Section Notes for PSTAT 213B

Haosheng Zhou

Jan, 2024

Contents

Week 1

Readers shall have a fundamental understanding in measure theory before taking this course, i.e. be familiar with concepts like $\pi - \lambda$ theorems, convergence theorems interchanging integrals and limits, L^p spaces, Radon-Nikodym derivatives etc. We will use those results from measure theory without providing any proofs.

Convergence Modes

The convergence modes we have learnt in the first week include almost sure convergence, convergence in probability, convergence in distribution and L^p convergence. The key takeaway here is the definitions of different convergence modes and the connection between them.

One of the mathematical perspective we can take to view those convergence modes is that if the convergence mode can be induced by a metric. That is to say, if there exists some metric (distance function) d on the space of certain random variables such that the convergence under d is exactly same as the convergence mode defined in the probabilistic setting.

Lemma 1 (L^p norm)**.** *Let $p \geq 1$, define $\left\|X\right\|_p = (\mathbb{E}|X^p|)^{\frac{1}{p}}$, show that $\left\|\cdot\right\|_p$ is a norm on the space of L^p random variables with the equality to be understood in the almost sure sense.*

Proof. Clearly $\forall c \in \mathbb{R}, \left\|cX\right\|_p = |c| \left\|X\right\|_p$ satisfies homogeneity. If $X = 0$ *a.s.* then $\left\|X\right\|_p = 0$. If $\left\|X\right\|_p = 0$, then $\mathbb{E}|X^p| = 0$ with $|X^p| \geq 0$ *a.s.* so $|X^p| = 0$ *a.s.*, and $X = 0$ *a.s.*

Finally, we prove the triangle inequality of this norm

$$\begin{aligned}
\left\|X+Y\right\|_p^p &= \mathbb{E}|X+Y| \cdot |X+Y|^{p-1} && (1)\\
&\leq \mathbb{E}|X| \cdot |X+Y|^{p-1} + \mathbb{E}|Y| \cdot |X+Y|^{p-1} && (2)\\
&= \left\||X| \cdot |X+Y|^{p-1}\right\|_1 + \left\||Y| \cdot |X+Y|^{p-1}\right\|_1 && (3)\\
&\leq \left\|X\right\|_p \left\||X+Y|^{p-1}\right\|_q + \left\|Y\right\|_p \left\||X+Y|^{p-1}\right\|_q && (4)
\end{aligned}$$

where we used Holder's inequality for Holder conjugate p, q such that $\frac{1}{p} + \frac{1}{q} = 1$. It's thus clear that $q = \frac{p}{p-1}$ and $\left\||X+Y|^{p-1}\right\|_q = (\mathbb{E}|X+Y|^p)^{\frac{p-1}{p}} = \left\|X+Y\right\|_p^{p-1}$, plug into the inequality above

$$\left\|X+Y\right\|_p^p \leq (\left\|X\right\|_p + \left\|Y\right\|_p) \cdot \left\|X+Y\right\|_p^{p-1} \tag{5}$$

proves the Minkowski inequality

$$\left\|X+Y\right\|_p \leq \left\|X\right\|_p + \left\|Y\right\|_p \tag{6}$$

and we argued that $\left\|\cdot\right\|_p$ is a norm under almost sure sense. □

Lemma 2 (Property of L^p convergence)**.** *1. Prove that $X_n \overset{L^p}{\to} X$ $(n \to \infty)$ implies the convergence of p-th moment $\mathbb{E}|X_n|^p \to \mathbb{E}|X|^p$ $(n \to \infty)$ for $p \geq 1$.*

2. *Suppose $X_n \xrightarrow{L^1} X$ $(n \to \infty)$, show that $\mathbb{E}X_n \to \mathbb{E}X$ $(n \to \infty)$. Is the converse true?*
3. *Suppose $X_n \xrightarrow{L^2} X$ $(n \to \infty)$, show that $Var(X_n) \to Var(X)$ $(n \to \infty)$.*

Proof. The first proof comes from Minkowski inequality of L^p norm proved above that $\|X_n\|_p \leq \|X_n - X\|_p + \|X\|_p$ so $|\,\|X_n\|_p - \|X\|_p\,| \leq \|X_n - X\|_p$. L^p convergence is equivalent to saying $\|X_n - X\|_p \to 0$, so $\|X_n\|_p \to \|X\|_p$ concludes the proof.

When the convergence is L^1, it's easy to see that $X_n^+ \xrightarrow{L^1} X^+$ $(n \to \infty)$. This is because

$$\mathbb{E}|X_n^+ - X^+| \leq \mathbb{E}|X_n - X| \to 0 \ (n \to \infty) \tag{7}$$

where $X^+ = \max\{X, 0\}$ is the positive part of X and $X^- = \max\{-X, 0\}$ is the negative part of X. Both the positive and negative parts are non-negative random variables. Apply the result proved above,

$$\mathbb{E}|X_n| \to \mathbb{E}|X|, \mathbb{E}X_n^+ \to \mathbb{E}X^+ \ (n \to \infty) \tag{8}$$

since $\mathbb{E}|X_n| = \mathbb{E}X_n^+ + \mathbb{E}X_n^-, \mathbb{E}X_n = \mathbb{E}X_n^+ - \mathbb{E}X_n^-$, it's clear that $\mathbb{E}X_n = 2\mathbb{E}X_n^+ - \mathbb{E}|X_n| \to 2\mathbb{E}X^+ - \mathbb{E}|X| = \mathbb{E}X$ $(n \to \infty)$. **Refer to the remark below for a much easier proof!**

However, the converse is not true. The counterexample can be constructed on the probability space $([0,1], \mathscr{B}_{[0,1]}, \lambda)$ with λ to be the Lebesgue measure. Set $X_n = n\mathbb{I}_{[0,\frac{1}{n}]}$ so $\forall n, \mathbb{E}X_n = n\frac{1}{n} = 1$ converges to 1 but X_n does not converge in L^1. To see this fact, we first observe that $X_n \xrightarrow{p} 0$ $(n \to \infty)$, since L^1 convergence implies convergence in probability and the limit under the convergence in probability is unique, the L^1 limit, if exists, must be 0. Let's check

$$\mathbb{E}|X_n - 0| = 1 \not\to 0 \tag{9}$$

proves that X_n does not converge in L^1. Actually the convergence of L^p norm and L^p convergence are equivalent under the uniform integrability condition shown by Vitali convergence theorem which we shall learn in the future.

When the convergence is L^2, from the conclusion proved above, $\mathbb{E}X_n^2 \to \mathbb{E}X^2$. Since L^2 convergence implies L^1 convergence, it's also true that $\mathbb{E}X_n \to \mathbb{E}X$, as a result, $Var(X_n) = \mathbb{E}X_n^2 - (\mathbb{E}X_n)^2 \to \mathbb{E}X^2 - (\mathbb{E}X)^2 = Var(X)$ $(n \to \infty)$.

□

Remark. *There is a much easier way to argue $X_n \xrightarrow{L^1} X$ $(n \to \infty)$ implies $\mathbb{E}X_n \to \mathbb{E}X$ $(n \to \infty)$ that from Jensen's inequality, since $|x|$ is convex,*

$$|\mathbb{E}X_n - \mathbb{E}X| \leq \mathbb{E}|X_n - X| \to 0 \ (n \to \infty) \tag{10}$$

I want to thank Sam for reminding me that.

Remark. *L^q convergence implies L^p convergence for $q > p$. Firstly, check that $\|X\|_q < \infty$ implies $\|X\|_p < \infty$*

through a simple application of Holder's inequality

$$\|X\|_p^p = \||X|^p\|_1 \leq \||X|^p\|_{\frac{q}{p}} \cdot \|1\|_{\frac{q}{q-p}} = \|X\|_q^p \tag{11}$$

with $\frac{p}{q} + \frac{q-p}{q} = 1$ *so* $\|X\|_p \leq \|X\|_q$. *Replace* X *with* $X_n - X$ *to see that* L^q *convergence implies* L^p *convergence.*

It's clear that L^p ***convergence is metric-induced****, the metric is induced by the norm that* $d(X, Y) = \|X - Y\|_p$.

Lemma 3 (Levy metric)**.** *For two distribution functions* F, G*, define*

$$d(F, G) = \inf\{\delta > 0 : \forall x \in \mathbb{R}, F(x - \delta) - \delta \leq G(x) \leq F(x + \delta) + \delta\} \tag{12}$$

show that d *defines a metric on the space of distribution functions (d.f.).*

Proof. Obviously for any F, G, $d(F, G) \geq 0$. First prove it's symmetric. If $\delta > 0$ is such that $\forall x \in \mathbb{R}, F(x-\delta) - \delta \leq G(x) \leq F(x+\delta)+\delta$, then set $x = y+\delta$ to see $\forall y \in \mathbb{R}, F(y) \leq G(y+\delta)+\delta$, set $x = z-\delta$ to see $\forall z \in \mathbb{R}, G(z-\delta)-\delta \leq F(z)$. Merge those two inequalities to see that such $\delta > 0$ satisfies $\forall x \in \mathbb{R}, G(x - \delta) - \delta \leq F(x) \leq G(x + \delta) + \delta$. Actually the fact holds vice versa. Through a same argument, one knows

$$\{\delta > 0 : \forall x \in \mathbb{R}, F(x - \delta) - \delta \leq G(x) \leq F(x + \delta) + \delta\} = \{\delta > 0 : \forall x \in \mathbb{R}, G(x - \delta) - \delta \leq F(x) \leq G(x + \delta) + \delta\} \tag{13}$$

taking inf on both sides gives $d(F, G) = d(G, F)$.

If $d(F, G) = 0$, it means that

$$\exists \delta_n \to 0 \ (n \to \infty), \forall x \in \mathbb{R}, \forall n, \delta_n > 0, G(x) \leq F(x + \delta_n) + \delta_n \tag{14}$$

set $n \to \infty$, due to right-continuity of d.f. F, $F(x + \delta_n) + \delta_n \to F(x)$ proves $\forall x \in \mathbb{R}, G(x) \leq F(x)$. Interchange the position of F, G, from the symmetricity of d, $\forall x \in \mathbb{R}, F(x) \leq G(x)$ holds. Hence $d(F, G) = 0$ implies $F = G$.

Finally we prove the triangle inequality. Denote $d(F, G) = a, d(G, H) = b$, we want to prove $d(F, H) \leq a + b$, it suffices to prove that

$$\forall x \in \mathbb{R}, F(x - a - b) - a - b \leq H(x) \leq F(x + a + b) + a + b \tag{15}$$

from $d(F, G) = a$ it's clear that

$$\exists \eta_n \to a \ (n \to \infty), \forall x \in \mathbb{R}, \forall n, a < \eta_n < a + \frac{1}{n}, F(x - \eta_n) - \eta_n \leq G(x) \leq F(x + \eta_n) + \eta_n \tag{16}$$

from $d(G, H) = b$ it's clear that

$$\exists \mu_n \to b \ (n \to \infty), \forall x \in \mathbb{R}, \forall n, b < \mu_n < b + \frac{1}{n}, G(x - \mu_n) - \mu_n \leq H(x) \leq G(x + \mu_n) + \mu_n \tag{17}$$

where the $\eta_n < a + \frac{1}{n}, \mu_n < b + \frac{1}{n}$ conditions can be ensured by taking a good enough subsequence. Combine two inequalities to see that

$$\begin{cases} \forall x \in \mathbb{R}, \forall n, F(x-\eta_n) - \eta_n \leq G(x) \leq H(x+\mu_n) + \mu_n \\ \forall x \in \mathbb{R}, \forall n, H(x-\mu_n) - \mu_n \leq G(x) \leq F(x+\eta_n) + \eta_n \end{cases} \tag{18}$$

set $x = y + \frac{1}{n}$

$$\forall y \in \mathbb{R}, F\left(y + \frac{1}{n} - \eta_n\right) - \eta_n \leq H\left(y + \frac{1}{n} + \mu_n\right) + \mu_n, H\left(y + \frac{1}{n} - \mu_n\right) - \mu_n \leq F\left(y + \frac{1}{n} + \eta_n\right) + \eta_n \tag{19}$$

the reason we are doing this is because $\eta_n - \frac{1}{n} < a$ so $\eta_n - \frac{1}{n} \to a^-$ $(n \to \infty)$ hence $\frac{1}{n} - \eta_n \to (-a)^+$ $(n \to \infty)$ approximates $-a$ from the right hand side. Similarly, $\frac{1}{n} - \mu_n \to (-b)^+$ $(n \to \infty)$. Set $n \to \infty$, the approximation from right hand side matches the right-continuity of F, H that

$$\forall y \in \mathbb{R}, F(y-a) - a \leq H(y+b) + b, H(y-b) - b \leq F(y+a) + a \tag{20}$$

concludes the proof. □

Lemma 4 (Convergence in distribution)**.** *Prove that convergence in distribution is equivalent to convergence under the Levy metric defined above.*

Proof. Denote F_n as d.f. of X_n, F as d.f. of X and $C(F)$ the set of all continuity points of F.

If $d(F_n, F) \to 0$ $(n \to \infty)$, $\forall \varepsilon > 0, \exists N, \forall n > N, d(F_n, F) < \varepsilon$, from the definition of Levy metric,

$$\forall \varepsilon > 0, \exists N, \forall n > N, \forall x \in \mathbb{R}, F(x-\varepsilon) - \varepsilon \leq F_n(x) \leq F(x+\varepsilon) + \varepsilon \tag{21}$$

set $n \to \infty$,

$$\forall \varepsilon > 0, \forall x \in \mathbb{R}, \liminf_{n\to\infty} F_n(x) \geq F(x-\varepsilon) - \varepsilon, \limsup_{n\to\infty} F_n(x) \leq F(x+\varepsilon) + \varepsilon \tag{22}$$

restrict ourselves to $\forall x \in C(F)$, set $\varepsilon \to 0$ to see

$$\forall x \in C(F), \liminf_{n\to\infty} F_n(x) \geq F(x), \limsup_{n\to\infty} F_n(x) \leq F(x) \tag{23}$$

proves $\forall x \in C(F), F_n(x) \to F(x)$ $(n \to \infty)$ hence $X_n \xrightarrow{d} X$ $(n \to \infty)$.

If $X_n \xrightarrow{d} X$ $(n \to \infty)$, then $\forall x \in C(F), F_n(x) \to F(x)$ $(n \to \infty)$. Since F is increasing, it has at most countably many discontinuities, hence on fixing $\varepsilon > 0$, we can figure out a compact concentration region of F, i.e. there exists $x_1, ..., x_k \in C(F), x_1 < x_2 < ... < x_k$ such that

$$F(x_1) < \varepsilon, F(x_k) > 1 - \varepsilon, x_i - x_{i-1} < \varepsilon \ (i = 2, 3, ..., k) \tag{24}$$

so the spaces between $x_1, ..., x_k$ are small enough and at most ε probability mass is missing at the left and right tail respectively. This compactness argument has its motivation coming from the definition of Levy metric that $F(x+\delta)+\delta$ allows δ difference in probability mass and δ difference in the variable, i.e. we shall use open intervals of radius δ to cover the compact set.

At each $x_i \in C(F)$, there exists $N_i, \forall n > N_i, |F_n(x_i) - F(x_i)| < \varepsilon$. Naturally take

$$N = \max_i \{N_i\} \tag{25}$$

so for $\forall n > N$, let's discuss where $\forall x \in \mathbb{R}$ is located.

If $x < x_1$,

$$F(x-2\varepsilon) - 2\varepsilon \le F(x_1) - 2\varepsilon < 0 \le F_n(x) \le F_n(x_1) \le F(x_1) + \varepsilon < 2\varepsilon \le F(x+2\varepsilon) + 2\varepsilon \tag{26}$$

if $x > x_k$,

$$F(x-2\varepsilon) - 2\varepsilon \le 1 - 2\varepsilon < F(x_k) - \varepsilon \le F_n(x_k) \le F_n(x) \le 1 < F(x_k) + 2\varepsilon \le F(x+2\varepsilon) + 2\varepsilon \tag{27}$$

if $x_1 \le x \le x_k$, then $x_{i-1} \le x \le x_i$ for some $i \in \{2, 3, ..., k\}$, in this case $x + \varepsilon \ge x_i$ and $x - \varepsilon \le x_{i-1}$

$$F(x-2\varepsilon) - 2\varepsilon \le F(x-\varepsilon) - \varepsilon \le F(x_{i-1}) - \varepsilon \le F_n(x_{i-1}) \le F_n(x) \le F_n(x_i) \le F(x_i) + \varepsilon \le F(x+\varepsilon) + \varepsilon \le F(x+2\varepsilon) + 2\varepsilon \tag{28}$$

everything we have used above is that F is increasing and takes value in $[0, 1]$. As a result, for fixed $\forall \varepsilon > 0$ and such N constructed above,

$$\forall n > N, \forall x \in \mathbb{R}, F(x-2\varepsilon) - 2\varepsilon \le F_n(x) \le F(x+2\varepsilon) + 2\varepsilon, d(F_n, F) < 2\varepsilon \tag{29}$$

as a result, $d(F_n, F) \to 0$ $(n \to \infty)$.

□

Remark. *From the lemmas prove above,* ***convergence in distribution is metric-induced****. To understand convergence in distribution which is essentially different from other convergence modes, notice that by saying* $X_n \overset{d}{\to} X$ $(n \to \infty)$*, we only care about the d.f. of* X_n *and* X*, which means that it's even possible that* $X_1, X_2, ..., X$ *are not in the same probability space. That's why the Levy metric is defined as a metric on the space of d.f. but not on the space of random variables. On the other hand, if* $X_1, X_2, ..., X$ *are not in the same probability space, almost sure convergence, convergence in probability and* L^p *convergence cannot be discussed.*

Remark. *Levy metric is defined above only on* $\mathbb{R}$ *but can we generalize it onto* $\mathbb{R}^d$ *or more general metric spaces? The answer is yes and it's called* ***Levy-Prokhorov metric****. Consider space* M *equipped with metric* ρ *and* σ*-field* $\mathscr{F}$*,* ν, μ *as two probability measures on* $(M, \mathscr{F})$*, the Levy-Prokhorov metric is defined as*

$$d_L(\mu, \nu) = \inf\left\{\delta > 0 : \forall A \in \mathscr{F}, \mu(A) \le \nu(A^\delta) + \delta, \nu(A) \le \mu(A^\delta) + \delta\right\} \tag{30}$$

where $A^\delta = \left\{x \in \mathbb{R}^d : \inf_{y\in A} \rho(x,y) < \delta\right\}$ is the δ-fattened version of A. Convergence in distribution on space M is still equivalent to the convergence under metric d_L.

Lemma 5 (Metric for convergence in probability)**.** *Show that*

$$d(X,Y) = \mathbb{E}\frac{|X-Y|}{1+|X-Y|} \tag{31}$$

*defines a metric on the space of certain random variables in the sense of almost sure equality, check that $d(X_n, X) \to 0$ $(n \to \infty)$ iff $X_n \xrightarrow{p} X$ $(n \to \infty)$. This shows that **convergence in probability is metric-induced**.*

Proof. Clearly $d(X,Y) \geq 0$, if $d(X,Y) = 0$, then since $\frac{|X-Y|}{1+|X-Y|} \geq 0$ *a.s.*, $|X-Y| = 0$ *a.s.* and $X = Y$ *a.s.* proves positivity. It's obvious that d is symmetric. Notice that $f(x) = \frac{x}{1+x}$ is increasing for $x \geq 0$ and $|X-Z| \leq |X-Y| + |Y-Z|$

$$d(X,Z) \leq \mathbb{E}\frac{|X-Y|+|Y-Z|}{1+|X-Y|+|Y-Z|} \leq d(X,Y) + d(Y,Z) \tag{32}$$

proves the triangle inequality.

If $X_n \xrightarrow{p} X$ $(n \to \infty)$, then $|X_n - X| \xrightarrow{p} 0$ $(n \to \infty)$, since $f(x) = \frac{x}{1+x}$ takes value in $[0,1)$ as $x \geq 0$, $\frac{|X_n-X|}{1+|X_n-X|} \xrightarrow{p} 0, \left|\frac{|X_n-X|}{1+|X_n-X|}\right| \leq 1$ *a.s.*, by bounded convergence theorem,

$$d(X_n, X) = \mathbb{E}\frac{|X_n - X|}{1+|X_n-X|} \to 0 \ (n \to \infty) \tag{33}$$

conversely, if $d(X_n, X) \to 0$ $(n \to \infty)$, by Markov inequality,

$$\forall \varepsilon > 0, \mathbb{P}\left(|X_n - X| \geq \varepsilon\right) = \mathbb{P}\left(\frac{|X_n-X|}{1+|X_n-X|} \geq \frac{\varepsilon}{1+\varepsilon}\right) \leq \frac{\mathbb{E}\frac{|X_n-X|}{1+|X_n-X|}}{\frac{\varepsilon}{1+\varepsilon}} \to 0 \ (n \to \infty) \tag{34}$$

proves $X_n \xrightarrow{p} X$ $(n \to \infty)$.

□

Lemma 6 (Almost sure convergence)**.** *Show that $X_n \xrightarrow{p} X$ $(n \to \infty)$ iff for every subsequence X_{n_k} there exists a further subsequence $X_{n_{k_q}}$ such that $X_{n_{k_q}} \xrightarrow{a.s.} X$ $(q \to \infty)$. Use this fact to show that **almost sure convergence is not metric-induced, actually it's even not topology-induced**.*

Proof. If for every subsequence X_{n_k} there exists a further subsequence $X_{n_{k_q}}$ such that $X_{n_{k_q}} \xrightarrow{a.s.} X$ $(q \to \infty)$, fix $\forall \varepsilon > 0$ and consider the sequence of real numbers $a_n = \mathbb{P}\left(|X_n - X| \geq \varepsilon\right)$. For every subsequence a_{n_k}, there exists a further subsequence $a_{n_{k_q}}$ such that $a_{n_{k_q}} \xrightarrow{a.s.} 0$ $(q \to \infty)$. This implies $a_n \to 0$ $(n \to \infty)$ so $X_n \xrightarrow{p} X$ $(n \to \infty)$.

On the other hand, if $X_n \xrightarrow{p} X$ $(n \to \infty)$, for every subsequence X_{n_k}, there exists its further subsequence n_{k_q} such that

$$\forall q \in \mathbb{N}, \mathbb{P}\left(|X_{n_{k_q}} - X| \geq \frac{1}{q}\right) \leq \frac{1}{q^2} \tag{35}$$

by Borel-Cantelli, since $\sum_{q=1}^{\infty}\mathbb{P}\left(|X_{n_{k_q}}-X|\geq\frac{1}{q}\right)<\infty$,

$$\mathbb{P}\left(|X_{n_{k_q}}-X|\geq\frac{1}{q}\ i.o.\right)=0 \tag{36}$$

which mean almost surely eventually $|X_{n_{k_q}}-X|<\frac{1}{q}$ so $X_{n_{k_q}}\overset{a.s.}{\to}X\ (q\to\infty)$.

It's clear that almost sure convergence implies convergence in probability but not vice versa. As a result, there exists $\{X_n\}$ such that for its every subsequence X_{n_k} there exists a further subsequence $X_{n_{k_q}}$ such that $X_{n_{k_q}}\overset{a.s.}{\to}X\ (q\to\infty)$ but $X_n\overset{a.s.}{\not\to}X\ (n\to\infty)$. This violates the property of metric-induced convergence, even topology-induced convergence. As a result, there exists no underlying metric and underlying topology inducing almost sure convergence. □

More Exercise on Convergence Mode

Slutsky's Theorem

Lemma 7 (Slutsky's Theorem). *Show that if $X_n \xrightarrow{d} X, Y_n \xrightarrow{p} c$ for some constant $c \in \mathbb{R}$, then $(X_n, Y_n) \xrightarrow{d} (X, c)$ $(n \to \infty)$.*

Use this conclusion to show that $X_n + Y_n \xrightarrow{d} X + c, X_nY_n \xrightarrow{d} cX$ $(n \to \infty)$.

Proof. Consider the joint CDF of X_n, Y_n

$$F_{(X_n,Y_n)}(x,y) = \mathbb{P}\left(X_n \le x, Y_n \le y\right) \tag{37}$$

and the joint CDF of (X, c) given by

$$F_{(X,c)}(x,y) = \begin{cases} 0 & y < c \\ \mathbb{P}\left(X \le x\right) & y \ge c \end{cases} \tag{38}$$

when $y < c$, $\forall \varepsilon > 0$,

$$\begin{aligned} F_{(X_n,Y_n)}(x,y) &= \mathbb{P}\left(X_n \le x, |Y_n - c| \ge \varepsilon, Y_n \le y\right) + \mathbb{P}\left(X_n \le x, |Y_n - c| < \varepsilon, Y_n \le y\right) && (39) \\ &\le \mathbb{P}\left(|Y_n - c| \ge \varepsilon\right) + \mathbb{P}\left(X_n \le x, |Y_n - c| < \varepsilon, Y_n \le y\right) && (40) \end{aligned}$$

with the first term on RHS converging to zero as $n \to \infty$, specify $0 < \varepsilon < c - y$ so that $\{|Y_n - c| < \varepsilon\}$ contradicts $\{Y_n \le y\}$, the second term on RHS is always zero, so

$$\forall y < c, F_{(X_n,Y_n)}(x,y) \to 0 \ (n \to \infty) \tag{41}$$

On the other hand, when $y > c$,

$$\begin{aligned} |F_{(X_n,Y_n)}(x,y) - F_{(X,c)}(x,y)| &= |\mathbb{P}\left(X_n \le x, Y_n \le y\right) - \mathbb{P}\left(X \le x\right)| && (42) \\ &\le |\mathbb{P}\left(X_n \le x, Y_n \le y\right) - \mathbb{P}\left(X_n \le x\right)| + |\mathbb{P}\left(X_n \le x\right) - \mathbb{P}\left(X \le x\right)| && (43) \end{aligned}$$

bound the first term on RHS that

$$|\mathbb{P}\left(X_n \le x, Y_n \le y\right) - \mathbb{P}\left(X_n \le x\right)| \le \mathbb{P}\left(X_n \le x, Y_n > y\right) \le \mathbb{P}\left(Y_n > y\right) \to 0 \ (n \to \infty) \tag{44}$$

from the convergence in probability of Y_n to $c < y$. The second term on RHS converges to zero as long as $x \in C(F_X)$.

The last case to discuss is when $y = c$. Notice that we only have to consider $(x, y) \in C(F_{(X,c)})$, so if (x, c) is a continuity point then $\mathbb{P}\left(X \le x\right) = 0, x \in C(F_X)$. Now that

$$\forall \varepsilon > 0, F_{(X_n,Y_n)}(x,c) \le F_{(X_n,Y_n)}(x,c+\varepsilon) \to F_{(X,c)}(x,c+\varepsilon) = 0 \ (n \to \infty) \tag{45}$$

from the case of $y > c$ shown above. From the definition of convergence in distribution, we proved that $(X_n, Y_n) \xrightarrow{d} (X, c)$ $(n \to \infty)$.

From continuous mapping theorem, for any continuous function $g : \mathbb{R}^2 \to \mathbb{R}$, $g(X_n, Y_n) \xrightarrow{d} g(X, c)$ $(n \to \infty)$. Apply this for $g(x, y) = x + y, g(x, y) = xy$ to conclude. □

Remark. *The reader shall check that* $Y_n \xrightarrow{d} c$ $(n \to \infty)$ *iff* $Y_n \xrightarrow{p} c$ $(n \to \infty)$*. This provides the final form of Slutsky's theorem.*

Check that Slutsky's theorem generally does not hold, e.g. when the limit in distribution of Y_n *is not constant. A counterexample:* $X_n = -Y_n, \forall n, X_n \sim N(0, 1)$*, then the limit of* X_n *and* Y_n *in distribution are both* $N(0, 1)$ *random variable but we can set the limit to be independent, i.e.* $X, Y \sim N(0, 1)$ *are independent. Then* $X_n + Y_n = 0$ *a.s. but* $X + Y \sim N(0, 2)$*.*

The next example illustrates why Slutsky's theorem is useful in statistics.

Lemma 8 (Asymptotic Normality of T-statistic)**.** *Prove that T-statistic* $T = \frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}}$ *for i.i.d. sample* $X_1, ..., X_n$ *where* $\mathbb{E}X_1 = \mu, Var(X_1) = \sigma^2$ *is asymptotically normal, i.e.* $T \xrightarrow{d} N(0, 1)$ $(n \to \infty)$*.*

Proof.

$$T = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \cdot \frac{\sigma}{S} \tag{46}$$

from CLT

$$\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{d} N(0, 1) \ (n \to \infty) \tag{47}$$

we think about using Slutsky's theorem stated above. It suffices to prove that

$$\frac{\sigma}{S} \xrightarrow{p} 1 \ (n \to \infty) \tag{48}$$

it's clear that

$$S^2 = \frac{n}{n-1}\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{49}$$

$$= \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\overline{X}\right)^2\right] \tag{50}$$

where

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} \mathbb{E}X_1^2 = \sigma^2 + \mu^2, \overline{X} \xrightarrow{p} \mathbb{E}X_1 = \mu \ (n \to \infty) \tag{51}$$

from WLLN, by continuous mapping theorem,

$$S^2 \to \sigma^2, \frac{\sigma}{S} \to 1 \ (n \to \infty) \tag{52}$$

concludes the proof. □

Convergence on the Space of Measure

When it comes to the convergence only w.r.t. the distribution of random variables, there are actually a lot of different notions of convergence available. To think about this, a random variable X induces a probability measure $\mathbb{P}\left(X \in \cdot\right)$ on the real line $\mathbb{R}$, so a sequence of random variables induce a sequence of probability measure $\mathbb{P}_n$ on the real line. If it's possible to establish a norm/metric on the space of probability measures

$$\mathscr{P} = \{\mathbb{P} : \mathbb{P}(\mathbb{R}) = 1\} \tag{53}$$

then it's possible to build a certain notion of convergence.

We have seen in the previous week that the Levy-Prokhorov metric $d_L(\mathbb{P}, \mathbb{Q})$ is an example of a metric on the space of probability measures and induces the convergence of distribution we are familiar with. It's thus natural to ask: if there exists any other possible metric on $\mathscr{P}$. Notice that $\mathscr{P}$ is not a vector space and only the convex combination of probability measures is guaranteed to be a probability measure.

One idea comes from thinking of a quantity that simultaneously contains the information in $\mathbb{P}$ and $\mathbb{Q}$, may also a comparison of those two measures. A natural idea comes from measure theory that we have the Radon-Nikodym derivative of two probability measures! That is to say, if $\mathbb{P} << \mathbb{Q}$ are two probability measures on the measurable space $(X, \mathscr{F})$, then $\frac{d\mathbb{P}}{d\mathbb{Q}}(\omega)$ is well-defined and $\mathbb{Q} - a.s.$ unique such that

$$\forall A \in \mathscr{F}, \mathbb{P}\left(A\right) = \int_A \frac{d\mathbb{P}}{d\mathbb{Q}}(\omega)\, \mathbb{Q}(d\omega) = \mathbb{E}_{\mathbb{Q}}\left(\frac{d\mathbb{P}}{d\mathbb{Q}} \cdot \mathbb{I}_A\right) \tag{54}$$

as an example to illustrate this idea, if X, Y are two continuous random variables inducing probability measure $\mathbb{P}, \mathbb{Q}$ on the real line, then

$$\mathbb{P} << \lambda, \mathbb{Q} << \lambda \tag{55}$$

with λ to be the Lebesgue measure and thus

$$p(x) = \frac{d\mathbb{P}}{d\lambda}(x), q(x) = \frac{d\mathbb{Q}}{d\lambda}(x) \tag{56}$$

are measure w.r.t. λ, i.e. are Borel measurable functions on $\mathbb{R}$. They satisfy the property that

$$\forall A \in \mathscr{B}_{\mathbb{R}}, \int_A p(x)\, dx = \mathbb{P}(A) = \mathbb{P}\left(X \in A\right), \int_A q(x)\, dx = \mathbb{Q}(A) = \mathbb{P}\left(Y \in A\right) \tag{57}$$

so those p, q are just the density functions! Actually **density functions are essentially Radon-Nikodym derivatives w.r.t. the Lebesgue measure**. Now if $\mathbb{P} << \mathbb{Q}$ holds and both measures are absolute continuous w.r.t. the Lebesgue measure, then the chain rule for Radon-Nikodym derivative tells us

$$\frac{d\mathbb{P}}{d\mathbb{Q}} = \frac{\frac{d\mathbb{P}}{d\lambda}}{\frac{d\mathbb{Q}}{d\lambda}} = \frac{p}{q} \tag{58}$$

is just the likelihood ratio! It should be obvious that likelihood ratio directly reflects the relationship between two probability measures, so this approach actually makes sense.

Remark. *There are some details hidden behind here. What about the case where all three of* $\mathbb{P} << \mathbb{Q}, \mathbb{P} << \lambda, \mathbb{Q} << \lambda$ *does not hold? The trick is to find a reference measure* $\frac{\mathbb{P}+\mathbb{Q}}{2}$ *still as a probability measure but now (check this fact)*

$$\mathbb{P} << \frac{\mathbb{P}+\mathbb{Q}}{2}, \mathbb{Q} << \frac{\mathbb{P}+\mathbb{Q}}{2} \tag{59}$$

so the definition of Radon-Nikodym derivative can be extended such that the chain rule still formally holds

$$\frac{d\mathbb{P}}{d\mathbb{Q}} = \frac{\frac{d\mathbb{P}}{d\frac{\mathbb{P}+\mathbb{Q}}{2}}}{\frac{d\mathbb{Q}}{d\frac{\mathbb{P}+\mathbb{Q}}{2}}} \tag{60}$$

but it remains to check if this definition is always well-defined (independent of the selection of the reference measure). In the use of our construction below, it can be verified that the quantities are always well-defined so we don't have to worry about those corner cases.

Now it's time to define a distance between two probability measures on $\mathscr{P}$ as a function of $\frac{d\mathbb{P}}{d\mathbb{Q}}$. For simplicity, we assume that $\mathbb{P}, \mathbb{Q}$ are induced by continuous random variables X, Y so that $\mathbb{P} << \lambda, \mathbb{Q} << \lambda$. The general definition of the **total variation** is given by

$$TV(X,Y) = TV(\mathbb{P},\mathbb{Q}) = \frac{1}{2}\mathbb{E}_{\mathbb{Q}}\left|\frac{d\mathbb{P}}{d\mathbb{Q}} - 1\right| \tag{61}$$

in our setting, the expression can be simplified to

$$TV(X,Y) = TV(\mathbb{P},\mathbb{Q}) = \frac{1}{2}\int_{\mathbb{R}}\left|\frac{p(x)}{q(x)} - 1\right| q(x)\,dx = \frac{1}{2}\int_{\mathbb{R}} |p(x) - q(x)|\,dx \tag{62}$$

written in terms of the density functions.

Lemma 9 (Total Variation as a Metric)**.** *Prove that total variation is a metric on the space of density functions under almost everywhere equality. Prove another representation*

$$TV(\mathbb{P},\mathbb{Q}) = \sup_{A\in\mathscr{B}_{\mathbb{R}}} |\mathbb{P}(A) - \mathbb{Q}(A)| \tag{63}$$

and conclude that it only takes values in $[0, 1]$.

Proof. Obviously it's non-negative and if $TV(\mathbb{P},\mathbb{Q}) = 0$, then

$$|p(x) - q(x)| = 0 \ a.e. \tag{64}$$

so $p = q \ a.e.$. Obviously it's symmetric so we only need to check the triangle inequality

$$\int_{\mathbb{R}} |p(x) - q(x)| \, dx + \int_{\mathbb{R}} |q(x) - r(x)| \, dx \geq \int_{\mathbb{R}} |p(x) - r(x)| \, dx \tag{65}$$

proves that it's a metric.

Now we prove that its value does not exceed 1.

$$\begin{aligned}
TV(\mathbb{P},\mathbb{Q}) &= \frac{1}{2}\int_{\mathbb{R}} |p(x) - q(x)| \, dx && (66)\\
&= \frac{1}{2}\int_{p<q} [q(x) - p(x)] \, dx + \frac{1}{2}\int_{p>q} [p(x) - q(x)] \, dx && (67)\\
&= -\frac{1}{2}\int_{\mathbb{R}} [q(x) - p(x)] \, dx + \frac{1}{2}\int_{p<q} [q(x) - p(x)] \, dx + \frac{1}{2}\int_{p>q} [p(x) - q(x)] \, dx && (68)\\
&= -\frac{1}{2}\int_{p>q} [q(x) - p(x)] \, dx + \frac{1}{2}\int_{p>q} [p(x) - q(x)] \, dx && (69)\\
&= \int_{p>q} [p(x) - q(x)] \, dx && (70)
\end{aligned}$$

so $\forall A \in \mathscr{B}_{\mathbb{R}}, TV(\mathbb{P},\mathbb{Q}) \geq \int_{A\cap\{p>q\}} [p(x) - q(x)] \, dx = \left|\int_{A\cap\{p>q\}} [p(x) - q(x)] \, dx\right|$. By symmetricity, switch the position of p, q to get $\forall A \in \mathscr{B}_{\mathbb{R}}, TV(\mathbb{P},\mathbb{Q}) \geq \left|\int_{A\cap\{q>p\}} [q(x) - p(x)] \, dx\right|$. As a result,

$$\forall A \in \mathscr{B}_{\mathbb{R}}, TV(\mathbb{P},\mathbb{Q}) \geq \left|\int_{A} [p(x) - q(x)] \, dx\right| = |\mathbb{P}(A) - \mathbb{Q}(A)| \tag{71}$$

concludes the proof. □

This metric induces the **convergence in total variation**.

Lemma 10 (Convergence in Total Variation)**.** *Denote $X_n \overset{TV}{\to} X \ (n \to \infty)$ if $TV(X_n, X) \to 0 \ (n \to \infty)$. Prove that for u bounded, $\mathbb{E}u(X_n) \to \mathbb{E}u(X) \ (n \to \infty)$. Prove that $X_n \overset{TV}{\to} X \ (n \to \infty)$ implies $X_n \overset{d}{\to} X \ (n \to \infty)$.*

Proof. The convergence means that if we denote f_n as the density function of X_n, f as the density function of X, then

$$\int_{\mathbb{R}} |f_n(x) - f(x)| \, dx \to 0 \tag{72}$$

now that $|u| \leq M$,

$$\begin{aligned} |\mathbb{E}u(X_n) - \mathbb{E}u(X)| &= \left| \int_{\mathbb{R}} u(x)[f_n(x) - f(x)]\, dx \right| & (73) \\ &\leq \int_{\mathbb{R}} |u(x)| \cdot |f_n(x) - f(x)|\, dx & (74) \\ &\leq M \int_{\mathbb{R}} |f_n(x) - f(x)|\, dx \to 0 \ (n \to \infty) & (75) \end{aligned}$$

concludes the proof.

For the second statement, use the second definition of total variation, convergence in total variation implies

$$\sup_{A \in \mathscr{B}_{\mathbb{R}}} |\mathbb{P}(X_n \in A) - \mathbb{P}(X \in A)| \to 0 \ (n \to \infty) \tag{76}$$

as a result, take $A = (-\infty, x]$ to get

$$\sup_{x \in \mathbb{R}} |F_{X_n}(x) - F_X(x)| \to 0 \ (n \to \infty) \tag{77}$$

the CDF converges uniformly on $\mathbb{R}$, which implies pointwise convergence, proved.

□

Remark. *The idea of using $\frac{d\mathbb{P}}{d\mathbb{Q}}$ to construct some function that can measure the distance between two probability measures is crucial in information theory. This gives rise to definition of Kullback-Leibler divergence, Chi-square divergence etc. and they are closely connected to the intrinsic complexity of a problem in statistics, e.g. proving Cramer-Rao bound, proving the optimality of an algorithm etc. You can check topics regarding f-divergence if interested.*

However, this is not the only way to construct the distance between two probability measures. In the literature of optimal transport, the Wasserstein distance is introduced to form another notion of convergence, which is defined in terms of coupling, a very important technique in probability.