Interplay Between Metric Entropy, Bracketing Entropy and Small Ball Probability

Fuchang Gao

University of Idaho

NSF/CBMS Conference, June 4-8, 2012 at UAH

Small Ball Probability

Let X be a random element. The small ball probability studies the asymptotic behavior of $\mathbb{P}(\|X\| < \varepsilon)$ as $\varepsilon \to 0^+$, where X is a random element. It is a typical rare event, and lies at the center of probability research.

Metric Entropy

Let K be a pre-compact compact in a space equipped with metric ρ. The metric entropy of K is defined to be the quantity $\log_2 N(\varepsilon, K, \rho)$ where $N(\varepsilon, K, \rho)$ is the minimal number of open balls of radius ε needed to cover K.

Example: If K is the set of all bounded increasing functions $f\colon [0,1] \mapsto [0,1]$, then $\log N(\varepsilon, K, \|\cdot\|_{L^p}) \asymp \varepsilon^{-1}$ for all $1 \le p < \infty$, but $\log N(\varepsilon, K, \|\cdot\|_{L^\infty}) = \infty$ for all $0 < \varepsilon < 1$.

Roughly speaking, metric entropy is a geometric quantification of the compactness of K under distance ρ. It is a term purely in analysis, approximation theory, convex geometry.

Bracketing Entropy

Bracketing entropy $\log N_{[]}(\varepsilon, \mathcal{P}, \rho)$:

$N(\varepsilon, \mathcal{P}, \rho) =$ minimum number of ε-brackets (under distance ρ) needed to cover $\mathcal{P}$,

An ε-bracket $[\underline{f}, \overline{f}]$ consists of all the functions $g \in \mathcal{P}$ such that $\underline{f} \leq g \leq \overline{f}$, where $\rho(\underline{f}, \overline{f}) \leq \varepsilon$.

Recall: Metric entropy $\log N(\varepsilon, \mathcal{P}, \rho)$:

$N(\varepsilon, \mathcal{P}, \rho) =$ minimum number of ε-balls (under distance ρ) needed to cover $\mathcal{P}$

Why Bracketing Entropy?

It is used in statistics to determined the convergence rate of non-parametric density estimation.

- Optimal Rate r_n is determined by (Le Cam (1973); Birgé (1983):)

$$nr_n^{-2} = \log N_{[]}(1/r_n, \mathcal{P}, h).$$

- MLE Achieved Rate r_n is determined by (Birgé and Massart (1993))

$$\sqrt{n}r_n^{-2} = \int_{cr_n^{-2}}^{r_n^{-1}} \log N_{[]}(\varepsilon, \mathcal{P}, h)d\varepsilon.$$

where h is the Hellinger distance h, defined by

$$h(\hat{p}_n, p_0) = \|\sqrt{\hat{p}_n} - \sqrt{p_0}\|_{L^2(Q)}.$$

(Why Hellinger? Because it does not depend on Q.

Note that if $\log N_{[]}(\varepsilon, \mathcal{P}, h) = \varepsilon^{-\alpha}$, for $\alpha < 2$, then both Optimal rate and MLE Achieved Rate $r_n = n^{1/(2+\alpha)}$.
Thus, in this case, MLE is the best (one of the best).

Relation Between the Two Entropies

Relation

$$N(\varepsilon, \mathcal{P}, \rho) \leq N_{[]}(2\varepsilon, \mathcal{P}, \rho).$$

The reverse is not necessary true, unless ρ is L^∞ distance.

Remarks

- The lower bound of metric entropy is typically more difficult than upper bound;
- A powerful method of estimating metric entropy upper bound is using Fourier series. This method can no longer be used to to estimate Bracketing entropy;
- Duality is no longer available for bracketing entropy;

- Convex hull relation for bracketing entropy is no longer available.
- Relations with other quantities are also lost.

Kuelbs-Li Connection

It was established in Kuelbs and Li (1993) and completed in Li and Linde (1999) that the behavior of $\mathbb{P}(\|X\| < \varepsilon)$ 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.

Zoom in

Theorem If $\mathcal{F}$ is the convex hull of the functions $K(\cdot, \omega)$, $\omega \in \Omega$, where $K(t, \cdot)$ are square-integrable functions on a bounded set Ω in $\mathbb{R}^d$, $d \geq 1$. Let $X(t) = \int_\Omega K(t, x) dB(x)$, $t \in T$, and $B(x)$ is the d-dimensional Brownian sheet on Ω. Then

$$\log N(\varepsilon, \mathcal{F}, \|\cdot\|_2) \geq C\varepsilon^{-\frac{2\alpha}{2+\alpha}} |\log \varepsilon|^{\frac{2\beta}{2+\alpha}};$$

for $\alpha > 0$ and $\beta \in \mathbb{R}$ if and only if

$$\log \mathbb{P}\left(\sup_{t \in T} |X(t)| < \varepsilon\right) \leq -C'\varepsilon^{-\alpha} |\log \varepsilon|^{\beta}$$

Furthermore, the relation also holds if " $\leq$ " and " $\geq$ " are reversed.

Zoom In—continue

$$\log N(\varepsilon, \mathcal{F}, \|\cdot\|_2) \geq C|\log\varepsilon|^{\beta}(\log|\log\varepsilon|)^{\gamma}.$$

for $\beta > 0$ and $\gamma \in \mathbb{R}$, if and only if

$$\log \mathbb{P}\left(\sup_{t\in T}|X(t)| < \varepsilon\right) \leq -C'|\log\varepsilon|^{\beta}(\log|\log\varepsilon|)^{\gamma}$$

Furthermore, the relation also holds if " $\leq$ " and " $\geq$ " are reversed.

A Simple Example

To experience the power of this connection, let us consider the case where $K(t,\cdot) = 1_{[0,t]}(\cdot)$. In the case $\mathcal{F}$ is the set of non-decreasing functions f on $[0,1]$ such that $f(0) = 0$ and $0 \le f \le 1$. It is not trivial to show that $\log N(\varepsilon, \mathcal{F}, \|\cdot\|_2) \asymp \varepsilon^{-1}$. However, the corresponding Gaussian process $X(t) = \int_0^1 1_{[0,t]}(s) dB(s) = B(t)$ is just the Brownian motion, and it is known that

$$\log \mathbb{P}(\sup_{t\in[0,1]} |B(t)| < \varepsilon) \asymp -\varepsilon^{-2}.$$

Applying the aforemention connection, we immediately obtain

$$\log N(\varepsilon, \mathcal{F}, \|\cdot\|_2) \asymp \varepsilon^{-1}.$$

A Challenging Example

For $t = (t_1, t_2, ..., t_d)$. If $K(t, \cdot) = 1_{[0,t_1]} \otimes 1_{[0,t_2]} \otimes \cdots \otimes 1_{[0,t_d]}$, then $\mathcal{F}$ is class of probability distributions on $[0,1]^d$, while $B(t)$ is the d-dimensional Brownian sheet. We have

$$\log N(\varepsilon, \mathcal{F}_d, \|\cdot\|_2) \asymp \varepsilon^{-1} |\log \varepsilon|^{\beta}$$

if and only if

$$\mathbb{P}(\sup_{t \in [0,1]^d} |B(t)| < \varepsilon) \asymp -\varepsilon^{-2} |\log \varepsilon|^{2\beta}.$$

Through the result on the small ball probability of Brownian sheets, we have

$$C_1 \varepsilon^{-1} |\log \varepsilon|^{d-1+\delta} \leq \log N(\varepsilon, \mathcal{F}_d, \|\cdot\|_2) \leq C_2 \varepsilon^{-1} |\log \varepsilon|^{d-1/2},$$

for some $\delta > 0$. (Talagrand 1994, Dunker et al 1999; Blei et al. 2007; Bylik and Lacey 2008). The exact rate is unknown and the problem is related to discrepancy and irregularity of distributions.

A more representative example

Let $\mathcal{F}$ be the class of bounded completely monotone function on $[0, \infty)$. That is, $(-1)^k f^{(k)} \geq 0$ for all $k \geq 0$. By Bernstein's Theorem, it is the Laplace transform of a bounded measure. Thus, $\mathcal{F}$ is the closed convex hull of $K(t, \cdot)$, where $K(t, s) = e^{-ts}$. The corresponding Gaussian process $X(t)$ on $(0, \infty)$ has covariance structure: $\mathbb{E}X(t)X(s) = \frac{1-e^{-t-s}}{t+s}$.

Note: It is the (unscaled) limit of m-times integrated Brownian motion as $m \to \infty$. The scaled limit of m-times integrated Brownian motion has covariance $\mathbb{E}X(t)X(s) = \frac{2st}{t+s}$.–which is related to real zeros of random polynomials—Dembo, Poonen, Shao and Zeitouni (2000)

Metric Entropy Upper Bound

The idea

Because the functions in $\mathcal{F}$ is bounded by 1, and non-negative, for every $f, g \in \mathcal{F}$

$$\int_0^{\varepsilon^2} |f(s) - g(s)|^2 ds \leq \varepsilon^2.$$

So, we only need to consider the interval $[\varepsilon^2, 1]$.

For $s \geq \varepsilon^2$, Because

$$\int_T^{\infty} e^{-st} d\mu(t) \leq e^{-T\varepsilon^2} \leq \varepsilon$$

for T large enough, we only need to consider $\int_0^T e^{-st} d\mu(t)$.

Write the Taylor series of $f \in \mathcal{F}$

$$f(s) = \sum_{k=0}^{\infty} \int_0^{\infty} (-1)^k \frac{(st)^k}{k!} d\mu(t)$$

Because it converges fast, we only need to consider the partial sum

$$\sum_{k=0}^{N} \int_0^{\infty} (-1)^k \frac{t^k}{k!} d\mu(t)\, s^k$$

For polynomials of fixed degree, we can construct an ε-net by hand. In order to do this correctly, we need to work on the interval $[2^k\varepsilon^2, 2^{k+1}\varepsilon^2]$, for each k.

It is merely a problem of counting. On each interval, we pick up $\exp(C|\log\varepsilon|^2)$.

There are roughly $|\log\varepsilon|$ such intervals, which gives us the final estimate $\exp(C|\log\varepsilon|^3)$.

How about Metric Entropy Lower Bound?

Metric entropy lower bound is typically difficult. However, we can estimate the upper bound of the small ball probability for the corresponding Gaussian process instead.

Upper Bound of Small Ball Probability

$$\begin{aligned}
\mathbb{P}\left(\sup_{t\geq 0}|X(t)|<\varepsilon\right) &\leq \mathbb{P}(\max_{1\leq i\leq n}|X(\delta_i)|<\varepsilon)\\
= \quad (2\pi)^{-n/2}(\det\Sigma)^{-1/2}&\int_{\max_{1\leq i\leq n}|y_i|\leq\varepsilon}\exp\left(-\langle y,\Sigma^{-1}y\rangle\right)d\\
\leq \quad (2\pi)^{-n/2}(\det\Sigma)^{-1/2}&(2\varepsilon)^n\\
= \quad (C\varepsilon)^n(\det\Sigma)^{-1/2}&.
\end{aligned}$$

where the covariance matrix

$$\Sigma=(\mathbb{E}X(\delta_i)X(\delta_j))_{1\leq i,j\leq n}=\left(\frac{1-e^{-\delta_i-\delta_j}}{\delta_i+\delta_j}\right)_{1\leq i,j\leq n}.$$

A Direct Calculation

Choose $\{\delta_i\} = \{\delta, 2\delta, ..., n\delta\}$, where $\delta = \sqrt{1/n}$. By direct calculation, we have

$$\det\left(\frac{1-e^{-i\delta-j\delta}}{i\delta+j\delta}\right)_{1\le i,j\le n} = D_n\delta^{-n}(1-e^{-\delta})^{n^2}\cdot\sum_{j=0}^{n}\binom{n}{j}^2 e^{-j\delta}$$

where

$$D_n = \det\left(\frac{1}{i+j}\right)_{1\le i,j\le n} = \frac{(1!2!\cdots(n-1)!)^3 n!}{(n+1)!\cdots(2n)!}.$$

For the optimal choice of n, we obtain

$$\varepsilon^n(\det\Sigma)^{-1/2} \asymp \exp(-C|\log\varepsilon|^2).$$

What can be improved?

The above calculation only gives us

$$\log \mathbb{P}\left(\sup_{t>0} |X(t)| < \varepsilon\right) \leq -C'|\log \varepsilon|^2.$$

not $|\log \varepsilon|^3$. What can be improved? Recall

$$\mathbb{P}\left(\sup_{t\geq 0} |Y(t)| < \varepsilon\right) \leq \mathbb{P}(\max_{1\leq i\leq n} |Y(\delta_i)| < \varepsilon)$$

$$\begin{aligned} &= (2\pi)^{-n/2}(\det \Sigma)^{-1/2} \int_{\max_{1\leq i\leq n}|y_i|\leq \varepsilon} \exp\left(-\langle y, \Sigma^{-1} y\rangle\right) d \\ &\leq (2\pi)^{-n/2}(\det \Sigma)^{-1/2}(2\varepsilon)^n \\ &= (C\varepsilon)^n (\det \Sigma)^{-1/2} \asymp \exp(-C|\log \varepsilon|^2). \end{aligned}$$

Optimal Choice of δ_i

It turns out that the first inequality needs to be improved, not the second one! We choose $\{\delta_i\}_{i=1}^n$ so that

$$\delta_{mp+q} = 4^{p+m}(m+q), \quad 0 \le p < m, 1 \le q \le m$$

for $n = m^2$. With such a choice of δ_i, we are unable to evaluate the determinant exactly. But a careful estimate gives us

$$\log \mathbb{P}\left(\sup_{t>0} |X(t)| < \varepsilon\right) \le -C''|\log \varepsilon|^3.$$

for the optimal choice of n.

Remark

To prove the lower bound of the small ball probability is more difficult. However, since it already follows from the upper bound of metric entropy, we no longer need to worry about it.
This example illustrate a typically use of the close connection between small ball probability and metric entropy in Hilbert space.

Results

Theorem (Gao, Li and Wellner 2010) Let $X(t)$, $t > 0$, be a Gaussian process with covariance $\mathbb{E}X(t)X(s) = (1 - e^{-t-s})/(t+s)$, then for $0 < \varepsilon < 1$

$$\log \mathbb{P}\left(\sup_{t>0} |X(t)| < \varepsilon\right) \asymp -|\log \varepsilon|^3.$$

Corollary: Let $\mathcal{F}$ be the class of completely monotone functions on $[0,1]$, then

$$\log N(\varepsilon, \mathcal{F}, \|\cdot\|_2) \asymp |\log \varepsilon|^3.$$

Zoom Out

The so-called convex hull problem asks the optimal estimate for $N(\varepsilon, \mathrm{conv}(K), \|\cdot\|)$ given the rate of $N(\varepsilon, K, \|\cdot\|)$. When $\|\cdot\|$ is the Hilbert space norm, by using the idea of Kuelbs-Li, together with duality of metric entropy and Khatri-Sidak inequality, etc, we can prove:

Theorem (Gao 2004, Corollary 2.1)

$$\log N(\varepsilon, \mathrm{cov}(K), \|\cdot\|_2) \leq C \inf_{\delta} \left(\frac{I^2(\delta)}{\varepsilon^2} + N(\delta, K, \|\cdot\|_2) \right),$$

where $I(x) := \int_0^x \sqrt{\log N(\varepsilon, K, \|\cdot\|_2)} dt$, provided that $N(\varepsilon/2, K, \|\cdot\|) \geq C N(\varepsilon, K, \|\cdot\|_2)$ for some $C > 1$.

Some Corollaries

- If $\log N(\varepsilon, T, \|\cdot\|_2) \leq C\varepsilon^{-\alpha}\log^{\beta}(1/\varepsilon)$ for some $0 < \alpha < 2$, and $\beta \in \mathbb{R}$, then

  $$\log N(\varepsilon, \mathrm{conv}(T), \|\cdot\|_2) \leq K\varepsilon^{-2}(\log(1/\varepsilon))^{1-2/\alpha}(\log\log(1$$

- By choosing $\delta = |\log\varepsilon|^{-1/3}$ we immediately obtain that if $\log N(\varepsilon, K, \|\cdot\|_2) = O(\varepsilon^{-2}|\log\varepsilon|^{-\beta})$ for some $\beta > 2$, then

  $$\log N(\varepsilon, \mathrm{conv}(K), \|\cdot\|_2) = O(\varepsilon^{-2}(\log|\log\varepsilon|)^{2-\beta}).$$

Remarks

- This result, that is, if $\log N(\varepsilon, K, \|\cdot\|_2) = O(\varepsilon^{-2}|\log\varepsilon|^{-\beta})$ for some $\beta > 2$, then

  $$\log N(\varepsilon, \mathrm{conv}(K), \|\cdot\|_2) = O(\varepsilon^{-2}(\log|\log\varepsilon|)^{2-\beta})$$

  was proved by Carl et al, 2012 and by Kley 2012 independently.

- These estimate are best possible. (Gao 2001, 2004, 2012)

Critical Case

When $\log N(\varepsilon, T, \|\cdot\|_2) = \varepsilon^{-2}|\log\varepsilon|^{-2}$, $I(x) := \int_0^x \sqrt{\log N(\varepsilon, K, \|\cdot\|_2)}dt = \infty$, the corresponding Gaussian process is no longer bounded. Consequently,

$$\log N(\varepsilon, \mathrm{cov}(T), \|\cdot\|_2) \leq C \inf_{\eta} \left(\frac{\eta^2}{\varepsilon^2} + N(I^{-1}(\eta), T, \|\cdot\|_2) \right)$$

no longer holds. Gao 2012 proved that the following CKP inequality of Carl et al 1999 is sharp in this case:

$$\sqrt{\log N(2\varepsilon, \mathrm{cov}(T), \|\cdot\|_2)} \leq \frac{C}{\varepsilon} \int_{\varepsilon/2}^{\infty} \sqrt{\log N(r, T, \|\cdot\|_2)}dr.$$

Remark Lifshits provided a simple proof of CKP inequality using Gaussian techniques, which suggests that even

Gaussian techniques remains powerful when the Dudley integral diverges.

Connection Between Metric Entropy and Bracketing Entropy

Relation

$$N(\varepsilon, \mathcal{P}, \rho) \leq N_{[]}(2\varepsilon, \mathcal{P}, \rho).$$

The reverse is not necessary true, unless ρ is L^∞ distance. However, when the functions are smooth, we have

Theorem Let $\mathcal{F}$ be a class of functions on $[0,1]$, and $\mathcal{G}$ be the class of function on $[0,1]$ defined by $\mathcal{G} = \{\int_0^x f(t)dt : f \in \mathcal{F}\}$. If $\log N(\varepsilon, \mathcal{F}, \|\cdot\|_1) \leq \phi(\varepsilon)$, then for any probability measure Q on $[0,1]$

$$\log N_{[]}\left(\frac{\varepsilon}{\phi(\varepsilon)}, \mathcal{G}, \|\cdot\|_{p,Q}\right) \leq C\phi(\varepsilon).$$

Applications and Remarks

- By repeatedly using this theorem, we have

  Theorem If $\mathcal{M}_m$ is the class of bounded m-monotone functions, i.e. $(-1)^k f^{(k)}(x) \geq 0$ for $x > 0$ and $0 \leq k \leq m$, $m > 1$, then

$$\log N_{[]}(\varepsilon, \mathcal{M}_m, \|\cdot\|_2) = \varepsilon^{1/m}.$$

- High dimensional generalization and fractional integral generalization are also available

- Not covered by the theorem: the class of high dimensional distribution function.

Bracketing entropy of high dimensional distributions

In comparison with

$$\log N(\varepsilon, \mathcal{F}_d, \|\cdot\|_2) \leq C_2 \varepsilon^{-1} |\log \varepsilon|^{d-1/2},$$

we have

Theorem

$$\log N_{[]}(\varepsilon, \mathcal{F}_d, \|\cdot\|_p) \leq C_2 \varepsilon^{-1} |\log \varepsilon|^{2d-2},$$

for all $1 \leq p < \infty$ and all $d > 1$.

Remark While I will working on the possible improvement, I believe in the case $d = 2$, the correct rate is $\varepsilon^{-1} |\log \varepsilon|^2$. In other words, there is a different discrepancy for bracketing entropy.

Open Questions

- Any connection between $\log N(\varepsilon, \mathcal{F}, \|\cdot\|_p)$ and some small ball probability?
- Any direct connection between $\log N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|_2)$ and some small ball probability?
- Any connection between $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ and $N_{[]}(\varepsilon, \mathrm{conv}(\mathcal{F}), \|\cdot\|)$
- Any duality theory on $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$?

References

-

Thank You!