Ten Lectures on Small Value Probabilities and Applications

L5: Metric Entropy and Small Ball Probability

Wenbo V. Li

University of Delaware
http://www.math.udel.edu/~wli
wli@math.udel.edu

CBMS Lectures at UAH, June 4-8, 2012

For a continues centered Gaussian processes, the generating linear operator is compact and so is the unit ball of the associated reproducing kernel Hilbert space. The fundamental links between small ball probability for Gaussian measure and the metric entropy are given and various far-reaching implications are explored. Several purely probabilistic results, obtained via the analytic connection without direct probabilistic proofs, are analyzed.

Covering Number, Metric Entropy and ε-nets

Let A be a compact subset in a metric space (E, ρ), and let $\varepsilon > 0$ be given. The *metric entropy* of A is denoted by $H(A, \rho, \varepsilon) = \log N(A, \rho, \varepsilon)$ where

$$\begin{aligned} N(A, \varepsilon) &= N(A, \rho, \varepsilon) = N(A, \varepsilon B_\rho) \\ &= \min \{ n \geq 1 \,:\, \exists x_1, \cdots, x_n \in A \\ &\qquad\quad \text{such that } A \subset \cup_{j=1}^{n} (x_j + \varepsilon B_\rho) \} \,, \end{aligned}$$

and $B_\rho(a; r) = \{x \,:\, \rho(x, a) < r\}$ is the open ball of radius r centered at a.

We also say a set $F \subset \mathbb{R}^d$ is an ε-net for A with respect to B if $A \subset \cup_{x \in F} (x + \varepsilon B)$. The smallest cardinality of an ε-net is denoted by $N(A, B, \varepsilon) = N(A, \varepsilon B)$.

•The metric entropy is a natural representation of how many bits you need to send in order to identify an element of a set up to precision ε. It is a tool heavily used in approximation theory, probability and statistics, learning theory, compressive sensing and random matrix theory

•Kolmogorov (1956), Asymptotic characteristic of some complete bounded metric spaces. Dokl. Akad, Nauk SSSR, 108, 585-589.
•A.N. Kolmogorov and V.M. Tikhomirov, ε-entropy and ε-capacity of sets in function spaces, Uspecki 14 (1959), 3-86.
•Lorentz (1966), Metric entropy and approximation, Bull. Amer. Math. Soc. 72, 903-937.
•Kuelbs, J. and Li, W.V. (1993). Metric entropy and the small ball problem for Gaussian measures. J. Funct. Anal. 116, 133-157.
•Li, W.V. and Linde, W. (1999). Approximation, metric entropy and small ball estimates for Gaussian measures. Ann. Probab. 27, 1556-1578.

ε-Distinguishable and Capacity

Points $h_1, h_2, \cdots, h_m$ of A are called ε-*distinguishable* if the distance between each two of them exceeds ε.
The *capacity* of A is denoted by $C(A, \varepsilon) = \log M(A, \varepsilon)$ where

$$\begin{aligned} M(A, \varepsilon) &= \max\{n \geq 1 : \exists h_1, \cdots, h_n \in A, \\ &\qquad \|h_i - h_j\| \geq \varepsilon \text{ for all } i \neq j\} \end{aligned}$$

The general facts about the metric entropy:

$$M(A, 2\varepsilon) \leq N(A, \varepsilon) \leq M(A, \varepsilon), \quad \text{i.e} \quad C(A, 2\varepsilon) \leq H(A, \varepsilon) \leq C(A, \varepsilon).$$

If E is a Banach space and $\lambda > 0$, then

$$H(\lambda A, \varepsilon) = H(A, \varepsilon\lambda^{-1}).$$

Notations: $f(x) \approx g(x)$ as $x \to a$ if

$$0 < \liminf_{x \to a} f(x)/g(x) \leq \limsup_{x \to a} f(x)/g(x) < \infty.$$

and $f(x) \preceq g(x)$ as $x \to a$ if $\limsup_{x \to a} f(x)/g(x) < \infty$.

Methods To Find Metric Entropy

•Construction/approximation, both upper and lower bounds;
•Volume arguments (lower bounds); Embedding (upper bounds); etc.
•Small ball probability, both upper and lower bounds;

Ex: $A = [0,1]^d$ = unit cube in $\mathbb{R}^d$ and ρ=usual Euclidean metric. Then as $\varepsilon \to 0$,

$$c_d \varepsilon^{-d} \le N(A, \varepsilon) \le C_d \varepsilon^{-d} \qquad \text{and} \quad H(A_1, \varepsilon) \sim d \log \frac{1}{\varepsilon}.$$

The constants α_d such that $N(A_1, \varepsilon) \sim \alpha_d \varepsilon^{-d}$ is almost impossible to obtain for $d \ge 3$.
For $d = 2$, it is easy to see that

$$\frac{1}{\pi \varepsilon^2} \le N(A_1, \varepsilon) \le \left([\varepsilon^{-1}] + 1\right)^2$$

Ex:

$$A_2 = \{f \in C[0,1] : f(0) = 0, |f(x) - f(y)| \le |x - y|^\alpha, \forall x, y \in [0,1]\}$$

for $0 < \alpha \le 1$ and $\|f\| = \sup_{0 \le x \le 1} |f(x)|$.
Then $H(A_2, \varepsilon) \approx (1/\varepsilon)^{1/\alpha}$ as $\varepsilon \to 0$.
Ex:

$$\begin{aligned} A_3 &= \{f \in C[0,1] : f(0) = 0, |f(x) - f(y)| \le |x - y|^\alpha, \\ &\quad \forall x, y \in [0,1] \text{and} \quad \mathrm{Var}(f, [0,1]) \le 1 \ \} \end{aligned}$$

where $0 < \alpha < 1$.
Then $H(A_3, \varepsilon) \approx (1/\varepsilon) \log(1/\varepsilon)$ as $\varepsilon \to 0$ (Clements, 1963).

Ex:

$$K = \left\{ g \in C[0,1] : g(0) = 0, g \text{ absol. cont.}, \int_0^1 |g'(s)|^2 ds \le 1 \right\}.$$

Note that $K \subset A_3$ when $\alpha = 1/2$ since

$$\begin{aligned} |g(t) - g(s)| &= \left| \int_s^t g'(u) du \right| \\ &\le (t-s)^{1/2} \left(\int_s^t \left| g'(u) \right|^2 du \right)^{1/2} \le (t-s)^{1/2} \end{aligned}$$

and $\mathrm{Var}(g) \le \int_0^1 |g'(t)|\, dt \le 1$

- Kolmogorov and Tihomirov (1961): $H(\varepsilon, K, \|\cdot\|_2) \approx 1/\varepsilon$.
- Birman and Solomjak (1967): $H(\varepsilon, K, \|\cdot\|_\infty) \approx 1/\varepsilon$.
- Kuelbs and Li (1993): As $\varepsilon \to 0$

$$(2-\sqrt{3})/4 \le \varepsilon \cdot H(K, \|\cdot\|_2, \varepsilon) \le 1$$
$$(2-\sqrt{3})\pi/4 \le \varepsilon \cdot H(K, \|\cdot\|_\infty, \varepsilon) \le \pi.$$

Note that K is the unit ball of the reproducing kernel Hilbert space generated by Brownian motion.

Metric Entropy of High Dimensional Distributions

Let $\mathcal{F}_d$ be the collection of all d-dimensional probability distribution functions on $[0,1]^d$, $d \geq 2$. The metric entropy of $\mathcal{F}_d$ under the $L_2([0,1]^d)$ norm arises naturally in non-parametric estimation in statistics; see van der Vaart and Wellner (1996). The following result is proved in Blei, Li and Gao (2007) by interplays among duality relations, metric entropy and small ball probability.

Thm: We have

$$\log N(\mathcal{F}_d, \|\cdot\|_2, \varepsilon) \approx \log N(K_d, \|\cdot\|_\infty, \varepsilon).$$

In particular, for $d \geq 2$ there exist constants $c_1, c_2 > 0$ depending only on d such that

$$c_1 \varepsilon^{-1}[\log(1/\varepsilon)]^{d-1} \leq \log N(\mathcal{F}_d, \|\cdot\|_2, \varepsilon) \leq c_2 \varepsilon^{-1}[\log(1/\varepsilon)]^{d-1/2}.$$

and for $d = 2$,

$$c_1 \varepsilon^{-1}[\log(1/\varepsilon)]^{3/2} \leq \log N(\mathcal{F}_2, \|\cdot\|_2, \varepsilon) \leq c_2 \varepsilon^{-1}[\log(1/\varepsilon)]^{3/2}.$$

•Small ball inequality on a lower bound on the L_1 norm of sums of Haar functions, see Talagrand (1994), Temlyakov (1998).
•van der Vaart and van Zanten (2008): Statistical applications for Gaussian priors based on Reproducing kernel Hilbert spaces of Gaussian priors.
•Bilyk and Lacey (2008): Small ball inequality in harmonic analysis and discrepancy theory.
•Gao, Li and Welner (2010): How many Laplace transforms of probability measures are there? Applications to bracket entropy in empirical processes theory and learning theory.

Links between Small Ball and Metric Entropy

As it was established in Kuelbs and Li (1993) and completed in Li and Linde (1999), the behavior of

$$\phi(\varepsilon) := \log \mathbb{P}\left(\|X\| \leq \varepsilon\right)$$

for Gaussian random element X is determined up to a constant by the metric entropy of the unit ball of the reproducing kernel Hilbert space associated with X, and vice versa.

- The Links can be formulated for entropy numbers of compact operator from Hilbert space to Banach space.
- This is a fundamental connection (both asymptotic and non-asymptotic) that has been used to solve important questions on both directions.

Ex: For the standard Brownian motion $W(t), 0 \le t \le 1$ on $C[0,1]$, the associated compact operate is the integration operator

$$uf(t) = \int_0^t f(s)ds.$$

The unit ball of the RKHS is

$$K = \left\{ g \in C[0,1] : g(0) = 0, g \,\text{absol. cont.}, \int_0^1 |g'(s)|^2 ds \le 1 \right\}.$$

- Kolmogorov and Tihomirov (1961): $H(\varepsilon, K, \|\cdot\|_2) \approx 1/\varepsilon$.
- Birman and Solomjak (1967): $H(\varepsilon, K, \|\cdot\|_\infty) \approx 1/\varepsilon$.
- Kuelbs and Li (1993): As $\varepsilon \to 0$

$$(2-\sqrt{3})/4 \le \varepsilon \cdot H(\varepsilon, K, \|\cdot\|_2) \le 1$$
$$(2-\sqrt{3})\pi/4 \le \varepsilon \cdot H(\varepsilon, K, \|\cdot\|_\infty) \le \pi.$$

The Small Ball Probability for Brownian Sheets

The standard Brownian sheet $W(t_1, \cdots t_d)$ on $[0,1]^d$ with values $C([0,1]^d)$ is associated with the integration operator

$$uf(t_1, \cdots t_d) = \int_0^{t_1} \cdots \int_0^{t_d} f(s_1, \cdots, s_d) ds_1 \cdots ds_d.$$

The associated compact set is

$$K_d = \{g \in C([0,1]^d) : g(t_1, \cdots, t_i = 0, \cdots, t_d) = 0, \\ \int_{[0,1]^d} \left| \frac{\partial^d}{\partial t_1 \cdots \partial t_d} g \right|^2 d\mathbf{t} \le 1\}.$$

• Lifshits and Tsyrelson (1986), Bass (1988), Talagrand (1994), Gao and Li (2007): For $d = 2$,

$$\log \mathbb{P}\Big(\sup_{\mathbf{t}\in[0,1]^2} |W(\mathbf{t})| \le \varepsilon \Big) \approx -\varepsilon^{-2} |\log \varepsilon|^3$$

or equivalently

$$\log N(\varepsilon, K_2, \|\cdot\|_\infty) \approx \varepsilon^{-1} (\log 1/\varepsilon)^{3/2}$$

by using the entropy link discovered in Kuelbs and Li (1993).

• The best known bounds for $d \geq 3$ are

$$-C_2\varepsilon^{-2}|\log\varepsilon|^{2d-1} \leq \log\mathbb{P}\Big(\sup_{\mathbf{t}\in[0,1]^d}|W(\mathbf{t})| \leq \varepsilon\Big) \leq -C_1\varepsilon^{-2}|\log\varepsilon|^{2d-2+\delta}$$

and it is conjectured that the lower bound is sharp.
•The upper bound without some $\delta > 0$ follows from simple L_2 estimates, see Li (1992) and the one with $\delta > 0$ is given in Bilyk and Lacey (2008) based on Harmonic analysis.
•The lower bound is proved in Dunker, Kuhn, Lifshits, and Linde (1999) based on combined techniques from entropy estimates and small ball estimates. The
•It is easy to reformulate the result as two-sided boundary crossing:

$$\log\mathbb{P}\Big(\sup_{t\in[0,T]^d}|W(t)| \leq 1\Big), \quad \text{as} \quad T \to \infty$$

for Gaussian random fields.

Precise links between small ball probability and metric entropy

The following is established in Kuelbs and Li (1993) and completed in Li and Linde (1999).

Thm: Let $J(x)$ be a slowly varying function at infinity such that $J(x) \approx J(x^\rho)$ as $x \to \infty$ for each $\rho > 0$.

(I). If $\phi(\varepsilon) \succeq \varepsilon^{-\alpha} J(\varepsilon^{-1})$, $\phi(2\varepsilon) \succeq \phi(\varepsilon)$, $\alpha > 0$, then

$$H(K_\mu, \varepsilon) \succeq \varepsilon^{-2\alpha/(2+\alpha)} J(1/\varepsilon)^{2/(2+\alpha)}.$$

(II). If $\phi(\varepsilon) \preceq \varepsilon^{-\alpha} J(\varepsilon^{-1})$, $\alpha > 0$, then

$$H(K_\mu, \varepsilon) \preceq \varepsilon^{-2\alpha/(2+\alpha)} J(1/\varepsilon)^{2/(2+\alpha)}.$$

(III). If $H(K_\mu, \varepsilon) \succeq \varepsilon^{-\alpha} J(1/\varepsilon)$, $0 < \alpha < 2$, then

$$\phi(\varepsilon) \succeq \varepsilon^{-2\alpha/(2-\alpha)} (J(1/\varepsilon))^{2/(2-\alpha)}.$$

(IV). If $H(K_\mu, \varepsilon) \preceq \varepsilon^{-\alpha} J(1/\varepsilon)$, $0 < \alpha < 2$ then

$$\phi(\varepsilon) \preceq \varepsilon^{-2\alpha/(2-\alpha)} (J(1/\varepsilon))^{2/(2-\alpha)}.$$

• As a consequence, for $\alpha > 0$ and $\beta \in \mathbb{R}$,

$$\phi(\varepsilon) \approx \varepsilon^{-\alpha}(\log 1/\varepsilon)^{\beta}$$

if and only if

$$H(K_\mu, \varepsilon) \approx \varepsilon^{-2\alpha/(2+\alpha)}(\log 1/\varepsilon)^{2\beta/(2+\alpha)}.$$

• In the notation of (dyadic) entropy numbers, the following are equivalent for $0 < \theta < 2$:
(i). $e_n(u) \approx n^{-1/\theta}(1 + \log n)^{\beta}$ as $n \to \infty$
(ii). $\log \mathbb{P}\left(\|X\|_E < \varepsilon\right) \approx -\varepsilon^{-2\theta/(2-\theta)}|\log \varepsilon|^{2\theta\beta/(2-\theta)}$ as $\varepsilon \to 0$,

• This is a fundamental connection that has been used to solve important questions on both directions.

• Many important problems are open, in particular, small ball or entropy number for tensors.

• Similar connections for other measures such as stable are studied. One direction is given in Li and Linde (2003).

Two Basic Relations for Centered Gaussian Measure μ

Let K_μ be the unit ball of reproducing kernel Hilbert spaces generated by μ in a Banach space with unit ball B. We have for any $\lambda > 0$ and $\varepsilon > 0$

$$\log N(\lambda K_\mu, 2\varepsilon B) \leq \lambda^2/2 - \log \mu(\varepsilon B)$$

and

$$\log N(\lambda K_\mu, \varepsilon B) + \log \mu(2\varepsilon B) \geq \log \Phi(\lambda + \alpha_\varepsilon)$$

where

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-u^2/2} du \quad \text{and} \quad \Phi(\alpha_\varepsilon) = \mu(\varepsilon B).$$

We outline the proof in the finite dimensional setting:

$$\begin{aligned} \mu &= \gamma_n \quad \text{(standard Gaussian measure on } \mathbb{R}^n) \\ K_\mu &= B_2^n \quad \text{(Euclidean ball)} \\ B &= K \quad (\text{norm } \|\cdot\| = \|\cdot\|_K) \end{aligned}$$

- Inequalities for shifted symmetric convex set K:

$$e^{-|h|_2^2/2} \cdot \mu(\varepsilon K) \leq \mu(\varepsilon K + h) \leq \mu(\varepsilon K).$$

The first relation uses full strength of the argument in proving dual Sudakov inequality. Assume that the points $x_1, \cdots, x_M \in B_2^n$ are 2ε-separated in $\|\cdot\|_K$. Then for any $j \neq l$,

$$(x_j + \varepsilon K) \cap (x_l + \varepsilon K) = \emptyset$$

Hence for any $\lambda > 0$

$$\begin{aligned} 1 &\geq \mu\left(\cup_{j=1}^M (\lambda x_j + \lambda\varepsilon K)\right) = \sum_{j=1}^M \mu(\lambda x_j + \lambda\varepsilon K) \\ &\geq \sum_{j=1}^M e^{-\lambda^2\|x_j\|^2/2} \cdot \mu(\lambda\varepsilon K) \geq M \cdot e^{-\lambda^2/2} \cdot \mu(\lambda\varepsilon K) \end{aligned}$$

This implies one key relation

$$\log N(B_2^n, 2\varepsilon K) - \lambda^2/2 + \log \mu(\lambda\varepsilon K) \leq 0.$$

For the second relation, by Gaussian isoperimetric inequality,

$$\begin{aligned} \Phi(\lambda + \Phi^{-1}(\mu(\varepsilon K)) &\leq \mu(\lambda B_2^n + \varepsilon K) \\ &\leq N(\lambda B_2^n + \varepsilon K, 2\varepsilon K) \cdot \mu(2\varepsilon K) \\ &\leq N(\lambda B_2^n, \varepsilon K) \cdot \mu(2\varepsilon K). \end{aligned}$$

Analytic (Entropy) Approach to Small Ball Probability

Advantages:

• Helpful algebraic properties of entropy numbers such as $e_{n+m-1}(u+v) \le e_n(u) + e_m(v)$.

• Tight relations between entropy numbers and other useful approximation quantities such as Kolmogorov widths and l_n-numbers.

• Powerful tools and methods to study entropy numbers.

• Enormous amount of deep and strong results about entropy numbers.

• Only known approach to some very useful small ball estimates.

Disadvantages:

• Analytic proofs give fewer insights into the structure of the Gaussian process.

• Results are never sharp w.r.t. the constants. Existence of small ball constants cannot be obtained.

• One relation to obtain $\phi(\varepsilon)$ from $e_n(u)$ is quite implicit. It is only useful in terms of asymptotic behavior rather than direct estimates, unless additional information on $\phi(\varepsilon)$ is known.

Probabilistic (SBP) Approach to Entropy

Advantages:

• Helpful algebraic properties such as

$$\mathbb{P}(\|X+Y\| \le \varepsilon) \ge \mathbb{P}(\|X\|+\|Y\| \le \varepsilon) \ge \mathbb{P}(\|X\| \le \lambda\varepsilon)\cdot\mathbb{P}(\|Y\| \le (1-\lambda)\varepsilon)$$

for independent random element X and Y, and any $0 < \lambda < 1$.

• Powerful Gaussian inequalities such as the weaker correlation inequality.

• Useful interactions with other areas of probability theory such as large deviations.

• Systematic probabilistic arguments such as Chaining and exponential Chebyshev's inequality.

• Direct but less sharp connections with the l_n numbers, the rate of approximation by "finite rank" processes.

• Reasonable constants for metric entropy can be obtained in nice cases.

Disadvantages:

• So far, only compact operator $u : H \to E$ with "good" rate can be studied this way.

The Winning Approach to Both Directions Is: Combining Them and Go Back and Forth

All of these enable the application of tools and results from functional analysis to small ball probabilities and vice versa.
•The following result in Li and Linde (1999) has no purely probabilistic proof to date.

Theorem

Let $Y = (Y(t))_{t\in[0,1]}$ be a centered Gaussian process with continuous sample path and assume that

$$\log \mathbb{P}\left(\sup_{0\le t\le 1} |Y(t)| \le \varepsilon\right) \succeq -\varepsilon^{-\alpha}\left(\log \frac{1}{\varepsilon}\right)^{\beta}$$

for $\alpha > 0$. If

$$X(t) = \int_0^1 K(t,s)Y(s)ds \tag{0.1}$$

with the kernel $K(t,s)$ satisfying the Hölder condition

$$\int_0^1 \left|K(t,s) - K(t',s)\right| ds \le c\left|t - t'\right|^{\lambda}, \quad t, t' \in [0,1], \tag{0.2}$$

for some $\lambda \in (0,1]$ and some $c > 0$, then

$$\log \mathbb{P}\left(\sup_{0\le t\le 1} |X(t)| \le \varepsilon\right) \succeq -\varepsilon^{-\alpha/(\alpha\lambda+1)}\left(\log \frac{1}{\varepsilon}\right)^{\beta/(\alpha\lambda+1)} .$$

- Note that the integrated kernel $K(t,s) = 1_{(0,t)}(s)$ satisfies the Hölder condition with $\lambda = 1$. So if $Y(t)$ is a fractional Brownian motion and $X(t)$ is the integrated fractional Brownian motion, then the lower bound given above is sharp.
- Other significant applications are for integrated Brownian sheets.