High Dimensional Probability and Statistics **2nd Semester, 2022-2023**

Lecture 6: Lower Bound of Suprema for Gaussian Process

Instructor: Ke Wei *Scribe: Ke Wei (Updated: 2023/04/06)*

Recap and motivation: In the last few lectures, we have studied the upper bound for $\mathbb{E}\left[\sup_{t\in T} X_t\right]$, especially by Dudley inequality:

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \lesssim \int_0^\infty \sqrt{\log N(T,d,\varepsilon)}d\varepsilon,$$

where d is defined through the increments of the process. In a reverse direction, we can interpret this result as measuring the complexity of a set via a random process, with Radamacher complexity and Gaussian complexity as special examples.

In this section we study the lower bound of $\mathbb{E}\left[\sup_{t\in T} X_t\right]$. It is clear that we cannot expect to obtain a nontrivial lower bound at the level of generality. For example, even in the case of finite maxima, we have seen that the additional assumption of independence is needed to obtain a meaningful lower bound. Otherwise, an extreme example would be $\mathbb{E}\left[\sup_{t\in T} X_t\right]$ with $X_t = X$ for all t. Therefore, in this lecture we will *restrict our attention to the Gaussian process*, whose additional properties enable us to establish lower bound of $\mathbb{E}\left[\sup_{t\in T} X_t\right]$ for certain random processes via Gaussian comparison theorems. As before, we will always assume X_t is centered (i.e., $\mathbb{E}\left[X_t\right] = 0$ for all t, unless stated otherwise).

Definition 6.1 (Gaussian process) *The random process $\{X_t\}_{t\in T}$ is called a centered Gaussian process if the random variables $\{X_{t_1}, \cdots, X_{t_n}\}$ are centered and jointly Gaussian*[1] *for all $n \geq 1$ and $t_1, \cdots, t_n \in T$.*

Recall that for the centered Gaussian random variable, its sub-Gaussian parameter is equal to its variance. Thus, if we define

$$d(t,s) = \sqrt{\mathbb{E}\left[(X_t - X_s)^2\right]} = \|X_t - X_s\|_{L_2}. \tag{6.1}$$

Then, a Gaussian process is a sub-Gaussian process on (T,d). Note d is usually referred to the *canonical metric* defined on T and it is indeed a pseudo-metric but it satisfies the triangle inequality. Gaussian process has additional properties that makes it easy to work with.

Agenda:

- Gaussian interpolation
- Gaussian comparison inequality
- Sudakov minoration inequality
- A short remark

[1] It is equivalent to that any linear combination of $\{X_{t_1}, \cdots, X_{t_n}\}$ is Gaussian. Note that it is possible to construct a set of random variables that are individually Gaussian but whose joint distribution is not Gaussian.

6.1 Gaussian Interpolation

The proof of the Gaussian comparison inequality in the next section relies on a technique known as Gaussian interpolation. First we have the multidimensional version of the Gaussian integration by parts.

Lemma 6.2 (Gaussian integration by parts) *Let $X \sim \mathcal{N}(0, \Sigma)$, where Σ is an $n \times n$ variance matrix. Then,*

$$\mathbb{E}\left[X_i f(X)\right] = \sum_{j=1}^{n} \Sigma_{ij} \mathbb{E}\left[\frac{\partial f}{\partial x_j}(X)\right].$$

Proof: In the special 1-d case when $X \sim \mathcal{N}(0,1)$, the claim of the lemma reduces to

$$\mathbb{E}\left[X f(X)\right] = \mathbb{E}\left[f'(X)\right],$$

which follows immediately after we apply the integration by part to

$$\mathbb{E}\left[f'(X)\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f'(x) e^{-\frac{x^2}{2}} dx.$$

In general, first note that letting $Z \sim \mathcal{N}(0, I_n)$, then X has the same distribution as $\Sigma^{1/2} Z$. Thus,

$$\mathbb{E}\left[X_i f(X)\right] = \sum_{k=1}^{n} \Sigma_{ik}^{1/2} \mathbb{E}\left[Z_k f(\Sigma^{1/2} Z)\right] = \sum_{k=1}^{n} \Sigma_{ik}^{1/2} \mathbb{E}\left[Z_k g(Z)\right],$$

where $g(z) = f(\Sigma^{1/2} z)$ and hence

$$\frac{\partial g}{\partial z_k}(z) = \sum_{j=1}^{n} \Sigma_{jk}^{1/2} \frac{\partial f}{\partial x_j}(\Sigma^{1/2} z).$$

Since the result for the special 1-d case implies (noting Z_k are independent)

$$\mathbb{E}\left[Z_k g(Z)\right] = \mathbb{E}\left[\frac{\partial g}{\partial z_k}(Z)\right] = \sum_{j=1}^{n} \Sigma_{jk}^{1/2} \mathbb{E}\left[\frac{\partial f}{\partial x_j}(\Sigma^{1/2} Z)\right],$$

we have

$$\begin{aligned}
\mathbb{E}\left[X_i f(X)\right] &= \sum_{k=1}^{n} \Sigma_{ik}^{1/2} \mathbb{E}\left[Z_k g(Z)\right] \\
&= \sum_{k=1}^{n} \Sigma_{ik}^{1/2} \sum_{j=1}^{n} \Sigma_{jk}^{1/2} \mathbb{E}\left[\frac{\partial f}{\partial x_j}(\Sigma^{1/2} Z)\right] \\
&= \sum_{j=1}^{n} \left(\sum_{k=1}^{n} \Sigma_{ik}^{1/2} \Sigma_{jk}^{1/2}\right) \mathbb{E}\left[\frac{\partial f}{\partial x_j}(\Sigma^{1/2} Z)\right] \\
&= \sum_{j=1}^{n} \Sigma_{ij} \mathbb{E}\left[\frac{\partial f}{\partial x_j}(X)\right],
\end{aligned}$$

as desired. ∎

Using the Gaussian integration by parts property, we are ready to present and prove the Gaussian interpolation result.

Lemma 6.3 *Let $X \sim \mathcal{N}(0, \Sigma^X)$ and $Y \sim \mathcal{N}(0, \Sigma^Y)$ be two independent n-dimensional Gaussian vectors. Define*

$$Z(t) = \sqrt{t}X + \sqrt{1-t}Y, \quad t \in [0,1].$$

Then for every smooth function f we have

$$\frac{d}{dt}\mathbb{E}\left[f(Z(t))\right] = \frac{1}{2}\sum_{i,j=1}^{n}\left(\Sigma_{ij}^X - \Sigma_{ij}^Y\right)\mathbb{E}\left[\frac{\partial^2 f}{\partial x_i \partial x_j}(Z(t))\right].$$

Proof: By the chain rule we have

$$\begin{aligned}\frac{d}{dt}\mathbb{E}\left[f(Z(t))\right] &= \sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial f}{\partial x_i}(Z(t))\frac{dZ_i}{dt}\right] \\ &= \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial f}{\partial x_i}(Z(t))\frac{X_i}{\sqrt{t}}\right] - \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial f}{\partial x_i}(Z(t))\frac{Y_i}{\sqrt{1-t}}\right].\end{aligned}$$

Considering the first term, as X and Y are independent, we can apply Lemma 6.2 to the $2n$-dimensional Gaussian random (X, Y) (**what is the covariance matrix?**) and obtain

$$\sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial f}{\partial x_i}(Z(t))\frac{X_i}{\sqrt{t}}\right] = \sum_{i=1}^{n}\sum_{j=1}^{n}\Sigma_{ij}^X\mathbb{E}\left[\frac{\partial^2 f}{\partial x_i \partial x_j}(Z(t))\right].$$

Since the second term can be bounded similarly, the proof is complete. ■

6.2 Gaussian Comparison Inequality

Theorem 6.4 (Sudakov-Fernique inequality) *Let $\{X_t\}_{t\in T}$ and $\{Y_t\}_{t\in T}$ be two mean zero separable Gaussian processes. Suppose*

$$\mathbb{E}\left[|X_t - X_s|^2\right] \geq \mathbb{E}\left[|Y_t - Y_s|^2\right] \quad \textit{for all } t, s \in T.$$

Then,

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \geq \mathbb{E}\left[\sup_{t\in T} Y_t\right].$$

This theorem is very intuitive: if $\{X_t\}_{t\in T}$ has larger pairwise variance than $\{Y_k\}_{t\in T}$, then $\mathbb{E}\left[\sup_{t\in T} X_t\right] \geq \mathbb{E}\left[\sup_{t\in T} Y_t\right]$. It is enough to establish the theorem for two Gaussian vectors $X \sim \mathcal{N}(0, \Sigma^X)$ and $Y \sim \mathcal{N}(0, \Sigma^Y)$. Moreover, we can assume X and Y are independent; otherwise we can consider an independent copy of one of them.

Proof: For any $\beta > 0$ define

$$f_\beta(x) = \frac{1}{\beta}\log\sum_{k=1}^{n} e^{\beta x_k}.$$

It is not hard to see that (**check this!**)

$$\max_{k=1,\cdots,n} x_k \le f_\beta(x) \le \max_{k=1,\cdots,n} x_k + \frac{\log n}{\beta}.$$

Thus, $f_\beta(x) \to \max_{k=1,\cdots,n} x_k$ as $\beta \to \infty$. Moreover,

$$\frac{\partial f}{\partial x_k} = \frac{e^{\beta x_k}}{\sum_{k=1}^n e^{\beta x_k}} =: p_k(x), \quad \frac{\partial^2 f}{\partial x_k \partial x_j} = \beta\left(\delta_{kj} p_k(x) - p_k(x) p_j(x)\right),$$

where δ_{kj} equals 1 if $k = j$ and equals 0 otherwise. It follows from Lemma 6.3 that

$$\begin{aligned}\frac{d}{dt}\mathbb{E}\left[f_\beta(Z(t))\right] &= \frac{1}{2}\sum_{k,j=1}^n \left(\Sigma^X_{kj} - \Sigma^Y_{kj}\right)\mathbb{E}\left[\frac{\partial^2 f_\beta}{\partial x_k \partial x_j}(Z(t))\right] \\ &= \frac{\beta}{2}\sum_{k=1}^n \left(\Sigma^X_{kk} - \Sigma^Y_{kk}\right)\mathbb{E}\left[p_k(Z(t))(1 - p_k(Z(t)))\right] - \frac{\beta}{2}\sum_{k\neq j}\left(\Sigma^X_{kj} - \Sigma^Y_{kj}\right)\mathbb{E}\left[p_k(Z(t))p_j(Z(t))\right].\end{aligned}$$

Noting that $1 - p_k(x) = \sum_{j\neq k} p_j(x)$, we have

$$\begin{aligned}\sum_{k=1}^n \left(\Sigma^X_{kk} - \Sigma^Y_{kk}\right)\mathbb{E}\left[p_k(Z(t))(1 - p_k(Z(t)))\right] &= \sum_{k\neq j}\left(\Sigma^X_{kk} - \Sigma^Y_{kk}\right)\mathbb{E}\left[p_k(Z(t))p_j(Z(t))\right] \\ &= \sum_{k\neq j}\left(\Sigma^X_{jj} - \Sigma^Y_{jj}\right)\mathbb{E}\left[p_k(Z(t))p_j(Z(t))\right].\end{aligned}$$

It follows that

$$\begin{aligned}\frac{d}{dt}\mathbb{E}\left[f_\beta(Z(t))\right] &= \sum_{k\neq j}\frac{\beta}{4}\left(\Sigma^X_{kk} - 2\Sigma^X_{kj} + \Sigma^X_{jj}\right)\mathbb{E}\left[p_k(Z(t))p_j(Z(t))\right] - \sum_{k\neq j}\frac{\beta}{4}\left(\Sigma^Y_{kk} - 2\Sigma^Y_{kj} + \Sigma^Y_{jj}\right)\mathbb{E}\left[p_k(Z(t))p_j(Z(t))\right] \\ &= \frac{\beta}{4}\sum_{k\neq j}\left(\mathbb{E}\left[|X_k - X_j|^2\right] - \mathbb{E}\left[|Y_k - Y_j|^2\right]\right)\mathbb{E}\left[p_k(Z(t))p_j(Z(t))\right] \\ &\ge 0,\end{aligned}$$

where in the last line we have used the assumption. Thus $f_\beta(Z(t))$ is increasing in t, yielding

$$\mathbb{E}\left[f_\beta(X)\right] \ge \mathbb{E}\left[f_\beta(Y)\right].$$

Letting $\beta \to \infty$ concludes the proof. ∎

There are also other types of Gaussian comparison inequalities such as the Slepian inequality or the Gordon inequality, which can be proved similarly, see for example [3]. The Gaussian comparison inequalities have many interesting applications. Here we give an example before presenting its application on the lower bound for the expectation of suprema of the Gaussian process.

Example 6.5 (Spectral norm of Gaussian matrices) *Let $W \in \mathbb{R}^{m\times n}$ be a random matrix with i.i.d $\mathcal{N}(0,1)$ entries. By the the finite approximation bound in Lecture 4, we have*

$$\mathbb{E}\left[\|W\|_2\right] \le C(\sqrt{m} + \sqrt{n}).$$

Next we can show that the bound can be sharpened to

$$\mathbb{E}\left[\|W\|_2\right] \leq \sqrt{m} + \sqrt{n}$$

by the Sudakov-Fernique inequality[2]*. We still begin with the variational form for* $\|W\|_2$*,*

$$\|W\|_2 = \sup_{u \in \mathbb{B}_2^m, v \in \mathbb{B}_2^n} u^T W v =: \sup_{u \in \mathbb{B}_2^m, v \in \mathbb{B}_2^n} Y_{uv}.$$

We have

$$\begin{aligned}
\mathbb{E}\left[|Y_{uv} - Y_{ts}|^2\right] &= \mathbb{E}\left[\left(\sum_{ij} W_{kj}(u_i v_j - t_i s_j)\right)^2\right] \\
&= \sum_{ij} (u_i v_j - t_i s_j)^2 \\
&= \|uv^T - ts^T\|_F^2 \\
&\leq \|u - t\|_2^2 + \|v - s\|_2^2.
\end{aligned}$$

If we construct another Gaussian process as follows,

$$X_{uv} = \langle g, u \rangle + \langle h, v \rangle, \quad g \sim \mathcal{N}(0, I_m),\ h \sim \mathcal{N}(0, I_n).$$

it is easy to see that $\mathbb{E}\left[|X_{uv} - X_{ts}|^2\right] = \|u - t\|_2^2 + \|v - s\|_2^2$*. Thus, applying the Sudakov-Fernique inequality yields*

$$\begin{aligned}
\mathbb{E}\left[\sup_{u \in \mathbb{B}_2^m, v \in \mathbb{B}_2^n} u^T W v\right] &\leq \mathbb{E}\left[\sup_{u \in \mathbb{B}_2^m, v \in \mathbb{B}_2^n} \langle g, u \rangle + \langle h, v \rangle\right] \\
&= \mathbb{E}\left[\sup_{u \in \mathbb{B}_2^m} \langle g, u \rangle\right] + \mathbb{E}\left[\sup_{v \in \mathbb{B}_2^n} \langle h, v \rangle\right] \\
&= \mathcal{G}(\mathbb{B}_2^m) + \mathcal{G}(\mathbb{B}_2^n) \\
&\leq \sqrt{m} + \sqrt{n}.
\end{aligned}$$

It is worth noting that this example is a special case of the Chevet theorem which considers the problem on a compact subsets of the unit spheres.

6.3 Sudakov Minoration Inequality

Theorem 6.6 (Sudakov minoration inequality) *Let* $\{X_t\}_{t \in T}$ *be a centered Gaussian process. Then*

$$\mathbb{E}\left[\sup_{t \in T} X_t\right] \gtrsim \sup_{\varepsilon > 0} \varepsilon \sqrt{\log N(T, d, \varepsilon)},$$

where d *is the canonical metric defined in* (6.1)*.*

[2]However, note that the finite approximation bound works for all the general sub-Gaussian matrices, not only the standard Gaussian matrices.

Proof: For any $\varepsilon > 0$, let P be ε-packing of T under the canonical metric with the packing number $P(T, d, \varepsilon)$. Let $X = \{X_t\}_{t\in P}$ and let $Y = \{Y_t\}_{t\in P}$ be a vector of length $P(T, d, \varepsilon)$ with i.i.d $\mathcal{N}(0, \frac{\varepsilon^2}{2})$ variables. Then,

$$\mathbb{E}\left[|X_t - X_s|^2\right] = d(t,s)^2 \geq \varepsilon^2 = \mathbb{E}\left[|Y_t - Y_s|^2\right].$$

Thus the Sudakov-Fernique inequality yields

$$\mathbb{E}\left[\sup_{t\in P} X_t\right] \geq \mathbb{E}\left[\sup_{t\in P} Y_t\right] \asymp \varepsilon\sqrt{\log P(T, d, \varepsilon)} \geq \varepsilon\sqrt{\log N(T, d, \varepsilon)},$$

where the last inequality follows from the relationship $N(T, d, \varepsilon) \leq P(T, d, \varepsilon)$. ■

Sudakov minoration inequality can be used in two different ways: converting lower bound of covering number into lower bound of the suprema of Gaussian process, and converting upper bound of the suprema of Gaussian process into upper bound of covering number.

Example 6.7 (Lower bound on suprema of i.i.d Gauss) *We have already shown in Lecture 4 that*

$$\mathbb{E}\left[\max_{k=1,\cdots,n} g_k\right] \gtrsim \sqrt{\log n}$$

for i.i.d standard Gaussian random variables g_k. The lower bound actually can also be established via Sudakov minoration inequality. First note that

$$\max_{k=1,\cdots,n} g_k = \max_{t\in T} \langle g, t\rangle,$$

where $T = \{e_1, \cdots, e_n\}$ and $g \in \mathcal{N}(0, I_n)$. Since for sufficiently small ε, $N(T, \|\cdot\|_2, \varepsilon) = n$, it follows immediately that $\mathbb{E}\left[\max_{k=1,\cdots,n} g_k\right] \gtrsim \sqrt{\log n}$.

Example 6.8 (Gaussian width of unit 2-norm ball $\mathbb{B}_2^d$) *In Lecture 4, we have seen that*

$$\mathcal{G}(\mathbb{B}_2^d) = \mathbb{E}\left[\sup_{t\in\mathbb{B}_2^d} \langle g, t\rangle\right] \asymp \sqrt{d},$$

where the lower bound is obtained via the comparison with the corresponding Rademacher complexity. Since $\langle g, t\rangle$ is a Gaussian process with the canonical metric given by

$$d(t,s) = \sqrt{\mathbb{E}\left[|\langle g, t-s\rangle|^2\right]} = \|t-s\|_2,$$

we can also use the Sudakov minoration inequality to get the lower bound,

$$\mathcal{G}(\mathbb{B}_2^d) \gtrsim \varepsilon\sqrt{\log N(\mathbb{B}_2^d, \|\cdot\|_2, \varepsilon)} \asymp \varepsilon\sqrt{d\log\frac{1}{\varepsilon}} \asymp \sqrt{d}$$

after choosing a proper ε.

Example 6.9 (Lower bound on spectral norm of Gaussian matrices) *In Example 6.5, we have seen that*

$$\mathbb{E}\left[\|W\|_2\right] \le \sqrt{m} + \sqrt{n}$$

for an $m \times n$ Gaussian random matrix. The Sudakov minoration inequality can be used to show that this bound is sharp in terms of the scaling. Recall that

$$\|W\|_2 = \sup_{u \in \mathbb{B}_2^m, v \in \mathbb{B}_2^n} u^T W v =: \sup_{u \in \mathbb{B}_2^m, v \in \mathbb{B}_2^n} Y_{uv}.$$

*The application of the Sudakov minoration inequality yields that (**complete the details!**)*

$$\begin{aligned}
\mathbb{E}\left[\|W\|_2\right] &\gtrsim \varepsilon \sqrt{\log N(\mathbb{B}_2^m \otimes \mathbb{B}_2^n, \|\cdot\|_F, \varepsilon)} \\
&\gtrsim \varepsilon \sqrt{\log \left(N(\mathbb{B}_2^m, \|\cdot\|_2, \varepsilon) \cdot N(\mathbb{B}_2^n, \|\cdot\|_2, \varepsilon)\right)} \\
&\asymp \varepsilon \sqrt{(m+n) \log \frac{1}{\varepsilon}} \\
&\gtrsim \sqrt{m} + \sqrt{n}
\end{aligned}$$

after choosing ε properly.

Example 6.10 (Metric entropy of unit 1-norm ball $\mathbb{B}_1^d$ under the Euclidean distance) *We have already seen that*

$$\mathcal{G}(\mathbb{B}_1^d) = \mathbb{E}\left[\sup_{t \in \mathbb{B}_1^d} \langle g, t \rangle\right] \asymp \sqrt{\log d}.$$

Together with the Sudakov minoration inequality, we have

$$\log N(\mathbb{B}_1^d, \|\cdot\|_2, \varepsilon) \lesssim \frac{1}{\varepsilon^2} \log d.$$

Up to constant, this result matches the bound for the covering number of a convex hull of a finite set due to Maurey. Noting that

$$\log N(\mathbb{B}_2^d, \|\cdot\|_2, \varepsilon) \asymp d \log \frac{1}{\varepsilon},$$

we can see in a different way that the unit 1-norm ball is much smaller than the unit 2-norm ball.

6.4 A Short Remark

Combining the Sudakov minoration inequality with the Dudley inequality/integral, we have for the Gaussian process $\{X_t\}_{t \in T}$

$$\sup_k 2^{-k} \sqrt{\log N(T, d, 2^{-k})} \lesssim \mathbb{E}\left[\sup_{t \in T} X_t\right] \lesssim \sum_k 2^{-k} \sqrt{\log N(T, d, 2^{-k})}.$$

In some situations, the upper and lower bounds are not as far apart as may appear at first sight because the term $2^{-k}\sqrt{\log N(T, d, 2^{-k})}$ behaves like a geometric sequence so that their sum is of the same order as the largest one (for example, consider Example 6.8). However, there are also cases where there is indeed a gap between these two bounds. It turns out the generic chaining bound is tight for Gaussian processes, i.e.,

$$\mathbb{E}\left[\sup_{t\in T} X_t\right] \asymp \gamma(T, d),$$

see Section 3 of Lecture 6 for the definition of $\gamma(T, d)$. This is the notable Talagrand majorizing measure theorem. We will omit the details, see for example [2] and [3]. For *stationary* Gaussian process, Dudley integral is also tight.

Reading Materials

[1] Martin Wainwright, *High Dimensional Statistics – A non-asymptotic viewpoint*, Chapter 5.4.

[2] Ramon van Handel, *Probability in High Dimension*, Chapter 6.1.

[3] Roman Vershynin, *High-Dimensional Probability: An introduction with applications in data science*, Chapter 7.