Stat 8112 Lecture Notes

Weak Convergence in Metric Spaces

Charles J. Geyer

January 23, 2013

1 Metric Spaces

Let X be an arbitrary set. A function $d : X \times X \to \mathbb{R}$ is called a *metric* if it satisfies the following properties

(i) $d(x, y) = d(y, x)$,

(ii) $d(x, y) \geq 0$,

(iii) $d(x, y) = 0$ if and only if $x = y$,

(iv) $d(x, z) \leq d(x, y) + d(y, z)$,

it being understood that these are identities holding for all values of the free variables. The identity (iv) is called the *triangle inequality.* A set equipped with a metric is called a *metric space.*

The most familiar metric spaces are the real numbers $\mathbb{R}$ with the metric

$$d(x, y) = |x - y|$$

and finite-dimensional vector spaces $\mathbb{R}^d$ with the metric

$$d(x, y) = \|x - y\| \tag{1}$$

where the norm $\| \cdot \|$ can be defined in a number of different ways

$$\|x\| = \sum_{i=1}^{d} |x_i| \tag{2a}$$

$$\|x\| = \left(\sum_{i=1}^{d} |x_i|^2 \right)^{1/2} \tag{2b}$$

$$\|x\| = \max_{1 \leq i \leq d} |x_i| \tag{2c}$$

which are called the L^1 norm, the L^2 norm (or the Euclidean norm), and the L^∞ norm (or the supremum norm), respectively.

2 Topology of Metric Spaces

A sequence x_n in a metric space X converges to x if

$$d(x_n, x) \to 0, \qquad \text{as } n \to \infty.$$

For any x and $\varepsilon > 0$, the set

$$B_\varepsilon(x) = \{\, y \in X : d(x, y) < \varepsilon \,\} \tag{3}$$

is called the *open ball* centered at x having radius ε. A subset of a metric space is called *open* if it is a union of open balls or if it is empty. A subset of a metric space is called *closed* if its complement is open. The following properties are easily shown.

- An arbitrary union of open sets is open.
- A finite intersection of open sets is open.
- A finite union of closed sets is closed.
- An arbitrary intersection of closed sets is closed.

For any set A the intersection of all closed sets containing A is a closed set called the *closure* of A, denoted $\operatorname{cl} A$. The complement of the closure is an open set called the *interior* of A, denoted $\operatorname{int} A$. The intersection of $\operatorname{cl} A$ and $\operatorname{cl}(A^c)$ is a closed set called the *boundary* of A, denoted $\operatorname{bdry} A$.

We have $x \in \operatorname{int} A$ if and only if there exists an $\varepsilon > 0$ such that

$$B_\varepsilon(x) \subset A.$$

We have $x \in \operatorname{cl} A$ if and only if for every $\varepsilon > 0$

$$B_\varepsilon(x) \cap A \neq \varnothing.$$

Hence we have $x \in \operatorname{cl} A$ if and only if there exists a sequence x_n contained in A such that $x_n \to x$.

If X and Y are metric spaces, having metrics d and e, respectively, then a function $f : X \to Y$ is said to be *continuous at a point* x if for every $\varepsilon > 0$ there exists a $\delta > 0$ (which may depend on ε and x) such that

$$e\bigl(f(x), f(y)\bigr) \le \varepsilon, \qquad \text{whenever } d(x, y) \le \delta. \tag{4}$$

A simpler characterization is

$$f(x_n) \to f(x), \qquad \text{whenever } x_n \to x.$$

A function is said to be *continuous* if it is continuous at each point of its domain. In (4) δ is allowed to depend on x. If the same δ can be used for all x, the function f is said to be *uniformly continuous*.

3 Polish Spaces

A sequence x_n in a metric space X is said to be *Cauchy* if for every $\varepsilon > 0$ there exists an integer N (which may depend on ε) such that

$$d(x_m, x_n) \le \varepsilon, \qquad \text{whenever } m \ge N \text{ and } n \ge N.$$

It follows from the triangle inequality that every convergent sequence is Cauchy. A metric space is said to be *complete* if every Cauchy sequence converges. Examples of complete metric spaces are $\mathbb{R}$ and $\mathbb{R}^d$.

A subset D of a metric space is said to be *dense* if every open ball contains an element of D. A metric space is said to be *separable* if it contains a countable dense set. Examples of separable metric spaces are $\mathbb{R}$ and $\mathbb{R}^d$. The rational numbers $\mathbb{Q}$ are a countable dense subset of $\mathbb{R}$, and $\mathbb{Q}^d$ is a countable dense subset of $\mathbb{R}^d$.

A complete separable metric space is also called a *Polish space*. The name comes from it being first studied by famous Polish topologists of the 1930's (Sierpinski, Kuratowski, Tarski and others).

3.1 Examples

An example of a Polish space that is not $\mathbb{R}^d$ is the space $C(0,1)$ of all continuous functions on the closed interval $[0,1]$ with norm defined by

$$\|f\| = \sup_{0 \le x \le 1} |f(x)| \tag{5}$$

and metric defined in terms of the norm by (1). It is complete because a uniform limit of continuous functions is continuous (Browder, 1996, Theorem 3.24). It is separable because the set of polynomials on $[0,1]$ is dense in $C(0,1)$ by the Weierstrass approximation theorem (Browder, 1996, Theorem 7.1), and the set of polynomials with rational coefficients is countable and dense in the set of all polynomials, hence also dense in $C(0,1)$.

It is a little off-topic, but $C(0,1)$ is also a vector space. Functions are like vectors in that they can be added

$$(f+g)(x) = f(x) + g(x), \qquad \text{for all } x$$

and multiplied by scalars

$$(af)(x) = af(x), \qquad \text{for all } x$$

and thus constitute a vector space. In fact our usual notion of vectors as n-tuples can also be thought of as functions-are-vectors, the function being

$i \mapsto x_i$. A complete normed vector space is called a *Banach space*. Thus $C(0, 1)$ is also a Banach space. Since separability is not part of the definition of Banach space, we have to say $C(0, 1)$ is a separable Banach space, if we want to mention separability.

It is clear that $C(0, 1)$ is not a finite-dimensional vector space. It is not isomorphic to $\mathbb{R}^n$ for any n.

3.2 Non-Examples

An example of a metric space that is not complete is the set $\mathbb{Q}$ of rational numbers, considered as a metric subspace of the set $\mathbb{R}$ of real numbers. The sequence of rational numbers 1.4, 1.41, 1.414, 1.4142, …, whose n-th element is $\sqrt{2}$ rounded to n decimal places, is a Cauchy sequence that converges in $\mathbb{R}$ to $\sqrt{2}$ but does not converge in $\mathbb{Q}$ because $\sqrt{2}$ is not rational (as was known in antiquity).

An example of a metric space that is not separable is the set $M(0, 1)$ of all nondecreasing functions on the closed interval $[0, 1]$ with the supremum norm (5) and metric defined in terms of the norm by (1). For each real number x the distribution function F_x of the distribution concentrated at x is an element of $M(0, 1)$. If $x \neq y$, then $\|F_x - F_y\| = 1$. Let A_x denote the open ball in $M(0, 1)$ of radius $1/2$ centered at F_x. If $x \neq y$, then A_x and A_y are disjoint. The real numbers are uncountable (as was proved by Cantor). Any set dense in $M(0, 1)$ must contain an element of each A_x. Hence $M(0, 1)$ is nonseparable.

4 Weak Convergence in Polish Spaces

A sequence X_n of random elements of a Polish space is said to converge *weakly* or *in law* to a random element X if

$$E\{f(X_n)\} \to E\{f(X)\}, \qquad \text{for every bounded continuous function } f.$$

We write

$$X_n \xrightarrow{w} X$$

or

$$X_n \xrightarrow{\mathcal{L}} X$$

to signify this.

Let $\mathcal{B}$ denote the Borel sigma-field of the metric space (the smallest sigma-field containing the open sets) and let P_n and P denote the laws of

X_n and X, respectively, defined by

$$P_n(B) = \Pr(X_n \in B), \qquad B \in \mathcal{B}$$

and

$$P(B) = \Pr(X \in B), \qquad B \in \mathcal{B}.$$

Then we also write $P_n \Rightarrow P$ to denote this convergence in law.

The standard reference for convergence in law in Polish spaces is Billingsley (1999, first edition 1968). In it we find Theorem 2.1, called the "portmanteau theorem," which we repeat here.

Theorem 1. *The following conditions are equivalent*

(i) $P_n \Rightarrow P$.

(ii) $\int f dP_n \to \int f dP$ *for all bounded uniformly continuous* f.

(iii) $\limsup_n P_n(F) \le P(F)$ *for all closed events* F.

(iv) $\liminf_n P_n(G) \ge P(G)$ *for all open events* G.

(v) $\lim_n P_n(A) = P(A)$ *for all events* A *such that* $P(\operatorname{bdry} A) = 0$.

Condition (ii) just repeats the definition of weak convergence with uniformly continuous f replacing continuous f because $\int f dP_n$ is another notation for $E\{f(X_n)\}$ when P_n is the law of X_n. The word "events" in (iii), (iv), and (v) refers to elements of the Borel sigma-field of the metric space. Of course, in (iii) every closed set is Borel and in (iv) every open set is Borel.

It is not necessary to check all Borel sets or all bounded continuous functions. The characteristic function convergence theorem provides an example where not all bounded continuous functions need be checked. The following theorem, which is Theorem 2.2 in Billingsley (1999), gives a criterion for weak convergence in which not all Borel sets are checked.

Theorem 2. *Suppose* $\mathcal{A}$ *is a family of Borel sets that is closed under finite intersections and each open set is a countable union of sets in* $\mathcal{A}$*. Then* $P_n(A) \to P(A)$ *for all* $A \in \mathcal{A}$ *implies* $P_n \Rightarrow P$.

Let $\mathcal{P}$ denote the set of all probability measures on some Polish space. Then there is a metric π for $\mathcal{P}$ that induces weak convergence, that is, $P_n \Rightarrow P$ if and only if $\pi(P_n, P) \to 0$. One such metric is called the *Prohorov metric,* which is defined as follows (Billingsley, 1999, pp. 7 and 72).

Let S be a metric space with metric d. For any $\varepsilon > 0$ and any Borel set A, define

$$A^\varepsilon = \bigcup_{x \in A} B_\varepsilon(x) \tag{6}$$

(called the *ε-dilation* of A), where $B_\varepsilon(x)$ is defined by (3). Let P and Q be probability measures on S. Then the Prohorov distance $\pi(P, Q)$ between P and Q is the infimum of the set of $\varepsilon > 0$ such that the two inequalities

$$P(A) \le Q(A^\varepsilon) + \varepsilon \qquad \text{and} \qquad Q(A) \le P(A^\varepsilon) + \varepsilon$$

hold for all events A. The assertion that $P_n \Rightarrow P$ if and only if $\pi(P_n, P) \to 0$ is in Theorem 7.6 in Billingsley (1999).

5 Convergence in Probability and Almost Surely

We say a sequence of random elements X_n of a metric space with metric d *converges in probability* to a random element X of the same metric space if for every $\varepsilon > 0$

$$\Pr\{d(X_n, X) > \varepsilon\} \to 0, \qquad \text{as } n \to \infty.$$

We say such a sequence *converges almost surely* if there exists a Borel set A such that $\Pr(A) = 1$ and

$$d\bigl(X_n(\omega), X(\omega)\bigr) \to 0, \qquad \omega \in A.$$

We know for sequences of random vectors almost sure convergence implies convergence in probability implies convergence in law (Ferguson, 1996, Theorem 1). This is also true for sequences of random elements of a Polish space. That almost sure convergence implies convergence in probability follows immediately from the dominated convergence theorem. That convergence in probability implies convergence in law is the corollary to Theorem 3.1 in Billingsley (1999).

6 An Empirical Process Law of Large Numbers

Suppose X_1, X_2, $\ldots$ is an independent and identically distributed (IID) sequence of random elements of a Polish space S. The *empirical measure* is a probability measure $\widehat{P}_n$ on S defined by

$$\widehat{P}_n(B) = \frac{1}{n} \sum_{i=1}^{n} I_B(X_i), \qquad B \in \mathcal{B},$$

where $\mathcal{B}$ is the Borel sigma-field for the Polish space and

$$I_A(x) = \begin{cases} 1, & x \in A \\ 0, & \text{otherwise} \end{cases}$$

is the indicator function of the event A. Since $\widehat{P}_n$ is a function of random elements, it too is random, a random element of the space $\mathcal{P}$ of all probability measures on S. The value of $\widehat{P}_n$ at the outcome ω is

$$\widehat{P}_n(\omega)(B) = \frac{1}{n} \sum_{i=1}^{n} I_B\bigl(X_i(\omega)\bigr).$$

The strong law of large numbers (SLLN), Theorem 4(c) in Ferguson (1996), says

$$\widehat{P}_n(B) \xrightarrow{\text{a.s.}} P(B), \qquad B \in \mathcal{B}, \tag{7}$$

where P is the law of X_1 and by IID of the rest of the X_i too. In (7) the set B is arbitrary, but the almost sure convergence is not simultaneous: the exception set where convergence fails may depend on B. That is, for each B there is an event N_B such that $\Pr(N_B) = 0$ and

$$\widehat{P}_n(\omega)(B) \to P(B), \qquad \omega \notin N_B.$$

The following theorem strengthens this statement.

Theorem 3. *If $\widehat{P}_n$ is the empirical measure constructed from a sequence of IID elements of a Polish space, P is the law of each of those elements, and π denotes the Prohorov metric on the space of all probability measures on this Polish space, then*

$$\pi(\widehat{P}_n, P) \xrightarrow{a.s.} 0. \tag{8}$$

We could also rewrite (8) as

$$\Pr(\widehat{P}_n \Rightarrow P) = 1.$$

Proof. Let S denote the Polish space, let D be a countable dense set for it, and let

$$\mathcal{B}_c = \{\, B_{1/n}(x) : x \in D,\ n = 1, 2, \ldots \,\}.$$

Then every open set in S is a union of elements of $\mathcal{B}_c$. Let $\mathcal{A}_c$ be the set of all finite intersections of elements of $\mathcal{B}_c$. Then $\mathcal{A}_c$ is also countable. By subadditivity of probability

$$\Pr\left(\bigcup_{A \in \mathcal{A}_c} N_A \right) \le \sum_{A \in \mathcal{A}_c} \Pr(N_A) = 0$$

Thus we have

$$\widehat{P}_n(\omega)(A) \to P(A), \qquad \omega \notin \bigcup_{A \in A_c} N_A.$$

Now apply Theorem 2. □

A statistical model is a family $\mathcal{Q}$ of probability measures, that is, a subset of the family $\mathcal{P}$ of all probability measures. The task of statistical inference is to say something about the true unknown distribution P of the data from data X_1, ..., X_n. One way to "say something" is to provide a point estimate $\widehat{Q}_n$, which is an element of $\mathcal{Q}$ that is a guess at the true unknown P. We say $\widehat{Q}_n$ is a *minimum Prohorov distance estimator* if

$$\pi(\widehat{Q}_n, \widehat{P}_n) \le \inf_{Q \in \mathcal{Q}} \pi(Q, \widehat{P}_n) + \frac{1}{n}$$

(the reason for the $1/n$ term being that the infimum may not be achieved).

Corollary 4. *If $\widehat{Q}_n$ is a minimum Prohorov distance estimator and P is the true unknown distribution of the data, which is assumed to lie in the statistical model $\mathcal{Q}$, then*

$$\pi(\widehat{Q}_n, P) \xrightarrow{a.s.} 0.$$

Proof. Since $P \in \mathcal{Q}$ we have

$$\pi(\widehat{Q}_n, \widehat{P}_n) \le \inf_{Q \in \mathcal{Q}} \pi(Q, \widehat{P}_n) + \frac{1}{n} \le \pi(P, \widehat{P}_n) + \frac{1}{n} \xrightarrow{\text{a.s.}} 0$$

by Theorem 3. □

This corollary is useless for practical applications because the Prohorov metric is so unwieldy that minimum Prohorov distance estimators are impossible to compute, but the theorem does show that the empirical measure is a strongly consistent point estimate of the true unknown measure (in the Prohorov metric) under no assumptions whatsoever. The corollary does at least show that consistent estimation is possible in principle under no assumptions whatsoever (in the Prohorov metric), even if we don't know how to do it in practice. This is something to remember when we look at other methods of estimation, which do need lots of assumptions.

7 Compactness

A subset of $\mathbb{R}$ or $\mathbb{R}^d$ is *compact* if it is closed and bounded (the Heine-Borel theorem), but this criterion does not carry over to general metric spaces. The reason is that boundedness doesn't do anything. If d is a metric,

$$e(x, y) = \frac{d(x, y)}{1 + d(x, y)}$$

is another metric that determines the same topology because it has the same open balls, since

$$d(x, y) < \varepsilon \quad \text{if and only if} \quad e(x, y) < \frac{\varepsilon}{1 + \varepsilon},$$

and hence has the same convergent sequences. We say d and e are *equivalent* metrics to indicate that they have the same convergent sequences. If we use e as the metric, then every set is bounded, so boundedness cannot imply anything. We need another concept.

A subset K of a metric space S is *totally bounded* if for every $\varepsilon > 0$ there exists a finite subset F of S such that $K \subset F^\varepsilon$, where ε-dilation is defined by (6). In other words, for every $\varepsilon > 0$ there exists a finite family of open balls of radius ε that cover K. A subset of a complete metric space is compact if and only if it is closed and totally bounded (Browder, 1996, Theorems 6.41 and 6.63).

A topological space is said to be *sequentially compact* if every sequence has a convergent subsequence. Subsets of metric spaces are compact if and only if they are sequentially compact (Browder, 1996, Theorem 6.67). That is, a subset K of a metric space is compact if and only if every sequence in K has a subsequence that converges to a limit in K.

8 Tightness and Boundedness in Probability

A random element X of a metric space is said to be *tight* if for every $\varepsilon > 0$ there exists a compact set K such that

$$\Pr(X \in K) > 1 - \varepsilon.$$

In a Polish space every random element is tight (Billingsley, 1999, Theorem 1.3).

A sequence X_1, X_2, ... of random elements of a metric space is said to be *tight* if for every $\varepsilon > 0$ there exists a compact set K such that

$$\Pr(X_n \in K) > 1 - \varepsilon, \qquad \text{for all } n \in \mathbb{N}.$$

In a metric space in which bounded means the same thing as totally bounded, the term *bounded in probability* is also used as a synonym of tight. A sequence X_1, X_2, $\ldots$ of random vectors is bounded in probability (or tight) if for every $\varepsilon > 0$ there exists an $M < \infty$ such that

$$\Pr(\|X_n\| \le M) > 1 - \varepsilon, \qquad \text{for all } n \in \mathbb{N}.$$

9 The Prohorov Theorem

Theorem 5. *Let X_1, X_2, $\ldots$ be a sequence of random elements of a Polish space. If the sequence is tight, then every subsequence contains a weakly convergent subsubsequence. Conversely, if the sequence weakly converges, then it is tight.*

For an example of use of this theorem, see Section 11.

10 The Subsequence Principle

Every topological notion of convergence satisfies the *Urysohn property*, which says that, given a sequence x_n, if every subsequence x_{n_k} has a convergent subsubsequence $x_{n_{k_l}}$ and every such subsubsequence has the same limit x, then the whole sequence converges ($x_n \to x$). A sequential notion of convergence satisfies the *Hausdorff property* if no sequence converges to more than one limit point. A Hausdorff sequential notion of convergence is topological if and only if it satisfies the Urysohn property (Beattie and Butzmann, 2002, Proposition 1.7.15).

Since weak convergence in Polish spaces is metrizable (by the Prohorov metric), it is topological, hence satisfies the Urysohn property. The term "Urysohn property" is not widely used in probability theory, so we will call this the "subsequence principle."

11 The Skorohod Theorem

Convergence in law does not imply almost sure convergence; the random elements of a weakly convergence sequence do not even need to be defined on the same probability space, but the random elements of an almost surely convergent sequence do.

However, the following theorem is almost as good. Write $\mathcal{L}(X) = \mathcal{L}(Y)$ to indicate that random elements X and Y have the same law.

Theorem 6. *Suppose X_n is a sequence of random elements of a Polish space converging in law to X, then there exists a sequence X_n^* of random elements of the same Polish space converging almost surely to X^*, and*

$$\mathcal{L}(X) = \mathcal{L}(X^*) \tag{9a}$$

$$\mathcal{L}(X_n) = \mathcal{L}(X_n^*), \qquad \textit{for all } n. \tag{9b}$$

This theorem can be used to provide trivial proofs of facts that otherwise require complicated arguments. Here is an example that shows how Skorohod-type proofs work. Let S be a Polish space and X_n a sequence of random elements of S converging to a random element X of S. Let f be a nonnegative valued function on S that is continuous except at a set of points D satisfying $\Pr(X \in D) = 0$. Then

$$E\{f(X)\} \le \liminf_{n \to \infty} E\{f(X_n)\}. \tag{10}$$

Why? From Skorohod's theorem we have $X_n^* \xrightarrow{\text{a.s.}} X^*$ with (9a) and (9b) holding. By Fatou's lemma

$$E\{f(X^*)\} \le \liminf_{n \to \infty} E\{f(X_n^*)\}. \tag{11}$$

Since expectations only depend on laws, not on the random variables having those laws, the analogous expectations in (10) and (11) are the same.

12 Optimization of Random Functions

The Skorohod theorem is very powerful when combined with the Prohorov theorem. Often this allows one to use known results from real analysis directly. Here is an example that shows that. It is a simplification of an argument we will go into much detail with later in the context of maximum likelihood estimation. But before we can do that we need to learn about another Polish space.

12.1 Uniform Convergence on Compact Sets

Let K be a compact subset of $\mathbb{R}^d$ and $C(K)$ the set of all continuous functions on K equipped with the uniform norm

$$\|f\| = \sup_{x \in K} |f(x)| \tag{12}$$

and metric derived from the norm by (1). Like its special case $C(0,1)$, this is a Polish space and a separable Banach space, and for much the same reasons as $C(0,1)$, a uniform limit of a sequence of continuous functions is continuous on any metric space (Browder, 1996, Theorem 6.69) and the Weierstrass approximation theorem works for $\mathbb{R}^d$ as well as $\mathbb{R}$ (Browder, 1996, Corollary 7.7).

Now we want to consider convergence in $C(K)$ for various K, so we now denote (12) by $\|f\|_K$ to distinguish the different norms. Let $C(\mathbb{R}^d)$ denote the set of continuous functions on $\mathbb{R}^d$. We say a sequence of elements f_n of $C(\mathbb{R}^d)$ converges uniformly on compact sets to another element f of $C(\mathbb{R}^d)$ if

$$\|f_n - f\|_K \to 0, \qquad \text{as } n \to \infty \text{ for every compact subset } K \text{ of } \mathbb{R}^d.$$

It is not immediately obvious that this notion of convergence is induced by a metric, but we claim that

$$d(f,g) = \sum_{n=1}^{\infty} \frac{2^{-n}\|f-g\|_{B_n}}{1 + \|f-g\|_{B_n}} \tag{13}$$

is a metric for $C(\mathbb{R}^d)$ and that convergence in this metric is uniform convergence on compact sets, where B_n denotes the closed ball centered at zero having radius n.

Clearly,

$$d(f,g) \geq \frac{2^{-n}\|f-g\|_{B_n}}{1 + \|f-g\|_{B_n}}$$

and we know the right-hand side is a metric for $C(B_n)$. Thus convergence in the metric (13) implies uniform convergence on any compact set contained in B_n. Since every compact set is contained in some B_n, that proves one direction of the equivalence of the two notions.

Conversely,

$$d(f,g) \leq 2^{-n} + \|f-g\|_{B_n}$$

for each n, because $x/(1+x)$ is an increasing function of x and the supremum over a larger set is larger, and this shows that convergence in the B_n norm for all n implies convergence in the metric (13), and that proves the other direction of the equivalence of the two notions.

Finally, let f_n be a Cauchy sequence in $C(\mathbb{R}^d)$ equipped with the metric (13) (and we will always give this space this metric). Then we know there is a function f such that f_n converges uniformly to f on every compact set

K by completeness of each $C(K)$, but this implies convergence in $C(\mathbb{R}^d)$. Hence $C(\mathbb{R}^d)$ is complete.

It is also separable. Again, polynomials with rational coefficients are a countable dense set, because $\|f - g\|_{B_n} < \varepsilon/2$ and $2^{-n} < \varepsilon/2$ imply $d(f, g) < \varepsilon$.

12.2 Continuous Convergence

Another kind of convergence of sequences of functions is called continuous convergence. This is defined as follows: f_n converges continuously to f, written $f_n \xrightarrow{c} f$ if for any convergent sequence $x_n \to x$ we have $f_n(x_n) \to f(x)$. It turns out that this is the same as uniform convergence on compact sets.

To show that uniform convergence on compact sets implies continuous convergence we use the triangle inequality. The whole sequence together with its limit point constitute a compact set K, Hence, by the triangle inequality,

$$\begin{aligned} |f_n(x_n) - f(x)| &\le |f_n(x_n) - f(x_n)| + |f(x_n) - f(x)| \\ &\le \|f_n - f\|_K + |f(x_n) - f(x)| \end{aligned}$$

and the first term on the right-hand side converges to zero by uniform convergence on compact sets and the second term on the right-hand side converges to zero by continuity of f.

Conversely, if f_n does not converge to f uniformly on compact sets, then there exists a compact set K and an $\varepsilon > 0$ such that $\|f_n - f\|_K > \varepsilon$ for all n. Hence there is an $x_n \in K$ such that $|f_n(x_n) - f(x_n)| > \varepsilon$ for all n. And because K is compact, this sequence has a convergent subsequence $x_{n_k} \to x$.

$$|f_n(x_n) - f(x_n)| \le |f_n(x_n) - f(x)| + |f(x) - f(x_n)|$$

so

$$|f_n(x_n) - f(x)| \ge |f_n(x_n) - f(x_n)| - |f(x) - f(x_n)|$$

and the first term on the right-hand side is greater than ε for all n and the second term on the right-hand side converges to zero by continuity of f. Thus we do not have continuous convergence either.

12.3 A Theorem from Optimization Theory

The following is a special case of Theorem 1.10 in Attouch (1984), which is also stated and used in Geyer (1994).

Theorem 7. *Suppose $f_n : \mathbb{R}^d \to \mathbb{R}$ is a sequence of continuous functions and $f : \mathbb{R}^d \to \mathbb{R}$ is a continuous function. Suppose x_n is a sequence in $\mathbb{R}^d$ and x is a point in $\mathbb{R}^d$. If*

$$f_n \xrightarrow{c} f$$

and

$$x_n \to x$$

and

$$f_n(x_n) - \inf_{y \in \mathbb{R}^d} f_n(y) \to 0,$$

then

$$f_n(x_n) \to f(x)$$

and

$$f(x) = \inf_{y \in \mathbb{R}^d} f(y).$$

12.4 The Analogous Theorem for Random Functions

We use Skorohod and Prohorov to prove the analogous weak convergence analog. But before that we need a lemma.

Lemma 8. *Marginal tightness implies joint tightness, that is, if X_n is a tight sequence of random elements of a Polish space S, and Y_n is a tight sequence of random elements of a Polish space T, then (X_n, Y_n) is a tight sequence of $S \times T$.*

In order to make $S \times T$ a metric space we equip it with the metric defined by

$$d_{S \times T}[(u, v), (x, y)] = d_S(u, x) + d_T(v, y) \tag{14}$$

where d_S and d_T are the metrics for S and T. Then it is clear that $x_n \to x$ and $y_n \to y$ implies $(x_n, y_n) \to (x, y)$ in $S \times T$, which in turn implies the product of closed sets is closed. It is also clear that the product of totally bounded sets is totally bounded (if every point of A is within $\varepsilon/2$ of a finite set C and every point of B is within $\varepsilon/2$ of a finite set D then every point of $A \times B$ is within ε of $C \times D$ when (14) is used at the metric for the product space).

Corollary 9. *Suppose F_n is a sequence of random elements of the Polish space $C(\mathbb{R}^d)$ of all continuous functions from $\mathbb{R}^d$ to $\mathbb{R}$ with the metric of uniform convergence on compact sets* (13) *and F is another random element* *of that Polish space having the property that F has a unique global minimizer almost surely. Suppose X_n is a sequence of random vectors, If*

$$F_n \xrightarrow{w} F$$

and

$$X_n \text{ is bounded in probability}$$

and

$$F_n(X_n) - \inf_{y \in \mathbb{R}^d} F_n(y) \xrightarrow{w} 0,$$

then

$$X_n \xrightarrow{w} X,$$

where X is another random vector, and

$$F_n(X_n) \xrightarrow{w} F(X)$$

and

$$F(X) = \inf_{y \in \mathbb{R}^d} F(y), \qquad \text{almost surely.}$$

Proof. Define

$$Y_n = F_n(X_n) - \inf_{y \in \mathbb{R}^d} F_n(y).$$

We are going to use convergence in the product space $C(\mathbb{R}^d) \times \mathbb{R}^d \times R$, for which we need to know that the product of Polish spaces is a Polish space (Fristedt and Gray, 1996, Proposition 2 of Chapter 18).

We are going to use the subsequence principle so let n_k index a subsequence. By the converse part of Prohorov's theorem F_{n_k} and Y_{n_k} are tight sequences, hence by assumption and Lemma 8 the sequence $(F_{n_k}, X_{n_k}, Y_{n_k})$ is tight, hence by the direct part of Prohorov's theorem there exists a subsubsequence with indices n_{k_l} such that

$$(F_{n_{k_l}}, X_{n_{k_l}}, Y_{n_{k_l}}) \xrightarrow{w} (F, X, 0) \tag{15a}$$

where this denotes weak convergence in the metric space $C(\mathbb{R}^d) \times \mathbb{R}^d \times \mathbb{R}$. By Skorohod's theorem there exists a sequence

$$(F^*_{n_{k_l}}, X^*_{n_{k_l}}, Y^*_{n_{k_l}}) \xrightarrow{\text{a.s.}} (F^*, X^*, 0) \tag{15b}$$

with the respective parts of (15a) and (15b) having the same laws. Applying Theorem 7 to this result we see that

$$F^*_{n_{k_l}}(X^*_{n_{k_l}}) \xrightarrow{\text{a.s.}} F^*(X^*)$$

and

$$F^*(X^*) = \inf_{y \in \mathbb{R}^d} F^*(y), \qquad \text{almost surely.}$$

By assumption the law of F^* is the same as the law of F, hence also by assumption F^* has a unique minimizer almost surely, and the law of X^* is thereby uniquely determined. Since almost sure convergence implies convergence in law,

$$F^*_{n_{k_l}}(X^*_{n_{k_l}}) \xrightarrow{w} F^*(X^*)$$

and since convergence in law only depends on laws not random elements

$$F_{n_{k_l}}(X_{n_{k_l}}) \xrightarrow{w} F^*(X^*).$$

Since this is true regardless of which subsequence was chosen, by the subsequence principle the whole sequence converges

$$F_n(X_n) \xrightarrow{w} F^*(X^*).$$

By a similar use of the subsequence principle we have $X_n \xrightarrow{w} X^*$. □

13 Product Spaces

We have already introduced product spaces. Here we state two useful theorems about them.

Theorem 10. *Suppose (X_n, Y_n) is a sequence of random elements of a Polish space $S \times T$, and suppose X_n and Y_n are stochastically independent for each n. If*

$$X_n \xrightarrow{w} X, \qquad \textit{in } S \tag{16a}$$

$$Y_n \xrightarrow{w} Y, \qquad \textit{in } T \tag{16b}$$

then

$$(X_n, Y_n) \to (X, Y), \qquad \textit{in } S \times T, \tag{16c}$$

where the right-hand side denotes the random element of $S \times T$ having stochastically independent components, the first of which has the law of the right-hand side of (16a) *and the second of which has the law of the right-hand side of* (16b).

This is Example 3.2 in Billingsley (1999), which follows directly from his Theorem 2.8. It is frequently used in statistics, but not often cited.

To introduce our second theorem about product spaces, we start with a non-theorem. It is not true, in general, that (16a) and (16b) imply (16c) without the independence assumptions of the theorem. In other words, marginal convergence in law does not imply joint convergence in law. Here is a counterexample. Let U and V be independent standard normal random variables and define

$$\begin{aligned} X_n &= U \\ Y_n &= (-1)^n \cdot \frac{1}{2} \cdot U + \sqrt{\frac{3}{4}} \cdot V \end{aligned}$$

Then X_n and Y_n are both standard normal for all n, and hence trivially converge in law marginally. But

$$\operatorname{cov}(X_n, Y_n) = \frac{(-1)^n}{2}$$

for all n so the sequence (X_n, Y_n) of random vectors cannot converge in law.

Thus we see that Theorem 10 does not go without saying. Theorem 13 in the following section gives a different condition from independence under which marginal convergence in law implies joint convergence in law.

14 Mapping and Slutsky Theorems

The following theorem is called the "mapping theorem" by Billingsley (1999, p. 20) and the "continuous mapping theorem" by others.

Theorem 11. *Suppose $h : S \to T$ is a map between Polish spaces and X_n is a sequence of random elements of S converging weakly to another random element X. The set D_h of points at which h is not continuous is a Borel subset of S, and* $\Pr(X \in D_h) = 0$ *implies* $h(X_n) \xrightarrow{w} h(X)$.

This is Theorem 2.7 in Billingsley (1999), except for the assertion about measurability which is shown in Appendix M10 in Billingsley (1999). This theorem specialized to S and T being finite-dimensional vector spaces is Theorem 6(a) in Ferguson (1996).

The following theorem is Theorem 3.1 in Billingsley (1999). Specialized to finite-dimensional vector spaces, it is Theorem 6(b) in Ferguson (1996).

Theorem 12. *Suppose S is a Polish space with metric d and (X_n, Y_n) are random elements of $S \times S$. Suppose $X_n \xrightarrow{w} X$ and $d(X_n, Y_n) \xrightarrow{w} 0$, then $Y_n \xrightarrow{w} X$.*

The Polish space generalization of Theorem 6(c) in Ferguson (1996) does not seem to be in Billingsley (1999), but it is easy to prove from Theorem 12, the proof in Ferguson (1996) carrying over to the Polish space case with almost no changes.

Theorem 13. *Suppose (X_n, Y_n) is a sequence of random elements of a Polish space $S \times T$. If* (16a) *and* (16b) *hold with Y a constant random variable, then* (16c) *holds.*

Proof. As in (14) let d_S, d_T, and $d_{S \times T}$ be the respective metrics. Note that

$$d_{S \times T}\bigl((X_n, Y_n), (X_n, Y)\bigr) = d_T(Y_n, Y) \xrightarrow{w} 0$$

by assumption. So by Theorem 12 it is sufficient to show $(X_n, Y) \xrightarrow{w} (X, Y)$. But this is immediate from the definition of weak convergence: for any bounded continuous function f on $S \times T$, the function $x \mapsto f(x, Y)$ is a bounded continuous function on S because Y is constant. □

Theorem 14. *If X_n is a sequence of random elements of a Polish space and X is a constant random element of the same Polish space, then $X_n \xrightarrow{P} X$ if and only if $X_n \xrightarrow{w} X$.*

(See p. 27 in Billingsley (1999) for the trivial argument.) Thus Theorem 13 is often stated with $Y_n \xrightarrow{P} Y$ as the condition rather than $Y_n \xrightarrow{w} Y$, but the two conditions are the same.

Thus Theorem 13 is sometimes called Slutsky's lemma because when specialized to X_n and Y_n being sequences of random variables, the usual conclusions of Slutsky's theorem

$$\begin{aligned}
X_n + Y_n &\xrightarrow{\mathcal{L}} X + Y \\
X_n - Y_n &\xrightarrow{\mathcal{L}} X - Y \\
X_n * Y_n &\xrightarrow{\mathcal{L}} X * Y \\
X_n / Y_n &\xrightarrow{\mathcal{L}} X / Y, \qquad \text{provided } Y \neq 0
\end{aligned}$$

follow immediately from Theorems 11 and 13.

15 Uniform Integrability

This section is a bit out of place in this document because it doesn't involve metric spaces, but it isn't big enough for a document of its own. In Section 11 we used the Skorohod theorem to prove the following.

Theorem 15. *Suppose X_n is a sequence of random variables converging weakly to X. Then*

$$E\{|X|\} \leq \liminf_{n \to \infty} E\{|X_n|\}$$

This can also proved by other methods (Billingsley, 1999, Theorem 3.4).

The converse implication is not true in general but is true under certain conditions. A sequence X_n of random variables is *uniformly integrable* if

$$\lim_{c \to \infty} \sup_{n \in \mathbb{N}} E\{I_{(c,\infty)}(|X_n|)|X_n|\} = 0.$$

Theorem 16. *Suppose X_n is a uniformly integrable sequence of random variables converging weakly to X. Then*

$$X_n \xrightarrow{\mathcal{L}} X \quad \textit{implies} \quad E(X_n) \to E(X)$$

(Billingsley, 1999, Theorem 3.5).

References

Attouch, H. (1984). *Variational Convergence of Functions and Operators.* Pitman, Boston.

Beattie, R. and H.-P. Butzmann (2002). *Convergence Structures and Applications to Functional Analysis.* Dordrecht: Kluwer Academic Publishers.

Billingsley, P. (1999). *Convergence of Probability Measures*, second edition. New York: Wiley.

Browder, A. (1996). *Mathematical Analysis: An Introduction.* New York: Springer.

Ferguson, T. S. (1996). *A Course in Large Sample Theory.* London: Chapman & Hall.

Fristedt, B. E. and Gray, L. F. (1996). *A Modern Approach to Probability Theory.* Boston: Birkhäuser.

Geyer, C. J. (1994). On the asymptotics of constrained M-estimation. *Annals of Statistics*, **22**, 1993–2010.