18.175 Theory of Probability
Fall 2008

Section 24

Empirical process and Kolmogorov's chaining.

Empirical process and the Kolmogorov-Smirnov test. In this sections we show how the Brownian bridge B_t arises in another central limit theorem on the space of continuous functions on $[0,1]$. Let us start with a motivating example from statistics. Suppose that $x_1, \ldots, x_n$ are i.i.d. uniform random variables on $[0,1]$. By the law of large numbers, for any $t \in [0,1]$, the *empirical* c.d.f. $n^{-1}\sum_{i=1}^n \mathrm{I}(x_i \le t)$ converges to the true c.d.f. $\mathbb{P}(x_1 \le t) = t$ almost surely and, moreover, by the CLT,

$$X_t^n = \sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^n \mathrm{I}(x_i \le t) - t\Big) \to \mathcal{N}(0, t(1-t)).$$

The stochastic process X_t^n is called the *empirical process.* The covariance of this process,

$$\mathbb{E}X_t^n X_s^n = \mathbb{E}(\mathrm{I}(x_1 \le t) - t)(\mathrm{I}(x_1 \le s) - s) = s - ts - ts + ts = s(1-t),$$

is the same as the covariance of the Brownian bridge and, by the multivariate CLT, finite dimensional distributions of the empirical process converge to f.d. distributions of the Brownian bridge,

$$\mathcal{L}\Big((X_t^n)_{t\in F}\Big) \to \mathcal{L}\Big((B_t)_{t\in F}\Big). \tag{24.0.1}$$

However, we would like to show the convergence of X_t^n to B_t in some stronger sense that would imply weak convergence of continuous functions of the process on the space $(C[0,1], \|\cdot\|_\infty)$.

The Kolmogorov-Smirnov test in statistics provides one possible motivation. Suppose that i.i.d. $(X_i)_{i\ge1}$ have continuous distribution with c.d.f. $F(t) = \mathbb{P}(X_1 \le t)$. Let $F_n(t) = n^{-1}\sum_{i=1}^n \mathrm{I}(X_i \le t)$ be the empirical c.d.f. It is easy to see the equality in distribution

$$\sup_{t\in\mathbb{R}} \sqrt{n}|F_n(t) - F(t)| \overset{d}{=} \sup_{t\in[0,1]} |X_t^n|$$

because $F(X_i)$ have uniform distribution on $[0,1]$. In order to test whether $(X_i)_{i\ge1}$ come from the distribution with c.d.f. F, the statisticians need to know the distribution of the above supremum or, as approximation, the distribution of its limit. Equation (24.0.1) suggests that

$$\mathcal{L}(\sup_t |X_t^n|) \to \mathcal{L}(\sup_t |B_t|). \tag{24.0.2}$$

Since B_t is sample continuous, its distribution is the law on the metric space $(C[0,1], \|\cdot\|_\infty)$. Even though X_t^n is not continuous, its jumps are of order $n^{-1/2}$ so it has a "close" continuous version Y_t^n. Since $\|\cdot\|_\infty$ is a continuous functional on $C[0,1]$, (24.0.2) would hold if we can prove weak convergence $\mathcal{L}(Y_t^n) \to \mathcal{L}(B_t)$ on the space $(C[0,1], \|\cdot\|_\infty)$. Lemma 36 in Section 18 shows that we only need to prove uniform tightness of $\mathcal{L}(Y_t^n)$

because, by Lemma 45, (24.0.1) already identifies the law of the Brownian motion as the unique possible limit. Thus, we need to address the question of uniform tightness of $(\mathcal{L}(X_t^n))$ on the complete separable space $(C[0,1], ||\cdot||_\infty)$ or equivalently, by the result of the previous section, the equicontinuity of X_t^n,

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\mathbb{P}\Big(m(X^n,\delta)>\varepsilon\Big)=0.$$

By Chebyshev's inequality,

$$\mathbb{P}\Big(m(X^n,\delta)>\varepsilon\Big)\le\frac{1}{\varepsilon}\mathbb{E}m(X^n,\delta)$$

and we need to learn how to control $\mathbb{E}m(X^n,\delta)$. The modulus of continuity of X^n can be written as

$$\begin{aligned}m(X^n,\delta)=\sup_{|t-s|\le\delta}|X_t^n-X_s^n| &= \sqrt{n}\sup_{|t-s|\le\delta}\Big|\frac{1}{n}\sum_{i=1}^n \mathrm{I}(s<x_i\le t)-(t-s)\Big| \\ &= \sqrt{n}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n f(x_i)-\mathbb{E}f\Big|, \qquad (24.0.3)\end{aligned}$$

where we introduced the class of functions

$$\mathcal{F}=\Big\{f(x)=\mathrm{I}(s<x\le t):\,|t-s|<\delta\Big\}. \qquad (24.0.4)$$

We will develop one approach to control the expectation of (24.0.3) for general classes of functions $\mathcal{F}$ and we will only use the specific definition (24.0.4) at the very end. This will be done in several steps.

Symmetrization. At the first step, we will replace the empirical process (24.0.3) by a symmetrized version, called Rademacher process, that will be easier to control. Let $x_1',\ldots,x_n'$ be independent copies of $x_1,\ldots,x_n$ and let $\varepsilon_1,\ldots,\varepsilon_n$ be i.i.d. *Rademacher* random variables, such that $\mathbb{P}(\varepsilon_i=1)=\mathbb{P}(\varepsilon_i=-1)=1/2$. Let us define

$$\mathbb{P}_n f=\frac{1}{n}\sum_{i=1}^n f(x_i)\quad\text{and}\quad\mathbb{P}_n' f=\frac{1}{n}\sum_{i=1}^n f(x_i').$$

Notice that $\mathbb{E}\mathbb{P}_n' f=\mathbb{E}f$. Consider the random variables

$$Z=\sup_{f\in\mathcal{F}}\big|\mathbb{P}_n f-\mathbb{E}f\big|\quad\text{and}\quad R=\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n\varepsilon_i f(x_i)\Big|.$$

Then, using Jensen's inequality and then triangle inequality, we can write

$$\begin{aligned}\mathbb{E}Z &= \mathbb{E}\sup_{f\in\mathcal{F}}\Big|\mathbb{P}_n f-\mathbb{E}f\Big|=\mathbb{E}\sup_{f\in\mathcal{F}}\Big|\mathbb{P}_n f-\mathbb{E}\mathbb{P}_n' f\Big| \\ &\le \mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n(f(x_i)-f(x_i'))\Big|=\mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n\varepsilon_i(f(x_i)-f(x_i'))\Big| \\ &\le \mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n\varepsilon_i f(x_i)\Big|+\mathbb{E}\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^n\varepsilon_i f(x_i')\Big|=2\mathbb{E}R.\end{aligned}$$

Equality in the second line holds because switching $x_i\leftrightarrow x_i'$ arbitrarily does not change the expectation, so the equality holds for any fixed (ε_i) and, therefore, for any random (ε_i).

□

Hoeffding's inequality. The first step to control the supremum in R is to control the sum $\sum_{i=1}^n\varepsilon_i f(x_i)$ for a fixed function f. Consider an arbitrary sequence $a_1,\ldots,a_n\in\mathbb{R}$. Then the following holds.

Theorem 57 *(Hoeffding) For $t\ge 0$,*

$$\mathbb{P}\Big(\sum_{i=1}^n\varepsilon_i a_i\ge t\Big)\le\exp\Big(-\frac{t^2}{2\sum_{i=1}^n a_i^2}\Big).$$

Proof. Given $\lambda > 0$, by Chebyshev's inequality,

$$\mathbb{P}\Big(\sum_{i=1}^n \varepsilon_i a_i \geq t\Big) \leq e^{-\lambda t}\mathbb{E}\exp\Big(\lambda\sum_{i=1}^n \varepsilon_i a_i\Big) = e^{-\lambda t}\prod_{i=1}^n \mathbb{E}\exp\big(\lambda\varepsilon_i a_i\big).$$

Using the inequality $(e^x + e^{-x})/2 \leq e^{x^2/2}$, we get

$$\mathbb{E}\exp\big(\lambda\varepsilon_i a_i\big) = \frac{e^{\lambda a_i} + e^{-\lambda a_i}}{2} \leq \exp\Big(\frac{\lambda^2 a_i^2}{2}\Big).$$

Hence,

$$\mathbb{P}\Big(\sum_{i=1}^n \varepsilon_i a_i \geq t\Big) \leq \exp\Big(-\lambda t + \frac{\lambda^2}{2}\sum_{i=1}^n a_i^2\Big)$$

and minimizing over $\lambda > 0$ gives the result.

□

Covering numbers, Kolmogorov's chaining and Dudley's entropy integral. To control $\mathbb{E}R$ for general classes of functions $\mathcal{F}$, we will need to use some measures of complexity of $\mathcal{F}$. First, we will show how to control the Rademacher process R conditionally on $x_1, \ldots, x_n$.

Definition. Suppose that (F, d) is a totally bounded metric space. For any $u > 0$, a u-packing number of F with respect to d is defined by

$$D(F, u, d) = \max\operatorname{card}\Big\{F_u \subseteq F :\ d(f,g) > u \text{ for all } f, g \in F_u\Big\}$$

and a u-covering number is defined by

$$N(D, u, d) = \min\operatorname{card}\Big\{F_u \subseteq F : \forall f \in F\ \exists\ g \in F_u \text{ such that } d(f,g) \leq u\Big\}.$$

□

Both packing and covering numbers measure how many points are needed to approximate any element in the set F within distance u. It is a simple exercise to show that

$$N(F, u, d) \leq D(F, u, d) \leq N(F, u/2, d)$$

and, in this sense, packing and covering numbers are closely related. Let F be a subset of the cube $[-1, 1]^n$ equipped with a scaled Euclidean metric

$$d(f, g) = \Big(\frac{1}{n}\sum_{i=1}^n (f_i - g_i)^2\Big)^{1/2}.$$

Consider the following Rademacher process on F,

$$\mathcal{R}(f) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f_i.$$

Then we have the following version of the classical Kolmogorov's chaining lemma.

Theorem 58 *(Kolmogorov's chaining) For any $u > 0$,*

$$\mathbb{P}\Big(\forall f \in F,\ \mathcal{R}(f) \leq 2^{9/2}\int_0^{d(0,f)} \log^{1/2} D(F, \varepsilon, d)d\varepsilon + 2^{7/2} d(0,f)\sqrt{u}\Big) \geq 1 - e^{-u}.$$

Proof. Without loss of generality, assume that $0 \in F$. Define a sequence of subsets

$$\{0\} = F_0 \subseteq F_1 \ldots \subseteq F_j \subseteq \ldots \subseteq F$$

such that F_j satisfies

1. $\forall f, g \in F_j,\ d(f,g) > 2^{-j}$,
2. $\forall f \in F$ we can find $g \in F_j$ such that $d(f,g) \le 2^{-j}$.

F_0 obviously satisfies these properties for $j = 0$. To construct F_{j+1} given F_j:

- Start with $F_{j+1} := F_j$.
- If possible, find $f \in F$ such that $d(f,g) > 2^{-(j+1)}$ for all $g \in F_{j+1}$.
- Let $F_{j+1} := F_{j+1} \cup \{f\}$ and repeat until you cannot find such f.

Define projection $\pi_j : F \to F_j$ as follows:

$$\text{for } f \in F \text{ find } g \in F_j \text{ with } d(f,g) \le 2^{-j} \text{ and set } \pi_j(f) = g.$$

Any $f \in F$ can be decomposed into the telescopic series

$$\begin{aligned} f &= \pi_0(f) + (\pi_1(f) - \pi_0(f)) + (\pi_2(f) - \pi_1(f)) + \dots \\ &= \sum_{j=1}^{\infty} (\pi_j(f) - \pi_{j-1}(f)). \end{aligned}$$

Moreover,

$$\begin{aligned} d(\pi_{j-1}(f), \pi_j(f)) &\le d(\pi_{j-1}(f), f) + d(f, \pi_j(f)) \\ &\le 2^{-(j-1)} + 2^{-j} = 3 \cdot 2^{-j} \le 2^{-j+2}. \end{aligned}$$

As a result, the jth term in the telescopic series for any $f \in F$ belongs to a finite set of possible *links*

$$L_{j-1,j} = \Big\{ f - g : f \in F_j, g \in F_{j-1}, d(f,g) \le 2^{-j+2} \Big\}.$$

Since $\mathcal{R}(f)$ is linear,

$$\mathcal{R}(f) = \sum_{j=1}^{\infty} \mathcal{R}(\pi_j(f) - \pi_{j-1}(f)).$$

We first show how to control $\mathcal{R}$ on the set of all links. Assume that $\ell \in L_{j-1,j}$. By Hoeffding's inequality,

$$\mathbb{P}\Big(\mathcal{R}(\ell) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i \ell_i \ge t \Big) \le \exp\Big(-\frac{t^2}{2n^{-1}\sum_{i=1}^{n} \ell_i^2} \Big) \le \exp\Big(-\frac{t^2}{2 \cdot 2^{-2j+4}} \Big).$$

If $|F|$ denotes the cardinality of the set F then

$$|L_{j-1,j}| \le |F_{j-1}| \cdot |F_j| \le |F_j|^2$$

and, therefore,

$$\mathbb{P}\Big(\forall \ell \in L_{j-1,j},\ \mathcal{R}(\ell) \le t \Big) \ge 1 - |F_j|^2 \exp\Big(-\frac{t^2}{2^{-2j+5}} \Big) = 1 - \frac{1}{|F_j|^2} e^{-u}$$

after making a change of variables

$$t = \Big(2^{-2j+5} (4 \log |F_j| + u) \Big)^{1/2} \le 2^{7/2} 2^{-j} \log^{1/2} |F_j| + 2^{5/2} 2^{-j} \sqrt{u}.$$

Hence,

$$\mathbb{P}\Big(\forall \ell \in L_{j-1,j},\ \ \mathcal{R}(\ell) \le 2^{7/2} 2^{-j} \log^{1/2} |F_j| + 2^{5/2} 2^{-j} \sqrt{u}\ \Big) \ge 1 - \frac{1}{|F_j|^2} e^{-u}.$$

If $F_{j-1} = F_j$ then we can define $\pi_{j-1}(f) = \pi_j(f)$ and, since in this case $L_{j-1,j} = \{0\}$, there is no need to control these links. Therefore, we can assume that $|F_{j-1}| < |F_j|$ and taking a union bound for all steps,

$$\mathbb{P}\Big(\forall j \geq 1\ \forall \ell \in L_{j-1,j},\quad \mathcal{R}(\ell) \leq 2^{7/2}2^{-j}\log^{1/2}|F_j| + 2^{5/2}2^{-j}\sqrt{u}\ \Big)$$
$$\geq 1 - \sum_{j=1}^{\infty}\frac{1}{|F_j|^2}e^{-u} \geq 1 - \sum_{j=1}^{\infty}\frac{1}{(j+1)^2}e^{-u} = 1 - (\pi^2/6 - 1)e^{-u} \geq 1 - e^{-u}.$$

Given $f \in F$, let integer k be such that $2^{-(k+1)} < d(0, f) \leq 2^{-k}$. Then in the above construction we can assume that $\pi_0(f) = \ldots = \pi_k(f) = 0$, i.e. we will project f on 0 if possible. Then with probability at least $1 - e^{-u}$,

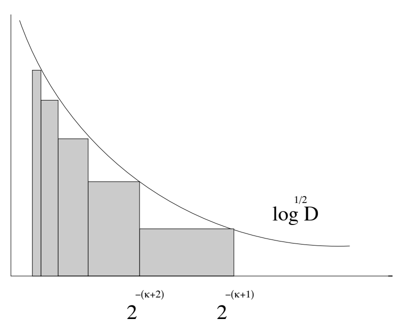
$$\begin{aligned}
\mathcal{R}(f) &= \sum_{j=k+1}^{\infty}\mathcal{R}(\pi_j(f) - \pi_{j-1}(f)) \\
&\leq \sum_{j=k+1}^{\infty}\Big(2^{7/2}2^{-j}\log^{1/2}|F_j| + 2^{5/2}2^{-j}\sqrt{u}\Big) \\
&\leq 2^{7/2}\sum_{j=k+1}^{\infty}2^{-j}\log^{1/2}D(F, 2^{-j}, d) + 2^{5/2}2^{-k}\sqrt{u}.
\end{aligned}$$

Note that $2^{-k} < 2d(f, 0)$ and $2^{5/2}2^{-k} < 2^{7/2}d(f, 0)$. Finally, since packing numbers $D(F, \varepsilon, d)$ are decreasing in ε, we can write (see figure 24)

$$\begin{aligned}
2^{9/2}\sum_{j=k+1}^{\infty}2^{-(j+1)}\log^{1/2}D(F, 2^{-j}, d) &\leq 2^{9/2}\int_0^{2^{-(k+1)}}\log^{1/2}D(F, \varepsilon, d)d\varepsilon \\
&\leq 2^{9/2}\int_0^{d(0,f)}\log^{1/2}D(F, \varepsilon, d)d\varepsilon \qquad (24.0.5)
\end{aligned}$$

since $2^{-(k+1)} < d(0, f)$. This finishes the proof.

□



The integral in (24.0.5) is called *Dudley's entropy integral.* We would like to apply the bound of the above theorem to

$$\sqrt{n}R = \sup_{f \in \mathcal{F}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i f(x_i)\Big|$$

for a class of functions $\mathcal{F}$ in (24.0.4). Suppose that $x_1, \ldots, x_n \in [0, 1]$ are fixed and let

$$F = \Big\{(f_i)_{1\leq i\leq n} = \Big(\mathrm{I}(s < x_i \leq t)\Big)_{1\leq i\leq n} :\ |t - s| \leq \delta \text{ and } t, s \in [0, 1]\Big\} \subseteq \{0, 1\}^n.$$

Then the following holds.

Lemma 47 *$N(F,u,d) \le Ku^{-4}$ for some absolute $K>0$ independent of the points $x_1,\ldots,x_n$.*

Proof. We can assume that $x_1 \le \ldots \le x_n$. Then the class F consists of all vectors of the type

$$(0\ldots 1\ldots 1\ldots 0),$$

i.e. the coordinates equal to 1 come in blocks. Given u, let F_u be a subset of such vectors with blocks of 1's starting and ending at the coordinates $k\lfloor nu\rfloor$. Given any vector $f\in F$, let us approximate it by a vector in $f'\in F_u$ by choosing the closest starting and ending coordinates for the blocks of 1's. The number of different coordinates will be bounded by $2\lfloor nu\rfloor$ and, therefore, the distance between f and f' will be bounded by

$$d(f,f') \le \sqrt{2n^{-1}\lfloor nu\rfloor} \le \sqrt{2u}.$$

The cardinality of F_u is, obviously, of order u^{-2}. This proves that $N(F,\sqrt{2u},d)\le Ku^{-2}$. Making the change of variables $\sqrt{2u}\to u$ proves the result.

□

To apply the Kolmogorov chaining bound to this class F let us make a simple observation that if a random variable $X\ge 0$ satisfies $\mathbb{P}(X\ge a+bt)\le Ke^{-t^2}$ for all $t\ge 0$ then

$$\mathbb{E}X = \int_0^\infty \mathbb{P}(X\ge t)dt \le a + \int_0^\infty \mathbb{P}(X\ge a+t)dt \le a + K\int_0^\infty e^{-\frac{t^2}{b^2}}dt \le a+Kb \le K(a+b).$$

Theorem 58 then implies that

$$\mathbb{E}_\varepsilon \sup_F \Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f_i\Big| \le K\Big(\int_0^{D_n}\sqrt{\log\frac{K}{u}}du + D_n\Big) \tag{24.0.6}$$

where $\mathbb{E}_\varepsilon$ is the expectation with respect to (ε_i) only and

$$\begin{aligned} D_n^2 = \sup_F d(0,f)^2 &= \sup_{\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(x_i)^2 = \sup_{|t-s|\le\delta}\frac{1}{n}\sum_{i=1}^n \mathrm{I}(s<x_i\le t) \\ &= \sup_{|t-s|\le\delta}\left|\frac{1}{n}\sum_{i=1}^n \mathrm{I}(x_i\le t) - \frac{1}{n}\sum_{i=1}^n \mathrm{I}(x_i\le s)\right|. \end{aligned}$$

Since the integral on the right hand side of (24.0.6) is concave in D_n, by Jensen's inequality,

$$\mathbb{E}\sup_{\mathcal{F}}\Big|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i f(x_i)\Big| \le K\Big(\int_0^{\mathbb{E}D_n}\sqrt{\log\frac{K}{u}}du + \mathbb{E}D_n\Big).$$

By the symmetrization inequality, this finally proves that

$$\mathbb{E}m(X^n,\delta) \le K\Big(\int_0^{\mathbb{E}D_n}\sqrt{\log\frac{K}{u}}du + \mathbb{E}D_n\Big).$$

The strong law of large numbers easily implies that

$$\sup_{t\in[0,1]}\left|\frac{1}{n}\sum_{i=1}^n \mathrm{I}(x_i\le t) - t\right| \to 0 \quad \text{a.s.}$$

and, therefore, $D_n^2\to\delta$ a.s. and $\mathbb{E}D_n\to\sqrt{\delta}$. This implies that

$$\limsup_{n\to\infty}\mathbb{E}m(X_t^n,\delta) \le K\Big(\int_0^{\sqrt{\delta}}\sqrt{\log\frac{K}{u}}du + \sqrt{\delta}\Big).$$

The right-hand side goes to zero as $\delta\to 0$ and this finishes the proof of equicontinuity of X^n. As a result, for any continuous function Φ on $(C[0,1],\|\cdot\|_\infty)$ the distibution of $\Phi(X_t^n)$ converges to the distribution of $\Phi(B_t)$. For example,

$$\sqrt{n}\sup_{0\le t\le 1}\Big|\frac{1}{n}\sum \mathrm{I}(x_i\le t) - t\Big| \to \sup_{0\le t\le 1}|B_t|$$

in distribution. We will find the distribution of the right hand side in the next section.