Chapter 6: Maximal Inequalities and Chaining

Empirical Processes (MATH-522)

Myrto Limnios

April, 2025

This chapter focuses on providing an upperbound of $\mathbb{E}\left[\sup_{t\in T}|X_t|\right]$ and shows why it measures the *size* of a generic process $\{X_t\}_{t\in T}$, when the index set is considered to be infinite. We will present a general method, namely *chaining method*, for obtaining sharp bounds of the quantity $\mathbb{E}\left[\sup_{t\in T}|X_t|\right]$ called *maximal inequalities*. We see that if the *size* of the index set T can be analyzed w.r.t. a distance based on the process X as in the Chapter 4, then we can control the *worst* deviation of the process uniformly on T.

Notations. We will use the notations of Chapter 5 without further notice. Recall that we considered an i.i.d. samples $X_1, \dots X_n$ defined on the p.s. $(\Omega, \mathcal{A}, \mathbb{P})$, valued in a measurable space $\mathcal{X} \subseteq \mathbb{R}^d$, $d \in \mathbb{N}^*$, of probability distribution P and of empirical p.d. P_n.

1 Introduction to the Chaining Method

1.1 Finite Index Set

Suppose the set T to be finite. We want to upperbound the maximum of a finite number of r.v.s. Notice that

$$\mathbb{E}[\sup_{t\in T} X_t] \leq \mathbb{E}[\sum_{t\in T}|X_t|] \leq |T| \sup_{t\in T}\mathbb{E}|X_t| \ .$$

Remark 1.1. *1. Controlling the magnitude of each of the r.v.s X-s seems unsatisfactory, and we want to take advantage of possibly some tail assumption of the r.v.s.*

2. The bound grows linearly with the size of T that is, again, unsatisfactory. It seems that we cannot get any good conclusion from this.

Suppose now that the r.v.s X have bounded p-moment, then Jensen's inequality helps us understand a refined control

$$\mathbb{E}[\sup_{t\in T} X_t] \leq \mathbb{E}[\sup_{t\in T}|X_t|^p]^{1/p} \leq |T|^{1/p} \sup_{t\in T}\mathbb{E}[|X_t|^p]^{1/p} \ .$$

Remark 1.2. *This bound is more interesting: if p is large, then the tails of the r.v.s are vanishing implying a smaller value for the expectation. In addition, the larger p and the slower growth in terms of $|T|$.*

We prove a first *maximal inequality* to see how to use more general functionals related to Cramér-Chernoff method resulting in sharper bounds. Before stating the first result, recall an important definition.

Definition 1.1. A process $(X_t)_{t\in T}$ defined on a metric space (T,d) is centered $(\nu>0)$-sub-Gaussian if $PX_t=0$, if, for all $\lambda>0$, for all $t\in T$,

$$\psi_X(\lambda)=\log\mathbb{E}[e^{\lambda X_t}]\leq\frac{\lambda^2\nu}{2}\ .$$

Lemma 1.2. *Consider a finite collection of elements $|T|<\infty$, and the process X_t be centered ν-sub-Gaussian. Suppose that we observe an i.i.d. sequence of X_t of size n of empirical measure P_n. Then it holds true that*

$$\mathbb{E}\left[\max_{t\in T}|P_nt|\right]\leq\sqrt{\frac{\nu\log(2|T|)}{n}}\ . \tag{1}$$

Remark 1.3.

1. *The result is interesting as soon as we choose $n\geq\sqrt{\nu\log(2|T|)}$.*
2. *We study the max because the class of functions is finite, thus the supnorm is attainable.*
3. *Notice that the size of the class enters into play similarly as the square-root of the entropy.*
4. *This is the strongest result we can have, and it is sharp: If we know the best constant v upperbounding the variance of the process, then we cannot obtain better upperbound.*
5. *Notice that it can be related to Massart's inequality (Massart (2000), Lemma 5.2). Notice that we consider the absolute valued here, yielding a 2 in the log.*

Proof. By Jensen's inequality:

$$\begin{aligned}\mathbb{E}\left[\max_{t\in T}|P_nt|\right] &= \frac{1}{\lambda}\mathbb{E}\left[\log\exp\lambda\max_{t\in T}|P_nt|\right]\\ &\leq \frac{1}{\lambda}\log\mathbb{E}\left[\exp\lambda\max_{t\in T}|P_nt|\right]\\ &= \frac{1}{\lambda}\log\mathbb{E}\left[\max_{t\in T}\exp\lambda|P_nt|\right]\\ &\leq \frac{1}{\lambda}\log\sum_{t\in T}\mathbb{E}\left[\exp\lambda|P_nt|\right]\ .\end{aligned}$$

Notice that $e^{|x|}\leq e^x+e^{-x}$, so that for $\lambda\geq0$, and using the sub-Gaussianity assumption yields

$$\mathbb{E}[\exp\{\lambda|P_nt|\}]\leq\mathbb{E}[\exp\{\lambda P_nt\}]+\mathbb{E}[\exp\{-\lambda P_nt\}]\leq2e^{\lambda^2\nu/2n}\ .$$

Thus

$$\mathbb{E}\left[\max_{t\in T}|P_nt|\right]\leq\frac{\log(2|T|)}{\lambda}+\frac{\lambda\nu}{2n}\ .$$

Minimizing w.r.t. $\lambda^*=\sqrt{2n\log2|T|/\nu}$ yields the result.

□

We can see this result similarly to the Chernoff bound that we studied in Chapter 2, i.e., if $\log\mathbb{E}[e^{\lambda X_t}]\leq\psi(\lambda)$, then $\mathbb{P}(X_t\geq u)\leq e^{-\psi^*(u)}$, for all $u\geq0$ and $t\in T$. We thus formulate the maximal tail inequality.

Lemma 1.3 (Maximal tail inequality)**.** *Consider the process X_t as defined in Lemma 1.2, then for all $\delta\in(0,1)$, with probability at least $1-\delta$*

$$\max_{t\in T}|P_nt|\leq\frac{\sqrt{\nu\log(2N)}}{n}+\frac{\sqrt{\nu\log(1/\delta)}}{n}\ .$$

Remark 1.4. *This result is again sharp, as soon as all the X-s are independent. Notice that if there is dependence, and for example all r.v.s are equal for all t, then essentially $\max_{i\leq N}|P_n t_j| = |P_n t_1|$ and thus the bound is far from being optimal.*

The next section provides a generic method for obtaining similar rate of convergence, when the index set is uncountable.

1.2 One-step Discretization Chain under an Entropic Condition

From that simple derivation when considering the set T to be finite/countable, we saw that we can replace the supremum by the maximum, to then invoke the union bound. The idea is to approximate the supremum over T by a maximum of increments over an increasing sequence of covering sets with accuracy ε, plus an approximation error depending on ε and converging to 0. We will see that the sup can be upperbounded by the maximum of increments that depend on their size and number. Consider now an index class that can be approximated with an ε-cover w.r.t. the stochastic distance based d_X based on the process X_t. A key tool for the following results is to be able to control the size of the increments of the process $X_s - X_t$ in terms of the distance between the two points s and t, formulated below.

Definition 1.4. A process $(X_t)_{t\in T}$ defined on a metric space (T,d) is said to satisfy the **increment condition** if, for all $u>0$, for all $s,t\in T$,

$$\mathbb{P}(|X_s - X_t| \geq u d(s,t)) \leq 2\exp\left(-\frac{u^2}{2}\right) .$$

We say that the diameter of (T,d) is defined by $D(T) = \sup_{s,t\in T} d(s,t)$.

Reminder 1.5. *A centered process $(X_t)_{t\in T}$ defined on a metric space (T,d) is said to be sub-Gaussian iff*

$$\mathbb{E}[e^{\lambda(X_s - X_t)}] \leq e^{\lambda^2 d_X(s,t)^2/2} ,$$

for all $\lambda\in\mathbb{R}$, and for all $s,t\in T$.

Remark 1.6. *Sub-Gaussian processes satisfy the increment condition w.r.t. a stochatstic metric d_X that it can be a pseudometric. It is typically $d_X(s,t) = \mathbb{E}[|X_s - X_t|^2]^{1/2}$. Gaussian or Rademacher processes indexed on $[0,1]$, we consider the Euclidean metric $d(s,t) = \|s-t\|_2$.*

Lemma 1.5. *Suppose $(X_t)_{t\in T}$ to be a sub-Gaussian centered process w.r.t d_X. Then for any $\varepsilon\in[0,D(T)]$, such that $N(\varepsilon,T,d_X)\geq c$, with $c>0$ universal constant, it holds true that*

$$\mathbb{E}\left[\sup_{t,t'\in T}(X_t - X_{t'})\right] \leq 2\mathbb{E}\left[\sup_{t,t'\in T,\, d_X(t,t')\leq\varepsilon}(X_t - X_{t'})\right] + 4D(T)\sqrt{\log N(\varepsilon,T,d_X)} .$$

Proof. The idea of the proof is to use a cover of T to approximate the increment $(X_t - X_{t'})$ by the increments based on the centers of the covering sequence, with an additional approximation error.

Let $\varepsilon>0$. Define by $t^1,\dots,t^N$ the centers of the ε-cover of T. Then, for any $t\in T$, there exists an index $i\leq N$, such that $d_X(t,t^i)\leq\varepsilon$. Hence

$$\begin{aligned} X_t - X_{t^1} &= \underbrace{X_t - X_{t^i}}_{\text{increment in } X \text{ between } t \text{ and its closest center } t^i} \\ &+ \underbrace{X_{t^i} - X_{t^1}}_{\text{increment in } X \text{ between the } \textit{best} \text{ approximation of } t \text{ and any center of the cover}} \end{aligned}$$

Notice that

$$X_t - X_{t^i} \leq \sup_{t,t' \in T,\ d_X(t,t') \leq \varepsilon} (X_t - X_{t'})$$

because both t and t^i are in the ith element of the cover, thus $d_X(t, t^i) \leq \varepsilon$ by construction. And

$$X_{t^i} - X_{t^1} \leq \max_{i \leq N} |X_{t^i} - X_{t^1}|$$

because both t^i and t^1 are centers of the cover of T. Thus

$$X_t - X_{t^1} \leq \sup_{t,t' \in T,\ d_X(t,t') \leq \varepsilon} (X_t - X_{t'}) + \max_{i \leq N} |X_{t^i} - X_{t^1}| \ . \tag{2}$$

It holds true for any point $t' \in T$ as well (the cover is independent of t), so that we can add both bounds to obtain

$$\sup_{t,t' \in T} (X_t - X_{t'}) \leq 2 \sup_{t,t' \in T,\ d_X(t,t') \leq \varepsilon} (X_t - X_{t'}) + 2 \max_{i \leq N} |X_{t^i} - X_{t^1}| \ .$$

Notice that because the r.v.s are sub-Gaussian, then each increament is centered sub-Gaussian as well, with at most $d_X(t^i, t^1) \leq D(T)$ for the index set of the max. Lemma 1.2 applies, to the maximum on the right

$$\mathbb{E}\left[\max_{i \leq N} |X_{t^i} - X_{t^1}|\right] \leq 2D(T)\sqrt{\log N} \ .$$

We can select the optimal size of the cover to be $N = N(\varepsilon, T, d_X)$, and the result is proved. □

Remark 1.7. *It is not clear why we consider increments until we point out the following fact. For any fixed $t' \in T$, then*

$$\mathbb{E}\left[\sup_{t \in T} X_t\right] = \mathbb{E}\left[\sup_{t \in T} (X_t - X_{t'})\right]$$

and it is evident now that

$$\mathbb{E}\left[\sup_{t \in T} X_t\right] \leq \mathbb{E}\left[\sup_{t,t' \in T} (X_t - X_{t'})\right] \ .$$

We consider some examples.

Example 1.8 (Lipschitz processes)**.** *Consider the assumptions from Lemma 1.5, and that there exists a r.v. L such that X_t is L-Lipschitz, then*

$$\mathbb{E}\left[\sup_{t,t' \in T} (X_t - X_{t'})\right] \leq 2 \inf_{\varepsilon \in [0, D(T)]} \{\varepsilon \mathbb{E}[L] + 4D(T)\sqrt{\log N(\varepsilon, T, d_X)}\} \ .$$

Example 1.9 (Localized Canonical Gaussian Complexity)**.** *Recall that $\mathcal{G}(T) = \mathbb{E}[\sup_{t \in T} G_t] = \mathbb{E}[\sup_{t \in T} \langle \eta, t \rangle]$, with $T \subset \mathbb{R}^d$. Suppose $0 \in T$, and consider the ℓ_2-ball of radius ε by $T(\varepsilon) = \{t - t' \in \mathbb{R}^d,\ \|t - t'\|_2 \leq \varepsilon\}$. The natural metric d_X is the Euclidean $\|\cdot\|_2$. Thus Lemma 1.5 shows that we can control the Gaussian complexity of T by its localized complexity based on the ball $T(\varepsilon)$*

$$\mathcal{G}(T) \leq \inf_{\varepsilon \in [0, D(T)]} \{\mathcal{G}(T(\varepsilon)) + 2D(T)\sqrt{\log N_2(\varepsilon, T)}\} \ ,$$

with $N_2(\varepsilon, T)$ being the ε-covering number of T w.r.t. the Euclidean norm. Now, using Example 3.2 of Chapter 4, we have the explicit bound $\mathcal{G}(T(\varepsilon)) \leq \varepsilon\sqrt{d}$. It remains to compute an explicit upperbound of $N_2(\varepsilon, T)$ (left as exercise). Notice that we can get rid of the constant 2 in that case.

Example 1.10 (VC-classes of functions)**.** *Suppose $\mathcal{H}$ to be a* VC*-class of functions with finite* VC*-dimension. By Lemma 4.6, we have that $\sqrt{\log N(\varepsilon, T, d_X)} \leq C\sqrt{\mathcal{V} \log(1/\varepsilon)}$.*

1.3 Generic chaining based on covering sets - Dudley's entropy integral

Important intuition of the chaining method. Before, the supremum was approximated by a finite maximum over an ε-cover with an additional approximation error. We will now write the supremum as a finite sum of maxima indexed by successively refined sets.

Definition 1.6. Let (T, d) be a pseudometric space, and consider an ε-cover of finite covering number $N(\varepsilon, T, d)$ such that *Dudley's entropy integral* is well-defined by

$$\mathcal{J}(\varepsilon, D(T)) = \int_{\varepsilon}^{D(T)} \sqrt{\log N(\varepsilon, T, d)} \ ,$$

where $D(T) = \sup_{s,t\in T} d(s,t)$ is the diameter of T.

Theorem 1.7 (Wainwright , Theorem 5.22)**.** *Consider X_t a sub-Gaussian centered process w.r.t. the induced pseudometric d_X on T. Then, for any $\varepsilon \in [0, D(T)]$,*

$$\mathbb{E}\left[\sup_{t,t'\in T}(X_t - X_{t'})\right] \leq 2\mathbb{E}\left[\sup_{t,t'\in T,\ d_X(t,t')\leq\varepsilon}(X_t - X_{t'})\right] + 32\int_{\varepsilon/4}^{D(T)} \sqrt{\log N(\varepsilon, T, d_X)} \ .$$

Corollary 1.8 (Dudley's Entropy Integral)**.** *Consider X_t a sub-Gaussian centered process w.r.t. the induced pseudometric d_X on T.*

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \leq 32\int_{0}^{D(T)} \sqrt{\log N(\varepsilon, T, d_X)} \ .$$

Proof. Board.

□

1.4 Generic chaining based on admissible partitions