Advanced Probability

==
(2nd November, Friday)

Chapter 5. Weak Convergence

5.1. Definitions

Let E be a metric space. Whenever we are talking about a metric space, the σ-algebra is given by the Borel σ-algebra. Write $C_b(E)$ for the set of bounded continuous functions on E.

- Let $(\mu_n : n \in \mathbb{N})$ be a sequence of probability measures and let μ be another probability measure on E. We say that $\mu_n \to \mu$ **weakly** (as $n \to \infty$) if $\mu_n(f) \to \mu(f)$ for all $f \in C_b(\mathbb{R})$.

Theorem 5.1.1) The following are equivalent.

(a) $\mu_n \to \mu$ weakly on E

(b) $\liminf_{n\to\infty} \mu_n(U) \geq \mu(U)$ for all U open

(c) $\limsup_{\mu(F)} \leq \mu(F)$ for all F closed.

(d) $\mu_n(B) \to \mu(B)$ for all $B \in \mathcal{B}$ such that $\mu(\partial B) = 0$.(Boundary is the set of limit points of B that are not contained in B.)

proof) Exercise.

For an example, consider a sequence $(x_n)_n \subset \mathbb{R}$ such that $x_n \to 0$ as $n \to \infty$. We want to have $\delta_{x_n} \to \delta_0$. Indeed, this is true in the weak sense. However, the sequence has $\delta_{x_n}(\{0\}) = 0$ for all n, hence we should have inequality in condition (c).

We have a similar version of the theorem for the real line.

Proposition 5.1.2) Consider the case $E = \mathbb{R}$. TFAE

(a) $\mu_n \to \mu$ weakly for some probability measure μ.

(b) $F_n(x) \to F(x)$ for all $x \in \mathbb{R}$ such that $F(x^-) = F(x)$. (Here, $F(x) = \mu((\infty, x])$ is the **distribution function** of μ.) (Sometimes called convergence of distributions)

(c) There exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and random variables X_n, X on Ω such that $X_n \sim \mu_n$, $X \sim \mu$ and $X_n \to X$ almost surely.

proof) See probability and measure notes.

5.2. Prohorov's Theorem

When does a sequence of probability measures has a converging subsequence?

Let E be a metric space and $(\mu_n : n \in \mathbb{N})$ be a sequence of probability measures on E.

- We say that $(\mu_n)_n$ is **tight** if for all $\epsilon > 0$, there is a compact set $K \subset E$ such that

$$\mu_n(E \backslash K) \leq \epsilon \quad \forall n \in \mathbb{N}$$

  For example, the sequence $(\delta_n)_n$ is *not* tight.

Theorem 5.2.1) Let $(\mu_n : n \in \mathbb{N})$ be a sequence of probability measures on a metric space E and suppose that $(\mu_n : n \in \mathbb{N})$ is tight. Then there exists a subsequence $(n_k)_k \subset \mathbb{N}$ and probability measure μ on E such that $\mu_{n_k} \to \mu$ weakly as $k \to \infty$.

This gives a version of weakly sequential compactness of probability measures. We are only going to prove this for $\mathbb{R}$. This theorem is hard to prove in general.(e.g. there is a method using Monge-Kantorovich metric defined for Polish spaces. For this method, see "Topics in Optimal Transport", C.Villani, Ame.Soc.Math. For the general version, see the attached note)

proof for $E = \mathbb{R}$) By a diagonal argument and by passing to a subsequence, it suffices to consider the case where $F_n(x) \to g(x)$ as $n \to \infty$ for all $x \in \mathbb{Q}$ for some $g(x) \in [0,1]$, where F_n is the distribution function of F_n. Now $g : \mathbb{Q} \to [0,1]$ is non-decreasing so g has a non-decreasing extension $G : \mathbb{R} \to [0,1]$, i.e.

$$G(x) = \lim_{q \searrow x, q \in \mathbb{Q}} g(q)$$

which has only countably many discontinuities.(because there should be a rational number in each discontinuity). Now we must have

$$F_n(x) \to G(x) \quad \forall x \text{ s.t. } G\text{is continuous at } x$$

Set $F(x) = G(x^+)$, then F and G have same points of continuity, so $F_n(x) \to F(x)$ for all $x \in \mathbb{R}$.

We are only left to check that $G(x) \to 1$ as $x \to \infty$ using tightness condition.

Since $(\mu_n : n \in \mathbb{N})$ is tight, given $\epsilon > 0$, there exists $R < \infty$ such that $\mu_n(\mathbb{R} \backslash (-R, R)) \leq \epsilon$ for all ϵ so $F_n(-R) \leq \epsilon$, $F_n(R) \geq 1 - \epsilon$. So

$$\begin{aligned} F(x) &\to 0 \quad \text{as } x \to -\infty \\ F(x) &\to 1 \quad \text{as } x \to \infty \end{aligned}$$

So F is distribution function. So there exists a probability measure μ such that $\mu((-\infty, x]) = F(x)$. Then $\mu_n \to \mu$ by **Prop 5.1.2.**

(End of proof) □

5.3. Weak Convergence and Characteristic Functions

Take $E = \mathbb{R}^d$.

- For a probability measures *mu* on $\mathbb{R}^d$, define its **characteristic function** $\phi : \mathbb{R}^d \to \mathbb{C}$ by

$$\phi(u) = \int_{\mathbb{R}^d} e^{i\langle u, x \rangle} \mu(dx)$$

Lemma 5.3.1) Fix $d = 1$. For all $\lambda \in (0, \infty)$,

$$\mu(\mathbb{R} \backslash (-\lambda, \lambda)) \leq C\lambda \int_0^\lambda (1 - \text{Re}(\phi(u)))du$$

where $C = (1 - \sin(1))^{-1} < \infty$.

proof) Consider for $t \geq 1$. Let $A(t) = t^{-1}\int_0^t (1-\cos v)dv$. Then

$$A(t) \geq A(0) = 1 - \sin(t)$$

(to see this, observe that $A(t)$ is the average of $(1-\cos(v))$ on interval $(0,t)$ and divide the cases $|t| \leq \pi/2$ and $|t| \geq \pi/2$)
So $Ct^{-1}\int_0^t (1-\cos(v))dv \geq 1$. Substitute $v = uy$, $u = v/y$,

$$Ct^{-1}\int_0^{t/y} (1-\cos(uy))ydu \geq 1$$

Put $t/y = 1/\lambda$, $\lambda = y/t$, $t = y/\lambda \geq 1$ to see

$$C\lambda \int_0^{1/\lambda} (1-\cos(uy))du \geq 1$$

whenever $t = y/\lambda \geq 1$(this was the assumption we started with). Now for general $y \in \mathbb{R}$, has

$$C\lambda \int_0^{1/\lambda} (1-\cos(uy))du \geq 1_{|y|\geq\lambda}$$

Now integrate with respect to μ and use Fubini.

$$\begin{aligned}
\mu(\mathbb{R}\backslash(-\lambda,\lambda)) &\leq C\lambda \int_{\mathbb{R}}\int_0^{1/\lambda} (1-\cos(uy))du\mu(dy) \\
&= C\lambda \int_0^{1/\lambda}\int_{\mathbb{R}} (1-\cos(uy))du\mu(dy)
\end{aligned}$$

(End of proof) $\square$

==
(5th November, Monday)

Theorem 5.3.2) Let μ_n, μ be probability measures on $\mathbb{R}^d$ with characteristic functions ϕ_n, ϕ. Then the following are equivalent

(a) $\mu_n \to \mu$ weakly on $\mathbb{R}^d$.

(b) $\phi_n(u) \to \phi(u)$ for all $u \in \mathbb{R}^d$.

We will prove only for the case $d = 1$.

proof) It is clear that (a) implies (b). Suppose (b) holds. We prove via a 'compactness argument'. We aim to show that the sequence $(\mu_n)_n$ tight, and therefore has a converging subsequence, and show that the converging point is in fact μ.

Note that $\phi(0) = 1$ and ϕ is continuous. Given $\epsilon > 0$, there exists $\lambda < \infty$ such that

$$C\lambda \int_0^{1/\lambda} (1 - \mathrm{Re}(\phi(u)))du \leq \epsilon/2$$

with $C = (1-\sin(1))^{-1} < \infty$. By dominated convergence,

$$\int_0^{1/\lambda} (1-\mathrm{Re}(\phi_n(u)))du \xrightarrow{n\to\infty} \int_0^{1/\lambda} (1-\mathrm{Re}(\phi(u)))du$$

so for sufficiently large n, by **Lemma 5.3.1,**,

$$\mu_n(\mathbb{R}\backslash(-\lambda,\lambda)) \leq C\lambda \int_0^{1/\lambda} (1-\mathrm{Re}(\phi_n(u)))du \leq \epsilon$$

Since ϵ was arbitrary, we see that $(\mu_n : n \in \mathbb{N})$ is tight. By Prohorov's theorem, we have a converging subsequence $\mu_{n_k} \to \nu$ for some probability measure ν.

Suppose for a contradiction that $\nu \neq \mu$. Therefore, there exists $\epsilon > 0$, and $f \in C_b(\mathbb{R}^n)$ such that

$$|\mu_{n_k}(f) - \mu(f)| \geq \epsilon \quad \forall k$$

By above argument, we have $\mu_{n_k} \to \nu$. But then, since e^{inx} is a bounded continuous function,

$$\int_{\mathbb{R}} e^{inx}\nu(dx) = \lim_{k\to\infty} \phi_{n_k}(n) = \phi(n)$$

which indicates $\mu = \nu$ by uniqueness of characteristic functions (see PM notes), a contradiction.

(End of proof) □

In fact, the proof of the theorem implies a slightly stronger statement, which is less useful.

Theorem 5.3.3) *(Lévy's continuity theorem for characteristic functions)* Let $(\mu_n : n \in \mathbb{N})$ be a sequence of probability measures on $\mathbb{R}^n$ with characteristic functions ϕ_n. Suppose $\phi_n(u) \to \phi(u)$ for all u for some function ϕ (not necessarily a characteristic function) such that ϕ is continuous at 0. Then ϕ is the characteristic function of some probability measure μ on $\mathbb{R}^d$ and $\mu_n \to \mu$ weakly on $\mathbb{R}^d$.

6. Large Deviations

6.1. Cramérs theorem

Theorem 6.1.1) Let $(X_n : n \in \mathbb{N})$ be a sequence of integrable *i.i.d.* random variables in $\mathbb{R}$. Set $m = \mathbb{E}(X_1)$, $S_n = X_1 + \cdots + X_n$. We know $S_n/n \to \delta_m$ in probability, so if $(m-\epsilon, m+\epsilon) \cap B = \phi$ then $\mathbb{P}(S_n/n \in B) \to 0$ as $n \to \infty$. Then in fact the convergence rate is given by $\sim \exp(-n\alpha(B))$ for some α. To be precise, for all $a \geq m = \mathbb{E}(X_1)$, as $n \to \infty$,

$$\frac{1}{n} \log \mathbb{P}(S_n \geq na) \to -\psi^*(a)$$

where ψ^* is the *Legendre transform* of the *cumulant generating function* $\psi(\lambda) = \log(\mathbb{E}(e^{\lambda X_1}))$, where Legendre transform is given by

$$\psi^*(x) = \sup_{\lambda\in\mathbb{R}}\{\lambda x - \psi(\lambda)\}$$

In particular, for n sufficiently large and in case $\psi^*(a) < \infty$, we get

$$-\psi^*(a) - \epsilon \leq \frac{1}{n}\log(\mathbb{P}(S_n \geq a)) \leq -\psi^*(a) + \epsilon$$

and therefore

$$e^{-n(\psi^*(a)+\epsilon)} \leq \mathbb{P}(S_n \geq na) \leq e^{-n(\psi^*(a)-\epsilon)}.$$

Note : ψ is always a convex function, so ψ^* is also a convex function.

Examples :

(i) $X_1 \sim N(0,1)$, then $\mathbb{E}(e^{\lambda X_1}) = e^{\lambda^2/2}$, $\psi(\lambda) = \lambda^2/2$ and $\psi^*(x) = x^2/2$. Hence

$$\frac{1}{n}\log(\mathbb{P}(S_n \geq a)) \to -\frac{a^2}{2} \quad \forall a \geq 0$$

Can check this directly, using the fact that $S_n \sim N(0,n)$ in this case.

(ii) $X_1 \sim \text{Exp}(1)$, then

$$\mathbb{E}(e^{\lambda X_1}) = \int_0^\infty e^{\lambda x}e^{-x}dx = \begin{cases} \infty & \text{if } \lambda \geq 1 \\ \frac{1}{1-\lambda} & \text{if } \lambda < 1 \end{cases}$$

so $\psi(\lambda) = -\log(1-\lambda)$ if $\lambda < 1$ and ∞ otherwise, and $\psi^*(x) = x - 1 - \log(x)$ for $x > 0$. Cramér's theorem implies that

$$\frac{1}{n} \log \mathbb{P}(S_n \geq na) \to -(a - 1 - \log(a)) \quad \forall a \geq 1$$

On the other hand, $\mathrm{Var}(X_1) = 1 < \infty$, so $\frac{S_n - n}{\sqrt{n}} \to N(0,1)$ by CLT. So

$$\mathbb{P}(S_n \geq n + a\sqrt{n}) \to \int_a^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

so Cramér's theorem gives a result of a different flavour from CLT for distributions with bounded variation : while CLT provides a description for distribution near the average, Cramér gives an explanation of tail distribution of S_n.

preparation for proof of Cramér's theorem) Let $\mu(B) = \mathbb{P}(X_1 \in B)$. Exclude the easy case where $\mu = \delta_m$. Define for $\lambda \geq 0$ with $\psi(\lambda) < \infty$, the **tilted distribution** μ_λ by

$$\mu_\lambda(dx) \propto e^{\lambda x} \mu(dx)$$

For $K \geq m = \mathbb{E}(X_1)$, define the conditional distribution by

$$\mu_K(dx | x \leq K) \propto 1_{\{x \leq K\}} \mu(dx)$$

The CGF(cumulant generating function) of μ_K is then given by

$$\psi_K(\lambda) = \log(\mathbb{E}(e^{\lambda X_1} | X_1 \leq K))$$

==

(7th November, Wednesday)

We now start proving the following theorem.

Theorem 6.1.1) Let $(X_n : n \in \mathbb{N})$ be a sequence of integrable *i.i.d.* random variables in $\mathbb{R}$. Set $m = \mathbb{E}(X_1)$, $S_n = X_1 + \cdots + X_n$. Then for all $a \geq m = \mathbb{E}(X_1)$, as $n \to \infty$,

$$\frac{1}{n} \log \mathbb{P}(S_n \geq na) \to -\psi^*(a)$$

where $\psi(\lambda) = \log(\mathbb{E}(e^{\lambda X_1}))$, and $\psi^*(x) = \sup_{\lambda \in \mathbb{R}} \{\lambda x - \psi(\lambda)\}$.

proof) *(Upper bound)* For all $\lambda \geq 0$ and $n \geq 1$

$$\mathbb{P}(S_n \geq na) = \mathbb{P}(e^{\lambda S_n} \geq e^{\lambda na}) \leq e^{-\lambda na} \mathbb{E}(e^{\lambda S_n}) = e^{-(\lambda a - \psi(\lambda))n}$$

so $\frac{1}{n} \log \mathbb{P}(S_n \geq na) \leq -(\lambda a - \psi(\lambda))$ and

$$\frac{1}{n} \log \mathbb{P}(S_n \geq na) \leq -\psi^*(a)$$

(Lower bound) It remains to show the lower bound. That is, we aim to prove

$$\liminf_{n \to \infty} \frac{1}{n} \log \mathbb{P}(S_n \geq na) \geq -\psi^*(a)$$

Consider first the case where $\mathbb{P}(X_1 \leq a) = 1$. Then

$$\mathbb{E}(e^{\lambda(X_1 - a)}) \xrightarrow{\lambda \to \infty} \mathbb{P}(X_1 = a)$$

Call $p = \mathbb{P}(X_1 = a)$, so $\lambda a - \psi(\lambda) \to -\log(p)$. So in particular,

$$\psi^*(a) \geq -\log(p)$$

Now $\mathbb{P}(S_n \geq na) = p^n$ so

$$\frac{1}{n}\log \mathbb{P}(S_n \geq na) = \log(p) \geq -\psi^*(a)$$

hence we can eliminate the case $\mathbb{P}(X_1 \leq a) = 1$.

Next consider the case $\psi(\lambda) < \infty$ for all $\lambda \geq 0$ and $\mathbb{P}(X_1 > a) > 0$. Fix $\epsilon > 0$ and set $b = a + \epsilon$, $c = a + 2\epsilon$, choosing ϵ small enough so $\mathbb{P}(X_1 > b) > 0$. Then there exists λ such that $\psi'(\lambda) = b$ - where the differentiability and the existence is justified in the following proposition :

Proposition 6.1.2) Suppose X is integrable and not a.s. constant. Then

$$\begin{aligned} \psi_K(\lambda) = \log \mathbb{E}(e^{\lambda X_1}|X_1 \leq K) < \infty \quad \forall K < \infty \\ \textit{and} \quad \psi_K(\lambda) \nearrow \psi(\lambda) \quad \text{as } K \to \infty \end{aligned}$$

Moreover in the case $\psi(\lambda) < \infty$ for all $\lambda \geq 0$, ψ has a continuous derivative on $[0, \infty)$ and is C^2 on $(0, \infty)$ with

$$\begin{aligned} \psi'(\lambda) &= \int_{\mathbb{R}} x\mu_\lambda(dx) \\ \psi''(\lambda) &= \mathrm{Var}(\mu_\lambda) > 0 \end{aligned}$$

and ψ' is a homeomorphism from $[0, \infty)$ to $[m, \sup(\mathrm{supp}(\mu))$.

proof) (Exercise)

Now we use the idea of tilting the probability measure. Define a new probability measure $\mathbb{P}_\lambda$ by $d\mathbb{P}_\lambda = e^{\lambda S_n - n\psi(\lambda)}d\mathbb{P}$. Then observe that under $\mathbb{P}_\lambda$ the random variables $X_1, \cdots, X_n$ are independent with distributions μ_λ and that $\mathbb{E}_\lambda(X_1) = b$. Consider the event

$$A_n = \{\left|\frac{S_n}{n} - b\right| \leq \epsilon\} = \{(b-\epsilon)n = an \leq S_n \leq (b+\epsilon)n = cn\}$$

By the weak law of large numbers, $\mathbb{P}_\lambda(A_n) \to 1$. So

$$\begin{aligned} \mathbb{P}(S_n \geq na) \geq \mathbb{P}(A_n) &= \mathbb{E}_\lambda\Big(1_{A_n} e^{-\lambda S_n + \psi(\lambda)n}\Big) \\ &\geq e^{-\lambda cn + \psi(\lambda)n}\mathbb{P}_\lambda(A_n) \end{aligned}$$

So

$$\frac{1}{n}\log \mathbb{P}(S_n \geq na) \geq -\lambda c + \psi(\lambda) + \frac{\log(\mathbb{P}_\lambda(A_n))}{n}$$

and

$$\liminf_{n\to\infty} \frac{1}{n}\log \mathbb{P}(S_n \geq na) \geq -(\lambda c - \psi(\lambda)) \geq -\psi^*(c)$$

Now ψ^* is continuous at a (recall, ψ^* is a Legendre transform of a convex function so is convex, and therefore continuous. Or, see **Lemma 6.1.3**) and $\epsilon > 0$ is arbitrary so the desired lower bound follows on letting $\epsilon \to 0$.

Finally, consider the general case $\mathbb{P}(X_1 > a) > 0$ but allowing $\psi(\lambda) = \infty$ for some $\lambda \geq 0$. For $K > a$, we have $\mathbb{P}(X_1 > a|X_1 \leq K) > 0$ and $\psi_K(\lambda) < \infty$ for all $\lambda \geq 0$. So preceding argument shows

$$\liminf_{n\to\infty} \frac{1}{n}\log \mathbb{P}_K(S_n > na) \geq -\psi_K^*(a)$$

where $\mathbb{P}_K$ is the probability measure given by

$$d\mathbb{P}_K^{(n)} \propto 1_{\{X_1 \leq K, \cdots, X_n \leq K\}}d\mathbb{P}$$

(To see this, note, under $\mathbb{P}_K$, random variables $X_1, \cdots X_n$ are independent with distribution $\mu(\cdot|x \leq K)$). But

$$\mathbb{P}(S_n \geq na) \geq \mathbb{P}(S_n \geq na|X_1 \leq K, \cdots, X_n \leq K) = \mathbb{P}_K(S_n \geq na)$$

and $\psi_K^*(a) \searrow \psi^*(a)$ as $K \to \infty$ (by **Lemma 6.1.3**) so we see

$$\liminf_{n\to\infty} \frac{1}{n}\log \mathbb{P}(S_n \geq na) \geq -\psi_K^*(a) \nearrow -\psi^*(a)$$

(End of proof) $\square$

One different way to see that ψ^* is continuous at a is presented in the following lemma.

Lemma 6.1.3) For all $a \geq m$, with $\mathbb{P}(X_1 > 0) > 0$ we have $\psi_K^*(a) \searrow \psi^*(a)$ as $K \to \infty$. Moreover in the case $\psi(\lambda) < \infty$ for all $\lambda \geq 0$, ψ^* is continuous at a and we have $\psi^*(a) = \lambda^* a - \psi(\lambda^*)$ where λ^* is uniquely determined by $\psi'(\lambda^*) = a$.

proof) Consider first the later case where $\psi(\lambda) < \infty$ for $\lambda \geq 0$. Then by **Proposition 6.1.2** wee see that

$$\psi^*(a) = \lambda^* a - \psi(\lambda^*)$$

where $a = \psi'(\lambda^*)$ and ψ^* is continuous at a with $\lambda^* = (\psi')^{-1}(a)$.

For the first part, note that ψ_K^* is non-increasing in K. For K sufficiently large, we have

$$\mathbb{P}(X_1 > a | X_1 \leq K) > 0$$

and $a \geq m \geq m_K$ (where $m_K = \mathbb{E}(X_1 | \leq X_1 \leq K)$) and $\psi_K(\lambda) < \infty$ for all $\lambda \geq 0$, so we may apply the preceding argument to μ_K to see that

$$\psi_K^*(a) = \lambda_K^* a - \psi_K(\lambda_K^*)$$

where $\lambda_K^* \geq 0$ is determined by $\psi_K'(\lambda_K^*) = a$. Now $\psi_K'(\lambda)$ is non-decreasing in K and λ, so $\lambda_K^* \searrow \lambda^*$ for some $\lambda^* \geq 0$. Also $\psi_K'(\lambda) \geq m_K$ for all $\lambda \geq 0$ so

$$\psi_K(\lambda_K^*) \geq \psi_K(\lambda^*) + m_K(\lambda_K^* - \lambda^*)$$

Then

$$\psi_K^*(a) = \lambda_K^* a - \psi_K(\lambda_K^*) \leq \lambda_K^* a - \psi_K(\lambda^*) - m_K(\lambda_K^* - \lambda^*) \to \lambda^* a - \psi(\lambda^*) \leq \psi^*(a)$$

So $\psi_K^*(a) \searrow \psi^*(a)$ as $K \to \infty$ as claimed.

(End of proof) $\square$