The Method of Chaining

MA5249 Project Report Part 1

Pan Jing Bin
A0199300H

1 Introduction

In high-dimensional probability theory, a major area of research is the development of new tools to control the regularity of random processes $\{X_t\}_{t\in T}$. A task that is of central importance in many real-world applications is establishing lower and upper bounds on the two closely related quantities

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \qquad \text{and} \qquad P\left(\sup_{t\in T} X_t \geq x\right).$$

One simple example to keep in mind is the operator norm of random matrices

$$\|M\|_{\text{op}} := \sup_{\|v\|_2=1} \|Mv\|_2 = \sup_{v\in\mathbb{S}^n} M_v.$$

Throughout this report, we will encounter numerous other examples. Unfortunately, there is no hope that such a problem can be handled in its full generality and additional assumptions will be needed to develop a meaningful theory.

Definition 1.1. Let $\sigma \in \mathbb{R}_{\geq 0}$. A random variable X is $\boldsymbol{\sigma^2}$**-subgaussian** if for all $\lambda \in \mathbb{R}$, we have

$$\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}}.$$

In this two-part report, we will develop, from the ground up, one such tool to control the supremum of subgaussian processes. While seemingly a very strong assumption, such processes arise naturally in many situations (especially those with many independent sources of "randomness").

We will start with very crude methods as our basic ingredient and develop the core of the general theory in increasingly sophisticated ways. The end result is a powerful weapon, known as the method of chaining, that yield remarkably sharp bounds in many situations.

Remark 1.2. Definition 1.1 is equivalent up to constants to the more suggestive condition that $\mathbb{E}[X] = 0$ and X has Gaussian type decay, i.e. there exists $C, v \in \mathbb{R}_{>0}$ such that

$$P\left(|X| \geq \lambda\right) \leq Ce^{-v\lambda^2} \qquad \text{for all } \lambda \in \mathbb{R}.$$

Remark 1.3. A technical subtlety here is that $\sup_{t\in T} X_t$ may not be measurable if the inedx set T is uncountably infinite. We will ignore this issue for now and rectify it only in the second part of the report after introducing the notion of separable processes.

Throughout this report, we use log to denote the natural logarithm and $X \lesssim Y$ to denote $|X| \leq CY$ for some absolute constant C. The notation $X \approx Y$ denotes both $X \lesssim Y$ and $Y \lesssim X$.

2 Finite Maxima

The first step in the development of our theory is to tackle the simplest possible situation: the maximum of a finite number of random variables. An naive approach is to bound the supremum by the sum

$$\max_{t\in T} X_t \leq \sum_{t\in T} |X_t|.$$

While the idea that such a crude method can produce meaningful estimates may seem quite absurd, a cleaver trick here is to leverage on concavity. For example, if each X_t has finite variance, then the concavity of the square root gives

$$\mathbb{E}\left[\max_{t\in T} X_t\right] \leq \mathbb{E}\left[\max_{t\in T} |X_t|^2\right]^{1/2} \leq |T|^{1/2} \max_{t\in T} \left\{\mathbb{E}\left[\left|X_t\right|^2\right]^{1/2}\right\}.$$

Observe that we have already significantly improved the dependence on $|T|$. Taking this philosophy one step further using the exponential function, we obtain the elementary upper bound

Lemma 2.1. (Maximal inequality) Let $\{X_t\}_{t\in T}$ be a random process. Supoose that X_t is σ^2-subgaussian for each $t \in T$. Then we have

$$\mathbb{E}\left[\max_{t\in T} X_t\right] \leq \sqrt{2\sigma^2 \log |T|}.$$

Proof. For any $\lambda \in \mathbb{R}_{>0}$, Jensen's inequality gives

$$\mathbb{E}\left[\max_{t\in T} X_t\right] \leq \frac{1}{\lambda} \log\left(\mathbb{E}\left[e^{\lambda \max_{t\in T} X_t}\right]\right) \leq \frac{1}{\lambda} \log\left(\sum_{t\in T} \mathbb{E}\left[e^{\lambda X_t}\right]\right) \leq \frac{\log |T|}{\lambda} + \frac{\lambda\sigma^2}{2}.$$

The conclusion follows by optimizing in λ and choosing $\lambda = \dfrac{\sqrt{2\log|T|}}{\sigma}$. □

Analogously, we may establish elementary upper bounds on the tail probability using the union bound.

Lemma 2.2. (Maximal tail inequality) Let $\{X_t\}_{t\in T}$ be a random process. Suppose that X_t is σ^2-subgaussian for each $t \in T$. Then for any $x \in \mathbb{R}_{\geq 0}$, we have

$$P\left(\max_{t\in T} X_t \geq \sqrt{2\sigma^2 \log |T|} + x\right) \leq e^{-\frac{x^2}{2\sigma^2}}.$$

Proof. For any $\lambda \in \mathbb{R}_{>0}$, a direct application of the union bound and the Chernoff bound gives

$$P\left(\max_{t\in T} X_t \geq x\right) = P\left(\bigcup_{t\in T} \{X_t \geq x\}\right) \leq \sum_{t\in T} P(X_t \geq x) \leq \sum_{t\in T} \frac{\mathbb{E}\left[e^{\lambda X_t}\right]}{e^{\lambda x}} \leq e^{\log|T| + \frac{\lambda^2\sigma^2}{2} - \lambda x}.$$

Here, $\lambda = \dfrac{x}{\sigma^2}$ minimizes the right-hand side. Replacing x by $\sqrt{2\sigma^2 \log |T|} + x$ then gives

$$P\left(\max_{t\in T} X_t \geq \sqrt{2\sigma^2 \log |T|} + x\right) \leq \exp\left(\log|T| - \frac{2\sigma^2\log|T| + 2x\sqrt{2\sigma^2\log|T|} + x^2}{2\sigma^2}\right) \leq e^{-\frac{x^2}{2\sigma^2}}.$$

□

Both techniques, while seemingly crude at first glance, in fact forms the backbone behind the development of the theory.

3 Covering and Packing

To generalise the results from the previous section to the case in which T is infinite, we require additional structure on the process $\{X_t\}_{t\in T}$. In many real-world processes, the random variables X_t depend on the index t in an intricate manner. Hence the index set T itself has a very rich structure. For example, the following identity for a (discrete) homogeneous Markov chain

$$P\big(X_r = k \;\big|\; X_i = x_i \text{ for } i = 0, \cdots, r-1, r+1, \cdots, n\big) = P\big(X_r = k \;\big|\; X_{r-1} = x_{r-1}, X_{r+1} = x_{r+1}\big)$$

suggests that each variable X_r is most dependent on its adjacent neighbours and less dependent on the random variables that are further away.

More generally, if the index set T is endowed with the structure of a metric space (say $\mathbb{R}$) and the map $t \mapsto X_t$ is continuous in a suitable sense, then the condition $\lim_{t\to s} X_t = X_s$ implies that X_t and X_s are highly dependent variables when s is close to t. This will be our guiding philosophy behind the development of the theory in this section.

Definition 3.1. (Lipschitz process) A random process $\{X_t\}_{t\in T}$ is **Lipschitz** for a metric d on T if there exists a random variable C such that

$$|X_s - X_t| \le C d(s,t) \qquad \text{for all } s, t \in T.$$

Given a Lipschitz process, our goal is to first approximate the supremum over T by a finite set N and then apply the inequalities in the previous section to N. Hence to obtain a good upper bound, we must first minimize the cardinality of the set N.

Definition 3.2. (ϵ-net and covering number) A set N is called an **ϵ-net** for (T,d) if for every $t \in T$, there exists $\pi(t) \in N$ such that $d(t, \pi(t)) \le \epsilon$. The smallest cardinality of an ϵ-net for (T,d) is called the **covering number**

$$N(T,d,\epsilon) := \inf\big\{|N| : N \text{ is an } \epsilon\text{-net for } (T,d)\big\}.$$

In a manner analogous to the study of compact topological spaces, we now develop the first non-trivial upper bound on Lipschitz random processes.

Lemma 3.3. (Lipschitz maximal inequality) Let $\{X_t\}_{t\in T}$ be a Lipschitz random process. Suppose that X_t is σ^2-subgaussian for each $t \in T$. Then

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \le \inf_{\epsilon>0}\left\{\epsilon\mathbb{E}[C] + \sqrt{2\sigma^2 \log N(T,d,\epsilon)}\right\}.$$

Proof. Fix $\epsilon \in \mathbb{R}_{>0}$. As per the previous discussion, we choose an ϵ-net N satisfying $|N| = N(T,d,\epsilon)$ and perform the following decomposition:

$$\sup_{t\in T} X_t \le \sup_{t\in T}\big\{X_t - X_{\pi(t)}\big\} + \sup_{t\in T} X_{\pi(t)} \le C\epsilon + \max_{t\in N} X_t.$$

Taking expectation and applying Lemma 2.1 gives

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \le \epsilon\mathbb{E}[C] + \sqrt{2\sigma^2 \log N(T,d,\epsilon)}.$$

The desired conclusion follows by taking infimum over all $\epsilon > 0$. □

Clearly, the covering number plays a key role in the effectiveness of this technique. To choose a small ϵ-net, we should (intuitively) choose the points in N to be as far apart as possible. This motivates the following definition.

Definition 3.4. (ϵ-packing and packing number) A set $N \subseteq T$ is an **ϵ-packing** of (T, d) if $d(t, t') > \epsilon$ for all $t, t' \in N$ such that $t \neq t'$. The largest cardinality of an ϵ-packing of (T, d) is called the **packing number**

$$D(T, d, \epsilon) := \sup \big\{ |N| : N \text{ is an } \epsilon\text{-packing of } (T, d) \big\}.$$

The relationship between the covering number and packing number is given by the following lemma, which follows from a routine application of the triangle inequality.

Lemma 3.5. (Duality between covering and packing) For every $\epsilon \in \mathbb{R}_{>0}$,

$$D(T, d, 2\epsilon) \leq N(T, d, \epsilon) \leq D(T, d, \epsilon).$$

In general, establishing upper bounds and lower bounds require fundamentally different sets of tools, even if the type of quantity that is being studied is the same. However, here we have a dual approach. Lemma 3.3 works best when the covering number is small. Thus if the underlying space T itself is "large" with respect to the metric and small ϵ-nets cannot be found (i.e the packing number is large), then we should expect the supremum to be large as well. This concept will be further developed in the last section of the next report.

Before we develop our theory even further, we first demonstrate the effectiveness of what we have built so far with the following example.

Example 3.6. (Wasserstein law of large numbers) Suppose that one has i.i.d random variables $X_1, X_2, \cdots$ taking values in the interval $[0, 1]$. Further suppose that we have a bounded function $f : [0, 1] \to \mathbb{R}$. Then the law of large numbers dictate that

$$\mathbb{E}\left[\left| \sum_{i=1}^{n} \frac{f(X_i)}{n} - \mu_f \right| \right] \lesssim n^{-1/2}$$

where $\mu_f := \mathbb{E}\big[f(X_1)\big]$. The rate of $n^{-1/2}$ is asymptotically optimal. In this example, we will establish an upper bound for the rate of convergence that is also uniform in f.

Definition 3.7. (Lipschitz functions) Let (X, d) be a metric space. A function $f : X \to \mathbb{R}$ is ***L*-Lipschitz** if $|f(x) - f(y)| \leq L d(x, y)$ for all $x, y \in X$. The family of all 1-Lipschitz functions is denoted by $\mathrm{Lip}(X)$.

Define

$$X_f := \sum_{i=1}^{n} \frac{f(X_i)}{n} - \mu_f \qquad \text{and} \qquad \mathcal{F} := \big\{ f \in \mathrm{Lip}([0,1]) : 0 \leq f \leq 1 \big\}.$$

In our current setting, our goal is to give an upper bound for the quantity

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} X_f \right].$$

Firstly, the triangle inequality gives

$$\big| X_f - X_g \big| \leq \left| \sum_{i=1}^{n} \frac{(f-g)(X_i) - \mu_{f-g}}{n} \right| \leq 2 \left\| f - g \right\|_\infty$$

and hence the process $\{X_f\}_{f\in\mathcal{F}}$ is Lipschitz with respect to the supremum norm on $\mathcal{F}$. On the other hand, a routine application of the Azuma-Hoeffding's inequality reveals that the random variable X_f is $\frac{1}{n}$-subgaussian for every $f \in \mathcal{F}$. We then use Lemma 3.3 to obtain the estimate

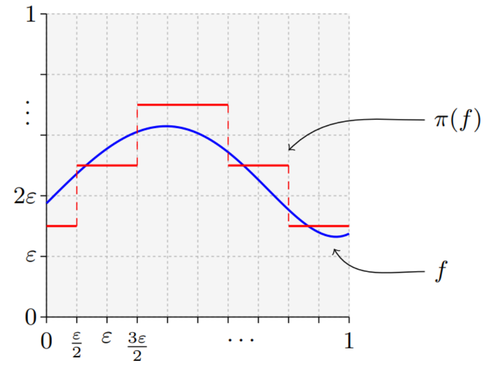
$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} X_f\right] \leq \inf_{\epsilon>0}\left\{2\epsilon + \sqrt{\frac{2}{n}\log N(\mathcal{F}, \|\cdot\|_\infty, \epsilon)}\right\}. \tag{1}$$

It remains to obtain a good upper bound on the covering number $N = (\mathcal{F}, \|\cdot\|_\infty, \epsilon)$.

Lemma 3.8. There exists a constant $c \in \mathbb{R}$ such that

$$N(\mathcal{F}, \|\cdot\|_\infty, \epsilon) \leq e^{c/\epsilon} \text{ for } \epsilon < \frac{1}{2}, \qquad N(\mathcal{F}, \|\cdot\|_\infty, \epsilon) = 1 \text{ for } \epsilon \geq \frac{1}{2}.$$

Proof. (Sketch) For a fixed $\epsilon \in \mathbb{R}_{>0}$ and $f \in \mathcal{F}$, we will use the 1-Lipschitz property of f to approximate it by a piecewise function $\pi(f)$. The key idea is illustrated by the following diagram:



Source: R. van Handel (2016, p.127)

More formally, we partition the horizontal axis into consecutive non-overlapping intervals $I_1, \cdots, I_{\lceil 2/\epsilon\rceil}$ of size $\epsilon/2$ and the vertical axis into consecutive non-overlapping intervals $J_1, \cdots, J_{\lceil 1/\epsilon\rceil}$ of size ϵ. We then define

$$\pi(f)(x) := m_\ell, \qquad \text{for } x \in I_k, \quad f(\inf(I_k)) \in J_\ell$$

where $m_\ell = \dfrac{\sup J_\ell + \inf J_\ell}{2}$ denotes the midpoint of the interval J_ℓ.

By construction, the set $N = \{\pi(f) : f \in \mathcal{F}\}$ is an ϵ-net. By considering all possible piecewise functions of the above form, a naive upper bound for $|N|$ is $\lceil 1/\epsilon\rceil^{\lceil 2/\epsilon\rceil}$. However, the 1-Lipschitz property of f allows us to improve on this bound in the following way: if $\pi(f)(I_k) = m_\ell$, then in the subsequent interval I_{k+1}, we have that $\pi(f)(I_{k+1})$ can take on at most three possible values $m_{\ell-1}, m_\ell, m_{\ell+1}$. Hence we obtain the tighter bound

$$N(\mathcal{F}, \|\cdot\|_\infty, \epsilon) \leq |N| \leq \lceil 1/\epsilon\rceil 3^{\lceil 2/\epsilon\rceil - 1} < e^{c/\epsilon}$$

for some absolute constant $c \in \mathbb{R}$ (that is uniform in ϵ). This completes the proof of the lemma. □

Returning back to the problem at hand, equation (1) reduces to

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} X_f\right] \leq \inf_{\epsilon>0}\left\{2\epsilon + \sqrt{\frac{2c}{\epsilon n}}\right\} \lesssim n^{-1/3}$$

which is unfortunately not sharp. In the second part of the report, we will see how sharper bounds can be obtained once we have improved our tool further.

References

M. Talagrand. *The generic chaining. Upper and lower bounds of stochastic processes.* Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2005.

R. van Handel. *Probability in High Dimension.* Course notes, Princeton University, 21 Dec 2016. Retrieved from https://web.math.princeton.edu/∼rvan/APC550.pdf