

Chapter 2

Empirical and Rademacher Processes

The empirical process is defined as

$$Z_n := n^{1/2}(P_n - P)$$

and it can be viewed as a random measure. However, more often, it has been viewed as a stochastic process indexed by a function class \mathcal{F} :

$$Z_n(f) = n^{1/2}(P_n - P)(f), f \in \mathcal{F}$$

(see Dudley [59] or van der Vaart and Wellner [148]).

The Rademacher process indexed by a class \mathcal{F} was defined in Sect. 1.3 as

$$R_n(f) := n^{-1} \sum_{i=1}^n \varepsilon_i f(X_i), f \in \mathcal{F},$$

$\{\varepsilon_i\}$ being i.i.d. Rademacher random variables (that is, ε_i takes the values $+1$ and -1 with probability $1/2$ each) independent of $\{X_i\}$.

It should be mentioned that certain measurability assumptions are required in the study of empirical and Rademacher processes. In particular, under these assumptions, such quantities as $\|P_n - P\|_{\mathcal{F}}$ are properly measurable random variables. We refer to the books of Dudley [59], Chap. 5 and van der Vaart and Wellner [148], Sect. 1.7 for precise formulations of these measurability assumptions. Some of the bounds derived and used below hold even without the assumptions of this nature, if the expectation is replaced by the outer expectation, as it is often done, for instance, in [148]. Another option is to “define”

$$\mathbb{E}\|P_n - P\|_{\mathcal{F}} := \sup \left\{ \mathbb{E}\|P_n - P\|_{\mathcal{G}} : \mathcal{G} \subset \mathcal{F}, \mathcal{G} \text{ is finite} \right\},$$

which provides a simple way to get around the measurability difficulties. Such an approach has been frequently used by Talagrand (see, e.g., [140]). In what follows, it will be assumed that the measurability problems have been resolved in one of these ways.

2.1 Symmetrization Inequalities

The following important inequality reveals close relationships between empirical and Rademacher processes.

Theorem 2.1. *For any class \mathcal{F} of P -integrable functions and for any convex function $\Phi : \mathbb{R}_+ \mapsto \mathbb{R}_+$*

$$\mathbb{E}\Phi\left(\frac{1}{2}\|R_n\|_{\mathcal{F}_c}\right) \leq \mathbb{E}\Phi\left(\|P_n - P\|_{\mathcal{F}}\right) \leq \mathbb{E}\Phi\left(2\|R_n\|_{\mathcal{F}}\right),$$

where $\mathcal{F}_c := \{f - Pf : f \in \mathcal{F}\}$. In particular,

$$\frac{1}{2}\mathbb{E}\|R_n\|_{\mathcal{F}_c} \leq \mathbb{E}\|P_n - P\|_{\mathcal{F}} \leq 2\mathbb{E}\|R_n\|_{\mathcal{F}}.$$

Proof. Assume that the random variables X_1, \dots, X_n are defined on a probability space $(\bar{\Omega}, \bar{\Sigma}, \bar{\mathbb{P}})$. We will also need two other probability spaces: $(\tilde{\Omega}, \tilde{\Sigma}, \tilde{\mathbb{P}})$ and $(\Omega_\varepsilon, \Sigma_\varepsilon, \mathbb{P}_\varepsilon)$. The main probability space on which all the random variables are defined will be denoted $(\Omega, \Sigma, \mathbb{P})$ and it will be the product space

$$(\Omega, \Sigma, \mathbb{P}) = (\bar{\Omega}, \bar{\Sigma}, \bar{\mathbb{P}}) \times (\tilde{\Omega}, \tilde{\Sigma}, \tilde{\mathbb{P}}) \times (\Omega_\varepsilon, \Sigma_\varepsilon, \mathbb{P}_\varepsilon).$$

The corresponding expectations will be denoted by $\bar{\mathbb{E}}, \tilde{\mathbb{E}}, \mathbb{E}_\varepsilon$ and \mathbb{E} . Let $(\tilde{X}_1, \dots, \tilde{X}_n)$ be an independent copy of (X_1, \dots, X_n) . Think of random variables $\tilde{X}_1, \dots, \tilde{X}_n$ as being defined on $(\tilde{\Omega}, \tilde{\Sigma}, \tilde{\mathbb{P}})$. Denote \tilde{P}_n the empirical measure based on $(\tilde{X}_1, \dots, \tilde{X}_n)$ (it is an independent copy of P_n). Then $\tilde{\mathbb{E}}\tilde{P}_n f = Pf$ and, using Jensen's inequality,

$$\begin{aligned} \mathbb{E}\Phi\left(\|P_n - P\|_{\mathcal{F}}\right) &= \bar{\mathbb{E}}\Phi\left(\|P_n - \tilde{\mathbb{E}}\tilde{P}_n\|_{\mathcal{F}}\right) = \bar{\mathbb{E}}\Phi\left(\|\tilde{\mathbb{E}}(P_n - \tilde{P}_n)\|_{\mathcal{F}}\right) \\ &\leq \tilde{\mathbb{E}}\tilde{\mathbb{E}}\Phi\left(\|P_n - \tilde{P}_n\|_{\mathcal{F}}\right) = \tilde{\mathbb{E}}\tilde{\mathbb{E}}\Phi\left(\left\|n^{-1} \sum_{j=1}^n (\delta_{X_j} - \delta_{\tilde{X}_j})\right\|_{\mathcal{F}}\right). \end{aligned}$$

Since $X_1, \dots, X_n, \tilde{X}_1, \dots, \tilde{X}_n$ are i.i.d., the distribution of $(X_1, \dots, X_n, \tilde{X}_1, \dots, \tilde{X}_n)$ is invariant with respect to all permutations of the components. In particular, one can switch any couple X_j, \tilde{X}_j . Because of this,

$$\tilde{\mathbb{E}}\tilde{\mathbb{E}}\Phi\left(\left\|n^{-1}\sum_{j=1}^n(\delta_{X_j} - \delta_{\tilde{X}_j})\right\|_{\mathcal{F}}\right) = \tilde{\mathbb{E}}\tilde{\mathbb{E}}\Phi\left(\left\|n^{-1}\sum_{j=1}^n\sigma_j(\delta_{X_j} - \delta_{\tilde{X}_j})\right\|_{\mathcal{F}}\right),$$

for an arbitrary choice of $\sigma_j = +1$ or $\sigma_j = -1$. Define now i.i.d. Rademacher random variables on $(\Omega_\varepsilon, \Sigma_\varepsilon, \mathbb{P}_\varepsilon)$ (thus, independent of $(X_1, \dots, X_n, \tilde{X}_1, \dots, \tilde{X}_n)$). Then, we have

$$\tilde{\mathbb{E}}\tilde{\mathbb{E}}\Phi\left(\left\|n^{-1}\sum_{j=1}^n(\delta_{X_j} - \delta_{\tilde{X}_j})\right\|_{\mathcal{F}}\right) = \mathbb{E}_\varepsilon\tilde{\mathbb{E}}\tilde{\mathbb{E}}\Phi\left(\left\|n^{-1}\sum_{j=1}^n\varepsilon_j(\delta_{X_j} - \delta_{\tilde{X}_j})\right\|_{\mathcal{F}}\right)$$

and the proof can be completed as follows:

$$\begin{aligned} \mathbb{E}\Phi\left(\|P_n - P\|_{\mathcal{F}}\right) &\leq \mathbb{E}_\varepsilon\tilde{\mathbb{E}}\tilde{\mathbb{E}}\Phi\left(\left\|n^{-1}\sum_{j=1}^n\varepsilon_j(\delta_{X_j} - \delta_{\tilde{X}_j})\right\|_{\mathcal{F}}\right) \\ &\leq \frac{1}{2}\mathbb{E}_\varepsilon\tilde{\mathbb{E}}\Phi\left(2\left\|n^{-1}\sum_{j=1}^n\varepsilon_j\delta_{X_j}\right\|_{\mathcal{F}}\right) + \frac{1}{2}\mathbb{E}_\varepsilon\tilde{\mathbb{E}}\Phi\left(2\left\|n^{-1}\sum_{j=1}^n\varepsilon_j\delta_{\tilde{X}_j}\right\|_{\mathcal{F}}\right) \\ &= \mathbb{E}\Phi\left(2\|R_n\|_{\mathcal{F}}\right). \end{aligned}$$

The proof of the lower bound is similar. \square

The upper bound is called the *symmetrization inequality* and the lower bound is sometimes called the *desymmetrization inequality*. The desymmetrization inequality is often used together with the following elementary lower bound (in the case of $\Phi(u) = u$)

$$\begin{aligned} \mathbb{E}\|R_n\|_{\mathcal{F}_c} &\geq \mathbb{E}\|R_n\|_{\mathcal{F}} - \sup_{f \in \mathcal{F}} |Pf| \mathbb{E}|R_n(1)| \geq \\ &\geq \mathbb{E}\|R_n\|_{\mathcal{F}} - \sup_{f \in \mathcal{F}} |Pf| \mathbb{E}^{1/2} |n^{-1} \sum_{j=1}^n \varepsilon_j|^2 = \mathbb{E}\|R_n\|_{\mathcal{F}} - \frac{\sup_{f \in \mathcal{F}} |Pf|}{\sqrt{n}}. \end{aligned}$$

2.2 Comparison Inequalities for Rademacher Sums

Given a set $T \subset \mathbb{R}^n$ and i.i.d. Rademacher variables $\varepsilon_i, i = 1, 2, \dots$, it is of interest to know how the expected value of the sup-norm of Rademacher sums indexed by T

$$R_n(T) := \mathbb{E} \sup_{t \in T} \left| \sum_{i=1}^n t_i \varepsilon_i \right|$$

depends on the geometry of the set T . The following beautiful *comparison inequality* for Rademacher sums is due to Talagrand and it is often used to control $R_n(T)$ for more complex sets T in terms of similar quantities for simpler sets.

Theorem 2.2. *Let $T \subset \mathbb{R}^n$ and let $\varphi_i : \mathbb{R} \mapsto \mathbb{R}$, $i = 1, \dots, n$ be functions such that $\varphi_i(0) = 0$ and*

$$|\varphi_i(u) - \varphi_i(v)| \leq |u - v|, \quad u, v \in \mathbb{R}$$

(that is, φ_i are contractions). For all convex nondecreasing functions $\Phi : \mathbb{R}_+ \mapsto \mathbb{R}_+$,

$$\mathbb{E}\Phi\left(\frac{1}{2} \sup_{t \in T} \left| \sum_{i=1}^n \varphi_i(t_i) \varepsilon_i \right| \right) \leq \mathbb{E}\Phi\left(\sup_{t \in T} \left| \sum_{i=1}^n t_i \varepsilon_i \right| \right).$$

Proof. First, we prove that for a nondecreasing convex function $\Phi : \mathbb{R} \mapsto \mathbb{R}_+$ and for an arbitrary $A : T \mapsto \mathbb{R}$

$$\mathbb{E}\Phi\left(\sup_{t \in T} \left[A(t) + \sum_{i=1}^n \varphi_i(t_i) \varepsilon_i \right] \right) \leq \mathbb{E}\Phi\left(\sup_{t \in T} \left[A(t) + \sum_{i=1}^n t_i \varepsilon_i \right] \right). \quad (2.1)$$

We start with the case $n = 1$. Then, the bound is equivalent to the following

$$\mathbb{E}\Phi\left(\sup_{t \in T} [t_1 + \varepsilon \varphi(t_2)]\right) \leq \mathbb{E}\Phi\left(\sup_{t \in T} [t_1 + \varepsilon t_2]\right)$$

for an arbitrary set $T \subset \mathbb{R}^2$ and an arbitrary contraction φ . One can rewrite it as

$$\begin{aligned} & \frac{1}{2} \left(\Phi\left(\sup_{t \in T} [t_1 + \varphi(t_2)]\right) + \Phi\left(\sup_{t \in T} [t_1 - \varphi(t_2)]\right) \right) \\ & \leq \frac{1}{2} \left(\Phi\left(\sup_{t \in T} [t_1 + t_2]\right) + \Phi\left(\sup_{t \in T} [t_1 - t_2]\right) \right). \end{aligned}$$

If now $(t_1, t_2) \in T$ denotes a point where $\sup_{t \in T} [t_1 + \varphi(t_2)]$ is attained and $(s_1, s_2) \in T$ is a point where $\sup_{t \in T} [t_1 - \varphi(t_2)]$ is attained, then it is enough to show that

$$\Phi(t_1 + \varphi(t_2)) + \Phi(s_1 - \varphi(s_2)) \leq \Phi\left(\sup_{t \in T} [t_1 + t_2]\right) + \Phi\left(\sup_{t \in T} [t_1 - t_2]\right)$$

(if the suprema are not attained, one can easily modify the argument). Clearly, we have the following conditions:

$$t_1 + \varphi(t_2) \geq s_1 + \varphi(s_2) \text{ and } t_1 - \varphi(t_2) \leq s_1 - \varphi(s_2).$$

First consider the case when $t_2 \geq 0, s_2 \geq 0$ and $t_2 \geq s_2$. In this case, we will prove that

$$\Phi(t_1 + \varphi(t_2)) + \Phi(s_1 - \varphi(s_2)) \leq \Phi(t_1 + t_2) + \Phi(s_1 - s_2), \quad (2.2)$$

which would imply the bound. Indeed, for

$$a := t_1 + \varphi(t_2), b := t_1 + t_2, c := s_1 - s_2, d := s_1 - \varphi(s_2),$$

we have $a \leq b$ and $c \leq d$ since $\varphi(t_2) \leq t_2$, $\varphi(s_2) \leq s_2$ (by the assumption that φ is a contraction and $\varphi(0) = 0$). We also have that

$$b - a = t_2 - \varphi(t_2) \geq s_2 - \varphi(s_2) = d - c,$$

because again φ is a contraction and $t_2 \geq s_2$. Finally, we have

$$a = t_1 + \varphi(t_2) \geq s_1 + \varphi(s_2) \geq s_1 - s_2 = c.$$

Since the function Φ is nondecreasing and convex, its increment over the interval $[a, b]$ is larger than its increment over the interval $[c, d]$ ($[a, b]$ is longer than $[c, d]$ and $a \geq c$), which is equivalent to (2.2).

If $t_2 \geq 0, s_2 \geq 0$ and $s_2 \geq t_2$, it is enough to use the change of notations $(t, s) \mapsto (s, t)$ and to replace φ with $-\varphi$.

The case $t_2 \leq 0, s_2 \leq 0$ can be now handled by using the transformation $(t_1, t_2) \mapsto (t_1, -t_2)$ and changing the function φ accordingly.

We have to consider the case $t_2 \geq 0, s_2 \leq 0$ (the only remaining case $t_2 \leq 0, s_2 \geq 0$ would again follow by switching the names of t and s and replacing φ with $-\varphi$). In this case, we have $\varphi(t_2) \leq t_2$, $-\varphi(s_2) \leq -s_2$, which, in view of monotonicity of Φ , immediately implies

$$\Phi(t_1 + \varphi(t_2)) + \Phi(s_1 - \varphi(s_2)) \leq \Phi(t_1 + t_2) + \Phi(s_1 - s_2).$$

This completes the proof of (2.1) in the case $n = 1$.

In the general case, we have

$$\begin{aligned} & \mathbb{E}\Phi\left(\sup_{t \in T}\left[A(t) + \sum_{i=1}^n \varphi_i(t_i)\varepsilon_i\right]\right) \\ &= \mathbb{E}_{\varepsilon_1, \dots, \varepsilon_{n-1}} \mathbb{E}_{\varepsilon_n} \Phi\left(\sup_{t \in T}\left[A(t) + \sum_{i=1}^{n-1} \varphi_i(t_i)\varepsilon_i + \varepsilon_n \varphi(t_n)\right]\right). \end{aligned}$$

The expectation $\mathbb{E}_{\varepsilon_n}$ (conditional on $\varepsilon_1, \dots, \varepsilon_{n-1}$) can be bounded using the result in the case $n = 1$. This yields (after changing the order of integration)

$$\mathbb{E}\Phi\left(\sup_{t \in T}\left[A(t) + \sum_{i=1}^n \varphi_i(t_i)\varepsilon_i\right]\right) \leq \mathbb{E}_{\varepsilon_n} \mathbb{E}_{\varepsilon_1, \dots, \varepsilon_{n-1}} \Phi\left(\sup_{t \in T}\left[A(t) + \varepsilon_n t_n + \sum_{i=1}^{n-1} \varphi_i(t_i)\varepsilon_i\right]\right).$$

The proof of (2.1) can now be completed by an induction argument.

Finally, to prove the inequality of the theorem, it is enough to write

$$\begin{aligned}
& \mathbb{E} \Phi \left(\frac{1}{2} \sup_{t \in T} \left| \sum_{i=1}^n \varphi_i(t_i) \varepsilon_i \right| \right) \\
&= \mathbb{E} \Phi \left(\frac{1}{2} \left[\left(\sup_{t \in T} \sum_{i=1}^n \varphi_i(t_i) \varepsilon_i \right)_+ + \left(\sup_{t \in T} \sum_{i=1}^n \varphi_i(t_i) (-\varepsilon_i) \right)_+ \right] \right) \\
&\leq \frac{1}{2} \left[\mathbb{E} \Phi \left(\left(\sup_{t \in T} \sum_{i=1}^n \varphi_i(t_i) \varepsilon_i \right)_+ \right) + \mathbb{E} \Phi \left(\left(\sup_{t \in T} \sum_{i=1}^n \varphi_i(t_i) (-\varepsilon_i) \right)_+ \right) \right],
\end{aligned}$$

where $a_+ := a \vee 0$. Applying the inequality (2.1) to the function $u \mapsto \Phi(u_+)$, which is convex and nondecreasing, completes the proof. \square

We will frequently use a corollary of the above comparison inequality that provides upper bounds on the moments of the sup-norm of Rademacher process R_n on the class

$$\varphi \circ \mathcal{F} := \{\varphi \circ f : f \in \mathcal{F}\}$$

in terms of the corresponding moments of the sup-norm of R_n on \mathcal{F} and Lipschitz constant of function φ .

Theorem 2.3. *Let $\varphi : \mathbb{R} \mapsto \mathbb{R}$ be a contraction satisfying the condition $\varphi(0) = 0$. For all convex nondecreasing functions $\Phi : \mathbb{R}_+ \mapsto \mathbb{R}_+$,*

$$\mathbb{E} \Phi \left(\frac{1}{2} \|R_n\|_{\varphi \circ \mathcal{F}} \right) \leq \mathbb{E} \Phi \left(\|R_n\|_{\mathcal{F}} \right).$$

In particular,

$$\mathbb{E} \|R_n\|_{\varphi \circ \mathcal{F}} \leq 2 \mathbb{E} \|R_n\|_{\mathcal{F}}.$$

The inequality of Theorem 2.3 will be called *the contraction inequality* for Rademacher processes.

A simple rescaling of the class \mathcal{F} allows one to use the contraction inequality in the case of an arbitrary function φ satisfying the Lipschitz condition

$$|\varphi(u) - \varphi(v)| \leq L|u - v|$$

on an arbitrary interval (a, b) that contains the ranges of all the functions in \mathcal{F} . In this case, the last bound of Theorem 2.3 takes the form

$$\mathbb{E} \|R_n\|_{\varphi \circ \mathcal{F}} \leq 2L \mathbb{E} \|R_n\|_{\mathcal{F}}.$$

This implies, for instance, that

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{i=1}^n \varepsilon_i f^2(X_i) \right| \leq 4U \mathbb{E} \sup_{f \in \mathcal{F}} \left| n^{-1} \sum_{i=1}^n \varepsilon_i f(X_i) \right| \quad (2.3)$$

provided that the functions in the class \mathcal{F} are uniformly bounded by a constant U .

2.3 Concentration Inequalities

A well known, simple and useful concentration inequality for functions

$$Z = g(X_1, \dots, X_n)$$

of independent random variables with values in arbitrary spaces is valid under so called *bounded difference condition* on g : there exist constants c_j , $j = 1, \dots, n$ such that for all $j = 1, \dots, n$ and all $x_1, x_2, \dots, x_j, x'_j, \dots, x_n$

$$\left| g(x_1, \dots, x_{j-1}, x_j, x_{j+1}, \dots, x_n) - g(x_1, \dots, x_{j-1}, x'_j, x_{j+1}, \dots, x_n) \right| \leq c_j. \quad (2.4)$$

Theorem 2.4 (Bounded difference inequality). *Under the condition (2.4),*

$$\mathbb{P}\{Z - \mathbb{E}Z \geq t\} \leq \exp\left\{-\frac{2t^2}{\sum_{j=1}^n c_j^2}\right\}$$

and

$$\mathbb{P}\{Z - \mathbb{E}Z \leq -t\} \leq \exp\left\{-\frac{2t^2}{\sum_{j=1}^n c_j^2}\right\}.$$

A standard proof of this inequality is based on bounding the exponential moment $\mathbb{E}e^{\lambda(Z - \mathbb{E}Z)}$, using the following martingale difference representation

$$Z - \mathbb{E}Z = \sum_{j=1}^n \left[\mathbb{E}(Z | X_1, \dots, X_j) - \mathbb{E}(Z | X_1, \dots, X_{j-1}) \right],$$

then using Markov inequality and optimizing the resulting bound with respect to $\lambda > 0$.

In the case when $Z = X_1 + \dots + X_n$, the bounded difference inequality coincides with Hoeffding inequality for sums of bounded independent random variables (see Sect. A.2).

For a class \mathcal{F} of functions uniformly bounded by a constant U , the bounded difference inequality immediately implies the following bounds for $\|P_n - P\|_{\mathcal{F}}$, providing a uniform version of Hoeffding inequality.

Theorem 2.5. *For all $t > 0$,*

$$\mathbb{P}\left\{\|P_n - P\|_{\mathcal{F}} \geq \mathbb{E}\|P_n - P\|_{\mathcal{F}} + \frac{tU}{\sqrt{n}}\right\} \leq \exp\{-t^2/2\}$$

and

$$\mathbb{P}\left\{\|P_n - P\|_{\mathcal{F}} \leq \mathbb{E}\|P_n - P\|_{\mathcal{F}} - \frac{tU}{\sqrt{n}}\right\} \leq \exp\{-t^2/2\}.$$

Developing uniform versions of Bernstein's inequality (see Sect. A.2) happened to be a much harder problem that was solved in the famous papers by Talagrand [138, 139] on concentration inequalities for product measures and empirical processes.

Theorem 2.6 (Talagrand's inequality). *Let X_1, \dots, X_n be independent random variables in S . For any class of functions \mathcal{F} on S that is uniformly bounded by a constant $U > 0$ and for all $t > 0$*

$$\mathbb{P}\left\{\left|\left\|\sum_{i=1}^n f(X_i)\right\|_{\mathcal{F}} - \mathbb{E}\left\|\sum_{i=1}^n f(X_i)\right\|_{\mathcal{F}}\right| \geq t\right\} \leq K \exp\left\{-\frac{1}{K} \frac{t}{U} \log\left(1 + \frac{tU}{V}\right)\right\},$$

where K is a universal constant and V is any number satisfying

$$V \geq \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^n f^2(X_i).$$

Using symmetrization inequality and contraction inequality for the square (2.3), it is easy to show that in the case of i.i.d. random variables X_1, \dots, X_n with distribution P

$$\mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^n f^2(X_i) \leq n \sup_{f \in \mathcal{F}} P f^2 + 8U \mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}}. \quad (2.5)$$

The right hand side of this bound is a common choice of the quantity V involved in Talagrand's inequality. Moreover, in the case when $\mathbb{E}f(X) = 0$, the desymmetrization inequality yields

$$\mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} \leq 2 \mathbb{E} \left\| \sum_{i=1}^n f(X_i) \right\|_{\mathcal{F}}.$$

As a result, one can use Talagrand's inequality with

$$V = n \sup_{f \in \mathcal{F}} P f^2 + 16U \mathbb{E} \left\| \sum_{i=1}^n f(X_i) \right\|_{\mathcal{F}}$$

and the size of $\left\| \sum_{i=1}^n f(X_i) \right\|_{\mathcal{F}}$ is now controlled in terms of its expectation only.

This form of Talagrand's inequality is especially convenient and there have been considerable efforts to find explicit and sharp values of the constants in such inequalities. In particular, we will frequently use the bounds proved by Bousquet [33] and Klein [77] (in fact, Klein and Rio [78] provide an improved version of this inequality). Namely, for a class \mathcal{F} of measurable functions from S into $[0, 1]$ (by a simple rescaling $[0, 1]$ can be replaced by any bounded interval) the following bounds hold for all $t > 0$:

Bousquet bound

$$\mathbb{P}\left\{\|P_n - P\|_{\mathcal{F}} \geq \mathbb{E}\|P_n - P\|_{\mathcal{F}} + \sqrt{2\frac{t}{n}\left(\sigma_P^2(\mathcal{F}) + 2\mathbb{E}\|P_n - P\|_{\mathcal{F}}\right)} + \frac{t}{3n}\right\} \leq e^{-t}$$

Klein-Rio bound

$$\mathbb{P}\left\{\|P_n - P\|_{\mathcal{F}} \leq \mathbb{E}\|P_n - P\|_{\mathcal{F}} - \sqrt{2\frac{t}{n}\left(\sigma_P^2(\mathcal{F}) + 2\mathbb{E}\|P_n - P\|_{\mathcal{F}}\right)} - \frac{t}{n}\right\} \leq e^{-t}.$$

Here

$$\sigma_P^2(\mathcal{F}) := \sup_{f \in \mathcal{F}} \left(P f^2 - (P f)^2 \right).$$

We will also need a version of Talagrand's inequality for unbounded classes of functions. Given a class \mathcal{F} of measurable functions $f : S \mapsto \mathbb{R}$, denote by F an envelope of \mathcal{F} , that is, a measurable function such that $|f(x)| \leq F(x)$, $x \in S$, $f \in \mathcal{F}$. The next bounds follow from Theorem 4 of Adamczak [1]: for all $\alpha \in (0, 1]$ there exists a constant $K = K(\alpha)$ such that

Adamczak bound

$$\mathbb{P}\left\{\|P_n - P\|_{\mathcal{F}} \geq K \left[\mathbb{E}\|P_n - P\|_{\mathcal{F}} + \sigma_P(\mathcal{F}) \sqrt{\frac{t}{n}} + \left\| \max_{1 \leq j \leq n} F(X_j) \right\|_{\psi_\alpha} \frac{t^{1/\alpha}}{n} \right] \right\} \leq e^{-t}$$

and

$$\mathbb{P}\left\{\mathbb{E}\|P_n - P\|_{\mathcal{F}} \geq K \left[\|P_n - P\|_{\mathcal{F}} + \sigma_P(\mathcal{F}) \sqrt{\frac{t}{n}} + \left\| \max_{1 \leq j \leq n} F(X_j) \right\|_{\psi_\alpha} \frac{t^{1/\alpha}}{n} \right] \right\} \leq e^{-t}.$$

Concentration inequalities can be also applied to the Rademacher process which can be viewed as an empirical process based on the sample $(X_1, \varepsilon_1), \dots, (X_n, \varepsilon_n)$ in the space $S \times \{-1, 1\}$ and indexed by the class of functions $\tilde{\mathcal{F}} := \{\tilde{f} : f \in \mathcal{F}\}$, where $\tilde{f}(x, u) := f(x)u$, $(x, u) \in S \times \{-1, 1\}$.

2.4 Exponential Bounds for Sums of Independent Random Matrices

In this section, we discuss very simple, but powerful noncommutative Bernstein type inequalities that go back to Ahlswede and Winter [4]. The goal is to bound the tail probability $\mathbb{P}\{\|X_1 + \cdots + X_n\| \geq t\}$, where X_1, \dots, X_n are independent Hermitian random $m \times m$ matrices with $\mathbb{E}X_j = 0$ and $\|\cdot\|$ is the operator norm.¹ The proofs of such inequalities are based on a matrix extension of the classical proof of Bernstein's inequality for real valued random variables, but they also rely on important matrix inequalities that have many applications in mathematical physics. In the case of sums of i.i.d. random matrices, it is enough to use the following well known *Golden-Thompson inequality* (see, e.g., Simon [133], p. 94):

Proposition 2.1. *For arbitrary Hermitian $m \times m$ matrices A, B*

$$\mathrm{tr}(e^{A+B}) \leq \mathrm{tr}(e^A e^B).$$

It is needed to control the matrix moment generating function

$$\mathbb{E} \mathrm{tr} \exp\{\lambda(X_1 + \cdots + X_n)\}.$$

This approach was used in the original paper by Ahlswede and Winter [4], but also in [70, 88, 124]. However, it does not seem to provide the correct form of “variance parameter” in the non i.i.d. case. We will use below another approach suggested by Tropp [142] that is based on the following classical result by Lieb [102] (Theorem 6).

Proposition 2.2. *For all Hermitian matrices A , the function*

$$G_A(S) := \mathrm{tr} \exp\{A + \log S\}$$

is concave on the cone of Hermitian positively definite matrices.

Given independent Hermitian random $m \times m$ matrices X_1, \dots, X_n with $\mathbb{E}X_j = 0$, denote

$$\sigma^2 := n^{-1} \left\| \mathbb{E}(X_1^2 + \cdots + X_n^2) \right\|.$$

Theorem 2.7. *1. Suppose that, for some $U > 0$ and for all $j = 1, \dots, n$, $\|X_j\| \leq U$. Then*

$$\mathbb{P}\left\{\|X_1 + \cdots + X_n\| \geq t\right\} \leq 2m \exp\left\{-\frac{t^2}{2\sigma^2 n + 2Ut/3}\right\}. \quad (2.6)$$

2. Let $\alpha \geq 1$ and suppose that, for some $U^{(\alpha)} > 0$ and for all $j = 1, \dots, n$,

¹For the notations used in this section, see Sect. A.4.

$$\left\| \|X_j\| \right\|_{\psi_\alpha} \vee 2\mathbb{E}^{1/2} \|X_j\|^2 \leq U^{(\alpha)}.$$

Then, there exists a constant $K > 0$ such that

$$\mathbb{P}\{\|X_1 + \dots + X_n\| \geq t\} \leq 2m \exp\left\{-\frac{1}{K} \frac{t^2}{n\sigma^2 + tU^{(\alpha)} \log^{1/\alpha}(U^{(\alpha)}/\sigma)}\right\}. \quad (2.7)$$

Inequality (2.6) is a direct noncommutative extension of classical Bernstein's inequality for sums of independent random variables. It is due to Ahlswede and Winter [4] (see also [70, 124, 142]). In inequality (2.7), the L_∞ -bound U on $\|X_j\|$ is replaced by a weaker ψ_α -norm. This inequality was proved in [88] in the i.i.d. case and in [89] in the general case. We follow the last paper below. Note that, when $\alpha \rightarrow \infty$, (2.7) coincides with (2.6) (up to constants).

Proof. Denote $Y_n := X_1 + \dots + X_n$ and observe that $\|Y_n\| < t$ if and only if $-tI_m < Y_n < tI_m$. It follows that

$$\mathbb{P}\{\|Y_n\| \geq t\} \leq \mathbb{P}\{Y_n \not\prec tI_m\} + \mathbb{P}\{Y_n \not\prec -tI_m\}. \quad (2.8)$$

The next bounds are based on a simple matrix algebra:

$$\mathbb{P}\{Y_n \not\prec tI_m\} = \mathbb{P}\{e^{\lambda Y_n} \not\prec e^{\lambda t I_m}\} \leq \mathbb{P}\left\{\text{tr}\left(e^{\lambda Y_n}\right) \geq e^{\lambda t}\right\} \leq e^{-\lambda t} \mathbb{E} \text{tr}(e^{\lambda Y_n}). \quad (2.9)$$

To bound the matrix moment generating function $\mathbb{E} \text{tr}(e^{\lambda Y_n})$, observe that

$$\mathbb{E} \text{tr}(e^{\lambda Y_n}) = \mathbb{E} \mathbb{E}_n \text{tr} \exp\{\lambda Y_{n-1} + \log e^{\lambda X_n}\} = \mathbb{E} \mathbb{E}_n G_{\lambda Y_{n-1}}(e^{\lambda X_n}).$$

where \mathbb{E}_n denotes the conditional expectation given X_1, \dots, X_{n-1} . Using Lieb's theorem (see Proposition 2.2), Jensen's inequality for the expectation \mathbb{E}_n and the independence of random matrices X_j , we get

$$\mathbb{E} \text{tr}(e^{\lambda Y_n}) \leq \mathbb{E} G_{\lambda Y_{n-1}}(\mathbb{E} e^{\lambda X_n}) = \mathbb{E} \text{tr} \exp\{\lambda Y_{n-1} + \log \mathbb{E} e^{\lambda X_n}\}.$$

Using the same conditioning trick another time, we get

$$\begin{aligned} \mathbb{E} \text{tr}(e^{\lambda Y_n}) &\leq \mathbb{E} \text{tr} \exp\{\lambda Y_{n-1} + \log \mathbb{E} e^{\lambda X_n}\} \\ &= \mathbb{E} \mathbb{E}_{n-1} \text{tr} \exp\{\lambda Y_{n-2} + \log \mathbb{E} e^{\lambda X_n} + \log e^{\lambda X_{n-1}}\} = \mathbb{E} \mathbb{E}_n G_{\lambda Y_{n-2} + \log \mathbb{E} e^{\lambda X_n}}(e^{\lambda X_{n-1}}) \end{aligned}$$

and another application of Lieb's theorem and Jensen's inequality yields

$$\mathbb{E} \text{tr}(e^{\lambda Y_n}) \leq \mathbb{E} \text{tr} \exp\{\lambda Y_{n-2} + \log \mathbb{E} e^{\lambda X_{n-1}} + \log \mathbb{E} e^{\lambda X_n}\}.$$

Iterating this argument, we get

$$\mathbb{E}\text{tr}(e^{\lambda Y_n}) \leq \text{tr} \exp\{\log \mathbb{E}e^{\lambda X_1} + \log \mathbb{E}e^{\lambda X_2} + \dots + \log \mathbb{E}e^{\lambda X_n}\}. \quad (2.10)$$

Next we have to bound $\mathbb{E}e^{\lambda X}$ for an arbitrary Hermitian random matrix with $\mathbb{E}X = 0$ and $\|X\| \leq U$. To this end, we use the Taylor expansion:

$$\begin{aligned} \mathbb{E}e^{\lambda X} &= I_m + \mathbb{E}\lambda^2 X^2 \left[\frac{1}{2!} + \frac{\lambda X}{3!} + \frac{\lambda^2 X^2}{4!} + \dots \right] \\ &\leq I_m + \lambda^2 \mathbb{E}X^2 \left[\frac{1}{2!} + \frac{\lambda \|X\|}{3!} + \frac{\lambda^2 \|X\|^2}{4!} + \dots \right] \\ &= I_m + \lambda^2 \mathbb{E}X^2 \left[\frac{e^{\lambda \|X\|} - 1 - \lambda \|X\|}{\lambda^2 \|X\|^2} \right]. \end{aligned}$$

Under the assumption $\|X\| \leq U$, this yields

$$\mathbb{E}e^{\lambda X} \leq I_m + \lambda^2 \mathbb{E}X^2 \left[\frac{e^{\lambda U} - 1 - \lambda U}{\lambda^2 U^2} \right].$$

Denoting $\phi(u) := \frac{e^u - 1 - u}{u^2}$, we easily get

$$\log \mathbb{E}e^{\lambda X} \leq \lambda^2 \mathbb{E}X^2 \phi(\lambda U).$$

We will use this bound for each random matrix X_j and substitute the result in (2.10) to get

$$\begin{aligned} \mathbb{E}\text{tr}(e^{\lambda Y_n}) &\leq \text{tr} \exp\left\{\lambda^2 \mathbb{E}(X_1^2 + \dots + X_n^2) \phi(\lambda U)\right\} \\ &\leq m \exp\left\{\lambda^2 \mathbb{E}(X_1^2 + \dots + X_n^2) \phi(\lambda U)\right\}. \end{aligned}$$

In view of (2.9), it remains to follow the usual proof of Bernstein–Bennett type inequalities to obtain (2.6).

To prove (2.7), we bound $\mathbb{E}e^{\lambda X}$ in a slightly different way. We do it for an arbitrary Hermitian random matrix with $\mathbb{E}X = 0$ and

$$\left\| \|X\| \right\|_{\psi_\alpha} \vee 2\mathbb{E}^{1/2} \|X\|^2 \leq U^{(\alpha)}.$$

For all $\tau > 0$, we get

$$\begin{aligned}
\mathbb{E}e^{\lambda X} &\leq I_m + \lambda^2 \mathbb{E}X^2 \left[\frac{e^{\lambda \|X\|} - 1 - \lambda \|X\|}{\lambda^2 \|X\|^2} \right] \\
&\leq I_m + \lambda^2 \mathbb{E}X^2 \left[\frac{e^{\lambda \tau} - 1 - \lambda \tau}{\lambda^2 \tau^2} \right] + I_m \lambda^2 \mathbb{E}\|X\|^2 \left[\frac{e^{\lambda \|X\|} - 1 - \lambda \|X\|}{\lambda^2 \|X\|^2} \right] I(\|X\| \geq \tau).
\end{aligned}$$

Take $M := 2(\log 2)^{1/\alpha} U^{(\alpha)}$ and assume that $\lambda \leq 1/M$. It follows that

$$\begin{aligned}
\mathbb{E}\|X\|^2 \left[\frac{e^{\lambda \|X\|} - 1 - \lambda \|X\|}{\lambda^2 \|X\|^2} \right] I(\|X\| \geq \tau) &\leq M^2 \mathbb{E}e^{\|X\|/M} I(\|X\| \geq \tau) \leq \\
M^2 \mathbb{E}^{1/2} e^{2\|X\|/M} \mathbb{P}^{1/2}\{\|X\| \geq \tau\}.
\end{aligned}$$

Since, for $\alpha \geq 1$,

$$M = 2(\log 2)^{1/\alpha} U^{(\alpha)} \geq 2 \left\| \|X\| \right\|_{\psi_1}$$

(see Sect. A.1), we get $\mathbb{E}e^{2\|X\|/M} \leq 2$ and also

$$\mathbb{P}\{\|X\| \geq \tau\} \leq \exp \left\{ -2^\alpha \log 2 \left(\frac{\tau}{M} \right)^\alpha \right\}.$$

Therefore, the following bound holds:

$$\mathbb{E}e^{\lambda X} \leq I_m + \lambda^2 \mathbb{E}X^2 \left[\frac{e^{\lambda \tau} - 1 - \lambda \tau}{\lambda^2 \tau^2} \right] + 2^{1/2} \lambda^2 M^2 \exp \left\{ -2^{\alpha-1} \log 2 \left(\frac{\tau}{M} \right)^\alpha \right\} I_m.$$

Take now $\tau := M \frac{2^{1/\alpha-1}}{(\log 2)^{1/\alpha}} \log^{1/\alpha} \frac{M^2}{\sigma^2}$ and suppose that λ satisfies the condition $\lambda \tau \leq 1$. This yields the following bound

$$\mathbb{E}e^{\lambda X} \leq I_m + \frac{C_1}{2} \lambda^2 (\mathbb{E}X^2 + \sigma^2 I_m),$$

which implies that

$$\log \mathbb{E}e^{\lambda X} \leq \frac{C_1}{2} \lambda^2 (\mathbb{E}X^2 + \sigma^2 I_m)$$

with some constant $C_1 > 0$. We use the last bound for each random matrix X_j , $j = 1, \dots, n$ and deduce from (2.10) that, for some constants $C_1, C_2 > 0$ and for all λ satisfying the condition

$$\lambda U^{(\alpha)} \left(\log \frac{U^{(\alpha)}}{\sigma} \right)^{1/\alpha} \leq C_2, \quad (2.11)$$

we have

$$\mathbb{E} \text{tr}(e^{\lambda Y_n}) \leq \text{tr} \exp \left\{ \frac{C_1}{2} \lambda^2 (\mathbb{E} X_1^2 + \cdots + \mathbb{E} X_n^2 + n \sigma^2 I_m) \right\},$$

which further implies that

$$\mathbb{E} \text{tr}(e^{\lambda Y_n}) \leq m \exp \{ C_1 \lambda^2 n \sigma^2 \}.$$

Combining this bound with (2.8) and (2.9), we get

$$\mathbb{P}\{\|Y_n\| \geq t\} \leq 2m \exp \left\{ -\lambda t + C_1 \lambda^2 n \sigma^2 \right\}.$$

The last bound can be now minimized with respect to all λ satisfying (2.11), which yields that, for some constant $K > 0$,

$$\mathbb{P}\{\|Y_n\| \geq t\} \leq 2m \exp \left\{ -\frac{1}{K} \frac{t^2}{n \sigma^2 + t U^{(\alpha)} \log^{1/\alpha}(U^{(\alpha)}/\sigma)} \right\}.$$

This proves inequality (2.7). \square

The next bounds immediately follow from (2.6) and (2.7): for all $t > 0$, with probability at least $1 - e^{-t}$

$$\left\| \frac{X_1 + \cdots + X_n}{n} \right\| \leq 2 \left(\sigma \sqrt{\frac{t + \log(2m)}{n}} \vee U \frac{t + \log(2m)}{n} \right) \quad (2.12)$$

and, with some constant $C > 0$,

$$\begin{aligned} \left\| \frac{X_1 + \cdots + X_n}{n} \right\| &\leq C \left(\sigma \sqrt{\frac{t + \log(2m)}{n}} \vee \right. \\ &\quad \left. U^{(\alpha)} \left(\log \frac{U^{(\alpha)}}{\sigma} \right)^{1/\alpha} \frac{t + \log(2m)}{n} \right). \end{aligned} \quad (2.13)$$

Note that the size m of the matrices has only logarithmic impact on the bounds.

It is easy to derive Bernstein type exponential inequalities for rectangular $m_1 \times m_2$ random matrices from the inequalities of Theorem 2.7 for Hermitian matrices. This is based on the following well known isomorphism trick (sometimes called *Paulsen dilation*). Denote by $\mathbb{M}_{m_1, m_2}(\mathbb{R})$ the space of all $m_1 \times m_2$ matrices with real entries and by $\mathbb{H}_m(\mathbb{C})$ the space of all Hermitian $m \times m$ matrices. Define the following linear mapping

$$J : \mathbb{M}_{m_1, m_2}(\mathbb{R}) \mapsto \mathbb{H}_{m_1 + m_2}(\mathbb{C}), \quad \text{where } JS := \begin{pmatrix} O & S \\ S^* & O \end{pmatrix}.$$

Clearly,

$$(JS)^2 := \begin{pmatrix} SS^* & 0 \\ 0 & S^*S \end{pmatrix}.$$

Therefore,

$$\|JS\| = \|SS^*\|^{1/2} \vee \|S^*S\|^{1/2} = \|S\|$$

and, for independent random matrices X_1, \dots, X_n in $\mathbb{M}_{m_1, m_2}(\mathbb{R})$ with $\mathbb{E}X_j = 0$, we have

$$\begin{aligned} \sigma^2 &:= n^{-1} \left(\|\mathbb{E}(X_1 X_1^*) + \dots + \mathbb{E}(X_n X_n^*)\| \vee \|\mathbb{E}(X_1^* X_1) + \dots + \mathbb{E}(X_n^* X_n)\| \right) \\ &= n^{-1} \|\mathbb{E}((JX_1)^2 + \dots + (JX_n)^2)\|. \end{aligned}$$

The following statement immediately follows from Theorem 2.7 by applying it to the Hermitian random matrices JX_1, \dots, JX_n .

Corollary 2.1. *1. Let $m := m_1 + m_2$. Suppose that, for some $U > 0$ and for all $j = 1, \dots, n$, $\|X_j\| \leq U$. Then*

$$\mathbb{P} \left\{ \|X_1 + \dots + X_n\| \geq t \right\} \leq 2m \exp \left\{ -\frac{t^2}{2\sigma^2 n + 2Ut/3} \right\}. \quad (2.14)$$

2. Let $\alpha \geq 1$ and suppose that for some $U^{(\alpha)} > 0$ and for all $j = 1, \dots, n$,

$$\left\| \|X_j\| \right\|_{\psi_\alpha} \vee 2\mathbb{E}^{1/2} \|X\|^2 \leq U^{(\alpha)}.$$

Then, there exists a constant $K > 0$ such that

$$\mathbb{P} \{ \|X_1 + \dots + X_n\| \geq t \} \leq 2m \exp \left\{ -\frac{1}{K} \frac{t^2}{n\sigma^2 + tU^{(\alpha)} \log^{1/\alpha}(U^{(\alpha)}/\sigma_X)} \right\}. \quad (2.15)$$

2.5 Further Comments

Initially, the theory of empirical processes dealt with asymptotic problems: uniform versions of laws of large numbers, central limit theorem and laws of iterated logarithm. It started with the work by Vapnik and Chervonenkis (see [147] and references therein) on Glivenko-Cantelli problem and by Dudley [59] on the central limit theorem (extensions of Kolmogorov–Donsker theorems). Other early references include Koltchinskii [80], Pollard [122] and Giné and Zinn [69]. Since Talagrand [138, 139] developed his concentration inequalities, the focus of the theory has shifted to the development of bounds on sup-norms of empirical processes with applications to a variety of problems in statistics, learning theory, asymptotic geometric analysis, etc (see also [137]).

Symmetrization inequalities of Sect. 2.1 were introduced to the theory of empirical processes by Giné and Zinn [69] (an earlier form of Rademacher symmetrization was used by Koltchinskii [80] and Pollard [122]).

In Sect. 2.2, we follow the proof of Talagrand's comparison inequality for Rademacher sums given by Ledoux and Talagrand [101], Theorem 4.12.

Talagrand's concentration inequalities for product measures and empirical processes were proved in [138, 139]. Another approach to their proof, the entropy method based on logarithmic Sobolev inequalities, was introduced by Ledoux. It is discussed in detail in [100] and [107] (see also [30]). The bounded difference inequality based on the martingale method is well known and can be found in many books (e.g., [51, 107]).

Noncommutative Bernstein's inequality (2.6) was discovered by Ahlswede and Winter [4]. This inequality and its extensions proved to be very useful in the recent work on low rank matrix recovery (see Gross et al. [71], Gross [70], Recht [124], Koltchinskii [88]). Tropp [142] provides a detailed review of various inequalities of this type.

Oracle Inequalities in Empirical Risk Minimization and
Sparse Recovery Problems

École d'Été de Probabilités de Saint-Flour XXXVIII-2008

Koltchinskii, V.

2011, IX, 254 p., Softcover

ISBN: 978-3-642-22146-0