Expectation of suprema of empirical processes

December 3, 2018

Christophe F. Gallesco, UNICAMP

1 Subgaussian processes

We start with the notion of a subgaussian random variable.

Definition 1.1. *A random variable X is called σ^2-subgaussian if*

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] \leqslant e^{\frac{\sigma^2\lambda^2}{2}}$$

for all $\lambda \in \mathbb{R}$. σ^2 is called the variance proxy.

Notation: In the rest of these notes, we will write $X \sim \text{Subgaus}(\sigma^2)$ to express that X is a σ^2-subgaussian random variable.

Applying the Chernoff bound and the definition of a σ^2-subgaussian random variable it is not hard to obtain the following

Proposition 1.1. *If $X \sim Subgaus(\sigma^2)$ then for all $t > 0$ we have*

$$\mathbb{P}[X \geqslant \mathbb{E}[X] + t] \leqslant e^{-\frac{t^2}{2\sigma^2}},$$
$$\mathbb{P}[X \leqslant \mathbb{E}[X] - t] \leqslant e^{-\frac{t^2}{2\sigma^2}}.$$

Definition 1.2. *A real stochastic process $(X_t)_{t\in T}$ on a pseudo metric space (T, d) is called d-subgaussian if $\mathbb{E}[X_t] = 0$ and*

$$\mathbb{E}[e^{\lambda(X_t-X_s)}] \leqslant e^{\frac{\lambda^2 d(t,s)^2}{2}} \tag{1}$$

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

Remark: Observe that (1) already implies that $\mathbb{E}[X_t] = \mathbb{E}[X_s]$ for all $t, s \in T$, so $\mathbb{E}[X_t] = 0$ is just a convenient normalization.
Notation: In the rest of these notes, we will write $(X_t)_{t\in T} \sim \text{Subgaus}(d)$ to express that $(X_t)_{t\in T}$ is a d-subgaussian process.

Definition 1.3. *Let T be a topological space. A real stochastic process $(X_t)_{t\in T}$ is called separable if there is a countable dense set $S \subset T$ such that*

$$\mathbb{P}\Big[X_t \in \lim_{s\to t, s\in S} X_s, \ \text{for all } t \in T\Big] = 1.$$

Example: if $T = \mathbb{R}_+$ and $(X_t)_{t\in T}$ has continuous trajectories $\mathbb{P}$-a.s. then it is separable.

Under the hypothesis of separability $\sup_{t\in T} X_t = \sup_{t\in S} X_t$, $\mathbb{P}$-a.s., and therefore no measurability issues arise.

2 Finite maxima

In this section we consider a real stochastic process $(X_t)_{t\in T}$ where T is a finite set.

Lemma 2.1 (Maximal inequality)**.** *If $X_t \sim$ Subgaus(σ^2) for all $t \in T$ then we have*

$$\mathbb{E}\big[\max_{t\in T}(X_t - \mathbb{E}[X_t])\big] \leqslant \sqrt{2\sigma^2 \ln |T|}.$$

Proof. By Jensen's inequality, we have for all $\lambda > 0$

$$\begin{aligned}\mathbb{E}\big[\max_{t\in T}(X_t - \mathbb{E}[X_t])\big] &\leqslant \frac{1}{\lambda}\ln \mathbb{E}\big[e^{\lambda \max_{t\in T}(X_t - \mathbb{E}[X_t])}\big] \\ &\leqslant \frac{1}{\lambda}\ln \sum_{t\in T}\mathbb{E}\big[e^{\lambda(X_t - \mathbb{E}[X_t])}\big].\end{aligned}$$

Since $X_t \sim$ Subgaus(σ^2) for all $t \in T$, we obtain

$$\mathbb{E}\big[\max_{t\in T}(X_t - \mathbb{E}[X_t])\big] \leqslant \frac{\ln |T|}{\lambda} + \frac{\sigma^2\lambda}{2}$$

for all $\lambda > 0$. Optimizing in λ we deduce the desired result. □

Corollary 2.1. *If $X_t \sim$ Subgaus(σ^2) for all $t \in T$ then we have*

$$\mathbb{E}\big[\max_{t\in T}|X_t - \mathbb{E}[X_t]|\big] \leqslant \sqrt{2\sigma^2 \ln(2|T|)}.$$

Proof. Just observe that

$$\max_{t\in T}|X_t - \mathbb{E}[X_t]| = \max_{t\in T}(X_t - \mathbb{E}[X_t]) \vee \max_{t\in T}(\mathbb{E}[X_t] - X_t)$$

and that $\mathbb{E}[X_t] - X_t \sim$ Subgaus(σ^2) for all $t \in T$.

3 Chaining method

If the set T is infinite the maximal inequality of the previous section is useless. In this case, we need the following notion

Definition 3.1. *Let (T,d) be a pseudo metric space and $\varepsilon>0$. A set N is called an ε-net for (T,d) if for every $t\in T$, there exists $\beta(t)\in N$ such that $d(t,\beta(t))\leqslant\varepsilon$. The smallest cardinality of an ε-net for (T,d) is called the covering number*

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

Theorem 3.1 (Dudley's entropy integral)**.** *Let $(X_t)_{t\in T}$ be a separable and d-subgaussian process. Then, we have*

$$\mathbb{E}\Big[\sup_{t\in T}X_t\Big]\leqslant 12\int_0^{\mathrm{diam}(T)}\sqrt{\ln N(T,d,\varepsilon)}d\varepsilon \tag{2}$$

where $\mathrm{diam}(T)$ *is the diameter of T relatively to the pseudo metric d.*

Proof.
First step: We first prove the so called "chaining inequality":

$$\mathbb{E}\Big[\sup_{t\in T}X_t\Big]\leqslant 6\sum_{k\in\mathbb{Z}}2^{-k}\sqrt{\ln N(T,d,2^{-k})}. \tag{3}$$

Let us start by proving the result when T is a finite set. In this case, let k_0 be the largest integer such that $2^{-k_0}\geqslant\mathrm{diam}(T)$. Then, any singleton $\{t_0\}$ is trivially a 2^{-k_0}-net. For $k>k_0$, let N_k be a 2^{-k}-net such that $|N_k|=N(T,d,2^{-k})$. Then, we write for $n>k_0$,

$$\begin{aligned}\mathbb{E}\Big[\sup_{t\in T}X_t\Big]\leqslant\mathbb{E}[X_{t_0}]&+\sum_{k=k_0+1}^{n}\mathbb{E}\Big[\sup_{t\in T}\{X_{\beta_k(t)}-X_{\beta_{k-1}(t)}\}\Big]\\&+\mathbb{E}\Big[\sup_{t\in T}\{X_t-X_{\beta_n(t)}\}\Big].\end{aligned} \tag{4}$$

By assumption $\mathbb{E}[X_{t_0}]=0$. Moreover, since T is finite and $(X_t)_{t\in T}$ is d-subgaussian, we can choose n large enough so that $\sup_{t\in T}|X_t-X_{\beta_n(t)}|=0$, $\mathbb{P}$-a.s. Thus, for n large enough, we are left with

$$\mathbb{E}\Big[\sup_{t\in T}X_t\Big]\leqslant\sum_{k=k_0+1}^{n}\mathbb{E}\Big[\sup_{t\in T}\{X_{\beta_k(t)}-X_{\beta_{k-1}(t)}\}\Big].$$

Now, for the k-th term in the sum, observe that each supremum contains at most $|N_k||N_{k-1}| \leqslant |N_k|^2$ terms. Futhermore, we have that

$$d(\beta_k(t), \beta_{k-1}(t)) \leqslant d(t, \beta_k(t)) + d(t, \beta_{k-1}(t)) \leqslant 2^{-k} + 2^{-(k-1)} = 3 \times 2^{-k}.$$

Since, $X_{\beta_k(t)} - X_{\beta_{k-1}(t)} \sim \text{Subgaus}((d(\beta_k(t), \beta_{k-1}(t)))^2)$, Lemma 2.1 yields

$$\begin{aligned}\mathbb{E}\Big[\sup_{t\in T} X_t\Big] &\leqslant \sum_{k>k_0} d(\beta_k(t), \beta_{k-1}(t))\sqrt{2\ln|N_k|^2} \\ &\leqslant 6\sum_{k>k_0} 2^{-k}\sqrt{\ln|N_k|} \\ &= 6\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\ln|N_k|}.\end{aligned}$$

By construction, $|N_k| = N(T, d, 2^{-k})$, so the proof of (3) is complete in the case $|T| < \infty$.

In the case $|T| = \infty$, since $(X_t)_{t\in T}$ is separable, there exists a countable set $S \subset T$ such that $\mathbb{P}$-a.s., $\sup_{t\in T} X_t = \sup_{t\in S} X_t$. Then,

$$\mathbb{E}\Big[\sup_{t\in T} X_t\Big] = \mathbb{E}\Big[\sup_{t\in S} X_t\Big].$$

Now, fix an enumeration of the elements of S and let S_l the set formed by the first l elements of S. By the monotone convergence theorem we obtain that

$$\mathbb{E}\Big[\sup_{t\in T} X_t\Big] = \sup_{l\geqslant 1}\mathbb{E}\Big[\sup_{t\in S_l} X_t\Big].$$

Applying the chaining inequality (3) to each finite supremum and using the fact that $N(S_l, d, \varepsilon) \leqslant N(T, d, \varepsilon)$, for all $l \geqslant 1$, yields the desired result.

Second step: In this second part we show that

$$\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\ln N(T, d, 2^{-k})} \leqslant 2\int_0^\infty \sqrt{\ln N(T, d, \varepsilon)}d\varepsilon. \tag{5}$$

For this, just observe that

$$\begin{aligned}\sum_{k\in\mathbb{Z}} 2^{-k}\sqrt{\ln N(T,d,2^{-k})} &= 2\sum_{k\in\mathbb{Z}}\int_{2^{-k-1}}^{2^{-k}} \sqrt{\ln N(T,d,2^{-k})}d\varepsilon \\ &\leqslant 2\sum_{k\in\mathbb{Z}}\int_{2^{-k-1}}^{2^{-k}} \sqrt{\ln N(T,d,\varepsilon)}d\varepsilon \\ &= 2\int_0^{\infty} \sqrt{\ln N(T,d,\varepsilon)}d\varepsilon\end{aligned}$$

where in the second step we used that $N(T,d,\varepsilon)$ is non-increasing in ε.

Third step: Finally, gathering (3), (5) and noticing that we always have $N(T,d,\varepsilon)=1$ when $\varepsilon \geqslant \mathrm{diam}(T)$, we obtain (2). □

Theorem 3.2. *Let $(X_t)_{t\in T}$ be a separable and d-subgaussian process. Then, we have*

$$\mathbb{E}\Big[\sup_{t\in T}|X_t|\Big] \leqslant \mathbb{E}[|X_{t_0}|] + 12\int_0^{\mathrm{diam}(T)} \sqrt{\ln 2N(T,d,\varepsilon)}d\varepsilon \tag{6}$$

where t_0 is an arbitrary point of T and $\mathrm{diam}(T)$ is the diameter of T relatively to the pseudo metric d.

Proof. The proof of (6) is very similar to the proof of (2). We first use the following chaining decomposition

$$\begin{aligned}\mathbb{E}\Big[\sup_{t\in T}|X_t|\Big] \leqslant \mathbb{E}[|X_{t_0}|] &+ \sum_{k=k_0+1}^{n}\mathbb{E}\Big[\sup_{t\in T}|X_{\beta_k(t)}-X_{\beta_{k-1}(t)}|\Big] \\ &+ \mathbb{E}\Big[\sup_{t\in T}|X_t - X_{\beta_n(t)}|\Big]\end{aligned}$$

instead of (4) and then apply Corollary 2.1. □

Remark: In Theorems 3.1 and 3.2, it is important to notice that the separability property of the process $(X_t)_{t\in T}$ is not related to the pseudo metric d. On one hand we need a topology on T which gives us the separability property of $(X_t)_{t\in T}$ (and avoid measurability issues of $\sup_{t\in T} X_t$) and on the other hand the pseudo metric d is related to the "subgaussianity" of the process $(X_t)_{t\in T}$.

4 Empirical processes

Definition 4.1 (Empirical process)**.** *Let $X_1, \dots, X_n$ be i.i.d. random elements in $\mathcal{X}$ and $\mathcal{F}$ a class of measurable functions from $\mathcal{X} \to \mathbb{R}$. The empirical process G_n over the class $\mathcal{F}$ is defined as*

$$G_n(f) = \frac{1}{n}\sum_{k=1}^{n}\{f(X_k) - \mathbb{E}[f(X_k)]\}, \quad \textit{for all } f \in \mathcal{F}.$$

In this section we are interested in obtaining upper bounds for $\sup_{f\in\mathcal{F}}|G_n(f)|$. From now on, we assume that the process $(G_n(f))_{f\in\mathcal{F}}$ is separable so that $\sup_{f\in\mathcal{F}}|G_n(f)|$ is measurable.

Lemma 4.1 (Symmetrization lemma)**.** *Let $X_1, \dots, X_n$ be i.i.d. random elements in $\mathcal{X}$ and $\mathcal{F}$ a class of measurable functions from $\mathcal{X} \to \mathbb{R}$. We have*

$$\begin{aligned}\mathbb{E}\Big[\sup_{f\in\mathcal{F}}|nG_n(f)|\Big] &\leqslant \mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big|\sum_{k=1}^{n}\varepsilon_k\{f(X_k) - f(Y_k)\}\Big|\Big]\\ &\leqslant 2\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big|\sum_{k=1}^{n}\varepsilon_k\{f(X_k) - \mathbb{E}[f(X_k)]\}\Big|\Big]\end{aligned}$$

where $(Y_1, \dots, Y_n)$ is an independent copy of $(X_1, \dots, X_n)$ and $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. Rademacher random variables independent of $X_1, \dots, X_n, Y_1, \dots, Y_n$.

Proof. Observe that $\mathbb{E}[f(X_k)] = \mathbb{E}[f(Y_k) \mid X_1, \dots, X_n]$, for all $1 \leqslant k \leqslant n$. Thus, by Jensen's inequality for conditional expectation, we obtain

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}}|nG_n(f)|\Big] \leqslant \mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big|\sum_{k=1}^{n}\varepsilon_k\{f(X_k) - f(Y_k)\}\Big|\Big]$$

Now, since $f(X_k) - f(Y_k)$ is a symmetric random variable, it has the same law as $\varepsilon_k\{f(X_k) - f(Y_k)\}$. This implies that

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big|\sum_{k=1}^{n}\{f(X_k) - f(Y_k)\}\Big|\Big] = \mathbb{E}\Big[\sup_{f\in\mathcal{F}}\Big|\sum_{k=1}^{n}\varepsilon_k\{f(X_k) - f(Y_k)\}\Big|\Big].$$

This proves the first inequality. The second inequality is easily obtained using the triangular inequality and the fact that $\mathbb{E}[f(X_k)] = \mathbb{E}[f(Y_k)]$, for $1 \leqslant k \leqslant n$. □

The symmetrization lemma is useful to extract the subgaussianity of an empririal process. In the following we take advantage of this fact. Let us look at the process

$$Z_n(f) = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \varepsilon_k \{f(X_k) - \mathbb{E}[f(X_k)]\}.$$

The key observation is that, under the law $\mathbb{P}_\varepsilon[\,\cdot\,] := \mathbb{P}[\,\cdot \mid X_1, \ldots, X_n, Y_1, \ldots, Y_n]$, $(Z_n(f))_{f\in\mathcal{F}}$ is a subgaussian process with respect to the *random* pseudo metric d_n on $\mathcal{F}$ defined by

$$d_n(f,g) = \left[\frac{1}{n}\sum_{k=1}^{n}\{f(X_k) - g(X_k) - \mathbb{E}[f(X_k) - g(X_k)]\}^2\right]^{1/2}$$

for $f, g \in \mathcal{F}$.

Applying Theorem 3.2, we obtain that

$$\mathbb{E}_\varepsilon\Big[\sup_{f\in\mathcal{F}} |Z_n(f)|\Big] \leqslant \mathbb{E}_\varepsilon[|Z_n(f_0)|] + 12\int_0^{\mathrm{diam}(\mathcal{F})} \sqrt{\ln 2N(T, d_n, s)}ds \qquad (7)$$

where f_0 is an arbitrary element of $\mathcal{F}$ and $\mathrm{diam}(\mathcal{F})$ is the diameter of $\mathcal{F}$ relatively to d_n. Taking the expectation of both sides of (7) and applying Lemma 4.1 we obtain the following

Theorem 4.1. *It holds that*

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}} |\sqrt{n}G_n(f)|\Big] \leqslant 2\mathbb{E}[|Z_n(f_0)|] + 24\mathbb{E}\left[\int_0^{\mathrm{diam}(\mathcal{F})} \sqrt{\ln 2N(T, d_n, s)}ds\right].$$

5 Application

Let Σ be a locally compact and polish space and $g : \Sigma \times \Sigma \to \mathbb{R}$ measurable, separable and bounded. *Separable* means here that there exists a countable set $S \subset \Sigma$ such that for all $x \in \Sigma$, $g(x, \cdot)$ is completely determined by its values on S.

Let $(W_k)_{k\geqslant 1}$ be i.i.d. random functions on Σ with the following structure

$$W_k(x) = \sum_{i=1}^{T_k} Y_{k,i}\, g(Z_{k,i}, x), \quad \text{for all } x \in \Sigma$$

and some random variables T_k, $(Y_{k,i})_{i\geqslant 1}$ and $(Z_{k,i})_{i\geqslant 1}$ with finite second moments. We also assume that for all $x\in\Sigma$, $\mathbb{E}[W_k(x)]=0$.

In [2], one considered an empirical process of the form

$$G_n(x)=\frac{1}{n}\sum_{k=1}^{n}W_k(x)$$

for all $x\in\Sigma$. In the same paper, it was used an inequality due to Pollard based on bracketing numbers to estimate $\mathbb{E}\Big[\sup_{x\in\Sigma}|\sqrt{n}G_n(x)|\Big]$. Here, our approach is of course based on covering numbers and we apply Theorem 4.1, to obtain that

$$\mathbb{E}\Big[\sup_{x\in\Sigma}|\sqrt{n}G_n(x)|\Big]\leqslant 2\mathbb{E}[|\sqrt{n}G_n(x_0)|]+24\mathbb{E}\left[\int_0^{\mathrm{diam}\Sigma}\sqrt{\ln 2N(T,d_n,s)}ds\right] \tag{8}$$

where x_0 is an arbitrary point of Σ and d_n is the random pseudo metric on Σ defined by

$$d_n(x,y)=\left[\frac{1}{n}\sum_{k=1}^{n}\{W_k(x)-W_k(y)\}^2\right]^{1/2}.$$

We have that

$$d_n(x,y)\leqslant\sup_{z\in\Sigma}|g(z,x)-g(z,y)|\left[\frac{1}{n}\sum_{k=1}^{n}\Big\{\sum_{i=1}^{T_k}Y_{k,i}\Big\}^2\right]^{1/2}.$$

Let us call $U_k=\sum_{i=1}^{T_k}Y_{k,i}$, for $k\geqslant 1$, and $\mathbf{U}_n=(U_1,\dots,U_n)$. Thus, we can rewrite

$$d_n(x,y)\leqslant\sup_{z\in\Sigma}|g(z,x)-g(z,y)|\left[\frac{1}{n}\sum_{k=1}^{n}U_k^2\right]^{1/2}=2\frac{\|\mathbf{U}_n\|_2}{\sqrt{n}}\tilde{d}(x,y)$$

where $\tilde{d}(x,y):=\frac{1}{2}\sup_{z\in\Sigma}|g(z,x)-g(z,y)|$ is a pseudo metric on Σ. By the above inequality, we have that

$$N(T,d_n,s)\leqslant N\left(T,2\frac{\|\mathbf{U}_n\|_2}{\sqrt{n}}\tilde{d}(x,y),s\right)$$

for all $s > 0$. Therefore, we obtain that

$$\int_0^{\mathrm{diam}(\Sigma)} \sqrt{\ln 2N(T, d_n, s)}ds \leqslant \frac{\|\mathbf{U}_n\|_2}{\sqrt{n}} \int_0^{\widetilde{\mathrm{diam}}(\Sigma)} \sqrt{\ln 2N(T, \tilde{d}, s)}ds.$$

Now, since $\tilde{d}(x, y) \leqslant \|g\|_\infty$, for all $x, y \in \Sigma$, we have that $\widetilde{\mathrm{diam}}(\Sigma) \leqslant \|g\|_\infty$. Then, we obtain by Jensen's inequality

$$\begin{aligned}\mathbb{E}\left[\int_0^{\mathrm{diam}\Sigma} \sqrt{\ln 2N(T, d_n, s)}ds\right] &\leqslant 2\mathbb{E}\left[\frac{\|\mathbf{U}_n\|_2}{\sqrt{n}}\right] \int_0^{\|g\|_\infty} \sqrt{\ln 2N(T, \tilde{d}, s)}ds \\ &\leqslant 2(\mathbb{E}[U_1^2])^{\frac{1}{2}} \int_0^{\|g\|_\infty} \sqrt{\ln 2N(T, \tilde{d}, s)}ds.\end{aligned}$$

As for the first term of the right-hand side of (8), we can use that

$$\mathbb{E}[|\sqrt{n}G_n(x_0)|] \leqslant \sqrt{\mathrm{Var}[\sqrt{n}G_n(x_0)]} \leqslant \|g\|_\infty (\mathbb{E}[U_1^2])^{\frac{1}{2}}$$

since $\mathbb{E}[G_n(x_0)] = 0$. We finally deduce that

$$\begin{aligned}\mathbb{E}\Big[\sup_{x\in\Sigma} |\sqrt{n}G_n(x)|\Big] &\leqslant 2\|g\|_\infty (\mathbb{E}[U_1^2])^{\frac{1}{2}} + 48(\mathbb{E}[U_1^2])^{\frac{1}{2}} \int_0^{\|g\|_\infty} \sqrt{\ln 2N(T, \tilde{d}, s)}ds \\ &\leqslant 50\|U_1\|_{L^2} \int_0^{\|g\|_\infty} \sqrt{1 + \ln N(T, \tilde{d}, s)}ds.\end{aligned}$$

Final remark: Actually, we can restate Theorem 1 of [2] in a more general and elegant way using the last estimate. First, let us replace Assumption 2.2 in [2] by

Assumption 2.2': The density $p : \Sigma \times \Sigma \to \mathbb{R}_+$ is separable.

Now, we have

Theorem 5.1. *Under the Assumptions 2.2' and 2.3 (in [2]), there exists a universal positive constant K such that, for all $n \geqslant 1$, it holds that*

$$\mathrm{d_{TV}}(L_n^X, L_n^Y) \leqslant K \int_0^\varepsilon \sqrt{1 + \ln N(T, \tilde{d}, s)}ds.$$

References

[1] S. Boucheron, G. Lugosi, P. Massart (2013) Concentration inequalities. *Oxford university press.*

[2] D. de Bernardini, C. Gallesco, S. Popov (2018) On uniform closeness of local times of Markov chains and i.i.d. sequences. *Stoch. Proc. and their Appl.* **128**(10), p. 3221–3252.

[3] R. Van Handel (2016) Probability in high dimension. *Lecture notes.*