Mathematical foundations of infinite-dimensional statistical models

Chapter 2.5 and 2.6.1

Presented by Ilsang Ohn

August 17, 2018

Main result of Chapter 2.5

Concentration of the supremum of a Gaussian process about its expectation:

Theorem 2.5.8 (Borell-Sudakov-Tsirelson inequality) Let $X(t)$, $t \in T$ be a separable centered Gaussian process such that

$$\|X\|_\infty := \sup_{t \in T} |X(t)| < \infty \text{ a.s.}$$

Let $\sigma^2 = \sup_{t \in T} \mathbb{E}\left[X^2(t)\right]$. Then for $u > 0$

$$\mathbb{P}\left(\|X\|_\infty \geq \mathbb{E}\,\|X\|_\infty + u\right) \leq e^{-u^2/2\sigma^2},$$
$$\mathbb{P}\left(\|X\|_\infty \leq \mathbb{E}\,\|X\|_\infty - u\right) \leq e^{-u^2/2\sigma^2}.$$

- $\|X\|_\infty$ is measurable since X is separable.
- The supremum of GP has sub-Gaussian tails.
- The bound is independent of the size or complexity of the index set T.

Theorem 2.5.6 (Log-Sobolev inequality) Let γ be the standard Gaussian measure on $\mathbb{R}^n$, and let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function such that f^2 and $\|f'\|^2$ are γ-integrable. Then

$$\mathrm{Ent}_\gamma(f^2) \leq 2 \int \|f'\|_2^2 \, \mathrm{d}\gamma$$

Definition 2.5.1 (Entropy) The entropy of $f \geq 0$ with respect to a probability measure μ is defined as

$$\mathrm{Ent}_\mu(f) := \int f \log f \mathrm{d}\mu - \left(\int f \mathrm{d}\mu \right) \left(\log \int f \mathrm{d}\mu \right),$$

if $\int f \log(1+f)\mu < \infty$ and as ∞ otherwise.

Proof of Theorem 2.5.6

Proposition 2.5.3 (Tensorization of entropy) Let $P = \mu_1 \times \cdots \times \mu_n$ and let $f \geq 0$ on a product space. Then

$$\mathrm{Ent}_P(f) \leq \sum_{i=1}^{n} \int \mathrm{Ent}_{\mu_i}(f) \mathrm{d}P$$

where

$$\mathrm{Ent}_{\mu_i}(f) = \int f \log f \mathrm{d}\mu_i(x_i) - \left(\int f \mathrm{d}\mu_i(x_i) \right) \left(\log \int f \mathrm{d}\mu_i(x_i) \right)$$

which is the function of $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$.

If we have $\mathrm{Ent}_{\mu_i}(f) \leq 2 \int |\partial f / \partial x_i|^2 \mathrm{d}\mu_i$, then

$$\mathrm{Ent}_P(f^2) \leq \sum_{i=1}^{n} \int \mathrm{Ent}_{\mu_i}(f) \mathrm{d}P \leq 2 \sum_{i=1}^{n} \int |\partial f / \partial x_i|^2 \mathrm{d}\mu_i = 2 \int \left\| f' \right\|_2^2 \mathrm{d}P$$

where $f' := (\partial f / \partial x_1, \ldots, \partial f / \partial x_n)$. Hence it is enough to show the cast of $n = 1$. Apply Taylor's expansion to g_λ, where $g_1 = g$ where $f = e^g$.

Inequality for Lipschitz functions on $\mathbb{R}^n$

Theorem 2.5.7 Let F be a Lipschitz on $\mathbb{R}^n$, with

$$\|F\|_{\mathrm{Lip}} := \sup_{x \neq y} \frac{|F(x) - F(y)|}{\|x - y\|_2} \leq 1.$$

Let $X = (g_1, \ldots, g_n)$ with g_i independent standard normal random variables. Then for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}\left(e^{\lambda F(X)}\right) \leq e^{\lambda \mathbb{E}(F(X)) + \lambda^2/2}.$$

As a consequence

$$\mathbb{P}\left(F(X) \geq \mathbb{E}(F(X)) + t\right) \leq e^{-t^2}, \quad \mathbb{P}\left(F(X) \leq \mathbb{E}(F(X)) - t\right) \leq e^{-t^2}.$$

Proof of Theorem 2.5.7

PROOF. First consider the case that F is twice continuously differentiable. Let $H(\lambda) := \mathbb{E}\left(e^{\lambda F(X)}\right)$. Apply the log-Sobolev inequality to $f^2 = e^{\lambda F}$ to get

$$\mathrm{Ent}_\gamma(e^{\lambda F}) \leq 2\left(\frac{\lambda}{2}\right)^2 \int \|F'\|_2^2 e^{\lambda F} d\gamma \leq \frac{\lambda^2}{2} \int e^{\lambda F} d\gamma = \frac{\lambda^2}{2} H(\lambda).$$

It is easy to see that

$$\mathrm{Ent}_\gamma(e^{\lambda F}) = \lambda H'(\lambda) - H(\lambda) \log H(\lambda).$$

Let $K(\lambda) = \frac{1}{\lambda} \log H(\lambda)$. Then by the above inequality,

$$\frac{d}{d\lambda} K(\lambda) = \frac{d}{d\lambda}\left(\frac{1}{\lambda} \log H(\lambda)\right) = \frac{1}{\lambda}\frac{H'(\lambda)}{H(\lambda)} - \frac{1}{\lambda^2} \log H(\lambda) \leq \frac{1}{2}.$$

Since $K(0) = H'(0)/H(0) = \mathbb{E}(F(X))$, we have

$$K(\lambda) = K(0) + \int_0^\lambda K'(t) dt \leq \mathbb{E}(F(X)) + \lambda/2.$$

Therefore

$$H(\lambda) \leq e^{\lambda \mathbb{E}(F(X)) + \lambda^2/2}.$$

The case of F Lipschitz follows by an approximation argument.

Proof of Theorem 2.5.8

Theorem 2.5.8 (Borell-Sudakov-Tsirelson inequality) Let $X(t)$, $t \in T$ be a separable centered Gaussian process such that $\|X\|_\infty := \sup_{t\in T} |X(t)| < \infty$ a.s.. Let $\sigma^2 = \sup_{t\in T} \mathbb{E}\left[X^2(t)\right]$. Then

$$\mathbb{P}\left(\|X\|_\infty \geq \mathbb{E}\|X\|_\infty + u\right) \leq e^{-u^2/2\sigma^2}, \quad \mathbb{P}\left(\|X\|_\infty \leq \mathbb{E}\|X\|_\infty - u\right) \leq e^{-u^2/2\sigma^2}.$$

PROOF. First we consider finite $T = \{t_1, \ldots, t_k\}$. Then we can write $X = AZ$ where $Z = (Z_1, \ldots, Z_k)$ is a standard normal vector and A is the symmetric square root of the covariance matrix of X. By the Cauchy-Schwarz inequality,

$$\max_{1\leq i\leq k} |Az|_i \leq \max_{1\leq i\leq k} \|A_i\|_2 \|z\|_2 = \max_{1\leq i\leq k} \sigma_i^2 \|z\|_2$$

and so the function $F(z) := \max_{1\leq i\leq k} |Az|_i/(\max_{1\leq i\leq k} \sigma_i^2)$ is Lipschitz with $\|F\|_{\text{Lip}} \leq 1$. Hence Theorem 2.5.7 yields the desired result.

Next we consider separable T. Since T is separable, $\|X\|_\infty$ is a monotone limit of a sequence of finite suprema almost surely. Then $\mathbb{E}\|X\|_\infty < \infty$ a.s. by monotone convergence. Fatou's lemma leads to the desired result.

Reproducing kernel Hilbert space of Gaussian processes

- The support of a centered Gaussian process (the smallest closed set having probability one under the induced measure) is equal to the closure of the reproducing kernel Hilbert space (RKHS) of the covariance kernel of the process (Corollary 2.6.17).
- RKHS determines the concentration of the Gaussian process (e.g., Theorem 2.6.12, Corollary 2.6.18).

Reproducing kernel Hilbert space of Gaussian processes

- The finite-dimensional distributions of a centered Gaussian process $X(t)$, $t \in T$ are determined by the covariance kernel $C : T \times T \mapsto \mathbb{R}$, defined by $C(s,t) = \mathbb{E}(X(s)X(t))$.
- The reproducing kernel Hilbert space of X is the completion of the linear space of all functions

$$t \mapsto \sum_{i=1}^{k} \alpha_i C(t_i, t)$$

where $\alpha_i \in \mathbb{R}$, $t_i \in T$, $k < \infty$, with inner product

$$\left\langle \sum_{i=1}^{k} \alpha_i C(t_i, \cdot), \sum_{j=1}^{l} \beta_i C(s_i, \cdot) \right\rangle_H = \sum_{i=1}^{k} \sum_{j=1}^{l} \alpha_i \beta_j C(t_i, s_i).$$

- We can write

$$\sum_{i=1}^{k} \alpha_i C(t_i, t) = \mathbb{E}\left[\left(\sum_{i=1}^{k} \alpha_i X(t_i)\right) X(t)\right]$$

and

$$\sum_{i=1}^{k} \sum_{j=1}^{l} \alpha_i \beta_j C(t_i, s_i) = \mathbb{E}\left[\sum_{i=1}^{k} \alpha_i X(t_i) \sum_{j=1}^{l} \beta_i X(s_i)\right]$$

- Reproducing property: for $h = \sum_{i=1}^{k} \alpha_i C(t_i, \cdot)$

$$h(t) = \langle h, C(t, \cdot) \rangle_H .$$

Reproducing kernel Hilbert space of Gaussian processes

For a Gaussian process $X(t)$, $t \in T$, let F be the set of all linear combinations $\sum_i \alpha_i X(t_i)$ and $\bar{F}$ be its closure in $L^2(\Omega, \Sigma, \mathbb{P})$.

Definition 2.6.1 (RKHS of GP) The reproducing kernel Hilbert space (RKHS) of a centered Gaussian process $X(t)$, $t \in T$ is the set of all functions

$$t \mapsto \mathbb{E}(hX(t))$$

where $h \in \bar{F}$, with inner product

$$\langle \mathbb{E}(h_1 X), \mathbb{E}(h_2 X) \rangle_H := \mathbb{E}(h_1 h_2).$$

REMARK. In the previous slide we define the RKHS as

$$H = \text{completion}\left(\{\mathbb{E}(hX) : h \in F\}\right)$$

Since $\langle \mathbb{E}(h_1 X), \mathbb{E}(h_2 X) \rangle_H = \langle h_1, h_2 \rangle_{L^2(\Omega, \Sigma, \mathbb{P})}$, the map $\phi : h \to \mathbb{E}(hX)$ is an linear isometry between $(F, \|\cdot\|_{L_2})$ and $(\phi(F), \|\cdot\|_H)$. Hence the two definitions coincide.

Example 2.6.2 (RKHS of iid standard normal variables) Suppose that $T = \mathbb{N}$ and for $n \in \mathbb{N}$

$$g_n := X(n) \overset{\text{iid}}{\sim} \mathrm{N}(0,1).$$

Then $\bar{F} = \{\sum_{i=1}^{\infty} \alpha_i g_i : \sum_{i=1}^{\infty} \alpha_i^2 < \infty\}$. Furthermore if $h = \sum_{i=1}^{\infty} \alpha_i g_i \in \bar{F}$, then

$$\mathbb{E}(hX(n)) = \mathbb{E}\left\{\left(\sum_{i=1}^{\infty} \alpha_i g_i\right) g_n\right\} = \alpha_n.$$

Hence $\mathbb{E}(hX) = \{\alpha_n\}_{n=1}^{\infty} \in \ell_2$. That is, the RKHS of the standard Gaussian measure on $\mathbb{R}^{\mathbb{N}}$ is ℓ_2.

Review: Banach-valued random variables

- Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space, and let $(B, \|\cdot\|)$ be a separable Banach space, equipped with its Borel σ-algebra $\mathcal{B}$. Let $X : \Omega \mapsto B$ be a B-valued random variable.
- For simple $X = \sum_{i=1}^{n} x_i \mathbb{I}_{A_i}$ $x_i \in B$, $A_i \in \Sigma$, define $\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{P}(A_i) x_i$.
- X is *Bochner integrable* or *strongly integrable* if there is a sequence of simple functions X_n such that $\mathbb{E}\|X_n - X\| \to 0$. Then define $\mathbb{E}X = \lim_{n\to\infty} \mathbb{E}X_n$, which is well defined since $\{\mathbb{E}X_n\}$ is Cauchy.
- X is *Pettis integrable* or *weakly integrable* if $f(X) \in L_1(\mathbb{P})$ for all $f \in B^*$ and there exists $x \in B$ such that $\mathbb{E}f(X) = f(x)$, $f \in B^*$, where B^* is the dual space of B, the collection of continuous, linear maps $f : B \mapsto \mathbb{R}$.
- **Lemma 2.6.3** Let B be a separable Banach space, and let X be a B-valued random variable. Then X is Bochner integrable if and only if $\mathbb{E}\|X\| < \infty$. Moreover, if X is Bochner integrable, then X is also Pettis integrable and both integrals coincide.

Review: GP as a Banach-valued random variable

- Gaussian processes can be viewed as a tight Borel measurable map in a Banach space B, for instance, a space of continuous functions or an L_p space (Example 2.1.6).
- A B-valued random variable X is centered Gaussian if $f(X)$ is a normal random variable with mean 0 for every $f \in B^*$.
- By the assumption that the Gaussian process X takes its values in the Banach space $(B, \|\cdot\|)$, $\|X\|$ is finite a.s..
- Since $\|X\|$ has sub-Gaussian tails (Theorem 2.6.8), all moments $\mathbb{E}\,\|X\|^p$ are finite.

RKHS of a Banach-valued Gaussian variable

Let $\bar{F}$ be a closure of $\{f(X) : f \in B^*\}$ in $L_2(\Omega, \Sigma, \mathbb{P})$.

Definition 2.6.4 Let B be a separable Banach space, and let X be a B-valued centered Gaussian variable. The reproducing kernel Hilbert space H of X is the vector space

$$H = \left\{\mathbb{E}(h(X)X) : h(X) \in \bar{F}\right\} \subset B,$$

with inner product

$$\langle \mathbb{E}(h_1(X)X), \mathbb{E}(h_2(X)X)\rangle_H := \mathbb{E}(h_1(X)h_2(X)).$$

- $\mathbb{E}\|h(X)X\| \leq \left[\mathbb{E}(h^2(X))\right]^{1/2} \left[\mathbb{E}\|X\|^2\right]^{1/2} < \infty$, since $h(X) \in L_2(\mathbb{P})$ and $\|X\|$ is square integrable.
- Hence $h(X)X$ is Pettis integrable by Lemma 2.6.3.

Lemma 2.6.6 The map $\varphi : B^* \to H$ defined as $\varphi(h) = \mathbb{E}(h(X)X)$ is weak* sequentially continuous. Consequently, if B_0^* is sequentially dense in B^* for the weak* topology, H is the closure of $\varphi(B_0^*)$ for the norm of H, $\|\cdot\|_H$.

- The weak* topology of B^* is the topology of pointwise convergence over B, denoted by $f_n \to_{w*} f$ (iff $f_n(x) \to f(x)$ for all $x \in B$).
- B_0^* is sequentially dense in B^* if for any $h \in B^*$ there is a sequence $\{h_n\} \subset B_0^*$ such that $h_n \to_{w*} h$.
- $\varphi(h)$ is weak* sequentially continuous if $\|\varphi(h_n) - \varphi(h)\|_H \to 0$ when $h_n \to_{w*} h$.
- Proof For Gaussian random variables, almost sure convergence implies L_2 convergence.

Example 2.6.7 (The RKHS of Brownian motion.)

- Brownian motion on [0,1] is a centered sample continuous Gaussian process W whose covariance is $\mathbb{E}(W(s)W(t)) = s \wedge t,\ s, t \in [0, 1]$.
- It can be thought as a B-valued random variable where $B = C([0, 1])$ endowed with sup norm.
- Let B_0^* be the set of finite linear combinations of point masses $\sum_{i=1}^{n} a_k \delta_{t_i}$.
- B_0^* is weak***** sequentially dense in B^*, which is the space of finite signed measures in [0,1]
- For $h \in B_0^*$, $\varphi(h) = \mathbb{E}(h(W)W)$ is piecewise linear continuous functions on [0,1] with $h(0) = 0$ and $\|h\|_H^2 = \int_0^1 (h'(t))^2 \mathrm{d}t$.
- Since $h'(t)$ is a step function and step functions are dense in L_2, the closure of $\varphi(B_0^*)$ is the set of absolutely continuous functions on [0,1] which are zero at zero and whose derivative is in $L_2([0, 1])$.
- By Lemma 2.6.6,

$$H = \{f : f(0) = 0,\ f \text{ is absolutely continuous, } f' \in L_2([0, 1])\}$$

equipped with the norm

$$\|f\|_H^2 = \int_0^1 (f'(t))^2 \mathrm{d}t.$$

RKHS is a separable Hilbert space.

Proposition 2.6.9 Let X be a centered B-valued Gaussian variable and B a separable Banach space. Then H is a separable Hilbert space and a measurable subset of B. The embedding of H into B is continuous, and in fact, the unit ball $O_H = \{h \in H : \|h\|_H \leq 1\}$ is a compact subset of B).

PROOF. By Alaoglu's theorem, the unit ball B_1^* in B^* is compact for the weak* topology. Thus there is a countable set D which is weak* dense in B_1^*. Since $B^* = \cup_n(nB_1^*)$, $B_0^* := \cup_n(nD)$ is countable and weak* sequentially dense in B^*. Since H is closure of $\varphi(B_0^*)$, H is separable.
Let $h \in H$ and let $k(X) \in \bar{F}$ be such that $h = \mathbb{E}(k(X)X)$

$$\|h\| = \sup_{f \in B_1^*} |\mathbb{E}(k(X)f(X))| \leq [\mathbb{E}(k(X)^2)]^{1/2} \sup_{f \in B_1^*} [\mathbb{E}(f(X)^2)]^{1/2} = \sigma \|h\|_H$$

where $\sigma = \sup_{f \in B_1^*} [\mathbb{E}(f(X)^2)]^{1/2}$, which shows that the embedding of H into B is continuous.

REMARK. The RKHS-norm on H is stronger than the original norm on B, and therefore a $\|\cdot\|_H$-Cauchy sequence in $H_0 := \{\mathbb{E}(h(X)X) : h(X) \in F\}$ is a $\|\cdot\|$-Cauchy sequence in B. Hence the RKHS H, which is the completion of H_0 under $\|\cdot\|_H$, can be identified with a subset of B.

The Karhunen-Loeve expansion of the Gaussian process

Theorem 2.6.10 Let X be a centered B-valued Gaussian variable and B a separable Banach space, and let H be its RKHS. Let z_j, $j \in \mathbb{N}$ be a complete orthonormal system of H, and let $k_j(X) \in \bar{F}$ be such that $\mathbb{E}(k_j(X)X) = z_j$. Then the series $\sum_{j=1}^{\infty} \mathbb{E}(k_j(X)X)k_j(X)$ converges a.s. to X in the norm of B (and the series reduces to a finite sum if $\dim(H) < \infty$).

NOTE. $k_j(X)$, $j \in \mathbb{N}$ are iid N$(0,1)$ random variables, since $\|z_j\|_H^2 = \langle \mathbb{E}(k_j(X)X), \mathbb{E}(k_j(X)X) \rangle_H$ $= \mathbb{E}(k_j(X)k_j(X)) = 1$.

PROOF. Define $U : H \to L_2(\Omega, \Sigma, \mathbb{P})$ by $U(\mathbb{E}(h(X)X)) = h(X)$. For any $f \in B^*$, by the linearity of U,

$$f(X) = U(\mathbb{E}(f(X)X)) = U\left(\sum_j \langle z_j, \mathbb{E}(f(X)X) \rangle_H z_j\right) = U\left(\sum_j \mathbb{E}(f(X)k_j(X))z_j\right)$$
$$= \sum_j f(\mathbb{E}(k_j(X)X))U(z_j) = \sum_j f(\mathbb{E}(k_j(X)X))k_j(X)$$

where the series converges in $L_2(\mathbb{P})$. In other words, for any $f \in B^*$,

$$f\left(\sum_{j=1}^{n} \mathbb{E}(k_j(X)X)k_j(X)\right) = \sum_{j=1}^{n} f(\mathbb{E}(k_j(X)X))k_j(X) \to_{L_2(\mathbb{P})} f(X).$$

By Levy-Ito-Nisio theorem (convergence in distribution of all marginals $f(\sum_{i=1}^{n} X_i)$ to the marginals $f(X)$ of some Borel measurable map X in a separable Banach space B for any $f \in B^*$, implies the almost sure convergence of the series $\sum_{i=1}^{n} X_i$), $\sum_{j=1}^{n} \mathbb{E}(k_j(X)X)k_j(X) \to X$ a.s.