MATH 281C: Mathematical Statistics

Lecture 11

1 Bracketing Control

Our main empirical process bound so far is the following. Under the usual notation,

$$\mathbb{E}\sup_{f\in\mathcal{F}}|P_nf - Pf| \le \frac{C}{\sqrt{n}}\|F\|_{L^2(P)}J(F,\mathcal{F}), \tag{11.1}$$

where

$$J(F,\mathcal{F}) := \int_0^1 \sqrt{1+\log\sup_Q M(\epsilon\|F\|_{L^2(Q)},\mathcal{F},L^2(Q))}\,\mathrm{d}\epsilon.$$

Bracketing methods provide another upper bound for $\mathbb{E}\sup_{f\in\mathcal{F}}|P_nf - Pf|$ which we will describe next. This bound will be very similar to (11.1) except that $\sup_Q M(\epsilon\|F\|_{L^2(Q)},\mathcal{F},L^2(Q))$ will be replaced by the ϵ-bracketing number of $\mathcal{F}$ in $L^2(P)$. Before we state this result, let us first define the notion of bracketing numbers:

1. Given two real-valued functions ℓ and u on $\mathcal{X}$, the bracket $[\ell,u]$ is defined as the collection of all functions $f:\mathcal{X}\to\mathbb{R}$ for which $\ell(x)\le f(x)\le u(x)$ for all $x\in\mathcal{X}$.
2. Given a probability measure P on $\mathcal{X}$, the $L^2(P)$-size of a bracket $[\ell,u]$ is defines as $\|u-\ell\|_{L^2(P)}$.
3. Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$. For $\epsilon>0$, the bracketing number $N_{[]}(\epsilon,\mathcal{F},L^2(P))$ is defined as the smallest number of brackets each having $L^2(P)$-size at most ϵ such that every $f\in\mathcal{F}$ belongs to one of the brackets.

It is important to notice that the bracketing numbers are larger than covering numbers as shown below.

Lemma 1.1. For every $\epsilon>0$,

$$N_{\mathcal{F}}(\epsilon,\mathcal{F},L^2(P)) \le N_{\mathcal{F}_{\text{all}}}(\epsilon/2,\mathcal{F},L^2(P)) \le N_{[]}(\epsilon,\mathcal{F},L^2(P)).$$

Here $\mathcal{F}_{\text{all}}$ denotes the class of all real-valued functions on $\mathcal{X}$.

Proof. The first inequality is something we have already seen when discussing covering numbers. The second inequality is proved as follows. First, get brackets $[\ell_i,u_i]$, $i=1,\ldots,N$, each of $L^2(P)$-size ϵ which cover $\mathcal{F}$. Then it is obvious to see that the mid-point functions $(\ell_i+u_i)/2$, $i=1,\ldots,N$ form an $\epsilon/2$-net for $\mathcal{F}$ in the $L^2(P)$ metric. □

Next, we provide an example where the bracketing numbers can be explicitly computed.

Example 1.1. Let $\mathcal{F} := \{I_{(-\infty,t]} : t \in \mathbb{R}\}$ and let P be a fixed probability measure on $\mathbb{R}$. Then

$$N_{[]}(\epsilon, \mathcal{F}, L^2(P)) \leq 1 + \frac{1}{\epsilon^2} \quad \text{for every } \epsilon > 0. \tag{11.2}$$

Here is an argument for (11.2). Let $t_0 := -\infty$, and recursively define

$$t_i := \sup\{x > t_{i-1} : P(t_{i-1}, x] \leq \epsilon\}.$$

Then, for every $\delta > 0$ sufficiently small, $P(t_{i-1}, t_i - \delta] \leq \epsilon$. By letting $\delta \to 0$, we deduce that $P(t_{i-1}, t_i) \leq \epsilon$. Also, if $t_i < \infty$, then for every $\delta > 0$, we have $P(t_{i-1}, t_i + \delta] > \epsilon$ so that (by letting $\delta \downarrow 0$), $P(t_{i-1}, t_i] \geq \epsilon$.

Let $k \geq 1$ be the smallest integer for which $t_k = \infty$. By the above, we have $P(t_{i-1}, t_i] \geq \epsilon$ for $i = 1, \ldots, k-1$ so that

$$1 = \mathbb{P}(-\infty, \infty) = \sum_{i=1}^{k} P(t_{i-1}, t_i] \geq (k-1)\epsilon,$$

which gives $k \leq 1 + \epsilon^{-1}$. Now consider the brackets $[I_{(-\infty,t_{i-1}]}, I_{(-\infty,t_i]}]$ for $i = 1, \ldots, k$. These obviously cover $\mathcal{F}$, i.e., each function in $\mathcal{F}$ belongs to one of these brackets, and their $L^2(P)$-size is

$$\sqrt{P(t_{i-1}, t_i)} \leq \sqrt{\epsilon}.$$

We have thus proved that

$$N_{[]}(\sqrt{\epsilon}, \mathcal{F}, L^2(Q)) \leq 1 + \frac{1}{\epsilon}.$$

This, being true for all $\epsilon > 0$, is the same as (11.2).

Before stating the analogue of (11.1) involving bracketing numbers, let us first state and prove a simple classical asymptotic result, which shows that bracketing number bounds can be used to control $\mathbb{E} \sup_{f \in \mathcal{F}} |P_n f - Pf|$.

Proposition 1.1. Suppose $\mathcal{F}$ is a function class such that $N_{[]}(\epsilon, \mathcal{F}, L^2(P)) < \infty$ for every $\epsilon > 0$. Then

$$\sup_{f \in \mathcal{F}} |P_n f - Pf| \xrightarrow{\mathbb{P}} 0 \quad \text{as } n \to \infty. \tag{11.3}$$

Proof. Fix $\epsilon > 0$, and let $[\ell_i, u_i], i = 1, \ldots, N$ denote brackets of $L^2(P)$-size $\leq \epsilon$ which cover $\mathcal{F}$. We first argue that

$$\sup_{f \in \mathcal{F}} |P_n f - Pf| \leq \max_{1 \leq i \leq N} \max(|P_n u_i - Pu_i|, |P_n \ell_i - P\ell_i|) + \epsilon. \tag{11.4}$$

Let us first complete the proof of (11.3) assuming that (11.4) is true. To see this, note that the RHS above converges to 0 almost surely as $n \to \infty$. This is because, by the strong LLN, $|P_n u_i - Pu_i|$ and $|P_n \ell_i - P\ell_i|$ converge to zero almost surely as $n \to \infty$ for each i (note that the functions u_i and ℓ_i do

not change with n) and hence the finite maximum of these over $i = 1, \ldots, N$ also converges to zero. Thus, from (11.4), we deduce that

$$\limsup_{n\to\infty} \sup_{f\in\mathcal{F}} |P_n f - Pf| \le \epsilon \quad \text{almost surely for every } \epsilon > 0.$$

Applying this for each $\epsilon = 1/m$ and letting $m \to \infty$ proves (11.3).

It remains to prove (11.4). Fix $f \in \mathcal{F}$ and get a bracket $[\ell_i, u_i]$ which contains f. This means that $\ell_i(x) \le f(x) \le u_i(x)$ for every $x \in \mathcal{X}$. Write

$$\begin{aligned} P_n f - Pf &\le P_n u_i - Pu_i + Pu_i - Pf \\ &\le P_n u_i - Pu_i + Pu_i - P\ell_i \\ &\le P_n u_i - Pu_i + \|u_i - \ell_i\|_{L^2(P)} \le P_n u_i - Pu_i + \epsilon. \end{aligned}$$

It can similarly be proved that $P_n f - Pf \ge P_n \ell_i - P\ell_i - \epsilon$. Both these inequalities together imply (11.4), which completes the proof of Proposition 1.1. □

We will now state the analogue of (11.1) involving bracketing numbers. This will be our second main result for bounding the expected suprema of empirical processes (the first main result being (11.1)).

Theorem 1.1. Let F be an envelop for the class $\mathcal{F}$ such that $PF^2 < \infty$. Then

$$\mathbb{E} \sup_{f\in\mathcal{F}} n^{1/2} |P_n f - Pf| \le C \|F\|_{L^2(P)} J_{[]}(F, \mathcal{F}), \tag{11.5}$$

where

$$J_{[]}(F, \mathcal{F}) := \int_0^1 \sqrt{1 + \log N_{[]}(\epsilon \|F\|_{L^2(P)}, \mathcal{F}, L^2(P))}\, d\epsilon.$$

The bound (11.5) is very similar to (11.1) with the only difference being that the "uniform" packing numbers $\sup_Q M(\epsilon, \|F\|_{L^2(Q)}, \mathcal{F}, L^2(Q))$ are replaced by the bracketing numbers $N_{[]}(\epsilon \|F\|_{L^2(P)}, \mathcal{F}, L^2(P))$ with respect to $L^2(P)$. Importantly, note that there is supremum over Q in (11.1) while the bracketing numbers only involve the measure P.

Example 1.2. Suppose $X_1, \ldots, X_n$ are iid observations having CDF F, and let F_n be the empirical CDF. We have seen previously that

$$\mathbb{E} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \le \frac{C}{\sqrt{n}}$$

for every $n \ge 1$. We will prove this via (11.5). For $\mathcal{F} = \{I_{(-\infty,t]} : t \in \mathbb{R}\}$, we have obtained bounds for $N_{[]}(\epsilon, \mathcal{F}, L^2(P))$ in (11.2). We deduce from these and (11.5) that

$$\mathbb{E} \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \le \frac{C}{\sqrt{n}} \int_0^1 \sqrt{1 + \log(1 + 1/\epsilon^2)}\, d\epsilon \le \frac{C}{\sqrt{n}}.$$

The following presents a situation where bounding the bracketing numbers is much more tractable compared to bounding the uniform covering numbers.

Proposition 1.2. Let $\Theta \subseteq \mathbb{R}^d$ be contained in a ball of radius R. Let $\mathcal{F} = \{m_\theta : \theta \in \mathbb{R}$ be a function class indexed by Θ. Suppose there exists a function M with $\|M\|_{L^2(P)} < \infty$ such that

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq M(x)\|\theta_1 - \theta_2\| \tag{11.6}$$

for all $x \in \mathcal{X}$ and $\theta_1, \theta_2 \in \Theta$. Then, for every $\epsilon > 0$,

$$N_{[]}(\epsilon\|M\|_{L^2(P)}, \mathcal{F}, L^2(P)) \leq \left(1 + \frac{4R}{\epsilon}\right)^d. \tag{11.7}$$

Proof. Let $\theta_1, \ldots, \theta_N$ be a maximal $\epsilon/2$-packing set of Θ in the Euclidean metric. Consider the brackets $[m_{\theta_i} - \epsilon M/2, m_{\theta_i} + \epsilon M/2]$ for $i = 1, \ldots, N$. Note that

1. These brackets cover $\mathcal{F}$. For every $\theta \in \Theta$, there exists $1 \leq i \leq N$ such that $\|\theta - \theta_i\| \leq \epsilon/2$. By condition 11.6,

$$|m_\theta(x) - m_{\theta_i}(x)| \leq M(x)\|\theta - \theta_i\| \leq \frac{\epsilon M(x)}{2},$$

   which implies that m_θ lies in the bracket $[m_{\theta_i} - \epsilon M/2, m_{\theta_i} + \epsilon M/2]$.

2. The $L^2(P)$-size of these brackets is at most $\epsilon\|M\|_{L^2(P)}$.

Because of these two observations, $N_{[]}(\epsilon\|M\|_{L^2(P)}, \mathcal{F}, L^2(P))$ is bounded from above by the $\epsilon/2$-packing number of Θ which we bounded previously. The completes the proof of Proposition 1.2. □

2 M-estimation

We now come back to the first statistics topic of the course: M-estimation. The basic abstract setting is the following.

Let Θ be an abstract parameter space. Usually, it is a subset of $\mathbb{R}^d$ for parametric estimation problems or it is a function class for nonparametric estimation problems. We have two processes (one stochastic and one deterministic) that are indexed by $\theta \in \Theta$. The stochastic process will usually depend on a "sample" size n and will be denoted by $M_n(\theta), \theta \in \Theta$. The deterministic process will usually not depend on n and will simply be denoted by $M(\theta), \theta \in \Theta$. We expect M_n to be close to M for large n.

Let $\widehat{\theta_n}$ denote a maximizer of $M_n(\theta)$ over $\theta \in \Theta$, and let θ_0 be a maximizer of $M(\theta)$ over $\theta \in \Theta$. The goal in M-estimation is to study the behavior of $\widehat{\theta_n}$ in relation to θ_0.

Some concrete M-estimators are described below.

1. (Classical parametric estimation): The most classical M-estimator is the maximum likelihood estimator (MLE). Here one typically has data $X_1, \ldots, X_n$ in $\mathcal{X}$ that are iid having distribution P. One also has a class $\{p_\theta : \theta \in \Theta\}$ of densities over the space. The MLE minimizes $M_n(\theta) := -P_n \log p_\theta$ over $\theta \in \Theta$. The process $M(\theta)$ here is $M(\theta) := -P \log p_\theta$ and θ_0 can then be taken to be the parameter value in Θ for which p_θ is closest to P in terms of the Kullback-Leibler divergence.

2. (Least squares estimation in regression): In regression problems, one observes data $(X_1, Y_1), \ldots, (X_n, Y_n)$ with $X_i \in \mathcal{X}$ and $Y_i \in \mathbb{R}$, which can be modeled as iid observations having some (joint) distribution P. Let Θ be a class of functions from $\mathcal{X}$ to $\mathbb{R}$. The least squares estimator over the class Θ corresponds to minimizer of

$$M_n(\theta) = P_n\{y - \theta(x)\}^2$$

over $\theta \in \Theta$. It is natural to compare this $\widehat{\theta_n}$ to θ_0 which is the minimizer of

$$M(\theta) = P\{y - \theta(x)\}^2.$$

3. (Empirical risk minimization procedures in classification): Here one observes data $(X_1, Y_1), \ldots, (X_n, Y_n)$, where $X_i \in \mathcal{X}$ and $Y_i \in \{-1, 1\}$. We model the data as iid having a distribution P. Let Θ denote a class of functions from $\mathcal{X}$ to $\mathbb{R}$; we are thinking of the sign of $\theta(x)$ as the output of the classifier. It is natural to consider

$$M_n(\theta) := P_n I\{y \neq \text{sign}(\theta(x))\} \quad \text{and} \quad M(\theta) := \mathbb{P}PI\{y \neq \text{sign}(\theta(x))\}.$$

In this case, $\widehat{\theta_n}$ will be the empirical minimizer of the misclassification rate and θ_0 will be the minimizer of the test error, both in the class Θ. It is therefore natural to compare the performance of $\widehat{\theta_n}$ to that of θ_0.

Note that it is difficult to compute $\widehat{\theta_n}$ because the minimization of $M_n(\theta)$ is a combinatorial problem. For this, one also studies other choices of $M_n(\theta)$ in classification. To motivate these other choices, let us first rewrite the above $M_n(\theta)$ as

$$M_n(\theta) = P_n I\{y \neq \text{sign}(\theta(x))\} = P_n \phi_0(-y\theta(x)), \quad \text{where } \phi_0(t) := I(t \geq 0).$$

For computational considerations, one often replaces ϕ_0 by another loss function that is convex and continuous but similar to ϕ_0. Common choices of ϕ include (a) Hinge loss: $\phi(t) := (1 + t)_+$, (b) Exponential loss: $\phi(t) := e^t$, and (c) Logistic loss: $\phi(t) := \log(1 + e^t)$. Note that these three functions are convex on $\mathbb{R}$, and they are similar to ϕ_0 (they also satisfy $\phi(t) \geq \phi_0(t)$ for all t). We will study procedures $\widehat{\theta_n}$ which minimize

$$M_n(\theta) := P_n \phi(-y\theta(x)) \quad \text{over } \theta \in \Theta,$$

and compare their performance to θ_0.

The theory of M-estimation concerns itself usually with three questions: (a) Consistency, (b) Rate of Convergence, and (c) Limiting Behavior. Consistency asserts that the discrepancy between $\widehat{\theta_n}$ and θ_0 converges to zero as $n \to \infty$. Rate of convergence aims to characterize the precise rate of this convergence. The goal of the third question will be to give a precise characterization of the limiting distribution of the discrepancy in the asymptotic setting where $n \to \infty$.

Consistency usually always holds and we have already seen a theorem last week on consistency. We will mainly concentrate on the problem of rates of convergence. In many cases, a rate of convergence result automatically implies consistency. In other cases, one needs a preliminary consistency result so that attention can be focused in a local neighbourhood of θ_0 in order to determine the rate of convergence. In cases where preliminary consistency is required and our consistency theorem last week is not sufficient, we will provide a different argument for consistency. Let us ignore consistency for the time being and proceed directly to the rates. For studying limiting behavior, we need theory on uniform central limit theorems which we are yet to cover.

3 Rages of Convergence of M-estimators

Again, we work in the abstract setting where $\widehat{\theta}_n$ minimizes a stochastic process $M_n(\theta)$ over $\theta \in \Theta$ and θ_0 minimizes a deterministic process $M(\theta)$ over $\theta \in \Theta$. The argument for deriving rates starts from the following basic inequality:

$$M(\widehat{\theta}_n) - M(\theta_0) \leq M(\widehat{\theta}_n) - M_n(\widehat{\theta}_n) - \{M(\theta_0) - M_n(\theta_0)\}.$$

We have already seen this inequality multiple times and it is a consequence of the simple inequality $M_n(\widehat{\theta}_n) \leq M_n(\theta_0)$. For convenience, we denote the RHS above by $(M - M_n)(\widehat{\theta}_n - \theta_0)$, so that

$$M(\widehat{\theta}_n) - M(\theta_0) \leq (M_n - M)(\widehat{\theta}_n - \theta_0). \tag{11.8}$$

We will use this inequality to study rates of convergence of $\widehat{\theta}_n$ to θ_0. We need to first fix a measure of discrepancy between $\widehat{\theta}_n$ and θ_0. Let this be given by $d(\widehat{\theta}_n, \theta_0)$. In cases where Θ is a subset of $\mathbb{R}^d$, it is natural to take $d(\cdot, \cdot)$ as the usual Euclidean metric.

Note that the discrepancy measure $d(\cdot, \cdot)$ is somewhat external to the problem and, therefore, to understand the behavior of $d(\widehat{\theta}_n, \theta_0)$, we need to connect it to $M(\theta)$ or $M_n(\theta)$. The usual assumption for this is to assume that

$$M(\theta) - M(\theta_0) \gtrsim d^2(\theta, \theta_0). \tag{11.9}$$

Here the notation $a \gtrsim b$ means that $a \geq Cb$ for a universal constant C (the notation $a \lesssim b$ is defined analogously).

Let us assume that (11.9) is true for all $\theta \in \Theta$. In some situations, it is only true in a neighborhood of θ_0 (we will come back to this later). Combining (11.8) and (11.9), we obtain

$$d^2(\widehat{\theta}_n, \theta_0) \lesssim (M_n - M)(\widehat{\theta}_n - \theta_0).$$

Let $\widehat{\delta}_n := d(\widehat{\theta}_n, \theta_0)$. Then the above inequality simply implies

$$\widehat{\delta}_n^2 \lesssim \sup_{\theta \in \Theta : d(\theta, \theta_0) \leq \widehat{\delta}_n} (M_n - M)(\theta - \theta_0).$$

We will now rigorously find upper bounds for the rate of convergence of $d(\widehat{\theta}_n, \theta_0)$. Formally, we say that δ_n is a rate of convergence of $d(\widehat{\theta}_n, \theta_0)$ to zero if for every $\epsilon > 0$, there exists a constant C_ϵ such that

$$d(\widehat{\theta}_n, \theta_0) \leq C_\epsilon \delta_n \ \text{ with probability } \geq 1 - \epsilon. \tag{11.10}$$

Note that this is equivalent to

$$\mathbb{P}\{d(\widehat{\theta}_n, \theta_0) > 2^M \delta_n\} \to 0 \ \text{ as } M \to \infty. \tag{11.11}$$

It should be noted that (11.10) and (11.11) are nonasymptotic statements (they hold for each finite n). Then imply, in particular, the asymptotic rate statement: $d(\widehat{\theta}_n, \theta_0) = O_{\mathbb{P}}(\delta_n)$, which means the following: for every $\epsilon > 0$, there exists C_ϵ and an integer N_ϵ such that

$$\mathbb{P}\{d(\widehat{\theta}_n, \theta_0) \leq C_\epsilon \delta_n\} \geq 1 - \epsilon \quad \text{for all } n \geq N_\epsilon. \tag{11.12}$$

Let us now study the probability

$$\mathbb{P}\{d(\widehat{\theta}_n, \theta_0) > 2^M \delta_n\}$$

for fixed δ_n and large M. We need to understand for which δ_n does this probability become small as $M \to \infty$.

Consider the decomposition

$$\mathbb{P}\{d(\widehat{\theta}_n, \theta_0) > 2^M \delta_n\} = \sum_{j>M} \mathbb{P}\{2^{j-1}\delta_n < d(\widehat{\theta}_n, \theta_0) \le 2^j \delta_n\}.$$

Applying the basic inequality and condition (11.9), we obtain

$$d^2(\widehat{\theta}_n, \theta_0) \lesssim M(\widehat{\theta}_n) - M(\theta_0) \le (M_n - M)(\widehat{\theta}_n - \theta_0).$$

It follows that

$$\begin{aligned}
&\mathbb{P}\{2^{j-1}\delta_n < d(\widehat{\theta}_n, \theta_0) \le 2^j \delta_n\} \\
&\le \mathbb{P}\{(M_n - M)(\widehat{\theta}_n - \theta_0) \gtrsim 2^{2j-2}\delta_n^2,\ d(\widehat{\theta}_n, \theta_0) \le 2^j \delta_n\} \\
&\le \mathbb{P}\left\{ \sup_{\theta : d(\theta,\theta_0) \le 2^j \delta_n} (M_n - M)(\theta - \theta_0) \gtrsim 2^{2j-2}\delta_n^2 \right\} \\
&\lesssim \frac{1}{2^{2j-2}\delta_n^2} \mathbb{E}\left\{ \sup_{\theta : d(\theta,\theta_0) \le 2^j \delta_n} (M_n - M)(\theta - \theta_0) \right\}.
\end{aligned}$$

Suppose that the function $\phi_n(\cdot)$ is such that

$$\mathbb{E}\left\{ \sup_{\theta : d(\theta,\theta_0) \le u} (M_n - M)(\theta - \theta_0) \right\} \lesssim \phi_n(u) \quad \text{for every } \ u. \tag{11.13}$$

We thus get

$$\mathbb{P}\{2^{j-1}\delta_n < d(\widehat{\theta}_n, \theta_0) \le 2^j \delta_n\} \lesssim \frac{\phi_n(2^j \delta_n)}{2^{2j}\delta_n^2} \quad \text{for every } \ j.$$

As a consequence,

$$\mathbb{P}\{d(\widehat{\theta}_n, \theta_0) > 2^M \delta_n\} \lesssim \sum_{j>M} \frac{\phi_n(2^j \delta_n)}{2^{2j}\delta_n^2}.$$

The following assumption on $\phi_n(\cdot)$ is usually made to simplify the expression above: there exits $0 < \alpha < 2$ such that

$$\phi_n(cx) \le c^\alpha \phi_n(x) \quad \text{for all } c > 1 \ \text{ and } \ x > 0. \tag{11.14}$$

Under this assumption, we get

$$\mathbb{P}\{d(\widehat{\theta}_n, \theta_0) > 2^M \delta_n\} \lesssim \frac{\phi_n(\delta_n)}{\delta_n^2} \sum_{j>M} 2^{j(\alpha-2)}.$$

The quantity $\sum_{j>M} 2^{j(\alpha-2)}$ converges to zero as $M \to \infty$. Therefore, if δ_n is such that

$$\phi_n(\delta_n) \lesssim \delta_n^2,$$

then

$$d(\widehat{\theta}_n, \theta_0) \le 2^M \delta_n \quad \text{with probability at least } \ 1 - u_M,$$

where $u_M \to 0$ as $M \to \infty$.

This gives us the following nonasymptotic rate of convergence theorem.

Theorem 3.1. Assume Condition 11.9 holds, and there exists a function $\phi_n(\cdot)$ satisfying (11.13) and (11.14). Then, for every $M > 0$, we have

$$d(\widehat{\theta}_n, \theta_0) \le 2^M \delta_n \quad \text{with probability at least } \ 1 - u_M \ \text{ provided that } \phi_n(\delta_n) \lesssim \delta_n^2.$$

Here $u_M := \sum_{j>M} 2^{j(\alpha-2)} \to 0$ as $M \to \infty$.

Next time we will give two examples to which the above general theorem can be applied.