\sup_{t \in T} X_t \leqslant \max_{t \in N_\epsilon} X_t + \epsilon \sup_{t \in T} X_t \ , \tag{6.1}$$

so

$$\sup_{t \in T} X_t \leqslant \frac{1}{1 - \epsilon} \max_{t \in N_\epsilon} X_t \ .$$

In particular, if X_t has sub-Gaussian increments, we deduce that, for any $z > 0$ and any $t_0 \in T$,

$$\begin{aligned} \mathbb{P}(\sup_{t \in T} X_t - X_{t_0} > z) &\leqslant \mathbb{P}(\max_{t \in N_\epsilon} X_t - X_{t_0} > (1 - \epsilon)t) \\ &\leqslant |\mathcal{N}(T, d, \epsilon)| \exp\left(- \frac{(1 - \epsilon)^2 z^2}{K^2 \max_{t \in T} d(t_0, t)^2} \right) \ . \end{aligned}$$

Rearranging the terms and integrating gives

$$\mathbb{E}[\sup_{t \in T} X_t] \lesssim \frac{K \mathrm{diam}(T)}{1 - \epsilon} \sqrt{\log(|\mathcal{N}(T, d, \epsilon)|)} \ ,$$

where $\mathrm{diam}(T) = \sup_{t \in T} d(t, t_0)$. Dudley's inequality extends this bound to processes that do not necessarily satisfy a key inequality like (6.1) and possibly refines it by considering the decomposition of T at all scales $\epsilon > 0$.

Theorem 53 (Dudley's inequality)**.** *Let $\{X_t, t \in T\}$ denote a centered process with sub-Gaussian increments, then*

$$\mathbb{E}[\sup_{t\in T}|X_t|] \leqslant CK\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\log \mathcal{N}(T,d,2^{-k})}$$
$$\leqslant CK\int_0^{+\infty}\sqrt{\log \mathcal{N}(T,d,\epsilon)}\mathrm{d}\epsilon \ .$$

Remark 54. *The second result follows from the first one and the comparison between series and integral that holds by monotonicity of the map $\epsilon \mapsto \mathcal{N}(T,d,\epsilon)$.*

Remark 55. *Dudley's inequality gives a result in expectation. It is interesting, as an exercise, to adapt slightly the proof to show the following version of the result: for any $z > 0$,*

$$\mathbb{P}\left(\sup_{s,t\in T}|X_t - X_s| > CK\left(\int_0^{+\infty}\sqrt{\log \mathcal{N}(T,d,\epsilon)}\mathrm{d}\epsilon + diam(T)z\right)\right) \leqslant \exp(-z^2) \ ,$$

where $diam(T) = \sup_{s,t\in T} d(s,t)$.

Proof. We prove the first bound. Let k_1 and k_0 be respectively the smallest k such that $\mathcal{N}(T,d,2^{-k}) = |T|$ and the largest k such that $\mathcal{N}(T,d,2^{-k}) = 1$. Let $t_0 \in T$ such that $B(t_0, 2^{-k_0}) \supset T$. For any $k \in \{k_0, \dots, k_1\}$, let T_k denote a set with cardinality $\mathcal{N}(T,d,2^{-k})$ such that $\cup_{u\in T_k} B(u, 2^{-k}) \supset T$. For any $t \in T$ and $k \in \{k_0, \dots, k_1\}$, let $\pi_k(t) \in T_k$ such that $d(t, \pi_k(t)) \leqslant 2^{-k}$. In particular, $X_{\pi_{k_1}(t)} = t$.

As $\mathbb{E}[X_{t_0}] = 0$, we have

$$\mathbb{E}[\sup_{t\in T} X_t] = \mathbb{E}[\sup_{t\in T}(X_t - X_{t_0})] \ .$$

The "chaining argument" is to write the difference $X_t - X_{t_0}$ as a chain

$$X_t - X_{t_0} = \sum_{k=k_0}^{k_1-1} X_{\pi_{k+1}(t)} - X_{\pi_k(t)} \ .$$

Using this decomposition, we obtain

$$\sup_{t\in T}(X_t - X_{t_0}) \leqslant \sum_{k=k_0}^{k_1-1}\sup_{t\in T}(X_{\pi_{k+1}(t)} - X_{\pi_k(t)}) \ . \tag{6.2}$$

Now, there is at most $\mathcal{N}(T,d,2^{-k})\mathcal{N}(T,d,2^{-(k+1)}) \leqslant \mathcal{N}(T,d,2^{-(k+1)})^2$ random variables $(X_{\pi_{k+1}(t)} - X_{\pi_k(t)})$ when t describes T and all these variables satisfy

$$\|X_{\pi_{k+1}(t)} - X_{\pi_k(t)}\|_{\psi_2} \leqslant \|X_{\pi_{k+1}(t)} - X_t\|_{\psi_2} + \|X_t - X_{\pi_k(t)}\|_{\psi_2} \leqslant CK2^{-(k+1)} \ .$$

Using a union bound, it follows therefore that, with probability at least $1 - 2\exp(-u)$,

$$\sup_{t\in T}(X_{\pi_{k+1}(t)} - X_{\pi_k(t)}) \leqslant CK2^{-(k+1)}\sqrt{\log(\mathcal{N}(T,d,2^{-(k+1)})) + u} \ .$$

Integrating this upper bound shows that

$$\mathbb{E}\left[\sup_{t\in T}(X_{\pi_{k+1}(t)} - X_{\pi_k(t)})\right] \leqslant CK2^{-(k+1)}\sqrt{\log(\mathcal{N}(T,d,2^{-(k+1)}))} \ .$$

It follows therefore from (6.2) that

$$\begin{aligned}\mathbb{E}\left[\sup_{t\in T}(X_t - X_{t_0})\right] &\leqslant CK\sum_{k=k_0}^{k_1-1} 2^{-(k+1)}\sqrt{\log\mathcal{N}(T,d,2^{-(k+1)})} \ . \\ &\leqslant CK\sum_{k\in\mathbb{Z}} 2^{-(k+1)}\sqrt{\log\mathcal{N}(T,d,2^{-(k+1)})} \ .\end{aligned}$$

We conclude by saying that (i), $\mathbb{E}\big[\sup_{t\in T}(X_t - X_{t_0})\big] = \mathbb{E}\big[\sup_{t\in T} X_t\big]$ and (ii) that the same argument can can be used to bound $\mathbb{E}\big[\sup_{t\in T}|X_t|\big] = \mathbb{E}\big[\sup_{t\in T}\max\{X_t, -X_t\}\big]$. □

6.2 VC dimension

In this section, we apply Dudley's inequality to processes indexed by Boolean functions and link Dudley's integral with the perhaps more familiar notion of Vapnik Chervonenkis complexity for these classes of functions. Let thus $\mathcal{F}$ denote a class of Boolean functions $f : \Omega \to \{0,1\}$ and, for the sake of completeness, recall the definition of VC dimension of $\mathcal{F}$.

Definition 56. *A set $\Lambda \subset \Omega$ is said* shattered *by $\mathcal{F}$ if any boolean function $f : \Lambda \to \{0,1\}$ can be obtained as a restriction of some $g \in \mathcal{F}$.*

The VC-dimension of $\mathcal{F}$ is the largest cardinality of a set $\Lambda \subset \Omega$ shattered by $\mathcal{F}$.

6.2.1 Examples

Let us recall here the classical example of Half spaces encoded in the set of Boolean functions

$$\mathcal{F} = \{x \in \mathbb{R}^p \mapsto \mathbf{1}_{\{\langle u,x\rangle > 0\}}, u \in \mathbb{S}_p\} \ .$$

We will show that

$$\mathrm{VC}(\mathcal{F}) = p \ .$$

To prove this result, we show first that there exists a set of p vectors in $\mathbb{R}^p$ that is shattered by $\mathcal{F}$. Let $e_1, \ldots, e_p$ denote the canonical basis of $\mathbb{R}^p$ and let $f : \{e_1, \ldots, e_p\} \to \{0, 1\}$. We consider the vector

$$u = \frac{1}{\sqrt{p}} \sum_{i=1}^{p} (2f(e_i) - 1)e_i \in \mathbb{S}_p \ .$$

Then, $\langle u, e_i \rangle = (2f(e_i) - 1)/\sqrt{p} > 0$ iff $f(e_i) = 1$, therefore, for any $i \in \{1, \ldots, p\}$, $f(e_i) = \mathbf{1}_{\{\langle u, e_i \rangle > 0\}}$, so f is the restriction of $x \mapsto \mathbf{1}_{\{\langle u, x \rangle > 0\}}$ on $\Lambda = \{e_1, \ldots, e_p\}$. This set is shattered by $\mathcal{F}$, so $\mathrm{VC}(\mathcal{F}) \geqslant p$.

Next, we have to show that any set of $p + 1$ vectors in $\mathbb{R}^p$ cannot be shattered by $\mathcal{F}$. Let $u_1, \ldots, u_{p+1}$ denote $p + 1$ vectors in $\mathbb{R}^p$. The family is linearly dependent, so w.l.o.g., we can assume that

$$u_{p+1} = \sum_{i=1}^{p} \beta_i u_i \ .$$

In this case, we define the boolean function

$$f(u_i) = \mathbf{1}_{\{\beta_i \leqslant 0\}}, \quad \forall i \in \{1, \ldots, p\}, \qquad f(u_{p+1}) = 1 \ .$$

If $u_1, \ldots, u_{p+1}$ was shattered by $\mathcal{F}$, there would exist $x \in \mathbb{R}^p$ such that f is the restriction of $u \mapsto \mathbf{1}_{\{\langle u, x \rangle > 0\}}$ to $u_1, \ldots, u_{p+1}$. In particular, for any $i \in \{1, \ldots, p\}$, $\langle u_i, x \rangle > 0$ iff $\beta_i \leqslant 0$, so

$$\forall i \in \{1, \ldots, p\}, \qquad \beta_i \langle x, u_i \rangle \leqslant 0 \ ,$$

and therefore

$$\langle u_{p+1}, x \rangle = \sum_{i=1}^{p} \beta_i \langle u_i, x \rangle \leqslant 0 \ .$$

On the other hand, as f is the restriction of $u \mapsto \mathbf{1}_{\{\langle u, x \rangle > 0\}}$ and $f(u_{p+1}) = 1$, we should have

$$\langle u_{p+1}, x \rangle > 0 \ .$$

This is absurd, so $u_1, \ldots, u_{p+1}$ cannot be shattered by $\mathcal{F}$. As this is true for any set of $p + 1$ vectors, we can conclude that $\mathrm{VC}(\mathcal{F}) \leqslant p$.

6.2.2 Pajor's lemma

As $\mathcal{F}$ shatters a set $\Lambda \subset \Omega$ with cardinality $\mathrm{VC}(\mathcal{F})$, it follows that $|\mathcal{F}| \geqslant 2^{\mathrm{VC}(\mathcal{F})}$.

Pajor's Lemma provides an upper bound on $|\mathcal{F}|$ using the number of sets shattered by $\mathcal{F}$ when Ω is finite.

Lemma 57 (Pajor's Lemma)**.** *Let $\mathcal{F}$ denote a set of Boolean functions defined on a finite set Ω. Then,*

$$|\mathcal{F}| \leqslant |\{\Lambda \subset \Omega : \Lambda \textit{ is shattered by } \mathcal{F}\}| \ .$$

Proof. We proceed by induction on the cardinality of Ω. As the result is trivial when $|\Omega| = 0$ (the empty set is always shattered by $\mathcal{F}$), we can assume that Pajor's Lemma is true for any set of cardinality n and consider a set Ω with cardinality $n+1$. Let $x_0 \in \Omega$ and define

$$\mathcal{F}_0 = \{f \in \mathcal{F} : f(x_0) = 0\}, \qquad \mathcal{F}_1 = \{f \in \mathcal{F} : f(x_0) = 1\} \ .$$

Then, we obviously have $|\mathcal{F}| = |\mathcal{F}_0| + |\mathcal{F}_1|$.

Let $\mathcal{S} = \{\Lambda \subset \Omega : \Lambda \text{ is shattered by } \mathcal{F}\}$. By our induction hypothesis, for any $i \in \{0, 1\}$,

$$|\mathcal{F}_i| \leqslant |\mathcal{S}_i|, \qquad \text{where} \qquad \mathcal{S}_i = \{\Lambda \subset \Omega \setminus \{x_0\} : \Lambda \text{ is shattered by } \mathcal{F}_i\} \ .$$

We have $\mathcal{S}_0 \cup \mathcal{S}_1 \subset \mathcal{S}$ and

$$\sum_{i=0}^{1} |\mathcal{S}_i| = |\mathcal{S}_0 \cup \mathcal{S}_1| + |\mathcal{S}_0 \cap \mathcal{S}_1| \ .$$

We build now two injections, one from $\mathcal{S}_0 \cup \mathcal{S}_1$ to $\mathcal{S}$ and the other one from $\mathcal{S}_0 \cap \mathcal{S}_1$ to $\mathcal{S}$, with disjoint images, from which we can conclude that $|\mathcal{S}_0 \cup \mathcal{S}_1| + |\mathcal{S}_0 \cap \mathcal{S}_1| \leqslant |\mathcal{S}|$. We call $\varphi_\cup$ the first injection and $\varphi_\cap$ the second.

1. If $\Lambda \in \mathcal{S}_0 \setminus \mathcal{S}_1$ or $\Lambda \in \mathcal{S}_1 \setminus \mathcal{S}_0$, let $\varphi_\cup(\Lambda) = \Lambda$.

2. If $\Lambda \in \mathcal{S}_0 \cap \mathcal{S}_1$, for any $g : \Lambda \to \{0, 1\}$, there exists $f_0 \in \mathcal{S}_0$ and $f_1 \in \mathcal{S}_1$ such that g is the restriction of both f_0 and f_1. It follows that both Λ and $\Lambda \cup \{x_0\}$ belong to $\mathcal{S}$. We define therefore $\varphi_\cup(\Lambda) = \Lambda$ and $\varphi_\cap(\Lambda) = \Lambda \cup \{x_0\}$.

As anticipated, we conclude therefore that

$$|\mathcal{S}| \geqslant |\mathcal{S}_0 \cup \mathcal{S}_1| + |\mathcal{S}_0 \cap \mathcal{S}_1| = |\mathcal{S}_0| + |\mathcal{S}_1| \geqslant |\mathcal{F}_0| + |\mathcal{F}_1| = |\mathcal{F}| \ .$$

□

We conclude this section with Sauer-Shelah Lemma, that gives a bound on the growth of cardinality of sets of Boolean functions on finite sets with fixed VC dimensions.

Lemma 58 (Sauer-Shelah Lemma)**.** *Let $\mathcal{F}$ denote a set of Boolean functions on an n-points set Ω, with $VC(\mathcal{F}) = d$. Then*

$$|\mathcal{F}| \leqslant \sum_{k=0}^{d} \binom{n}{k} \leqslant \left(\frac{en}{d}\right)^d \ .$$

Proof. By Pajor's Lemma

$$\begin{aligned}|\mathcal{F}| &\leqslant |\{\Lambda \subset \Omega : \Lambda \text{ is shattered by } \mathcal{F}\}| \\ &= \sum_{k=0}^{n} |\{\Lambda \subset \Omega : \Lambda \text{ is shattered by } \mathcal{F} \text{ and } |\Lambda| = k\}| \ .\end{aligned}$$

Now by definition of definition of VC dimension, the cardinalities in the last bound are null for any $k > d$. Hence,

$$\begin{aligned}|\mathcal{F}| &= \sum_{k=0}^{d} |\{\Lambda \subset \Omega : \Lambda \text{ is shattered by } \mathcal{F} \text{ and } |\Lambda| = k\}| \\ &\leqslant \sum_{k=0}^{d} |\{\Lambda \subset \Omega : |\Lambda| = k\}| \\ &= \sum_{k=0}^{d} \binom{n}{k} \ .\end{aligned}$$

For the second inequality, as $d/n \leqslant 1$, we have

$$\sum_{k=0}^{d} \binom{n}{k} \left(\frac{d}{n}\right)^d \leqslant \sum_{k=0}^{d} \binom{n}{k} \left(\frac{d}{n}\right)^k \leqslant \sum_{k=0}^{n} \binom{n}{k} \left(\frac{d}{n}\right)^k = \left(1 + \frac{d}{n}\right)^n \leqslant \mathrm{e}^d \ .$$

□

6.3 Covering numbers and VC dimension

In this section, we prove an upper bound on covering numbers of sets of Boolean functions using the VC dimension of this set. We apply this bound to derive a classical bound for ERM in classification.

6.3.1 Bounding covering numbers by VC dimension

The covering numbers of sets of Boolean functions can be bounded using the VC dimension as shown by the following result.

Theorem 59 (Covering via VC dimension)**.** *Let $\mathcal{F}$ be a class of Boolean functions on a probability space Ω, μ with VC dimension d. Then, for every $\epsilon \in (0, 1)$,*

$$\mathcal{N}(\mathcal{F}, L^2(\mu), \epsilon) \leqslant \left(\frac{2}{\epsilon}\right)^{Cd} \ .$$

The strength of this result is that it holds *for any probability measure μ.* The proof will be based on the following lemma.

Lemma 60 (Dimension reduction lemma)**.** *Let $\mathcal{F}$ denote a set of N Boolean functions on a probability space (Ω, μ). Assume that, for any $f, g \in \mathcal{F}$, $\|f - g\|_{L^2(\mu)} > \epsilon$. Then, there exist n points $\{x_1, \dots, x_n\}$ in Ω such that*

$$\forall f \neq g \in \mathcal{F}, \quad \frac{1}{n} \sum_{i=1}^{n} (f-g)^2(x_i) > (\epsilon/2)^2, \qquad n \leqslant C\epsilon^{-4} \log N \ .$$

Proof of the dimension reduction lemma. Let $X, X_1, \dots, X_n$ denote i.i.d. random variables with common distribution $\mathbb{P}$. Fix $f \neq g$ in $\mathcal{F}$ and let $h = (f-g)^2$. We have

$$\|h(X) - \mathbb{E}[h(X)]\|_{\psi_2} \leqslant C\|h(X)\|_{\psi_2} \leqslant C\|h\|_\infty \leqslant C \ .$$

By Theorem 9, it follows that

$$\| \sum_{i=1}^{n} (h(X_i) - \mathbb{E}[h(X)]) \|_{\psi_2} \leqslant C\sqrt{n} \ ,$$

thus

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} (h(X_i) - \mathbb{E}[h(X)]) \right| > \frac{\epsilon^2}{4} \right) \leqslant \exp(-Cn\epsilon^4) \ .$$

By a union bound, it follows therefore that

$$\mathbb{P}\left(\exists f \neq g \in \mathcal{F} : \ \left| \frac{1}{n} \sum_{i=1}^{n} ((f-g)^2(X_i) - \|f-g\|^2_{L^2(\mu)}) \right| > \frac{\epsilon^2}{4} \right) \leqslant N^2 \exp(-Cn\epsilon^4) \ .$$

If $n = C\epsilon^{-4} \log N$ for a large enough constant C, it follows therefore that

$$\mathbb{P}\left(\exists f \neq g \in \mathcal{F} : \ \left| \frac{1}{n} \sum_{i=1}^{n} ((f-g)^2(X_i) - \|f-g\|^2_{L^2(\mu)}) \right| > \frac{\epsilon^2}{4} \right) < 1 \ .$$

In other words, there exists at least a configuration of the X_i such that, for any $f \neq g$ in $\mathcal{F}$,

$$\frac{1}{n} \sum_{i=1}^{n} ((f-g)^2(X_i) > \|f-g\|^2_{L^2(\mu)} - \frac{\epsilon^2}{4} \geqslant \frac{3\epsilon^2}{4} \ .$$

□

We can now turn to the proof of the covering via VC dimension's theorem.

Proof of the covering via VC dimension's theorem. Let $\epsilon > 0$ and $\mathcal{F}_\epsilon$ denote an ϵ-separated set in $\mathcal{F}$ with maximal cardinality. Then, $\mathcal{F}_\epsilon$ is an ϵ-net of $\mathcal{F}$ (otherwise, $\mathcal{F}_\epsilon$ would not have maximal size), therefore $N = |\mathcal{F}_\epsilon| \geqslant |\mathcal{N}(\mathcal{F}, L^2, \epsilon)|$.

The dimension reduction lemma applied to $\mathcal{F}_\epsilon$ shows that there exist $n = C\epsilon^{-4}\log N$ points $\{x_1, \dots, x_n\}$ and n distinct Boolean functions in $\mathcal{F}_\epsilon$ obtained as restrictions of functions in $\mathcal{F}$. By Sauer-Shellah lemma applied to $\mathcal{F}_\epsilon$ and $\Omega = \{x_1, \dots, x_n\}$, we get, with $d_\epsilon = \mathrm{VC}(\mathcal{F}_\epsilon)$,

$$N \leqslant \left(\frac{en}{d_\epsilon}\right)^{d_\epsilon} \leqslant \left(\frac{C\epsilon^{-4}\log N}{d_\epsilon}\right)^{d_\epsilon} \leqslant \left(2C\epsilon^{-4}\right)^{d_\epsilon}\sqrt{N} \ .$$

To get the last inequality, we used that

$$\frac{\log N}{2d_\epsilon} = \log(N^{1/2d_\epsilon}) \leqslant N^{1/2d_\epsilon} \ .$$

We finally get, as $d_\epsilon \leqslant d$,

$$|\mathcal{N}(\mathcal{F}, L^2, \epsilon)| \leqslant N \leqslant \left(2C\epsilon^{-4}\right)^{2d_\epsilon} \leqslant 2^{Cd}\epsilon^{-8d} \ .$$

□

Let us now consider processes $(X_f)_{f\in\mathcal{F}}$ indexed by a separable set $\mathcal{F}$ of Boolean functions. If this process has sub-Gaussian increments, we can bound $\mathbb{E}[\sup_{f\in\mathcal{F}} X_f]$ using Dudley's integral. Then, we can use the upper bound on the covering numbers by VC dimension to bound Dudley's integral using VC dimension. The precise result is gathered in the following theorem.

Theorem 61. *Let $\mathcal{F}$ denote a set of Boolean functions and $(X_f)_{f\in\mathcal{F}}$ denote a random process indexed by $\mathcal{F}$ such that, for any f and g in $\mathcal{F}$,*

$$\|X_f - X_g\|_{\psi_2} \leqslant K\|f - g\|_{L^2(\mu)} \ ,$$

for some measure μ. Then

$$\mathbb{E}[\sup_{f\in\mathcal{F}} X_f] \leqslant CK\sqrt{VC(\mathcal{F})} \ .$$

Proof. By Dudley's integral bound, we have

$$\mathbb{E}[\sup_{f\in\mathcal{F}} X_f] \leqslant CK\int_0^{+\infty} \sqrt{\log(\mathcal{N}(\mathcal{F}, L^2(\mu), \epsilon)}\mathrm{d}\epsilon \ .$$

Thus by the covering via VC dimension theorem,

$$\mathbb{E}[\sup_{f\in\mathcal{F}} X_f] \leqslant CK\sqrt{\mathrm{VC}(\mathcal{F})}\int_0^1 \sqrt{\log(1/\epsilon)}\mathrm{d}\epsilon \ .$$

□

6.3.2 Application to ERM for classification

In this section, we consider binary classification where we observe i.i.d. couples $(x_i, y_i) \in \mathcal{F} \times \{0,1\}$, $i \in \{1, \dots, n\}$ and, given a set $\mathcal{F}$ of classifiers $f : \mathcal{X} \to \{0,1\}$, we are interested in the ERM

$$\hat{f} \in \operatorname{argmin}_{f \in \mathcal{F}} P_n \ell_f, \qquad \ell_f(x,y) = \mathbb{I}\{y \neq f(x)\} \ .$$

A very rough analyse of this estimator yields, for $f^* \in \operatorname{argmin}_{f \in \mathcal{F}} P\ell_f$,

$$P(\ell_{\hat{f}} - \ell_{f^*}) \leqslant (P - P_n)(\ell_{\hat{f}} - \ell_{f^*}) \leqslant \sup_{f \in \mathcal{F}} (P - P_n)(\ell_f - \ell_{f^*}) \ .$$

The random variables $\ell_f(x_i, y_i) - \ell_{f^*}(x_i, y_i)$ being independent and taking values in $[-1, 1]$, the bounded difference inequality shows that, with probability $1 - \delta$,

$$\sup_{f \in \mathcal{F}} (P - P_n)(\ell_f - \ell_{f^*}) \leqslant \mathbb{E}[\sup_{f \in \mathcal{F}} (P - P_n)(\ell_f - \ell_{f^*})] + C\sqrt{\frac{\log(1/\delta)}{n}} \ .$$

To bound this expectation, we use the symmetrization trick.

Lemma 62 (Symmetrization)**.** *Let $Z_1, \dots, Z_n \in \mathbb{R}^p$ denote i.i.d. random vectors. Then, if $\epsilon_1, \dots, \epsilon_p$ are i.i.d. Rademacher random variables, independent from $Z_1, \dots, Z_n$,*

$$\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^n Z_i - \mathbb{E}[Z_i]\|_\infty] \leqslant 2\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^n \epsilon_i Z_i\|_\infty] \ .$$

Proof of the symmetrization lemma. Let $Z_1', \dots, Z_n'$ denote independent copies of $Z_1, \dots, Z_n$ so $\mathbb{E}[Z_i] = \mathbb{E}[Z_i' | Z_1, \dots, Z_n]$. We have

$$\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^n Z_i - \mathbb{E}[Z_i]\|_\infty] \leqslant \mathbb{E}\big[\|\frac{1}{n}\sum_{i=1}^n [Z_i - Z_i']\|_\infty\big] \ .$$

Now the vector $\sum_{i=1}^n [Z_i - Z_i']$ has the same distribution as $\sum_{i=1}^n \epsilon_i [Z_i - Z_i']$ as shown by a straightforward computation of their Fourier transform. Therefore,

$$\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^n Z_i - \mathbb{E}[Z_i]\|_\infty] \leqslant \mathbb{E}\big[\|\frac{1}{n}\sum_{i=1}^n \epsilon_i [Z_i - Z_i']\|_\infty\big] \leqslant 2\mathbb{E}[\|\frac{1}{n}\sum_{i=1}^n \epsilon_i Z_i\|_\infty] \ .$$

□

We can assume as usual that $\mathcal{F}$ is finite. By the symmetrization lemma,

$$\mathbb{E}[\sup_{f \in \mathcal{F}} (P - P_n)(\ell_f - \ell_{f^*})] = \mathbb{E}[\sup_{f \in \mathcal{F}} (P - P_n)\ell_f] \leqslant 2\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i \ell_f(x_i, y_i)\right|\right] \ .$$

Now conditioning on $\mathcal{D}_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$ and let, for any $f \in \mathcal{F}$, let $Z_f = n^{-1} \sum_{i=1}^n \epsilon_i \ell_f(x_i, y_i)$, we have, by Theorem 9,

$$\|Z_f - Z_g\|_{\psi_2} \leqslant \frac{C}{n} \sqrt{\sum_{i=1}^n (\ell_f - \ell_g)^2(x_i)} = \frac{C}{\sqrt{n}} \|f - g\|_{L^2(\mu_n)} \ ,$$

where μ_n is the uniform distribution on $x_1, \dots, x_n$. By Theorem 61, therefore,

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i \ell_f(x_i, y_i)) \right| | \mathcal{D}_n \right] \lesssim \sqrt{\frac{\mathrm{VC}(\mathcal{F})}{n}} \ .$$

Integrating with respect to the distribution of $\mathcal{D}_n$ yields finally the risk bound for ERM in classification: If $\mathcal{F}$ is a set of classifiers with finite VC dimension and $\hat{f} \in \operatorname{argmin}_{f \in \mathcal{F}} P_n \ell_f$ is the ERM for the $0-1$ loss, for any $\delta \in (0,1)$, with probability $1 - \delta$,

$$P(\ell_{\hat{f}} - \ell_{f^*}) \lesssim \sqrt{\frac{\mathrm{VC}(\mathcal{F})}{n}} + \sqrt{\frac{\log(1/\delta)}{n}} \ .$$

6.4 Generic chaining bound

Definition 63. *Let (T, d) denote a metric space. A sequence $(T_k)_{k \geqslant 0}$ of finite subsets of T is called* admissible *if $|T_k| \leqslant 2^{2^k}$.*

The γ_2 functional of T is defined as

$$\gamma_2(T, d) = \inf_{T_k} \sup_{t \in T} \sum_{k=0}^{+\infty} 2^{k/2} d(t, T_k) \ ,$$

where the infimum is taken over all admissible sequences T_k.

Talagrand's generic chaining theorem shows that the supremum of a random process over a set T with sub-Gaussian increments does not deviate much from the γ_2 functional of T.

Theorem 64 (Talagrand's deviation inequality)**.** *Let $(X_t)_{t \in T}$ denote a centered random process with sub-Gaussian increments with respect to a distance d on T:*

$$\|X_t - X_s\|_{\psi_2} \leqslant K d(s, t), \qquad \forall s, t \in T \ .$$

Let $\Delta(T) = \sup_{s,t \in T} d(s, t)$ and $t_0 \in T$. Then, for any $u > 0$, with probability at least $1 - 2\exp(-u^2)$,

$$\sup_{t \in T} (X_t - X_{t_0}) \leqslant CK(\gamma_2(T, d) + u\Delta(T)) \ .$$

Proof. The proof relies on refinements of Dudley's chaining inequality. First, we can assume that $K = 1$ and T is finite, without loss of generality. Then, we consider an admissible sequence $(T_k)_{k\geqslant 0}$, with $T_0 = \{t_0\}$, and denote, for any $k \geqslant 0$ and $t \in T$, by $\pi_k(t) \in T_k$ a point such that $d(t, \pi_k(t)) = d(t, T_k) := \inf_{u\in T_k} d(t, u)$.

We first fix $t \in T$ and build an increasing sequence k_i such that $k_0 = 0$ and, for any $i \geqslant 1$, k_{i+1} is the first moment such that $d(t, \pi_{k_{i+1}}(t)) \leqslant d(t, \pi_{k_i}(t))/2$. Then we write as in the proof of Dudley's bound

$$X_t - X_{t_0} = \sum_{i=0}^{I} X_{\pi_{k_{i+1}}(t)} - X_{\pi_{k_i}(t)} \ .$$

For any fixed i and t, we have, with probability $1 - 2\exp(-C(2^{k_{i+1}} + u^2))$,

$$|X_{\pi_{k_{i+1}}(t)} - X_{\pi_{k_i}(t)}| \leqslant Cd(\pi_{k_{i+1}}(t), \pi_{k_i}(t))(u + 2^{k_{i+1}/2}) \ .$$

Taking a union bound over all $t \in T$ shows that the same holds for any $t \in T$ with probability

$$1 - 2|T^{k_i}||T^{k_{i+1}}|\exp(-C(2^{k_{i+1}} + u^2)) \geqslant 1 - 2^{2^{k_{i+1}+1}+1}\exp(-C(2^{k_{i+1}} + u^2)) \ .$$

Finally, a union bound over i shows that the same holds for any i and t with probability

$$1 - C\sum_{k=0}^{+\infty} 2^{2^k}\exp(-C(2^k + u^2)) \geqslant 1 - C\exp(-Cu^2) \ .$$

It follows that, with probability $1 - C\exp(-Cu^2)$,

$$\begin{aligned}
|X_t - X_{t_0}| &\leqslant \sum_{i=0}^{I} |X_{\pi_{k_{i+1}}(t)} - X_{\pi_{k_i}(t)}| \\
&\leqslant \sum_{i=0}^{I} Cd(\pi_{k_{i+1}}(t), \pi_{k_i}(t))(u + 2^{k_{i+1}/2}) \\
&\leqslant C\left(d(t, t_0)u + \sum_{i=0}^{I} 2^{k_{i+1}/2} d(t, \pi_{k_i}(t))\right) \\
&\leqslant C\left(\Delta(T)u + \sum_{i=0}^{I} 2^{k_{i+1}/2} d(t, \pi_{k_i}(t))\right) \ .
\end{aligned}$$

To conclude, it remains to show that

$$2^{k_{i+1}/2} d(t, \pi_{k_i}(t)) \leqslant C \sum_{k=k_i}^{k_{i+1}-1} 2^{k/2} d(t, T_k) \ .$$

We have, for any $k \leqslant k_{i+1} - 1$, $d(t, T_k) \geqslant d(t, T_{k_i})/2$, so

$$\sum_{k=k_i}^{k_{i+1}-1} 2^{k/2} d(t, T_k) \geqslant \frac{d(t, T_{k_i})}{2} \frac{2^{k_{i+1}/2} - 2^{k_i/2}}{\sqrt{2} - 1} \geqslant \frac{2^{k_{i+1}/2} d(t, T_{k_i})}{2\sqrt{2}} \ .$$

This concludes the proof of Talagrand's inequality. □

The strength of Talagrand's inequality can be appreciated thanks to Talagrand's majorizing measure theorem, which shows that the γ_2 functional is of the order of the Gaussian width of T.

Theorem 65 (Talagrand's majorizing measure theorem)**.** *Let $(X_t)_{t\in T}$ denote a centered Gaussian process and let $d(s,t) = \|X_s - X_t\|_2$. Then,*

$$c\gamma_2(T, d) \leqslant \mathbb{E}[\sup_{t\in T} X_t] \leqslant C\gamma_2(T, d) \ .$$

This theorem is proved in Chapter 7. We refer the interested reader to it for the details. The majorizing measure theorem directly implies the following reformulation of Talagrand's inequality.

Theorem 66 (Talagrand's deviation inequality in $\mathbb{R}^p$)**.** *Let $(X_t)_{t\in T}$ denote a centered random process with sub-Gaussian increments with respect to the Euclidean distance on $T \subset \mathbb{R}^p$:*

$$\|X_t - X_s\|_{\psi_2} \leqslant K\|s - t\|_2, \qquad \forall s, t \in T \ .$$

Let $\Delta(T) = \sup_{s,t\in T} \|s - t\|_2$ and $t_0 \in T$. Then, for any $u > 0$, with probability at least $1 - 2\exp(-u^2)$,

$$\sup_{t\in T}(X_t - X_{t_0}) \leqslant CK(w(T) + u\Delta(T)) \ .$$

6.5 Application to linear SVM estimators

Consider the binary classification setting where we observe i.i.d. couples (x_i, y_i), $i \in \{1, \dots, n\}$, with $x \sim \mathrm{N}(0, \mathbf{I})$ is a Gaussian vector in $\mathbb{R}^d$, and $y \in \{-1, 1\}$. The linear SVM estimator is defined as

$$\hat{\theta}_\lambda \in \mathrm{argmin}_{\theta\in\mathbb{R}^d} P_n\varphi(-y\langle\theta, x\rangle) + \lambda\|\theta\|_2^2, \qquad \varphi(z) = \max(0, 1 + z) \ .$$

Following the agenda of the first lecture, the analysis of this estimator can be reduced to the one of the supremum of the empirical process

$$\sup_{\theta\in\mathcal{E}}(P_n - P)(\ell_{\theta^*} - \ell_\theta) \ ,$$

where $\mathcal{E}$ is the ellipsoid $\mathcal{E} = \{\theta \in \mathbb{R}^d : \|\theta - \theta^*\|_{\mathbf{S}}^2 \leqslant 1\}$, for some positive symmetric matrix $\mathbf{S}$. In this section, we show Talagrand's theorem can be used to bound from above this process.

Step 1: We first prove that the process

$$X_\theta = (P_n - P)(\ell_{\theta^*} - \ell_\theta) \ ,$$

has sub-Gaussian increments. We have

$$X_\theta - X_{\theta'} = (P_n - P)(\ell_{\theta'} - \ell_\theta) \ .$$

For any $s \in \mathbb{R}$, as φ is 1-Lipschitz,

$$\exp(s^2(\ell_{\theta'}(x,y) - \ell_\theta(x,y))^2) \leqslant \exp(s^2 \left\langle \theta' - \theta, x \right\rangle^2) \ .$$

As $\langle u, x \rangle \sim \mathrm{N}(0,1)$, with $u = (\theta' - \theta)/\|\theta - \theta'\|_2$, we have, for any s for which it makes sense

$$\mathbb{E}[\exp(s^2 \left\langle \theta' - \theta, X \right\rangle^2)] = \frac{1}{1 - 2s^2\|\theta - \theta'\|_2^2} \ .$$

It follows that

$$\|\ell_{\theta'}(x,y) - \ell_\theta(x,y)\|_{\psi_2} \leqslant 2\|\theta - \theta'\|_2 \ .$$

Therefore, by centering,

$$\|\ell_{\theta'}(x,y) - \ell_\theta(x,y) - P(\ell_{\theta'} - \ell_\theta)\|_{\psi_2} \leqslant C\|\theta - \theta'\|_2 \ .$$

By general Hoeffding's inequality, it follows that

$$\|X_\theta - X_{\theta'}\|_{\psi_2} \leqslant \frac{C}{\sqrt{n}}\|\theta - \theta'\|_2 \ . \tag{6.3}$$

Step 2: The second step is to apply Talagrand's deviation inequality in $\mathbb{R}^d$, we deduce that, for any $z > 0$, with probability larger than $1 - 2\exp(-z^2)$,

$$\sup_{\theta \in \mathcal{E}} X_\theta - X_{\theta^*} \leqslant \frac{C}{\sqrt{n}}(w(\mathcal{E}) + z\Delta(\mathcal{E})) \ .$$

Step 3: The last step is to evaluate the geometric quantities appearing in the previous bound. Let $\mathbf{S} = \sum_{i=1}^d \lambda_i u_i u_i^T$ denote the eigenvalue decomposition of $\mathbf{S}$, with $\lambda_1 \geqslant \ldots \geqslant \lambda_d$. By definition $\mathcal{E}$ is the set of vectors $\theta = \theta^* + u$, where

$$\sum_{i=1}^d \lambda_i \left\langle u, u_i \right\rangle^2 \leqslant 1 \ .$$

Thus, for any $u \in \mathcal{E}$,

$$\|u\|_2^2 = \sum_{i=1}^n \left\langle u, u_i \right\rangle^2 \leqslant \frac{1}{\lambda_d} = \|\mathbf{S}^{-1}\| \ .$$

Hence, $\Delta(\mathcal{E}) \leqslant \sqrt{\|\mathbf{S}^{-1}\|}$. Regarding the Gaussian width, let X denote a standard Gaussian vector in $\mathbb{R}^d$, we have, for any $u = \sum_{i=1}^d \langle u, u_i\rangle\, u_i \in \mathcal{E}$,

$$\begin{aligned}\langle u, X\rangle &\leqslant \sum_{i=1}^d \sqrt{\lambda_i}\,\langle u, u_i\rangle\,\frac{\langle u_i, X\rangle}{\sqrt{\lambda_i}} \\ &\leqslant \sqrt{\sum_{i=1}^d \frac{\langle u_i, X\rangle^2}{\lambda_i}} \quad \text{Cauchy-Schwarz} \ ,\end{aligned}$$

so, by Cauchy-Schwarz inequality,

$$w(\mathcal{E}) \leqslant \sqrt{\sum_{i=1}^d \frac{\mathbb{E}[\langle u_i, X\rangle^2]}{\lambda_i}} = \sqrt{\mathrm{Tr}(\mathbf{S}^{-1})} \ .$$

Conclusion: The conclusion of this paragraph is that, for any $z > 0$,

$$\mathbb{P}\left(\sup_{\theta\in\mathcal{E}} (P_n - P)(\ell_{\theta^*} - \ell_\theta) > C\frac{\sqrt{\mathrm{Tr}(\mathbf{S}^{-1})} + z\sqrt{\|\mathbf{S}^{-1}\|}}{\sqrt{n}} \right) \leqslant 2\exp(-z^2) \ .$$

Chapter 7

Gaussian Processes

7.1 Setting

In this chapter, we provide tools to bound Gaussian processes. Hereafter, $(X_t)_{t\in T}$ denote a Gaussian process, that is, a collection of random variables indexed by T such that

$$\forall t_1,\dots,t_k \in T, \forall a_1,\dots,a_k, \qquad \sum_{i=1}^{k} a_i X_{t_i} \text{ is a Gaussian random variable } .$$

Without loss of generality, we also assume that $\mathbb{E}[X_t] = 0$ for all $t \in T$.

The distribution of the Gaussian vectors $(X_{t_i})_{i\in\{1,\dots,k\}}$ is characterized by the covariance matrix $\Sigma = (\mathrm{Cov}(X_{t_i}, X_{t_j}))_{1\leqslant i,j\leqslant k}$, thus the distribution of the Gaussian process is entirely characterized by the covariance function $\Sigma = (\Sigma_{s,t})_{s,t\in T}$, where

$$\Sigma_{s,t} = \mathrm{Cov}(X_s, X_t) = \mathbb{E}[X_s X_t] \ .$$

Similarly, the distribution is characterized by the values

$$\forall s,t \in T, \qquad \Sigma_{t,t} = \mathbb{E}[X_t^2], \qquad d(s,t) = \sqrt{\mathbb{E}[(X_s - X_t)^2]} \ .$$

7.2 Examples

7.2.1 Canonical Gaussian process on $\mathbb{R}^n$

Let g denote a standard Gaussian vector on $\mathbb{R}^n$. The canonical Gaussian process on $\mathbb{R}^n$ is then defined by

$$\forall t \in \mathbb{R}^n, \qquad X_t = \langle g, t\rangle \ .$$

One can easily check that this is a Gaussian process on $\mathbb{R}^n$, and that it is the one such that

$$\Sigma_{s,t} = \langle s,t\rangle\,, \qquad d(s,t) = \|s-t\|_2 \ .$$

If $G \sim \mathrm{N}(0, \Sigma)$ is a Gaussian vector on $\mathbb{R}^n$, its covariance matrix Σ is symmetric positive, so it can be written

$$\Sigma = \sum_{i=1}^{n} \lambda_i u_i u_i^T \ ,$$

where u_i is an orthonormal basis of $\mathbb{R}^n$. The matrix $A = \sum_{i=1}^n \sqrt{\lambda_i} u_i u_i^T$ satisfies $A = A^T$, $A^T A = \Sigma$, so its columns $t_1, \ldots, t_n$ satisfy $\langle t_i, t_j \rangle = \Sigma_{i,j}$. Hence, the Gaussian vector $(X_{t_i})_{i=1,\ldots,n}$, where $X_{t_i} = \langle g, t_i \rangle$ satisfies

$$\mathbb{E}[X_{t_i}] = 0, \qquad \mathbb{E}[X_{t_i} X_{t_j}] = \langle t_i, t_j \rangle = \Sigma_{i,j} \ ,$$

so $(X_{t_i})_{i=1,\ldots,n}$ is distributed as G.

7.2.2 Canonical Gaussian vector on Hilbert spaces

Let ℓ_2 denote the set of sequences of real numbers $t = (t_n)_{n \geqslant 1}$ such that $\sum_{n \geqslant 1} t_n^2 < \infty$, endowed with the inner product $\langle s, t \rangle_{\ell_2} = \sum_{n \geqslant 1} s_i t_i$. Let $(g_n)_{n \geqslant 1}$ denote a sequence of independent Gaussian random variables. The canonical Gaussian process on ℓ_2 is defined by

$$X_t = \sum_{n \geqslant 1} t_i g_i \ .$$

It is easy to check that it is the Gaussian process on ℓ_2 such that

$$\Sigma_{s,t} = \langle s, t \rangle_{\ell_2} \, , \qquad d(s,t) = \|s - t\|_{\ell_2} \ .$$

7.3 Bounding suprema

We are concerned in this chapter in

$$\mathbb{E}[\sup_{t \in T} X_t] \ .$$

To avoid measurability issues, we assume that T is finite, so $\sup_{t \in T} X_t = \max_{t \in T} X_t$. Without loss of generality, we can furthermore assume that X_t is a particular instance of the standard Gaussian process on $\mathbb{R}^{|T|}$, $X_t = \langle g, \gamma_t \rangle$, where γ_t are chosen such that $\Sigma_{s,t} = \langle \gamma_s, \gamma_t \rangle$.

Hereafter, we are therefore focusing on bounding the Gaussian width of finite subsets $\Gamma \in \mathbb{R}^n$

$$w(\Gamma) = \mathbb{E}[\sup_{\gamma \in \Gamma} \langle g, \gamma \rangle] \ .$$

It is easy to check that $w(\Gamma) = w(\Gamma - \gamma_0)$ for any $\gamma_0 \in \mathbb{R}^n$, so we can assume that $0 \in \Gamma$ without loss of generality and as a consequence $\sup_{\gamma \in \Gamma} \langle g, \gamma \rangle$ is a non-negative random variable, so

$$w(\Gamma) = \int_0^{+\infty} \mathbb{P}\Big(\sup_{\gamma \in \Gamma} \langle g, \gamma \rangle > z \Big) \mathrm{d}z \ .$$

A first idea would be to use a union bound to write

$$\mathbb{P}\left(\sup_{\gamma\in\Gamma} \langle g, \gamma\rangle > z \right) \leqslant \sum_{\gamma\in\Gamma} \mathbb{P}(\langle g, \gamma\rangle > z) \ .$$

Then we can use the estimate

$$\mathbb{P}(\langle g, \gamma\rangle > z) \leqslant \exp(-z^2/2\|\gamma\|^2) \ .$$

This yields the bound

$$\mathbb{P}\left(\sup_{\gamma\in\Gamma} \langle g, \gamma\rangle > z \right) \leqslant \sum_{\gamma\in\Gamma} \exp(-z^2/2\|\gamma\|^2) \ .$$

This bound can be made smaller than any $\delta \in (0,1)$ if we choose

$$z \geqslant 2 \max_{\gamma\in\Gamma}\{\|\gamma\|\}\sqrt{\log(|\Gamma|) + \log(\delta^{-1})} \ .$$

This bound cannot be improved without further assumptions on Γ. Indeed, if Γ is an orthonormal basis of $\mathbb{R}^n$, it yields

$$\mathbb{P}\left(\max_{i=1,\dots,n} g_i > 2\sqrt{\log n + \log 1/\delta} \right) \leqslant \delta \ , \tag{7.1}$$

where g_i are independent Gaussian random variables, which the correct order of magnitude given by the Gaussian concentration inequality.

On the other hand, if e_i is the canonical basis of $\mathbb{R}^n$, $\epsilon \in (0, 1/2)$ is a small number, and $\gamma_i = e_1 + \epsilon e_i$, we have

$$\sup_{\gamma\in\Gamma} \langle g, \gamma\rangle = \langle g, e_1\rangle + \epsilon \max_{i=1,\dots,n} \langle g, e_i\rangle \ . \tag{7.2}$$

Each γ_i has norm $\leqslant 3/2$, so our generic upper bound gives then

$$\mathbb{P}\left(\sup_{\gamma\in\Gamma} \langle g, \gamma\rangle > 3\sqrt{\log n + \log 1/\delta} \right) \leqslant \delta \ .$$

Integrating this bound then yields

$$w(\gamma) \leqslant 3\sqrt{\log n} + C \ .$$

On the other hand, taking advantage of our decomposition (7.2), we see that

$$\mathbb{P}\left(\sup_{\gamma\in\Gamma} \langle g, \gamma\rangle > z + \alpha \right) \leqslant \mathbb{P}(\langle g, e_1\rangle > z) + \mathbb{P}(\epsilon \max_{i=1,\dots,n} \langle g, e_i\rangle > \alpha) \ .$$

Now, for the first term to be smaller than $\delta/2$, we pick $z = 2\sqrt{\log 2/\delta}$ by the standard estimate of the tails of Gaussian random variable, and, by (7.1),

the second term does not exceed $\delta/2$ if $\alpha = 2\epsilon\sqrt{\log n + \log 2/\delta}$. Putting together these informations yields

$$\mathbb{P}\bigg(\sup_{\gamma\in\Gamma} \langle g, \gamma\rangle > 2\sqrt{\log 2/\delta} + 2\epsilon\sqrt{\log n + \log 2/\delta}\bigg) \leqslant \delta \ ,$$

Integrating this new bound then yields

$$w(\Gamma) \leqslant 2\epsilon\sqrt{\log n} + C \ .$$

If $\epsilon < 1/\sqrt{\log n}$, this new bound shows that $w(\Gamma)$ is bounded **independently of the size of n of Γ** while our upper bound derived from a rough shows that $w(\Gamma)$ grows with the size n of Γ, so it's not tight in situations as the second case, where the random variables $\langle g, \gamma\rangle$ are highly correlated. The generic chaining bound, that we develop in the following, intends to provide a generic bound on $w(\Gamma)$ that scales correctly with n in each situation.

7.4 The generic chaining bound

At a very general level, the idea of generic chaining is to cluster points at several scales and take union bounds over each cluster.

7.4.1 Hierarchical clustering

Recall that $0 \in \Gamma$ and let $\Gamma_0 = \{0\}$. Then, consider a growing sequence $\Gamma_0 \subset \Gamma_1 \subset \ldots \subset \Gamma_k = \Gamma$. For any $n \in \{0, \ldots, k\}$ and any $\gamma \in \Gamma$, we let $\pi_n(\gamma) \in \Gamma_n$ denote any point such that $d(\gamma, \Gamma_n) = d(\gamma, \pi_n(\gamma))$.

We have clearly, for any $\gamma \in \Gamma$, $\pi_0(\gamma) = 0$, $\pi_k(\gamma) = \gamma$. Therefore, the following **chaining equality** holds, which is at the heart of the generic chaining bound:

$$\langle g, \gamma\rangle = \sum_{n=1}^{k} \langle g, (\pi_n(\gamma) - \pi_{n-1}(\gamma)\rangle$$

This shows that, whatever the sequence z_n,

$$\mathbb{P}\bigg(\sup_{\gamma\in\Gamma} \langle g, \gamma\rangle > \sum_{n=1}^{k} z_n\bigg) \leqslant \sum_{n=1}^{k} |\Gamma_n||\Gamma_{n+1}|\mathbb{P}\big(\langle g, (\pi_n(\gamma) - \pi_{n-1}(\gamma)\rangle > z_n\big) \ .$$

7.4.2 How do we choose Γ_n?

To understand this choice, recall that we have, by the standard estimate on the tails of a Gaussian random variable:

$$\forall z > 0, \qquad \mathbb{P}(\langle g, (\pi_n(\gamma) - \pi_{n-1}(\gamma)\rangle > z) \leqslant \exp\left(-\frac{z^2}{2\|\pi_n(\gamma) - \pi_{n-1}(\gamma)\|_2^2}\right) \ .$$

Plugging this estimate in our bound yields

$$\mathbb{P}\bigg(\exists\gamma\in\Gamma,\quad \langle g,\gamma\rangle > \sum_{n=1}^{k} z_n(\gamma)\bigg) \leqslant \sum_{n=1}^{k} |\Gamma_n||\Gamma_{n-1}| \exp\left(-\inf_{\gamma\in\Gamma} \frac{z_n^2(\gamma)}{2\|\pi_n(\gamma)-\pi_{n-1}(\gamma)\|_2^2}\right) .$$

Now we bound, as the sequence Γ_n is growing:

$$|\Gamma_n||\Gamma_{n-1}| \leqslant |\Gamma_n|^2, \qquad \|\pi_n(\gamma)-\pi_{n-1}(\gamma)\|_2^2 \leqslant 4d(\gamma,\Gamma_{n-1})^2 .$$

Now, for any sequence β_n such that $\sum_{n=1}^k \beta_n = 1$, we take $z_n(\gamma) = 8d(\gamma,\Gamma_{n-1})\sqrt{\log(|\Gamma_n|^2/\beta_n)+z}$

We derive that, for any growing sequence Γ_n,

$$\mathbb{P}\bigg(\sup_{\gamma\in\Gamma}\langle g,\gamma\rangle > \sup_{\gamma\in\Gamma}\sum_{n=1}^{k} 8d(\gamma,\Gamma_{n-1})\sqrt{\log(|\Gamma_n|^2/\beta_n)+z}\bigg) \leqslant \exp(-z) .$$

Now, we take Γ_n any sequence such that $|\Gamma_n| = 2^{2^{\ell_n}}$, where ℓ_n is chosen such that $d(\Gamma,\Gamma_n) = d(\Gamma,\Gamma_{n-1})/2$ and let Γ'_ℓ denote a sequence such that $\Gamma'_{\ell_n} = \Gamma_n$. We also let $\beta_n = C/|\Gamma_n|$, we have

$$\sum_{n=1}^{k} d(\gamma,\Gamma_{n-1})\sqrt{\log(|\Gamma_n|^2/\beta_n)+z} \leqslant C\bigg(\sum_{\ell=1}^{\ell_k} d(\gamma,\Gamma'_\ell)2^{\ell/2} + \sup_{\gamma\in\Gamma} d(\gamma,\{0\})\sqrt{z}\bigg) .$$

This last bound suggests to introduce the following quantities:

$$\gamma_2(\Gamma) = \inf_{\Gamma'_n:|\Gamma'_n|=2^{2^n}} \sup_{\gamma\in\Gamma}\sum_{\ell=0}^{+\infty} d(\gamma,\Gamma'_\ell)2^{\ell/2}, \qquad \Delta(\Gamma) = \sup_{\gamma\in\Gamma} d(\gamma,0) .$$

Indeed, we have proved that, for any $\Gamma\subset\mathbb{R}^d$, we have

$$\forall z>0, \qquad \mathbb{P}\bigg(\sup_{\gamma\in\Gamma}\langle g,\gamma\rangle > C(\gamma_2(\Gamma)+\Delta(\Gamma)\sqrt{z})\bigg) \leqslant \exp(-z) .$$

Therefore, we have in particular

$$w(\Gamma) \lesssim \gamma_2(\Gamma)\vee\Delta(\Gamma) .$$

As the sequences Γ'_n satisfy $\Gamma'_0 = \{0\}$, we have $\gamma_2(\Gamma) \geqslant \Delta(\Gamma)$, so finally, we get the estimate

$$w(\Gamma) \lesssim \gamma_2(\Gamma) .$$

A remarkable result, due to Talagrand, that we will prove now, is that we can conversely show that

$$\gamma_2(\Gamma) \lesssim w(\Gamma) .$$

Hence, $\gamma_2(\Gamma)$ provides the correct order of magnitude for the Gaussian width $w(\Gamma)$.

7.5 The majorizing measure theorem

The purpose of this section is to show that the generic chaining upper bound proved in the previous section is tight in the sense that

$$\gamma_2(\Gamma) \lesssim w(\Gamma) \ .$$

We prove this using classical tools for Gaussian processes that may be of independent interest.

7.5.1 Another look at $\gamma_2(\Gamma)$

Recall the definition of γ_2:

$$\gamma_2(\Gamma) = \inf_{\Gamma'_n : |\Gamma'_n| = 2^{2^n}} \sup_{\gamma \in \Gamma} \sum_{\ell=0}^{+\infty} d(\gamma, \Gamma'_\ell) 2^{\ell/2},$$

For any sequence Γ'_n, we define $\mathcal{A}_0 = \Gamma$ and for any $n \geqslant 1$, let $\mathcal{A}_n$ denote a partition of Γ, which is a refinement of $\mathcal{A}_{n-1}$, such that $|\mathcal{A}_n| = 2^{2^n}$ and each element $A_n \in \mathcal{A}_n$ contains exactly one element of Γ'_n. Then it is clear that, if $A_n(\gamma)$ denotes the element of $\mathcal{A}_n$ containing γ and $\mathrm{diam}(A_n(\gamma)) = \sup_{\gamma', \gamma'' \in A_n(\gamma)} \|\gamma' - \gamma''\|_2$ denotes its diameter,

$$d(\gamma, \Gamma'_\ell) \leqslant \mathrm{diam}(A_\ell(\gamma)) \ ,$$

thus, denoting by ℓ_0 the first integer such that $\Gamma'_n = \Gamma$,

$$\gamma_2(\Gamma) \leqslant \inf_{\mathcal{A}_n : |\mathcal{A}_n| = 2^{2^n}} \sup_{\gamma \in \Gamma} \sum_{\ell=0}^{\ell_0} \mathrm{diam}(A_\ell(\gamma)) 2^{\ell/2} \ .$$

We will prove that there exists a numerical constant c such that

$$w(\Gamma) \geqslant c \inf_{\mathcal{A}_n : |\mathcal{A}_n| = 2^{2^n}} \sup_{\gamma \in \Gamma} \sum_{\ell=0}^{\ell_0} \mathrm{diam}(A_\ell(\gamma)) 2^{\ell/2} \ .$$

This proves that

$$w(\Gamma) \asymp \gamma_2(\Gamma) \asymp \inf_{\mathcal{A}_n : |\mathcal{A}_n| = 2^{2^n}} \sup_{\gamma \in \Gamma} \sum_{\ell=0}^{\ell_0} \mathrm{diam}(A_\ell(\gamma)) 2^{\ell/2} \ .$$

7.5.2 Gaussian Calculus

We start with elementary results on Gaussian random variables and vectors.

Lemma 67 (Gaussian integration by part)**.** *Let $X \sim \mathrm{N}(0,1)$ denote a standard Gaussian random variable and let $f : \mathbb{R} \to \mathbb{R}$ denote a regular function. Then,*

$$\mathbb{E}[f'(X)] = \mathbb{E}[X f(X)] \ .$$

Proof. Assume first that f has bounded support. We write

$$\mathbb{E}[f'(X)] = \int_{-\infty}^{+\infty} f'(x)\varphi(x)\mathrm{d}x \ .$$

We integrate by part, as $\varphi'(x) = -x\varphi(x)$, we deduce

$$\mathbb{E}[f'(X)] = -\int_{-\infty}^{+\infty} f(x)\varphi'(x)\mathrm{d}x = \mathbb{E}[Xf(X)] \ .$$

The formula is correct for functions with bounded support. We conclude the proof using standard approximation arguments. □

This result can easily be extended to more general centered Gaussian random variables: If $X = \sigma Y \sim \mathrm{N}(0, \sigma^2)$, then

$$\mathbb{E}[f'(X)] = \mathbb{E}[f'(\sigma Y)] = \frac{1}{\sigma}\mathbb{E}[\sigma f'(\sigma Y)] = \frac{1}{\sigma}\mathbb{E}[Yf(\sigma Y)] = \frac{1}{\sigma^2}\mathbb{E}[Xf(X)] \ .$$

Moreover, if $X = \Sigma^{1/2}Y \sim \mathrm{N}(0, \Sigma)$ is a centered Gaussian vector in $\mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable, we have, $\forall i \in \{1, \ldots, d\}$,

$$\begin{aligned}
\mathbb{E}[X_i f(X)] &= \sum_{j=1}^{d} \Sigma_{i,j}^{1/2} \mathbb{E}[Y_j f(\Sigma^{1/2}Y)] \\
&= \sum_{j=1}^{d} \Sigma_{i,j}^{1/2} \sum_{k=1}^{d} \Sigma_{k,j}^{1/2} \mathbb{E}[\partial_k f(\Sigma^{1/2}Y)] \\
&= \sum_{k=1}^{d} \left(\sum_{j=1}^{d} \Sigma_{i,j}^{1/2} \Sigma_{j,k}^{1/2} \right) \mathbb{E}[\partial_k f(X)] \\
&= \sum_{k=1}^{d} \Sigma_{i,k} \mathbb{E}[\partial_k f(X)] = \Big(\Sigma \mathbb{E}[\nabla f(X)] \Big)_i \ .
\end{aligned}$$

In the preceding computation, the second inequality is due to the Gaussian integration by part lemma, the third by symmetry of the square-root $\Sigma^{1/2}$ and the last one by standard matrix calculus.

Hence, if $X \sim \mathrm{N}(0, \Sigma)$,

$$\mathbb{E}[Xf(X)] = \Sigma \mathbb{E}[\nabla f(X)] \ .$$

More generally, if $\mathbf{h} = (h_1, \ldots, h_d)$ is a $\mathbb{R}^d$-valued function,

$$\mathbb{E}[\langle X, \mathbf{h}(X) \rangle] = \sum_{i=1}^{d} \mathbb{E}[X_i h_i(X)] = \sum_{i=1}^{d} \Big(\Sigma \mathbb{E}[\nabla h_i(X)] \Big)_i = \sum_{i,j=1}^{d} \Sigma_{i,j} \mathbb{E}[\partial_j h_i(X)] \ . \tag{7.3}$$

7.5.3 Slepian's and Sudakov-Fernique results

Slepian's lemma is a comparison result that allows to work with a particular instance of Gaussian process. It states the following:

Lemma 68. *Let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ denote two Gaussian processes such that*

$$\forall s,t, \qquad \mathbb{E}[(X_t - X_s)^2] \geqslant \mathbb{E}[(Y_t - Y_s)^2] \ .$$

Then, [Sudakov-Fernique],

$$\mathbb{E}[\sup_{t\in T} X_t] \geqslant \mathbb{E}[\sup_{t\in T} Y_t] \ .$$

If moreover,

$$\forall t, \qquad \mathbb{E}[X_t^2] = \mathbb{E}[Y_t^2] \ ,$$

then, [Slepian], for any $z \in \mathbb{R}$,

$$\mathbb{P}(\sup_{t\in T} X_t \leqslant z) \leqslant \mathbb{P}(\sup_{t\in T} Y_t \leqslant z) \ .$$

Proof. We prove the lemma in the case where T is finite, the general case follows by simple arguments. Let then X and Y denote two independent centered Gaussian vectors with respective covariance matrices Σ^X and Σ^Y such that

$$\Sigma^X_{i,j} \leqslant \Sigma^Y_{i,j} \ .$$

We define the Gaussian vectors, $\forall u \in [0,1]$, $Z_u = \sqrt{1-u}X + \sqrt{u}Y \sim \mathrm{N}(0, (1-u)\Sigma^X + u\Sigma^Y)$. For any twice differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, we have

$$\begin{aligned}\partial_u \mathbb{E}[f(Z_u)] &= \frac{1}{2}\mathbb{E}[\Big\langle \nabla f(Z_u), u^{-1/2}Y - (1-u)^{-1/2}X \Big\rangle] \\ &= \frac{1}{2\sqrt{u}}\mathbb{E}[\langle \nabla f(Z_u), Y\rangle] - \frac{1}{2\sqrt{1-u}}\mathbb{E}[\langle \nabla f(Z_u), X\rangle] \ .\end{aligned}$$

Start with the first term: By (7.3),

$$\mathbb{E}[\big\langle \nabla f(\sqrt{1-u}X + \sqrt{u}Y), Y\big\rangle \,|X] = \sqrt{u}\sum_{i,j=1}^{d} \Sigma^Y_{i,j}\mathbb{E}[\partial_{i,j} f(\sqrt{1-u}X + \sqrt{u}Y)|X]$$

Thus,

$$\frac{1}{2\sqrt{u}}\mathbb{E}[\langle \nabla f(Z_u), Y\rangle] = \frac{1}{2}\left\langle \Sigma^Y, \mathbb{E}[Hf(Z_u)]\right\rangle_F] \ .$$

Likewise,

$$\frac{1}{2\sqrt{1-u}}\mathbb{E}[\langle \nabla f(Z_u), X\rangle] = \frac{1}{2}\left\langle \Sigma^X, \mathbb{E}[Hf(Z_u)]\right\rangle_F] \ .$$

Hence,

$$\partial_u \mathbb{E}[f(Z_u)] = \frac{1}{2}\left\langle \Sigma^Y - \Sigma^X, \mathbb{E}[Hf(Z_u)]\right\rangle_F] \ .$$

Let $f(x_1, \ldots, x_d) = \frac{1}{\beta}\log\sum_{i=1}^d \exp(\beta x_i)$. Denote by

$$p_i(x) = \frac{\exp(\beta x_i)}{\sum_{j=1}^d \exp(\beta x_j)} \ .$$

We have

$$\partial_i f(x_1, \ldots, x_d) = p_i(x), \qquad \partial_{i,j} f(x_1, \ldots, x_d) = \begin{cases} -\beta p_i(x)p_j(x) & \text{if } i \neq j \ , \\ \beta p_i(x)(1-p_i(x)) & \text{if } i = j \ . \end{cases}$$

Write $\sigma_{i,j} = \Sigma^Y_{i,j} - \Sigma^X_{i,j}$, we have, as $\sum_{j=1}^d p_j(z) = 1$,

$$\begin{aligned} \frac{1}{\beta}\left\langle \Sigma^Y - \Sigma^X, Hf(z)\right\rangle_F] &= \sum_{i=1}^d \sigma_{i,i}p_i(z)(1-p_i(z)) - \sum_{i\neq j}\sigma_{i,j}p_i(z)p_j(z) \\ &= \sum_{i\neq j}\sigma_{i,i}p_i(z)p_j(z) - \sum_{i\neq j}\sigma_{i,j}p_i(z)p_j(z) \ . \end{aligned}$$

Symmetrically

$$\frac{1}{\beta}\left\langle \Sigma^Y - \Sigma^X, Hf(z)\right\rangle_F] = \sum_{i\neq j}\sigma_{j,j}p_i(z)p_j(z) - \sum_{i\neq j}\sigma_{i,j}p_i(z)p_j(z) \ .$$

Thus,

$$\frac{1}{2\beta}\left\langle \Sigma^Y - \Sigma^X, Hf(z)\right\rangle_F] = \sum_{i\neq j}(\sigma_{i,i} + \sigma_{j,j} - 2\sigma_{i,j})p_i(z)p_j(z) \ .$$

As

$$\sigma_{i,i} + \sigma_{j,j} - 2\sigma_{i,j} = \mathbb{E}[(Y_i - Y_j)^2] - \mathbb{E}[(X_i - X_j)^2] \leqslant 0 \ ,$$

it follows that $\left\langle \Sigma^Y - \Sigma^X, Hf(z)\right\rangle_F] \leqslant 0$, thus that $u \mapsto \mathbb{E}[f(Z_u)]$ is non-increasing, and therefore, that

$$\mathbb{E}[f(X)] = \mathbb{E}[f(Z_0)] \geqslant \mathbb{E}[f(Z_1)] = \mathbb{E}[f(Y)] \ .$$

As, for any fixed $x = (x_1, \ldots, x_d)$, we have $f(x)$ grows to $\max_{i=1,\ldots,d} x_i$ as $\beta \to +\infty$, we have, by the monotone convergence theorem,

$$\mathbb{E}[\max_i X_i] \geqslant \mathbb{E}[\max_i Y_i] \ ,$$

as desired.

Assume now moreover that

$$\Sigma^X_{i,i} = \Sigma^Y_{i,i} \ .$$

Let $z \in \mathbb{R}$ and κ denote a twice differenciable, non-increasing, non negative approximation of $\mathbf{1}_{x \leqslant z}$ and let $f(x_1, \ldots, x_d) = \prod_{i=1}^d h(x_i)$. The function f satisfies, for any $i \neq j$, $\partial_{i,j} f(x) = h'(x_i)h'(x_j) \prod_{k \neq i,j} h(x_k) \geqslant 0$, thus

$$\partial_u \mathbb{E}[f(Z_u)] = \frac{1}{2} \left\langle \Sigma^Y - \Sigma^X, \mathbb{E}[Hf(Z_u)] \right\rangle_F] \geqslant 0 \ .$$

Indeed, all terms $(\Sigma^Y_{i,j} - \Sigma^X_{i,j}) \partial_{i,j} f(\sqrt{1-u}X + \sqrt{u}Y) \geqslant 0$ when $i \neq j$ and are null when $i = j$ by assumption on Σ^X, Σ^Y. Therefore, the function $u \mapsto \mathbb{E}[f(Z_u)]$ is non-increasing, hence, $\mathbb{E}[f(X)] = \mathbb{E}[f(Z_0)] \geqslant \mathbb{E}[f(Z_1)] = \mathbb{E}[f(Y)]$. As this is true for any approximation function h, it follows that

$$\mathbb{P}(\max_i X_i \leqslant z) \leqslant \mathbb{P}(\max_i Y_i \leqslant z) \ .$$

□

A very nice corollary of Sudakov-Fernique's result is the following.

Theorem 69 (Sudakov's minoration)**.** *Assume that X is a Gaussian vector in $\mathbb{R}^d$ such that, for any $i \neq j$,*

$$\mathbb{E}[(X_i - X_j)^2] \geqslant \alpha^2 \ .$$

Then we have

$$\mathbb{E}[\max_{i \in \{1,\ldots,d\}} X_i] \geqslant \alpha \sqrt{\frac{\log d}{2}} \ .$$

Proof. Denote by $Y_i = \alpha g_i / \sqrt{2}$, where g is a standard Gaussian vector. We have

$$\forall i \neq j, \qquad \mathbb{E}[(Y_i - Y_j)^2] = \frac{\alpha^2}{2} \mathbb{E}[(g_i - g_j)^2] = \alpha^2 \leqslant \mathbb{E}[(X_i - X_j)^2] \ .$$

Therefore, from Sudakov-Fernique's bound,

$$\mathbb{E}[\max_{i \in \{1,\ldots,d\}} X_i] \geqslant \frac{\alpha}{\sqrt{2}} \mathbb{E}[\max_{i \in \{1,\ldots,d\}} g_i] \ .$$

Now we can use the standard result

$$\mathbb{E}[\max_{i \in \{1,\ldots,d\}} g_i] \geqslant \sqrt{\log d} \ ,$$

to conclude. □

7.5.4 Talagrand's recursive bound

Talagrand refined Sudakov's result to obtain a recursive bound that will allow to prove the majorizing measure theorem.

Lemma 70. *There exists an absolute constant $r > 0$ such that, for any $\alpha > 0$ and any α-separated subset of points $\{\gamma_1, \dots, \gamma_d\} \subset \Gamma$, that is, satisfying*

$$\forall i \neq j, \qquad \|\gamma_i - \gamma_j\|_2 \geqslant \alpha \ ,$$

we have

$$\mathbb{E}[\max_{\gamma\in\Gamma} \langle g, \gamma\rangle] \geqslant \frac{\alpha}{2}\sqrt{\log d} + \min_{i\in\{1,\dots,d\}} \mathbb{E}[\max_{\gamma\in\Gamma:\|\gamma-\gamma_i\|_2\leqslant\alpha/r} \langle g, \gamma\rangle] \ .$$

Proof. We write $B(\gamma, r) = \{\gamma' \in \Gamma : \|\gamma - \gamma'\|_2 \leqslant r\}$. We have

$$\max_{\gamma\in\Gamma} \langle g, \gamma\rangle \geqslant \max_{\gamma\in\cup_{i=1}^d B(\gamma_i,\alpha/r)} \langle g, \gamma\rangle \ .$$

Now, for any $\gamma \in \cup_{i=1}^d B(\gamma_i, \alpha/r)$, we write

$$\langle g, \gamma\rangle = \langle g, \gamma_i\rangle + \langle g, \gamma - \gamma_i\rangle \ ,$$

where γ_i is chosen such that $\|\gamma - \gamma_i\| \leqslant \alpha/r$, we deduce

$$\begin{aligned}\sup_{\gamma\in\cup_{i=1}^d B(\gamma_i,\alpha/r)} \langle g, \gamma\rangle \geqslant \max_i \langle g, \gamma_i\rangle + \min_i \mathbb{E}[\sup_{\gamma\in B(\gamma_i,\alpha/r)} \langle g, \gamma - \gamma_i\rangle] \\ - \max_i |\sup_{\gamma\in B(\gamma_i,\alpha/r)} \langle g, \gamma - \gamma_i\rangle - \mathbb{E}[\sup_{\gamma\in B(\gamma_i,\alpha/r)} \langle g, \gamma - \gamma_i\rangle]| \ .\end{aligned}$$

It comes from the Gaussian concentration inequality and Pisier-Massart's lemma that

$$\mathbb{E}\left[\max_i |\sup_{\gamma\in B(\gamma_i,\alpha/r)} \langle g, \gamma - \gamma_i\rangle - \mathbb{E}[\sup_{\gamma\in B(\gamma_i,\alpha/r)} \langle g, \gamma - \gamma_i\rangle]|\right] \leqslant C\frac{\alpha}{r}\sqrt{\log d} \ .$$

Therefore, the result follows from Sudakov-Fernique's bound on $\mathbb{E}[\max_i \langle g, \gamma_i\rangle]$. □

7.5.5 Proof of the Majorizing measure theorem

Recall that we intend to bound from bellow

$$w(\Gamma) = \mathbb{E}[\max_{\gamma\in\Gamma} \langle g, \gamma\rangle] \ ,$$

where Γ denote any finite subset $\Gamma \subset \mathbb{R}^n$. We use the approach mentionned in Section 7.5.1. For any sequence $\mathcal{A}_n$ of partitions of Γ such that $|\mathcal{A}_n| \leqslant 2^{2^n}$, for any $\gamma \in \Gamma$, we denote by $A_n(\gamma) \in \mathcal{A}_n$ the element containing γ.

The goal is to prove that there exists a sequence $\mathcal{A}_n$ of partitions such that

$$\forall \gamma \in \Gamma, \qquad \sum_n \operatorname{diam}(A_n(\gamma))2^{n/2} \leqslant Cw(\Gamma) \ .$$

We explain a possible partitioning scheme and give the main steps of its analysis in the end of this section.

Partitionning scheme. Let us first explain how to build a nice sequence of partitions whose elements A are weighted by $\alpha(A) > 0$. We proceed recursively, starting with

$\mathcal{A}_0 = \Gamma$ **and** $\alpha(\Gamma) = \Delta(\Gamma)$**.**

Suppose now that we have built $\mathcal{A}_n$ such that

$$|\mathcal{A}_n| \leqslant 2^{2^n} \ ,$$

and, for each $A \in \mathcal{A}_n$, an upper bound $\alpha(A) \geqslant \Delta(A)$.

To build the partition $\mathcal{A}_{n+1} \subset \mathcal{A}_n$, we split each $A \in \mathcal{A}_n$ into at most 2^{2^n} pieces, so

$$|\mathcal{A}_{n+1}| \leqslant \sum_{A \in \mathcal{A}_n} 2^{2^n} \leqslant 2^{2^n} * 2^{2^n} = 2^{2^n + 2^n} = 2^{2^{n+1}} \ .$$

Let r denote the real number in Lemma 70. Without loss of generality, we can assume that $r \geqslant 4$. Let

$$\gamma_1 \in \operatorname{argmax} w\left\{ B\left(\gamma, \frac{\alpha(A)}{r}\right) \cap A \right\} \ ,$$
$$A_1 = B\left(\gamma_1, \frac{\alpha(A)}{r}\right) \cap A, \quad \alpha(A_1) = \frac{\alpha(A)}{r}, \quad D_2 = A \setminus A_1 \ .$$

Then, as long as $\ell < 2^{2^n}$ and $D_\ell \neq \emptyset$,

$$\gamma_\ell \in \operatorname{argmax} w\left\{ B\left(\gamma, \frac{\alpha(A)}{r}\right) \cap D_\ell \right\} \ ,$$
$$A_\ell = B\left(\gamma_\ell, \frac{\alpha(A)}{r}\right) \cap D_\ell, \quad \alpha(A_\ell) = \frac{\alpha(A)}{r}, \quad D_{\ell+1} = D_\ell \setminus A_\ell \ .$$

If $D_{2^{2^n}} \neq \emptyset$, we state

$$A_{2^{2^n}} = D_{2^{2^n}}, \qquad \alpha(A_{2^{2^n}}) = \alpha(A) \ .$$

Let $m \leqslant 2^{2^n}$ denote the number A_ℓ. The centers $\gamma_1, \ldots, \gamma_m$ are $\alpha(A)/r$ separated by construction. Besides, the weights $\alpha(A_\ell)$ are all equal to $\alpha(A)/r$ except possibly the last one $\alpha(A_m)$ that may be $\alpha(A)$ if $m = 2^{2^n}$. Finally, the Gaussian width $w\big\{B\big(\gamma_i, \alpha(A)/r\big)\big\}$ are non increasing.

We continue until the partition $\mathcal{A}_n$ is only made of singletons.

Analysis 1: The tree. Let us now build a tree whose root is $\mathcal{A}_0$, and the children of each $A \in \mathcal{A}_n$ are given by the elements of its partition $A_1, \ldots, A_m$ described in the previous section. Each edge (A, A_i) is weighted $\alpha(A_i)2^{n/2}$ so, for each $\gamma \in \Gamma$,

$$\sum_n \operatorname{diam}(A_n(\gamma))2^{n/2} \leqslant \sum_n \alpha(A_n(\gamma))2^{n/2} := W(\gamma) \ .$$

$W(\gamma)$ is thus the weight of the path from the leaf γ in the tree to the root. We intend to prove that

$$\forall \gamma \in \Gamma, \qquad W(\gamma) \leqslant C w(\Gamma) \ .$$

Analysis 2: Decompositions of the sum. We break the sum defining $W(\gamma)$ into the moments n_k where $A_n(\gamma)$ is the last element of the partition, that is those where $\alpha(A_n(\gamma)) = \alpha(A_{n-1}(\gamma))$, and the instant in-between. We have

$$\begin{aligned} W(\gamma) &= \sum_k \alpha(A_{n_k}(\gamma)) 2^{n_k/2} + \sum_{n=n_k+1}^{n_{k+1}-1} \alpha(A_n(\gamma)) 2^{n/2} \\ &\leqslant \sum_k \alpha(A_{n_k}(\gamma)) 2^{n_k/2} + \sum_{j=1}^{+\infty} \frac{\alpha(A_{n_k}(\gamma))}{r^j} 2^{(n_k+j)/2} \\ &\leqslant 2 \sum_k \alpha(A_{n_k}(\gamma)) 2^{n_k/2} \ . \end{aligned}$$

The last inequality holds as $r \geqslant 4$. Besides, for any $\ell > 0$, if $n_{k+\ell} = n_k + \ell$, we have

$$\sum_{n_k}^{n_{k+\ell}} \alpha(A_n(\gamma)) 2^{n/2} \leqslant \alpha(A_{n_{k+\ell}}(\gamma)) 2^{n_{k+\ell}} \sum_j 2^{-j/2} \leqslant \frac{\sqrt{2}}{\sqrt{2}-1} \alpha(A_{n_{k+\ell}}(\gamma)) 2^{n_{k+\ell}} \ .$$

Hence, we can assume that $\alpha(A_{n_{k+1}}(\gamma)) \leqslant \alpha(A_{n_k}(\gamma))/r$.

Analysis 3: Talagrand's recursive lemma. To prove that this last upper bound can be bounded from above by the Gaussian width $w(\Gamma)$, we proceed recursively using Talagrand's recursive minoration lemma. We start with $k = 1$. The cell $A_{n_1-1}(\gamma)$ contained points $\gamma_1, \ldots, \gamma_{2^{2^{n_1}}} = \gamma$ that are $\alpha(A_{n_1}(\gamma))/r$ separated, so by Talagrand's recursive minoration lemma:

$$w(A_{n_1-1}(\gamma)) \geqslant \frac{\alpha(A_{n_1}(\gamma))}{2r} 2^{n_1/2} + w(B(\gamma, \alpha(A_{n_1-1}(\gamma))/r) \cap D_{2^{2^{n_1}}}) \ .$$

The min in Talagrand's minoration lemma is replaced here by the ball centered in γ by definition of the γ_i, which made the sequence $w(B(\gamma_i, \alpha(A_{n_1}(\gamma))/r) \cap D_i)$ non-increasing.

Now it follows from $\alpha(A_{n_2}(\gamma)) \leqslant \alpha(A_{n_1}(\gamma))/r$ that

$$A_{n_2-1}(\gamma) \subset B(\gamma, \alpha(A_{n_1-1}(\gamma))/r) \cap D_{2^{2^{n_1}}} \ ,$$

hence

$$w(A_{n_1-1}(\gamma)) \geqslant \frac{\alpha(A_{n_1}(\gamma))}{2r} 2^{n_1/2} + w(A_{n_2-1}(\gamma)) \ .$$

Proceeding recursively, it follows thus that

$$w(\Gamma) \geqslant w(A_{n_1}(\gamma)) \geqslant \frac{1}{2r} \sum_k \alpha(A_{n_k}(\gamma)) 2^{n_k/2} = \frac{\sqrt{2}-1}{4\sqrt{2}r} W(\gamma) \ .$$

As this is true for any $\gamma \in \Gamma$, this concludes the proof of the majorizing measure theorem.