Introduction to Empirical Processes and Semiparametric Inference
Lecture 10: Empirical Process Methods

Michael R. Kosorok, Ph.D.

Professor and Chair of Biostatistics

Professor of Statistics and Operations Research

University of North Carolina-Chapel Hill

Orlicz Norms and Maximal Inequalities

Maximal inequalities are a very powerful tool in empirical processes.

A very useful class of norms used in maximal inequalities are the *Orlicz norms* for a real random variable X, defined for a given nondecreasing, nonzero convex function $\psi : [0, \infty] \mapsto [0, \infty]$, with $\psi(0) = 0$.

The Orlicz norm of X, $\|X\|_\psi$, also called the ψ*-norm*, is

$$\|X\|_\psi \equiv \inf \left\{ c > 0 : \mathsf{E}\psi \left(\frac{|X|}{c} \right) \leq 1 \right\},$$

where the norm takes the value ∞ if no finite c exists with $\mathsf{E}\psi(|X|/c) \leq 1$.

Exercise 8.5.1 verifies that $\|\cdot\|_\psi$ is indeed a norm on the space of random variables with $\|X\|_\psi < \infty$.

When ψ is of the form $x \mapsto x^p$, where $p \geq 1$, the corresponding Orlicz norm is just the L_p-norm

$$\|X\|_p \equiv (\mathsf{E}|X|^p)^{1/p}.$$

For maximal inequalities, Orlicz norms defined with $\psi_p(x) \equiv e^{x^p} - 1$, for $p \geq 1$, are of greater interest because of their sensitivity to behavior in the tails.

Clearly, since $x^p \leq \psi_p(x)$, we have $\|X\|_p \leq \|X\|_{\psi_p}$.

Also, by the series representation for exponentiation, $\|X\|_p \leq (p!)^{1/p} \|X\|_{\psi_1}$ for all $p \geq 1$.

Orlicz norms based on ψ_p relate fairly precisely to the tail probabilities:

LEMMA 1. *For a real random variable X and any $p \in [1, \infty)$, the following are equivalent:*

(i) $\|X\|_{\psi_p} < \infty$.

(ii) There exist constants $0 < C, K < \infty$ such that

$$\text{pr}(|X| > x) \quad \leq \quad K e^{-Cx^p}, \quad \textit{for all } x > 0. \tag{1}$$

An important use for Orlicz norms is to control the behavior of maxima.

This control is somewhat of an extension of the following simple result for L_p-norms:

For any random variables $X_1, \ldots, X_m$,

$$\begin{aligned}
\left\| \max_{1 \le i \le m} X_i \right\|_p &\le \left(\mathsf{E} \max_{1 \le i \le m} |X_i|^p \right)^{1/p} \\
&\le \left(\mathsf{E} \sum_{i=1}^{m} |X_i|^p \right)^{1/p} \\
&\le m^{1/p} \max_{1 \le i \le m} \|X_i\|_p .
\end{aligned}$$

A similar result holds for certain Orlicz norms:

LEMMA 2. *Let* $\psi : [0, \infty) \mapsto [0, \infty)$ *be convex, nondecreasing and nonzero, with* $\psi(0) = 0$ *and*

$$\limsup_{x,y \to \infty} \frac{\psi(x)\psi(y)}{\psi(cxy)} < \infty$$

for some constant $c < \infty$*.*

Then, for any random variables $X_1, \ldots, X_m$*,*

$$\left\| \max_{1 \leq i \leq m} X_i \right\|_\psi \leq K \psi^{-1}(m) \max_{1 \leq i \leq m} \|X_i\|_\psi,$$

where the constant K *depends only on* ψ*.*

An important consequence of Lemma 2 is that maximums of random variables with bounded ψ-norm grow at the rate $\psi^{-1}(m)$.

Based on Exercise 8.5.4, ψ_p satisfies the conditions of Lemma 2 with $c = 1$, for any $p \in [1, \infty)$.

The implication is that the growth of maxima is at most logarithmic, since $\psi_p^{-1}(m) = (\log(m+1))^{1/p}$.

These results will prove quite useful in the next section.

Maximal Inequalities for Processes

The goals of this section are to first establish a general maximal inequality for *separable* stochastic processes and then specialize to *sub-Gaussian* processes.

A stochastic process $\{X(t), t \in T\}$ is separable when there exists a countable subset $T_* \subset T$ such that

$$\sup_{t \in T} \inf_{s \in T_*} |X(t) - X(s)| = 0$$

almost surely.

For example, any cadlag process indexed by a closed interval in $\mathbb{R}$ is separable because the rationals are a separable subset of $\mathbb{R}$.

The need for separability of certain processes in the Glivenko-Cantelli and Donsker theorems is hidden in other conditions of the involved theorems, and direct verification of separability is seldom required in statistical applications.

A stochastic process is sub-Gaussian when

$$P(|X(t) - X(s)| > x) \leq 2e^{-\frac{1}{2}x^2/d^2(s,t)},$$

for all $s, t \in T$, $x > 0$, (2)

for a semimetric d on T.

In this case, we say that X is sub-Gaussian with respect to d.

An important example of a separable sub-Gaussian stochastic process, the Rademacher process, will be presented later.

These processes will be utilized later in this chapter in the development of a Donsker theorem based on uniform entropy.

Another example of a sub-Gaussian process is Brownian motion on $[0, 1]$, which can easily be shown to be sub-Gaussian with respect to $d(s, t) = |s - t|^{1/2}$.

Because the sample paths are continuous, Brownian motion is also separable.

The conclusion of Lemma 2 above is not immediately useful for maximizing $X(t)$ over $t \in T$ since a potentially infinite number of random variables is involved.

However, a method called *chaining*, does make such maximization possible in some settings.

The technique depends on the *metric entropy* of the index set T based on the semimetric $d(s,t) = \|X(s) - X(t)\|_\psi$.

For an arbitrary semimetric space (T, d), the *covering number* $N(\epsilon, T, d)$ is the minimal number of closed d-balls of radius ϵ required to cover T.

The *packing number* $D(\epsilon, T, d)$ is the maximal number of points that can fit in T while maintaining a distance greater than ϵ between all points.

When the choice of index set T is clear by context, the notation will be abbreviated as $N(\epsilon, d)$ and $D(\epsilon, d)$, respectively.

The associated *entropy numbers* are the respective logarithms of the covering and packing numbers.

Taken together, these concepts define metric entropy.

For a semimetric space (T, d) and each $\epsilon > 0$,

$$N(\epsilon, d) \leq D(\epsilon, d) \leq N(\epsilon/2, d).$$

To see this, note that there exists a maximal subset $T_\epsilon \subset T$ such that the cardinality of $T_\epsilon = D(\epsilon, d)$ and the minimum distance between distinct points in T_ϵ is $> \epsilon$.

If we now place closed ϵ-balls around each point in T_ϵ, we have a covering of T.

If this were not true,

- there would exist a point $t \in T$ which has distance $> \epsilon$ from all the points in T_ϵ,
- but this would mean that $D(\epsilon, d) + 1$ points can fit into T while still maintaining a separation $> \epsilon$ between all points.
- But this contradicts the maximality of $D(\epsilon, d)$.

Thus $N(\epsilon, d) \leq D(\epsilon, d)$.

Now note that no ball of radius $\leq \epsilon/2$ can cover more than one point in T_ϵ, and thus at least $D(\epsilon, d)$ closed $\epsilon/2$-balls are needed to cover T_ϵ.

Hence $D(\epsilon, d) \leq N(\epsilon/2, d)$.

This discussion reveals that covering and packing numbers are essentially equivalent in behavior as $\epsilon \downarrow 0$.

However, it turns out to be slightly more convenient for our purposes to focus on packing numbers in this section.

Note that T is totally bounded if and only if $D(\epsilon, d)$ is finite for each $\epsilon > 0$.

The success of the upcoming maximal inequality depends on how fast $D(\epsilon, d)$ increases as $\epsilon \downarrow 0$.

THEOREM 1. *(General maximal inequality) Let ψ satisfy the conditions of Lemma 2, and let $\{X(t), t \in T\}$ be a separable stochastic process with $\|X(s) - X(t)\|_\psi \leq r d(s,t)$, for all $s, t \in T$, some semimetric d on T, and a constant $r < \infty$.*

Then for any $\eta, \delta > 0$,

$$\left\| \sup_{s,t \in T : d(s,t) \leq \delta} |X(s) - X(t)| \right\|_\psi \leq K \left[\int_0^\eta \psi^{-1}(D(\epsilon, d)) d\epsilon + \delta \psi^{-1}(D^2(\eta, d)) \right],$$

for a constant $K < \infty$ which depends only on ψ and r.

Moreover,

$$\left\| \sup_{s,t \in T} |X(s) - X(t)| \right\|_\psi \leq 2K \int_0^{\text{diam}\, T} \psi^{-1}(D(\epsilon, d)) d\epsilon,$$

where $\text{diam}\, T \equiv \sup_{s,t \in T} d(s,t)$ *is the diameter of* T*.*

Proof: Note that if the first integral were infinite, the inequalities would be trivially true.

Hence we can, without loss of generality, assume that the packing numbers and associated integral are bounded.

Construct a sequence of finite nested sets $T_0 \subset T_1 \subset \cdots \subset T$ such that for each T_j,

- $d(s,t) > \eta 2^{-j}$ for every distinct $s, t \in T_j$, and
- each T_j is "maximal" in the sense that no additional points can be added to T_j without violating the inequality.

Note that by the definition of packing numbers, the number of points in T_j is bounded above by $D(\eta 2^{-j}, d)$.

Now we will do the chaining part of the proof.

Begin by "linking" every point $t_{j+1} \in T_{j+1}$ to one and only one $t_j \in T_j$ such that $d(t_j, t_{j+1}) \leq \eta 2^{-j}$, for all points in T_{j+1}.

Continue this process to link all points in T_j with points in T_{j-1}, and so on, to obtain for every t_{j+1} ($\in T_{j+1}$) a chain $t_{j+1}, t_j, t_{j-1}, \ldots, t_0$ that connects to a point in T_0.

For any integer $k \geq 0$ and arbitrary points $s_{k+1}, t_{k+1} \in T_{k+1}$, the difference in increments along their respective chains connecting to s_0, t_0 can be bounded as follows:

$$\begin{aligned} &|\{X(s_{k+1}) - X(t_{k+1})\} - \{X(s_0) - X(t_0)\}| \\ &\qquad = \left| \sum_{j=0}^{k} \{X(s_{j+1}) - X(s_j)\} - \sum_{j=0}^{k} \{X(t_{j+1}) - X(t_j)\} \right| \\ &\qquad \leq 2 \sum_{j=0}^{k} \max |X(u) - X(v)|, \end{aligned}$$

where for fixed j the max is taken over all links (u, v) from T_{j+1} to T_j.

Hence the jth maximum is taken over at most the cardinality of T_{j+1} links, with each link having $\|X(u) - X(v)\|_\psi$ bounded by $rd(u, v) \leq r\eta 2^{-j}$.

By Lemma 2, we have for a constant $K_0 < \infty$ depending only on ψ & r,

$$\left\| \max_{s,t \in T_{k+1}} |\{X(s) - X(s_0)\} - \{X(t) - X(t_0)\}| \right\|_\psi \tag{3}$$

$$\begin{aligned} &\leq K_0 \sum_{j=0}^{k} \psi^{-1}(D(\eta 2^{-j-1}, d))\eta 2^{-j} \\ &= 4K_0 \sum_{j=0}^{k} \psi^{-1}(D(\eta 2^{-k+j-1}, d))\eta 2^{-k+j-2} \\ &\leq 4\eta K_0 \int_0^1 \psi^{-1}(D(\eta u, d))du \\ &= 4K_0 \int_0^\eta \psi^{-1}(D(\epsilon, d))d\epsilon. \end{aligned}$$

In this bound, s_0 and t_0 depend on s and t in that they are the endpoints of the chains starting at s and t, respectively.

The maximum of the increments $|X(s_{k+1}) - X(t_{k+1})|$, over all s_{k+1} and t_{k+1} in T_{k+1} with $d(s_{k+1}, t_{k+1}) < \delta$, is bounded by

- the left-hand-side of (3)
- plus the maximum of the discrepancies at the ends of the chains $|X(s_0) - X(t_0)|$ for those points in T_{k+1} which are less than δ apart.

For every such pair of endpoints s_0, t_0 of chains starting at two points in T_{k+1} within distance δ of each other, choose one and only one pair s_{k+1}, t_{k+1} in T_{k+1}, with $d(s_{k+1}, t_{k+1}) < \delta$, whose chains end at s_0, t_0.

By definition of T_0, this results in at most $D^2(\eta, d)$ pairs.

Now,

$$\begin{aligned} |X(s_0) - X(t_0)| &\leq |\{X(s_0) - X(s_{k+1})\} - \{X(t_0) - X(t_{k+1})\}| \\ &\quad + |X(s_{k+1}) - X(t_{k+1})|. \end{aligned} \tag{4}$$

Take the maximum of (4) over all pairs of endpoints s_0, t_0.

The maximum of the first term of the right-hand-side of (4) is bounded by the left-hand-side of (3).

The maximum of the second term of the right-hand-side of (4) is the maximum of $D^2(\eta, d)$ terms with ψ-norm bounded by $r\delta$.

By Lemma 2, this maximum is bounded by some constant C times $\delta\psi^{-1}(D^2(\eta, d))$.

Combining this with (3), we obtain

$$\left\| \max_{s,t \in T_{k+1}: d(s,t) < \delta} |X(s) - X(t)| \right\|_\psi \leq 8K_0 \int_0^\eta \psi^{-1}(D(\epsilon, d))d\epsilon + C\delta\psi^{-1}(D^2(\eta, d)).$$

By the fact that the right-hand-side does not depend on k, we can replace T_{k+1} with $T_\infty = \cup_{j=0}^\infty T_j$ by the monotone convergence theorem.

If we can verify that taking the supremum over T_∞ is equivalent to taking the supremum over T, then the first conclusion of the theorem follows with $K = (8K_0) \vee C$.

Since X is separable, there exists a countable subset $T_* \subset T$ such that $\sup_{t \in T} \inf_{s \in T_*} |X(t) - X(s)| = 0$ almost surely.

Let Ω_* denote the subset of the sample space of X for which this supremum is zero.

Accordingly pr$(\Omega_*) = 1$.

Now, for any point t and sequence $\{t_n\}$ in T, it is easy to see that $d(t, t_n) \to 0$ implies $|X(t) - X(t_n)| \to 0$ almost surely (see Exercise 8.5.5).

For each $t \in T_*$, let Ω_t be the subset of the sample space of X for which $\inf_{s \in T_\infty} |X(s) - X(t)| = 0$.

Since T_∞ is a dense subset of the semimetric space (T, d), pr$(\Omega_t) = 1$.

Letting $\tilde{\Omega} \equiv \Omega_* \cap (\cap_{t \in T_*} \Omega_t)$, we now have pr$(\tilde{\Omega}) = 1$.

This, combined with the fact that

$$\begin{aligned}\sup_{t\in T}\inf_{s\in T_\infty}|X(t)-X(s)| \quad &\leq \quad \sup_{t\in T}\inf_{s\in T_*}|X(t)-X(s)| \\ &\quad + \sup_{t\in T_*}\inf_{s\in T_\infty}|X(s)-X(t)|,\end{aligned}$$

implies that $\sup_{t\in T}\inf_{s\in T_\infty}|X(t)-X(s)|=0$ almost surely.

Thus taking the supremum over T is equivalent to taking the supremum over T_∞.

The second conclusion of the theorem follows from the previous result by setting $\delta = \eta =$ diam T and noting that, in this case, $D(\eta, d) = 1$.

Now we have

$$\begin{aligned} \delta\psi^{-1}(D^2(\eta, d)) &= \eta\psi^{-1}(D(\eta, d)) \\ &= \int_0^\eta \psi^{-1}(D(\eta, d))d\epsilon \\ &\leq \int_0^\eta \psi^{-1}(D(\epsilon, d))d\epsilon, \end{aligned}$$

and the second conclusion follows.□

As a consequence of Exercise 8.5.5 below, the conclusions of Theorem 1 show that X has d-continuous sample paths almost surely whenever the integral $\int_0^\eta \psi^{-1}(D(\epsilon, d))d\epsilon$ is bounded for some $\eta > 0$.

It is also easy to verify that the maximum of the process of X is bounded, since

$$\| \sup_{t \in T} X(t) \|_\psi \leq \|X(t_0)\|_\psi + \| \sup_{s,t \in T} |X(t) - X(s)| \, \|_\psi,$$

for any choice of $t_0 \in T$.

Thus X is tight and takes its values in $UC(T, d)$ almost surely.

These results will prove quite useful in later developments.

An important application of Theorem 1 is to *sub-Gaussian* processes:

COROLLARY 1. *Let* $\{X(t), t \in T\}$ *be a separable sub-Gaussian process with respect to* d*.*

Then for all $\delta > 0$*,*

$$\mathsf{E}\left(\sup_{s,t\in T: d(s,t)\leq\delta} |X(s) - X(t)|\right) \leq K \int_0^\delta \sqrt{\log D(\epsilon, d)}\, d\epsilon,$$

where K *is a universal constant.*

Also, for any $t_0 \in T$,

$$\mathsf{E}\left(\sup_{t\in T}|X(t)|\right) \le \mathsf{E}|X(t_0)| + K\int_0^{\operatorname{diam} T}\sqrt{\log D(\epsilon,d)}d\epsilon.$$

Proof. Apply Theorem 1 with $\psi = \psi_2$ and $\eta = \delta$.

Because $\psi_2^{-1}(m) = \sqrt{\log(1+m)}$,

$$\psi_2^{-1}(D^2(\delta,d)) \le \sqrt{2}\psi_2^{-1}(D(\delta,d)).$$

Hence the second term of the general maximal inequality can be replaced by

$$\sqrt{2}\delta\psi^{-1}(D(\delta,d)) \leq \sqrt{2}\int_0^\delta \psi^{-1}(D(\epsilon,d))d\epsilon,$$

and we obtain

$$\left\| \sup_{d(s,t)\leq\delta} |X(s)-X(t)| \right\|_{\psi_2} \leq K\int_0^\delta \sqrt{\log(1+D(\epsilon,d))}d\epsilon,$$

for an enlarged universal constant K.

Note that $D(\epsilon,d) \geq 2$ for all ϵ strictly less than diam T.

Since $(1+m) \leq m^2$ for all $m \geq 2$, the 1 inside of the logarithm can be removed at the cost of increasing K again, whenever $\delta <$ diam T.

Thus it is also true for all $\delta \leq$ diam T.

We are done with the first conclusion since $d(s,t) \leq$ diam T for all $s, t \in T$.

Since the second conclusion is an easy consequence of the first, the proof is complete.$\square$