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 $X_n \xrightarrow{L^p} X$ $(n \to \infty)$ 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 $\|X_n\|_p \le \|X_n - X\|_p + \|X\|_p$ so $|\,\|X_n\|_p - \|X\|_p\,| \le \|X_n - X\|_p$. 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

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

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,*

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

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 $x < x_1$,

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

if $x > x_k$,

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

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

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

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,

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

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 $X_n \xrightarrow{d} X \ (n \to \infty)$, 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*

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

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 $X_{n_{k_q}} \stackrel{a.s.}{\to} X \ (q \to \infty)$.

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 $X_{n_{k_q}} \stackrel{a.s.}{\to} X \ (q \to \infty)$ but $X_n \stackrel{a.s.}{\not\to} X \ (n \to \infty)$. 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. $\square$

Week 2

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

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

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

$$F_{(X,c)}(x, y) = \begin{cases} 0 & y < c \\ \mathbb{P}(X \leq x) & y \geq c \end{cases} \tag{38}$$

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

$$F_{(X_n,Y_n)}(x, y) = \mathbb{P}(X_n \leq x, |Y_n - c| \geq \varepsilon, Y_n \leq y) + \mathbb{P}(X_n \leq x, |Y_n - c| < \varepsilon, Y_n \leq y) \tag{39}$$

$$\leq \mathbb{P}(|Y_n - c| \geq \varepsilon) + \mathbb{P}(X_n \leq x, |Y_n - c| < \varepsilon, Y_n \leq y) \tag{40}$$

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 $\{Y_n \leq y\}$, 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 $y > c$,

$$|F_{(X_n,Y_n)}(x, y) - F_{(X,c)}(x, y)| = |\mathbb{P}(X_n \leq x, Y_n \leq y) - \mathbb{P}(X \leq x)| \tag{42}$$

$$\leq |\mathbb{P}(X_n \leq x, Y_n \leq y) - \mathbb{P}(X_n \leq x)| + |\mathbb{P}(X_n \leq x) - \mathbb{P}(X \leq x)| \tag{43}$$

bound the first term on RHS that

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

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 $\mathbb{P}(X \leq x) = 0, x \in C(F_X)$. Now that

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

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 $\mathbb{P}(X \in \cdot)$ 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

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

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

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

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 $X_n\stackrel{TV}{\to}X$ $(n\to\infty)$ 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 $X_n\stackrel{TV}{\to}X$ $(n\to\infty)$ implies $X_n\stackrel{d}{\to}X$ $(n\to\infty)$.*

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 $|u| \leq M$,

$$|\mathbb{E}u(X_n) - \mathbb{E}u(X)| = \left| \int_{\mathbb{R}} u(x)[f_n(x) - f(x)]\, dx \right| \tag{73}$$

$$\leq \int_{\mathbb{R}} |u(x)| \cdot |f_n(x) - f(x)|\, dx \tag{74}$$

$$\leq M \int_{\mathbb{R}} |f_n(x) - f(x)|\, dx \to 0 \ (n \to \infty) \tag{75}$$

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.

$\square$

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.

Week 3

Borel-Cantelli Lemma

The Borel-Cantelli lemma is one of the most important theorem in measure-based probability theory. It is heavily used in proving conclusions that have something to do with almost sure convergence. Let's first state a canonical framework of proving almost sure convergence below and then see some examples to get a feeling of how to use Borel-Cantelli lemma.

If one wants to prove $X_n \overset{a.s.}{\to} X$ $(n \to \infty)$, it suffices to prove that

$$\forall \varepsilon > 0, \sum_{n=1}^{\infty} \mathbb{P}\left(|X_n - X| \geq \varepsilon\right) < \infty \tag{78}$$

this is because through Borel-Cantelli we know

$$\forall \varepsilon > 0, \mathbb{P}\left(|X_n - X| \geq \varepsilon \textit{ i.o.}\right) = 0 \tag{79}$$

consider its complement

$$\forall \varepsilon > 0, \mathbb{P}\left(|X_n - X| < \varepsilon \text{ eventually}\right) = 1 \tag{80}$$

which implies

$$X_n - X \overset{a.s.}{\to} 0 \ (n \to \infty) \tag{81}$$

this is just the canonical framework of proving almost sure convergence that we shall keep in mind.

Lemma 11 (Order of *i.i.d.* exponential r.v. and its running maximum)**.** *$\{X_n\}$ i.i.d. $\sim \mathcal{E}(1)$. Show that*

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{X_n}{\log n} = 1\right) = 1 \tag{82}$$

define $M_n = \max\{X_1, ..., X_n\}$, show that

$$\frac{M_n}{\log n} \overset{a.s.}{\to} 1 \ (n \to \infty) \tag{83}$$

Proof. Let's first prove $\limsup_{n\to\infty} \frac{X_n}{\log n} \leq 1$ *a.s.*, just need to prove

$$\forall \varepsilon > 0, \mathbb{P}\left(\frac{X_n}{\log n} - 1 < \varepsilon \text{ eventually}\right) = 1 \tag{84}$$

which is equivalent to saying

$$\forall \varepsilon > 0, \mathbb{P}\left(\frac{X_n}{\log n} - 1 \geq \varepsilon \ i.o.\right) = 0 \tag{85}$$

implied by

$$\forall \varepsilon > 0, \sum_{n=1}^{\infty} \mathbb{P}\left(\frac{X_n}{\log n} - 1 \geq \varepsilon\right) < \infty \tag{86}$$

from Borel-Cantelli. Now estimate the probability

$$\begin{aligned}\forall \varepsilon > 0, \mathbb{P}\left(\frac{X_n}{\log n} - 1 \geq \varepsilon\right) &= \mathbb{P}\left(X_n \geq (1+\varepsilon)\log n\right) &(87)\\ &= e^{-(1+\varepsilon)\log n} = n^{-(1+\varepsilon)} &(88)\end{aligned}$$

and $\forall \varepsilon > 0, \sum_{n=1}^{\infty} n^{-(1+\varepsilon)} < \infty$ concludes the proof.

On the other hand, we want to argue that almost surely there exists a convergent subsequence with limit larger or equal to 1, which can be implied by

$$\mathbb{P}\left(\frac{X_n}{\log n} > 1 \ i.o.\right) = 1 \tag{89}$$

to prove this, it seems that we need the other part of Borel-Cantelli. Since $\{X_n\}$ are independent, only need to prove

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\frac{X_n}{\log n} > 1\right) = \infty \tag{90}$$

implied by

$$\mathbb{P}\left(\frac{X_n}{\log n} > 1\right) = \frac{1}{n} \tag{91}$$

concludes the proof.

For the second part of the conclusion for M_n, it's obvious that since $M_n \geq X_n \ a.s.$,

$$a.s. \ \limsup_{n\to\infty} \frac{M_n}{\log n} \geq 1 \tag{92}$$

to prove the other direction, we only need to prove that

$$\forall \varepsilon > 0, \mathbb{P}\left(\frac{M_n}{\log n} \leq 1 + \varepsilon \text{ eventually}\right) = 1 \tag{93}$$

which is equivalent to saying

$$\forall \varepsilon > 0, \mathbb{P}\left(\frac{M_n}{\log n} > 1 + \varepsilon \ i.o.\right) = 0 \tag{94}$$

implied by

$$\forall \varepsilon > 0, \sum_{n=1}^{\infty} \mathbb{P}\left(\frac{M_n}{\log n} > 1 + \varepsilon\right) < \infty \tag{95}$$

from Borel-Cantelli. As a result, let's calculate

$$\begin{aligned}
\mathbb{P}\left(\frac{M_n}{\log n} > 1 + \varepsilon\right) &= 1 - \mathbb{P}\left(\frac{M_n}{\log n} \leq 1 + \varepsilon\right) && (96)\\
&= 1 - \left[\mathbb{P}\left(X_1 \leq (1+\varepsilon)\log n\right)\right]^n && (97)\\
&= 1 - \left(1 - n^{-(1+\varepsilon)}\right)^n && (98)\\
&\sim 1 - e^{-n^{-\varepsilon}} && (99)\\
&\sim n^{-\varepsilon} && (100)
\end{aligned}$$

and it's quite obvious that $\sum_{n=1}^{\infty} n^{-\varepsilon} < \infty$ does not necessarily hold for $\forall \varepsilon > 0$, so Borel-Cantelli is not working well here.

This is an example of a case where we have to think about using other structures to prove almost sure convergence. Let's check that

$$\frac{M_n}{\log n} = \frac{\max_{1\leq m\leq n} X_m}{\log n} \tag{101}$$

with the numerator as a maximum of two parts, the part $\max_{1\leq m\leq N} X_m$ where m is not very large and the part $\max_{N+1\leq m\leq n} X_m$ when m is very large. When m is not large, it should be a maximum among finitely many terms with $\log n$ on the denominator, thus converging to zero. When m is large, the asymptotics of $\frac{X_n}{\log n}$ helps with the proof.

In detail, since we have proved that $\limsup_{n\to\infty} \frac{X_n}{\log n} = 1 \ a.s.$,

$$\exists A \in \mathscr{F}, \mathbb{P}(A) = 0, \forall \omega \notin A, \limsup_{n\to\infty} \frac{X_n(\omega)}{\log n} = 1 \tag{102}$$

it's clear that on fixing $\forall \varepsilon > 0$,

$$\exists N(\omega), \forall n > N(\omega), \frac{X_n(\omega)}{\log n} < 1 + \varepsilon \tag{103}$$

now that

$$\frac{M_n(\omega)}{\log n} = \frac{\max_{1\leq m\leq n} X_m(\omega)}{\log n} \tag{104}$$

$$= \max\left\{\max_{1\leq m\leq N(\omega)} \frac{X_m(\omega)}{\log n}, \max_{N(\omega)+1\leq m\leq n} \frac{X_m(\omega)}{\log n}\right\} \tag{105}$$

with $\max_{N(\omega)+1\leq m\leq n} \frac{X_m(\omega)}{\log n} < 1+\varepsilon$.

The first part has $N(\omega)$ to be finite, so the maximum is taken over finitely many terms and as $n\to\infty$ with no doubt the whole term converges to zero. In mathematical language,

$$\exists N'(\omega), \forall n > N'(\omega), \max_{1\leq m\leq N(\omega)} \frac{X_m(\omega)}{\log n} < \varepsilon \tag{106}$$

so for $\forall n > \max\left\{N_(\omega), N'(\omega)\right\}$, it's always true that

$$\frac{M_n(\omega)}{\log n} < 1+\varepsilon \tag{107}$$

which by definition implies that

$$\forall \omega \notin A, \limsup_{n\to\infty} \frac{M_n(\omega)}{\log n} \leq 1 \tag{108}$$

since $\mathbb{P}(A) = 0$, it's clear that

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{M_n}{\log n} \leq 1\right) = 1 \tag{109}$$

which concludes the proof.

□

Theorem 1 (WLLN for Triangular Arrays)**.** *Let $X_{n,k}$ be a lower triangular array of r.v. where $1\leq k\leq n$ and all r.v. in each row of the array are independent. For $b_n > 0$ such that $b_n\to\infty$ and $Y_{n,k} = X_{n,k}\mathbb{I}_{|X_{n,k}|\leq b_n}$ as the* ***truncation*** *of $X_{n,k}$, if the following conditions are satisfied that*

$$\sum_{k=1}^{n} \mathbb{P}(|X_{n,k}| > b_n) \to 0 \tag{110}$$

$$\frac{\sum_{k=1}^{n} \mathbb{E}Y_{n,k}^2}{b_n^2} \to 0 \tag{111}$$

as $n\to\infty$, then

$$\frac{S_n - a_n}{b_n} \xrightarrow{p} 0 \ (n\to\infty) \tag{112}$$

where

$$S_n = \sum_{k=1}^{n} X_{n,k}, a_n = \sum_{k=1}^{n} \mathbb{E}Y_{n,k} \tag{113}$$

Proof. Set

$$T_n = \sum_{k=1}^{n} Y_{n,k} \tag{114}$$

we first argue that T_n and S_n share the same asymptotic behavior. Since

$$\mathbb{P}\left(S_n \neq T_n\right) \leq \mathbb{P}\left(\bigcup_{k=1}^{n} \{X_{n,k} \neq Y_{n,k}\}\right) \tag{115}$$

$$\leq \sum_{k=1}^{n} \mathbb{P}\left(|X_{n,k}| > b_n\right) \to 0 \ (n \to \infty) \tag{116}$$

it's clear that

$$\forall \varepsilon > 0, \mathbb{P}\left(\left|\frac{S_n - a_n}{b_n}\right| \geq \varepsilon\right) \leq \mathbb{P}\left(S_n \neq T_n\right) + \mathbb{P}\left(\left|\frac{T_n - a_n}{b_n}\right| \geq \varepsilon\right) \tag{117}$$

so it suffices to prove

$$\frac{T_n - a_n}{b_n} \xrightarrow{p} 0 \ (n \to \infty) \tag{118}$$

Now apply Chebyshev inequality, since $\mathbb{E}T_n = a_n$,

$$\mathbb{P}\left(\left|\frac{T_n - a_n}{b_n}\right| \geq \varepsilon\right) \leq \frac{Var(T_n)}{b_n^2 \varepsilon^2} \tag{119}$$

$$\leq \frac{\sum_{k=1}^{n} \mathbb{E}Y_{n,k}^2}{b_n^2 \varepsilon^2} \to 0 \ (n \to \infty) \tag{120}$$

concludes the proof. □

Remark. *This truncation technique is especially useful when dealing with random variables whose expectation and variance does not exist. It's basically truncating the tail of r.v. and arguing that the tail is negligible when those two conditions above apply.*

Lemma 12 (St. Petersburg Paradox)**.** *Let $\{X_n\}$ be i.i.d. r.v. with*

$$\forall j \geq 1, \mathbb{P}\left(X_i = 2^j\right) = 2^{-j} \tag{121}$$

this is a game where one keeps tossing a coin and gets 2^j if one gets the first head at the j-th toss. X_i denotes the *money you get after playing a single game and* $S_n = \sum_{i=1}^n X_i$ *is the total amount of money you get after playing this game for* n *times.*

Prove that $\limsup_{n\to\infty} \frac{X_n}{n\log n} = \infty$ *a.s. thus*

$$\limsup_{n\to\infty} \frac{S_n}{n\log n} = \infty \text{ } a.s. \tag{122}$$

but

$$\frac{S_n}{n\log n} \xrightarrow{p} 1 \ (n\to\infty) \tag{123}$$

which is another evidence that convergence in probability does not imply convergence almost surely. (All log *in this question stands for the logarithm with base* 2*)*

Proof. We want to prove

$$\forall M > 0, \mathbb{P}\left(\frac{X_n}{n\log n} \geq M \ i.o.\right) = 1 \tag{124}$$

implied by

$$\forall M > 0, \sum_{n=1}^\infty \mathbb{P}\left(\frac{X_n}{n\log n} \geq M\right) = \infty \tag{125}$$

from Borel-Cantelli due to the independence of $\{X_n\}$. It suffices to estimate this probability

$$\mathbb{P}\left(X_n \geq Mn\log n\right) = \sum_{j:2^j\geq Mn\log n} 2^{-j} \tag{126}$$

$$= \sum_{j=\log(Mn\log n)}^{\infty} 2^{-j} \tag{127}$$

$$= \frac{2}{Mn\log n} \tag{128}$$

since $\sum_{n=1}^\infty \frac{1}{n\log n} = \infty$, the almost sure convergence part is proved by noticing $X_n \leq S_n$ *a.s.*.

For the convergence in probability, use the WLLN above. Specify $X_{n,k} = X_k$ with b_n to be specified later. We want b_n to satisfy the two conditions,

$$\sum_{k=1}^n \mathbb{P}\left(|X_{n,k}| > b_n\right) = n\mathbb{P}\left(|X_1| > b_n\right) \to 0 \ (n\to\infty) \tag{129}$$

naturally, we assume b_n has the form of the power of 2 that $b_n = 2^{m_n}$ so it's required that $n2^{-m_n} \to 0$ $(n\to\infty)$.

To satisfy the second condition,

$$\mathbb{E}Y_{n,k}^2 = \sum_{j:2^j \leq b_n} 2^{2j}2^{-j} = 2(b_n - 1) \leq 2b_n \tag{130}$$

so

$$\frac{\sum_{k=1}^n \mathbb{E}Y_{n,k}^2}{b_n^2} \to 0 \ (n \to \infty) \tag{131}$$

requires $\frac{n}{b_n} \to 0$.

Now we specify $m_n = \log n + \log\log n, b_n = n\log n$ that satisfies both conditions above, it's clear that

$$a_n = \sum_{k=1}^n \mathbb{E}X_k \mathbb{I}_{|X_k| \leq b_n} = n\mathbb{E}X_1\mathbb{I}_{|X_1|\leq b_n} \tag{132}$$

$$= n\sum_{2^j \leq b_n} 2^j 2^{-j} = nm_n = n\log n + n\log\log n \tag{133}$$

so

$$\frac{S_n - n\log n - n\log\log n}{n\log n} \xrightarrow{p} 0 \ (n \to \infty) \tag{134}$$

concludes the proof. □

Remark. *This game is called a paradox since the expected payoff of each single game $\mathbb{E}X_1 = \infty$ is infinite. However, one is definitely unwilling to pay an infinite amount to participate in this game since this game is too risky. This is one example where expectation fails to tell us something about the asymptotic behavior. We will see more interesting examples of this kind in the future.*

Finally, let me show an example where Borel-Cantelli can also be applied in other mathematical fields.

Lemma 13 (Application in Number Theory)**.** *For $\forall \varepsilon > 0$, show that for almost every $x \in [0,1]$, there only exists finitely many rational numbers $\frac{p}{q} \in (0,1)$ such that*

$$\left|x - \frac{p}{q}\right| < \frac{1}{q^2(\log q)^{1+\varepsilon}} \tag{135}$$

Proof. Consider probability space $([0,1], \mathscr{B}_{[0,1]}, \lambda)$ where λ is the Lebesgue measure. Consider the event

$$A_q = \left\{x \in [0,1] : \exists p \in \{1,2,...,q-1\}, \left|x - \frac{p}{q}\right| < \frac{1}{q^2(\log q)^{1+\varepsilon}}\right\} \tag{136}$$

it suffices to prove $\lambda(A_q \ i.o.) = 0$. This is because after fixing q, p always only has finitely many possible values to take, so the sequence of events shall be indexed in terms of q.

From Borel-Cantelli, it suffices to prove

$$\sum_{q=1}^{\infty}\lambda(A_q)<\infty \tag{137}$$

calculations tell us

$$\begin{aligned}\lambda(A_q)&=\lambda\left(\bigcup_{p=1}^{q-1}\left\{\left|x-\frac{p}{q}\right|<\frac{1}{q^2(\log q)^{1+\varepsilon}}\right\}\right) &(138)\\ &\le\sum_{p=1}^{q-1}\lambda\left(\left|x-\frac{p}{q}\right|<\frac{1}{q^2(\log q)^{1+\varepsilon}}\right) &(139)\\ &\le\frac{2}{q(\log q)^{1+\varepsilon}} &(140)\end{aligned}$$

with

$$\sum_{q=1}^{\infty}\frac{2}{q(\log q)^{1+\varepsilon}}<\infty \tag{141}$$

concludes the proof.

□

Remark. *The famous Dirichlet theorem tells us that for almost every $x\in[0,1]$, there exists infinitely many rational numbers $\frac{p}{q}\in(0,1)$ such that*

$$\left|x-\frac{p}{q}\right|<\frac{1}{q^2} \tag{142}$$

interestingly, adding a logarithm factor on the bound on the RHS completely changes the approximation theorem. Actually it's not too surprising that if the bound on RHS is changed to $\phi(q)$ then there exists finitely many rational numbers satisfying the approximation iff $\sum_{q=1}^{\infty}q\cdot\phi(q)<\infty$, the same form as Borel-Cantelli.

Week 4

This is the week before mid-term so we mainly review the contents we have learnt so far. The material is formed as several sections with illustrations and exercise problems included. You are welcome to attempt those problems on your own before reading my solution as the preparation for mid-term.

Convergence Mode

When it comes to convergence mode, the definition of different convergence modes and their characterizations are important to keep in mind. I will provide a diagram at the end of the quarter for the connection between different convergence modes after you have seen the concept of uniform integrability.

I would like to provide one exercise problem for convergence mode on the sequence of *i.i.d.* random variables.

Lemma 14. *Let $\{X_n\}$ be a sequence of i.i.d. random variables, find equivalent conditions for the following statements:*

(1): $\frac{X_n}{n} \xrightarrow{a.s.} 0\ (n \to \infty)$

(2): $\frac{\sup_{m\leq n} X_m}{n} \xrightarrow{a.s.} 0\ (n \to \infty)$

(3): $\frac{X_n}{n} \xrightarrow{p} 0\ (n \to \infty)$

(4): $\frac{\sup_{m\leq n} X_m}{n} \xrightarrow{p} 0\ (n \to \infty)$

Proof. (1): Let's characterize a.s. convergence with Borel-Cantelli. $\frac{X_n}{n} \xrightarrow{a.s.} 0\ (n \to \infty)$ is equivalent to saying

$$\forall \varepsilon > 0, \mathbb{P}\left(\left|\frac{X_n}{n}\right| \geq \varepsilon\ i.o.\right) = 0 \tag{143}$$

which can be implied by

$$\forall \varepsilon > 0, \sum_{n=1}^{\infty} \mathbb{P}\left(|X_n| \geq n\varepsilon\right) < \infty \tag{144}$$

conversely, if this condition does not hold, i.e.

$$\exists \varepsilon > 0, \sum_{n=1}^{\infty} \mathbb{P}\left(|X_n| \geq n\varepsilon\right) = \infty \tag{145}$$

since $\{X_n\}$ are independent, by Borel-Cantelli,

$$\exists \varepsilon > 0, \mathbb{P}\left(|X_n| \geq n\varepsilon\ i.o.\right) = 1 \tag{146}$$

which implies that

$$\exists \varepsilon > 0, \limsup_{n\to\infty} \frac{|X_n|}{n} \geq \varepsilon\ a.s. \tag{147}$$

and it contradicts $\frac{X_n}{n} \xrightarrow{a.s.} 0\ (n \to \infty)$.

As a result, we have found one equivalent condition for $\frac{X_n}{n} \stackrel{a.s.}{\to} 0\ (n \to \infty)$ that

$$\forall \varepsilon > 0, \sum_{n=1}^{\infty} \mathbb{P}\left(|X_n| \geq n\varepsilon\right) < \infty \tag{148}$$

can we further simplify this condition? The answer is yes, actually this is equivalent to saying $\mathbb{E}|X_1| < \infty$. Let's prove this argument below. If $\forall \varepsilon > 0, \sum_{n=1}^{\infty} \mathbb{P}\left(|X_n| \geq n\varepsilon\right) < \infty$, then specifying $\varepsilon = 1$ tells us

$$\sum_{n=1}^{\infty} \mathbb{P}\left(|X_n| \geq n\right) = \sum_{n=1}^{\infty} \mathbb{P}\left(|X_1| \geq n\right) < \infty \tag{149}$$

which implies $\mathbb{E}|X_1| < \infty$ (if you don't know why this holds, please refer to the remark below). Conversely, if $\mathbb{E}|X_1| < \infty$, then $\forall \varepsilon > 0, \mathbb{E}\frac{|X_1|}{\varepsilon} < \infty$, which implies

$$\forall \varepsilon > 0, \sum_{n=1}^{\infty} \mathbb{P}\left(|X_1| \geq n\varepsilon\right) = \sum_{n=1}^{\infty} \mathbb{P}\left(|X_n| \geq n\varepsilon\right) < \infty \tag{150}$$

at last, we see a very concise equivalent condition for $\frac{X_n}{n} \stackrel{a.s.}{\to} 0\ (n \to \infty)$ that $\mathbb{E}|X_1| < \infty$.

(2): Recall the same technique we have used last week to deal with the almost sure convergence of the running sup of *i.i.d.* random variables. Let's split the sup into two parts, the part containing finitely many terms $\sup\left\{\frac{X_1}{n}, ..., \frac{X_N}{n}\right\}$ and the part containing infinitely many terms $\sup\left\{\frac{X_{N+1}}{n}, ..., \frac{X_n}{n}\right\}$. It's obvious that the latter part causes the problem. However, if we have $\frac{X_n^+}{n} \stackrel{a.s.}{\to} 0$ to hold, then $\forall \varepsilon > 0, \mathbb{P}\left(\frac{X_n^+}{n} < \varepsilon \text{ eventually}\right) = 1$ and the latter part no longer causes any trouble! The motivation of considering the positive part of r.v. comes from the structure of the running sup that we actually don't care about how much negative value those X_n can take.

As a result, $\frac{\sup_{m \leq n} X_m}{n} \stackrel{a.s.}{\to} 0$ is equivalent to saying $\frac{X_n^+}{n} \stackrel{a.s.}{\to} 0$ and $\sup\left\{\frac{X_1}{n}, ..., \frac{X_N}{n}\right\} \stackrel{a.s.}{\to} 0$. By part (1) we have just proved, it seems that $\mathbb{E}X_1^+ < \infty$ should be enough. Let's prove that they are actually equivalent. If $\mathbb{E}X_1^+ < \infty$, then $\frac{X_n^+}{n} \stackrel{a.s.}{\to} 0$ from part (1) and

$$\sup\left\{\frac{X_1}{n}, ..., \frac{X_N}{n}\right\} \leq \sup\left\{\frac{X_1^+}{n}, ..., \frac{X_N^+}{n}\right\} \stackrel{a.s.}{\to} 0 \tag{151}$$

since $\mathbb{E}X_1^+ < \infty$ guarantees that $\forall n, X_n^+ < \infty\ a.s.$ and the sup of finitely many finite values is still finite. Conversely, if $\frac{X_n^+}{n} \stackrel{a.s.}{\to} 0$, from part (1), we have $\mathbb{E}X_1^+ < \infty$. In all, the equivalent condition for $\frac{\sup_{m \leq n} X_m}{n} \stackrel{a.s.}{\to} 0\ (n \to \infty)$ is that $\mathbb{E}X_1^+ < \infty$.

(3): From the definition, $\frac{X_n}{n} \stackrel{p}{\to} 0$ iff

$$\forall \varepsilon > 0, \mathbb{P}\left(|X_n| \geq n\varepsilon\right) = \mathbb{P}\left(|X_1| \geq n\varepsilon\right) \to 0 \tag{152}$$

specifying $\varepsilon = 1$ to see that $\mathbb{P}(|X_1| \geq n) \to 0$. From the continuity of probability measure,

$$\mathbb{P}(|X_1| = \infty) = \lim_{n\to\infty} \mathbb{P}(|X_1| \geq n) = 0 \tag{153}$$

proves that $|X_1| < \infty$ *a.s.*. Conversely, if $|X_1| < \infty$ *a.s.*, then $\forall \varepsilon > 0, \frac{|X_1|}{\varepsilon} < \infty$ *a.s.* and $\mathbb{P}\left(\frac{|X_1|}{\varepsilon} \geq n\right) = \mathbb{P}\left(\frac{|X_n|}{\varepsilon} \geq n\right) \to 0$ proves $\frac{X_n}{n} \xrightarrow{p} 0$. The equivalent condition is thus given by $|X_1| < \infty$ *a.s.*.

(4): The splitting technique for sup does not work any longer in the convergence in probability (think about the reason why it does not work). So let's start from definition that

$$\forall \varepsilon > 0, \mathbb{P}\left(\sup_{m\leq n} X_m \geq n\varepsilon\right) \to 0 \tag{154}$$

specifying $\varepsilon = 1$ (from the same argument as above, we know considering $\varepsilon = 1$ suffices) to see that it's equivalent to saying

$$\mathbb{P}\left(\sup_{m\leq n} X_m \geq n\right) \to 0 \tag{155}$$

consider its complement

$$\mathbb{P}\left(\sup_{m\leq n} X_m < n\right) = [\mathbb{P}(X_1 < n)]^n \to 1 \tag{156}$$

taking log to get

$$\frac{\log \mathbb{P}(X_1 < n)}{n} \to 0 \ (n \to \infty) \tag{157}$$

as the equivalent condition. □

Remark. *It's a useful conclusion that* $\mathbb{E}|X|$ *and* $\sum_{n=1}^{\infty} \mathbb{P}(|X| \geq n)$ *are both finite or both infinite. To see this fact*

$$\mathbb{E}|X| = \sum_{n=0}^{\infty} \mathbb{E}(|X|\mathbb{I}_{n\leq|X|<n+1}) \leq \sum_{n=0}^{\infty}(n+1)\mathbb{P}(n \leq |X| < n+1) = \sum_{n=0}^{\infty} \mathbb{P}(|X| \geq n) \tag{158}$$

the first equation follows from monotone convergence theorem and the second equation follows from Fubini's theorem. On the other hand,

$$\mathbb{E}|X| = \sum_{n=0}^{\infty} \mathbb{E}(|X|\mathbb{I}_{n\leq|X|<n+1}) \geq \sum_{n=0}^{\infty} n\mathbb{P}(n \leq |X| < n+1) = \sum_{n=1}^{\infty} \mathbb{P}(|X| \geq n) \tag{159}$$

proves the conclusion.

SLLN

SLLN tells us for $\{X_n\}$ to be a sequence of *i.i.d.* r.v., if $\mathbb{E}|X_1| < \infty$ then $\frac{\sum_{i=1}^n X_i}{n} \stackrel{a.s.}{\to} \mathbb{E}X_1$ $(n \to \infty)$. Actually, the converse of SLLN is also true, meaning that if this almost sure convergence holds then $\mathbb{E}|X_1| < \infty$. The proof of SLLN can be provided in multiple ways and each of them shows a different but important probabilistic perspective. Among all proofs, I recommend checking three of them if you are interested. The first proof starts from the convergence in probability and L^2 and then use subsequence techniques to lift it to almost sure convergence. The second proof makes use of Kronecker's lemma and the convergence of a series of independent r.v. (Kolmogorov's three series theorem). The third proof uses backward martingale (a martingale propagating backward in time) and recognizes the SLLN as a special case of the martingale convergence theorem.

Despite those theoretical interests, our main focus is the application of SLLN. I provide two problems below for SLLN, each having important implications in probability theory.

Lemma 15. *Catastrophes happen at time $T_1, T_2, ...$ where*

$$T_i = \sum_{j=1}^{i} X_j \tag{160}$$

and $\{X_n\}$ is a sequence of positive i.i.d. r.v. such that $\mathbb{P}(X_1 = 0) < 1$. Let

$$N_t = \max\{n : T_n \le t\} \tag{161}$$

be the number of catastrophes happened until time t. Prove that if $\mathbb{E}|X_1| < \infty$ then $\frac{N_t}{t} \stackrel{a.s.}{\to} \frac{1}{\mathbb{E}X_1}$ $(t \to \infty)$.

Proof. If $\mathbb{E}|X_1| < \infty$, obviously $\mathbb{E}X_1 > 0$ so due to SLLN

$$\frac{T_n}{n} \stackrel{a.s.}{\to} \mathbb{E}X_1 \ (n \to \infty) \tag{162}$$

let's now think about N_t, it's clear that whenever $t \in [T_k, T_{k+1})$ is between the occurence of catastrophes, $N_t = k$ is constant. It's possible to establish bounds for $\frac{N_t}{t}$ that

$$\forall k, \forall t \in [T_k, T_{k+1}), \frac{k}{k+1} \cdot \frac{k+1}{T_{k+1}} = \frac{k}{T_{k+1}} < \frac{N_t}{t} = \frac{k}{t} \le \frac{k}{T_k} \tag{163}$$

set $k \to \infty$, it's clear that $\frac{k}{T_k} \stackrel{a.s.}{\to} \frac{1}{\mathbb{E}X_1}, \frac{k}{k+1} \cdot \frac{k+1}{T_{k+1}} \stackrel{a.s.}{\to} \frac{1}{\mathbb{E}X_1}$ $(k \to \infty)$. Notice that $T_k \stackrel{a.s.}{\to} +\infty$ $(k \to \infty)$ from SLLN and that $\mathbb{E}X_1 > 0$ so as $k \to \infty$, we have $t \to \infty$. This proves that

$$\frac{N_t}{t} \stackrel{a.s.}{\to} \frac{1}{\mathbb{E}X_1} \ (t \to \infty) \tag{164}$$

$\square$

Remark. *This result is the fundamental result in renewal theory, saying that the frequency of renewal a.s. converges to the reciprocal of the expected lifespan, which is intuitively correct.*

Lemma 16. *Interval $[0,1]$ is partitioned into n disjoint intervals with length $p_1, ..., p_n$. The entropy of this partition is defined as*

$$h = -\sum_{i=1}^{n} p_i \log p_i \tag{165}$$

Now $\{X_n\}$ is a sequence of i.i.d. $U(0,1)$ r.v. and $Z_{m(i)}$ denotes the number among $X_1, ..., X_m$ which lie in the i-th interval of the partition above. Show that if we define

$$R_m = \prod_{i=1}^{n} p_i^{Z_{m(i)}} \tag{166}$$

then

$$\frac{\log R_m}{m} \overset{a.s.}{\to} -h \ (m \to \infty) \tag{167}$$

Proof. Let's try to formalize an expression for $Z_{m(i)}$. Denote the interval partition as $I_1, ..., I_n$ with length $p_1, ..., p_n$,

$$Z_{m(i)} = \sum_{j=1}^{m} \mathbb{I}_{X_j \in I_i} \tag{168}$$

so

$$\begin{aligned} \frac{\log R_m}{m} &= \frac{\sum_{i=1}^{n} Z_{m(i)} \log p_i}{m} && (169) \\ &= \sum_{i=1}^{n} \log p_i \frac{\sum_{j=1}^{m} \mathbb{I}_{X_j \in I_i}}{m} && (170) \end{aligned}$$

from SLLN, since $\mathbb{I}_{X_1 \in I_i}, ..., \mathbb{I}_{X_m \in I_i}$ are *i.i.d.* and integrable,

$$\frac{\sum_{j=1}^{m} \mathbb{I}_{X_j \in I_i}}{m} \overset{a.s.}{\to} \mathbb{P}(X_1 \in I_i) = p_i \ (m \to \infty) \tag{171}$$

as a result,

$$\frac{\log R_m}{m} \overset{a.s.}{\to} \sum_{i=1}^{n} p_i \log p_i = -h \ (m \to \infty) \tag{172}$$

□

Remark. *Entropy indicates how chaotic a probability distribution is. The more uncertainty contained in a probability distribution, the higher entropy it has. For example, for a single point mass, $p_1 = 1$ so the entropy is zero since there's no uncertainty at all.*

An interesting question to ask is that: does there exists any probability distribution with the maximum entropy?

The answer is yes and varies under different constraints. For example, if the mean and variance of a distribution is specified, Gaussian has the maximum entropy. If the mean of a distribution is specified as a positive number and the distribution is supported on $\mathbb{R}_+$*, exponential has the maximum entropy. This concept has close connection with information theory and statistics.*

Characteristic Function and Tightness

c.f. and d.f. has a one-to-one correspondence due to Levy's inversion formula. Convergence in distribution can also be characterized as the pointwise convergence of c.f. under the tightness condition. As a result, it's important to understand properties of c.f. and tightness.

The following problem shows some property of c.f.

Lemma 17. *Show that if $\phi(t) = 1 + o(t^2)$ $(t \to 0)$ then $\phi \equiv 1$.*

The symmetric α-stable law is a distribution with c.f. $\phi(t) = e^{-|t|^\alpha}$, use the fact proved above to derive a sufficient range of α such that $\phi(t)$ is a legal c.f.

Proof. Now that $\lim_{t\to 0} \frac{\phi(t)-1}{t^2} = 0$, it's clear that

$$\phi'(0) = \lim_{t\to 0} \frac{\phi(t)-1}{t} = 0 \tag{173}$$

so $\mathbb{E}|X| < \infty, \mathbb{E}X = 0$. On the other hand, recall the finite difference representation of the second derivative that

$$\phi''(0) = \lim_{t\to 0} \frac{\phi(t) - 2 + \phi(-t)}{t^2} = 0 \tag{174}$$

so $\mathbb{E}X^2 < \infty, \mathbb{E}X^2 = 0$ which implies that $X = 0$ *a.s.* and $\forall t \in \mathbb{R}, \phi(t) = 1$.

Obviously, c.f. has value 1 at $t = 0$ so $\alpha > 0$. On the other hand, α cannot be large enough. Through Taylor expansion,

$$e^{-|t|^\alpha} = 1 - |t|^\alpha + o(|t|^\alpha) \ (t \to 0) \tag{175}$$

and the r.v. degenerates if $|t|^\alpha = o(t^2)$, i.e. if $\alpha > 2$. As a result, $\alpha \in (0, 2]$(this is also necessary implied by Polya's criterion on c.f.). Notice that when $\alpha = 1$, it's the Cauchy distribution and when $\alpha = 2$, it's the Gaussian. □

Lemma 18. *If X_n has c.f. ϕ_n, prove that $X_n \xrightarrow{d} 0$ $(n \to \infty)$ iff*

$$\exists \delta > 0, \forall |t| \leq \delta, \phi_n(t) \to 1 \ (n \to \infty) \tag{176}$$

Proof. By Levy's continuity theorem, if $X_n \xrightarrow{d} 0$ $(n \to \infty)$ then $\forall t \in \mathbb{R}, \phi_n(t) \to 1$ $(n \to \infty)$.

Conversely, if $\exists \delta > 0, \forall |t| \leq \delta, \phi_n(t) \to 1$ $(n \to \infty)$, then if we denote $\phi(t) = \lim_{n\to\infty} \phi_n(t)$, $\phi(t)$ is constantly 1 in $[-\delta, \delta]$, continuous at $t = 0$. From the proof of Levy's continuity theorem, we know that this implies $\{X_n\}$ is tight so for any subsequence X_{n_k} there always exists a further subsequence $X_{n_{k_q}}$ converging in distribution. As a result, $\forall t \in \mathbb{R}, \phi_{n_{k_q}}(t) \to \phi(t)$ $(q \to \infty)$. Since this limit ϕ is a legal c.f. and it is constantly 1 in some neighborhood of $t = 0$, the last result tells us that $\phi \equiv 1$. As a result, for any subsequence of $\{X_n\}$, there always exists a further subseqeunce converging in distribution to 0, which concludes the proof. □

When it comes to tightness, we have seen Levy-Prokhorov's theorem stating the fact that a sequence of d.f. is tight iff there exists a weak converging subsequence, which we have used in the problem above. Other than that, it shall be clear to us the definition of tightness that $\{X_n\}$ is tight if

$$\forall\varepsilon > 0, \exists M > 0, \forall n, \mathbb{P}\left(|X_n| \geq M\right) < \varepsilon \tag{177}$$

so what kind of conditions imply tightness and how to interpret tightness intuitively?

The key point in the definition of tightness is that M is uniform in n, saying that given error tolerance ε, there exists a uniform bound M such that $[-M, M]$ only misses ε probability mass of any X_n. In other words, **tightness is saying that the probability mass has to stay in some compact set $[-M, M]$ and cannot escape to infinity!** To see an example of a violation of tightness, check the following problem.

Lemma 19. *$\{X_n\}$ is a sequence of independent r.v. with $X_n \sim U(-n, n)$, check if $\{X_n\}$ is tight and check if it the d.f. of X_n weakly converges to a legal d.f. as $n \to \infty$.*

Proof. The probability mass of X_1 stays on $[-1, 1]$, the probability mass of X_2 stays on $[-2, 2]$ but the support of r.v. is expanding to the whole real line. In other words, there is probability mass escaping to infinity as $n \to \infty$. As a result, we expect this sequence of r.v. to be not tight. Let's prove this fact below. Consider $\varepsilon = \frac{1}{2} > 0$,

$$\forall M > 0, \exists n = 2M, \mathbb{P}\left(|X_n| \geq M\right) = \mathbb{P}\left(|X_{2M}| \geq M\right) = \frac{1}{2} \geq \varepsilon \tag{178}$$

which is the negation of the definition of tightness.

When it comes to the d.f.,

$$F_{X_n}(x) = \begin{cases} 0 & x \leq -n \\ \frac{x+n}{2n} & -n < x < n \\ 1 & x \geq n \end{cases} \tag{179}$$

so $\forall x \in \mathbb{R}, F_{X_n}(x) \to \frac{1}{2}$ $(n \to \infty)$. It's clear that $\forall x \in \mathbb{R}, F(x) = \frac{1}{2}$ is not a legal d.f.

□

From the example above, we see that how the escape of probability mass may cause the failure of the limit of d.f. to be not a d.f. any longer. Naturally, we would ask: what conditions imply tightness? One of the useful conditions has already appeared in the proof of Levy's continuity theorem, which is that the pointwise limit of c.f. is continuous at $t = 0$. Naturally, if a sequence of r.v. $\{X_n\}$ is known to converge in distribution, it must be tight. Another frequently used condition is that $\exists \delta > 0, \sup_n \mathbb{E}|X_n|^\delta < \infty$. This is due to an application of Markov inequality that

$$\mathbb{P}\left(|X_n| \geq M\right) \leq \frac{\mathbb{E}|X_n|^\delta}{M^\delta} \leq \frac{\sup_n \mathbb{E}|X_n|^\delta}{M^\delta} \tag{180}$$

take M such that $\frac{\sup_n \mathbb{E}|X_n|^\delta}{M^\delta} \leq \varepsilon$, it's clear that M only depends on ε and is uniform in n. That's why $\{X_n\}$ is tight if it has some positive moments to be uniformly bounded.

Lemma 20. *$\{X_n\}$ is a sequence of non-negative r.v. such that $\mathbb{E}X_n^\alpha \to 1, \mathbb{E}X_n^\beta \to 1$ $(n \to \infty)$ for $0 < \alpha < \beta$. Show that $X_n \xrightarrow{p} 1$ $(n \to \infty)$.*

Proof. This problem seems easy and hard at the same time. It's easy since the two convergent moment sequences has some meaning of sandwiching and it's hard since we do not know how to make use of those conditions.

However, if we notice that the convergence of moment sequence implies that $\sup_n \mathbb{E}|X_n|^\alpha < \infty$, we know the sequence $\{X_n\}$ is tight. Due to Levy-Prokhorov theorem, for any subsequence $\{X_{n_k}\}$, there exists a further subsequence $X_{n_{k_q}} \xrightarrow{d} X$ $(q \to \infty)$. Now that

$$\sup_n \mathbb{E}|X_n|^\beta = \sup_n \mathbb{E}(|X_n|^\alpha)^{\frac{\beta}{\alpha}} < \infty, \frac{\beta}{\alpha} > 1 \tag{181}$$

we conclude that $\{|X_n|^\alpha\}$ is uniformly integrable and thus

$$\lim_{q\to\infty} \mathbb{E}X_{n_{k_q}}^\alpha = \mathbb{E}X^\alpha = 1 \tag{182}$$

since all those r.v. are non-negative, Fatou's lemma for convergence in distribution (proved by Skorokhod representation) tells us

$$\mathbb{E}X^\beta \le \liminf_{q\to\infty} \mathbb{E}X_{n_{k_q}}^\beta = 1 \tag{183}$$

written in terms of the L^p norm of random variables,

$$||X||_\alpha = 1, ||X||_\beta \le 1 \tag{184}$$

since $\forall p > q > 0, ||X||_p \ge ||X||_q$, it's clear that

$$\forall \gamma \in [\alpha, \beta], ||X||_\gamma = 1 \tag{185}$$

this implies that $X = 1$ *a.s.* (I will leave this proof to you, think about splitting using indicator w.r.t. if $|X| \ge 1$)

We have proved any subsequence of $\{X_n\}$ has a further subsequence converging in distribution to 1, so $X_n \xrightarrow{d} 1$ $(n \to \infty)$. Since the limit is a.s. constant, the convergence can be lifted to the convergence in probability $X_n \xrightarrow{p} 1$ $(n \to \infty)$.

□

Week 6

Martingale (MG)

The definition of a MG $\{X_n\}$ under filtration $\{\mathscr{F}_n\}$ consists of three main points, integrability at each time step, adaptiveness to $\{\mathscr{F}_n\}$ and the conditional expectation condition $\mathbb{E}(X_n|\mathscr{F}_{n-1}) = X_{n-1}$. Basically one can understand a MG as a "stationary"/"fair" sequence. Since conditional expectation can be understood as the orthogonal projection on L^2 space of r.v., $\mathbb{E}(X_n|\mathscr{F}_{n-1})$ is actually the best approximation (under mean square error) to X_n using the information till time $n-1$. As a result, a MG $\{X_n\}$ satisfies the property that the best approximation to its value at time n based on all information till time $n-1$ is just its value at time $n-1$. The "stationarity" of MG can thus be understood in this sense. Similar interpretations also hold for sub/super-MG.

As an example, stock price cannot be a MG, that's why it's always ridiculous to expect the stock price to stay close to its historical level and a trading strategy that works well in the past might not work well in the future. On the other hand, process with independent centered (mean zero) increments are MG (please check this fact).

The adaptiveness condition seems not that important and may be ignored by some of the readers. Actually this is quite an important condition restricting the flow of information we can have access to. In a later context, we shall see examples where a change in the filtration results in a process not maintaining its martingality any longer.

Lemma 21. *$\{X_n\}$ is a sequence of r.v. with finite means and $\mathbb{E}(X_{n+1}|X_0, ..., X_n) = aX_n + bX_{n-1}$ where $0 < a, b < 1, a + b = 1$. Find a value of α such that $S_n = \alpha X_n + X_{n-1}$ is a MG under the natural filtration $\{\mathscr{F}_n\}$ where $\mathscr{F}_n = \sigma(X_0, ..., X_n)$.*

Proof. Surely $\{S_n\}$ is adapted to $\{\mathscr{F}_n\}$ and is integrable for $\forall n$. Now let's compute the conditional expectation

$$\mathbb{E}(S_n|\mathscr{F}_{n-1}) = X_{n-1} + \alpha\mathbb{E}(X_n|\mathscr{F}_{n-1}) = (1 + \alpha a)X_{n-1} + \alpha b X_{n-2} \tag{186}$$

if we want to match it with $S_{n-1} = \alpha X_{n-1} + X_{n-2}$, it suffices to set $1 + \alpha a = \alpha, \alpha b = 1$ so $\alpha = \frac{1}{b}$. $\square$

MG Convergence

Since MG has the meaning of "stationarity", it's natural to ask if a MG always converges and what notion of convergence holds for MG. It should be easy for one to build up a MG $\{S_n\}$ such that $S_n \overset{a.s.}{\not\to} \infty$ $(n \to \infty)$. Consider a process $\{S_n\}$ with independent centered increments $X_1, ..., X_n$ and the filtration $\mathscr{F}_n = \sigma(X_1, ..., X_n)$. When the perturbation of X_n gets larger and larger when n gets larger and larger, it's natural to expect that $S_n \overset{a.s.}{\not\to} \infty$ $(n \to \infty)$ holds.

However, there also exists a MG $\{S_n\}$ such that $S_n \overset{a.s.}{\to} +\infty$ $(n \to \infty)$, which may be a little bit hard to construct since this MG is required to go toward the positive side almost surely when n gets large enough. Let's go through the construction of this counterexample since it helps us understand the "stationarity" of MG in a better way. Similar to above, we still consider $\{S_n\}$ as a process with independent centered increments, i.e.

$$S_n = \sum_{i=1}^{n} X_i \tag{187}$$

where $X_1, ..., X_n, ...$ are independent and $\forall n, \mathbb{E}X_n = 0$. Obviously, this guarantees that $\{S_n\}$ is a MG under the filtration generated by $\{X_n\}$.

How are we going to proceed to build up the trend of the MG going towards the positive side almost surely? The hard point is that any increment X_n shall have zero expectation while exhibiting some trend towards the positive side. This is exactly the place where Borel-Cantelli comes in. According to Borel-Cantelli, even if a sequence of random variables can take extreme values, as long as the probability of taking extreme values decays fast enough, its impact is asymptotically negligible. Let's split the distribution of each increment X_n into the positive part and the negative part. For the positive part, we don't want Borel-Cantelli to ignore the asymptotic impact so we are going to assign probability mass $\frac{1}{n}$ to extreme positive values of X_n (since $\sum_n \frac{1}{n} = \infty$). On the other hand, in order to ensure that X_n has zero expectation, we must add some extreme negative values. We want Borel-Cantelli to ignore the asymptotic impact of those extreme negative values so we are going to assign probability mass $\frac{1}{n^2}$ to extreme negative values of X_n (since $\sum_n \frac{1}{n^2} < \infty$).

As a result, a natural construction of X_n is provided as

$$X_n = \begin{cases} n & w.p. \frac{1}{n} \\ -n^2 & w.p. \frac{1}{n^2} \\ 0 & \text{else} \end{cases} \tag{188}$$

such that $\mathbb{E}X_n = 0$. From Borel-Cantelli, since $\{X_n\}$ are independent,

$$\mathbb{P}\left(X_n = n \ i.o.\right) = 1, \mathbb{P}\left(X_n = -n^2 \ i.o.\right) = 0 \tag{189}$$

in other words, the asymptotic behavior of S_n is the same as the asymptotic behavior of $\sum_{i=1}^{n} i \to +\infty \ (n \to \infty)$, i.e. $S_n \overset{a.s.}{\to} +\infty \ (n \to \infty)$.

Remark. *From the counterexample above, we realize that the "stationarity" of MG is under the sense of expectation, which means that extreme values can show up and mess things up as long as all extreme values eliminate the impact of each other under expectation.*

Lemma 22. *Consider $\{S_n\}$ as the stock price at time n with*

$$S_n = \prod_{i=1}^{n} X_i \tag{190}$$

where $X_1, X_2, ...$ are i.i.d. random variables taking value 1.1 with probability $\frac{1}{2}$ and taking value 0.9 with probability $\frac{1}{2}$. In other words, the stock price either rise by 10% or drop by 10% per time step with equal probability. Show that

S_n is a MG w.r.t. the filtration generated by $\{X_n\}$ that converges almost surely, identify the limit and show that it does not converge in L^1.

Proof. Obviously $\{S_n\}$ is adapted and $\mathbb{E}S_n = \prod_{i=1}^n \mathbb{E}X_i = 1 < \infty$ so it's integrable. Check that

$$\mathbb{E}(S_n|\mathscr{F}_{n-1}) = S_{n-1} \cdot \mathbb{E}(X_n|\mathscr{F}_{n-1}) = S_{n-1} \cdot \mathbb{E}X_n = S_{n-1} \tag{191}$$

so $\{S_n\}$ is a MG.

Notice that S_n has the special structure as a product of *i.i.d.* r.v. If it can be written as a sum of *i.i.d.* r.v., SLLN can help us prove the convergence and identify the limit! Naturally, we think of taking log.

$$\log S_n = \sum_{i=1}^n \log X_i \tag{192}$$

since $\log X_1$ is integrable, $\mathbb{E}\log X_1 = \frac{1}{2}\log 0.99$, apply SLLN

$$\frac{\log S_n}{n} \overset{a.s.}{\to} \frac{1}{2}\log 0.99 < 0 \ (n \to \infty) \tag{193}$$

we conclude that $\log S_n \overset{a.s.}{\to} -\infty \ (n \to \infty)$ so $S_n \overset{a.s.}{\to} 0 \ (n \to \infty)$ proves the convergence and identifies the limit.

If $S_n \overset{L^1}{\to} 0 \ (n \to \infty)$, then $\mathbb{E}S_n \overset{L^1}{\to} 0 \ (n \to \infty)$, however, $\forall n, \mathbb{E}S_n = 1$, a contradiction! This shows that L^1 convergence fails. $\square$

Unluckily, the almost sure convergence of MG cannot imply the L^1 convergence of MG due to a lack of uniform integrability that we will learn in a later context.

Now let's shift gear to check some corollary of Doob's maximal inequality which is an important tool for estimating the tail probability of a MG.

Lemma 23 (Kolmogorov's maximal inequality)**.** *Let $\{X_n\}$ be a sequence of independent centered r.v. with finite variance and $S_n = \sum_{i=1}^n X_i$, show that*

$$\forall \varepsilon > 0, \mathbb{P}\left(\sup_{j\leq n} |S_j| > \varepsilon\right) \leq \frac{\sum_{j=1}^n Var(X_j)}{\varepsilon^2} \tag{194}$$

Proof. When we see the tail probability contains the sup, Doob's maximal inequality immediately comes to our mind. To construct a MG, notice that we already have independent centered r.v. and they work as increments of $\{S_n\}$ so $\{S_n\}$ must be a MG under the filtration generated by $\{X_n\}$. As a result, $\{S_n^2\}$ is a sub-MG, by Doob's maximal inequality,

$$\forall \varepsilon > 0, \mathbb{P}\left(\sup_{j\leq n} |S_j| > \varepsilon\right) = \mathbb{P}\left(\sup_{j\leq n} |S_j|^2 > \varepsilon^2\right) \leq \frac{\mathbb{E}S_n^2}{\varepsilon^2} = \frac{\sum_{j=1}^n Var(X_j)}{\varepsilon^2} \tag{195}$$

concludes the proof. $\square$

Lemma 24. *Let $\{X_n\}$ be a sequence of independent centered r.v. with finite variance and $\sum_{n=1}^\infty \frac{Var(X_n)}{n^2} < \infty$, show that*

$$\sum_{i=1}^n \frac{X_i}{i} \stackrel{a.s.}{\to} Y \ (n \to \infty) \tag{196}$$

for some almost surely finite r.v. Y. Show further that

$$\frac{\sum_{i=1}^n X_i}{n} \stackrel{a.s.}{\to} 0 \ (n \to \infty) \tag{197}$$

Proof. Consider $S_n = \sum_{i=1}^n \frac{X_i}{i}$ as a MG under the filtration generated by $\{X_n\}$, by MG L^2 convergence theorem,

$$\sup_n \mathbb{E} S_n^2 = \sum_{n=1}^\infty \frac{Var(X_n)}{n^2} < \infty \tag{198}$$

so $S_n \stackrel{a.s.,L^2}{\to} S_\infty \ (n \to \infty)$ where $S_\infty < \infty \ a.s., S_\infty \in L^2$ proves the convergence.

To further prove the second fact, notice that $b_n = n$ is an increasing sequence with limit $+\infty$ and $a_n = X_n$ such that $\sum_{n=1}^\infty \frac{a_n}{b_n} = \sum_{n=1}^\infty \frac{X_n}{n} < \infty \ a.s.$ so by Kronecker's lemma

$$\frac{\sum_{i=1}^n X_i}{n} \stackrel{a.s.}{\to} 0 \ (n \to \infty) \tag{199}$$

concludes the proof. □

Remark. *This provides a proof for the weakened version of SLLN (for L^2 random variables) from the perspective of the convergence of the series of independent random variables.*

Just in case the reader is not clear with Kronecker's lemma, we state it here. If $\{a_n\}, \{b_n\}$ are real sequences with $b_n \nearrow \infty \ (n \to \infty)$ and $\sum_{n=1}^\infty \frac{a_n}{b_n} < \infty$, then $\frac{\sum_{i=1}^n a_i}{b_n} \to 0 \ (n \to \infty)$. The proof is a classical application of summation by parts and the reader can find it on any analysis textbook.

Week 7

Uniform Integrability (UI)

In last week, we show an example of a MG that converges almost surely but fails to converge in L^1, that is exactly due to the failure of UI. The definition of UI for a sequence of r.v. $\{X_n\}$ is

$$\sup_n \mathbb{E}|X_n|\mathbb{I}_{|X_n|\geq\lambda} \to 0 \ (\lambda \to \infty) \tag{200}$$

the expectation is taken only at the tail of X_n and the uniformity comes from the supreme in n. Directly speaking, UI condition is telling us that as λ grows large, the expectation of its tail part shrinks to zero uniformly. So why is it called "integrability"? Let's consider the case when $n = 1$, i.e. there's a single r.v. X, the condition becomes

$$\mathbb{E}|X|\mathbb{I}_{|X|\geq\lambda} \to 0 \ (\lambda \to \infty) \tag{201}$$

which is exactly the integrability of X, i.e. $X \in L^1$ by Cauchy principle.

UI condition is very useful due to **Vitali's convergence theorem** (we do not present the proof here but one is strongly encouraged to check the proof of this theorem). The theorem is saying that if $X_n \xrightarrow{p} X$ $(n \to \infty)$, then the convergence in L^p, i.e. $X_n \xrightarrow{L^p} X$ $(n \to \infty)$, is equivalent to the convergence of L^p norm, i.e. $\mathbb{E}|X_n|^p \xrightarrow{L^p} \mathbb{E}|X|^p$ $(n \to \infty)$, is equivalent to $\{|X_n|^p\}$ being UI.

In the class, we have learnt one equivalent condition for UI, which is $\{X_n\}$ being L^1 bounded, i.e. $\sup_n \mathbb{E}|X_n| < \infty$ and $\{X_n\}$ being uniformly absolute continuous. However, this equivalent condition is not at all useful since it's hard to check the uniformly absolute continuity. Instead, we use the following conditions to imply UI.

Lemma 25 (Sum of UI sequence is UI)**.** *$\{X_n\}, \{Y_n\}$ are two UI sequence, then $\{X_n + Y_n\}$ is UI.*

Proof. Notice that

$$\mathbb{E}|X_n + Y_n|\mathbb{I}_{|X_n+Y_n|\geq 2\lambda} \leq 2\mathbb{E}|X_n|\mathbb{I}_{|X_n|\geq\lambda} + 2\mathbb{E}|Y_n|\mathbb{I}_{|Y_n|\geq\lambda} \tag{202}$$

take sup w.r.t. n on both sides and set $\lambda \to \infty$ to see

$$\sup_n \mathbb{E}|X_n + Y_n|\mathbb{I}_{|X_n+Y_n|\geq 2\lambda} \to 0 \ (\lambda \to \infty) \tag{203}$$

proves that $\{X_n + Y_n\}$ is UI. $\square$

Lemma 26 (Uniformly bounded $1+\varepsilon$-moment implies UI)**.** *$g : \mathbb{R}_+ \to \mathbb{R}_+$ be such that $\frac{g(x)}{x} \to \infty$ $(x \to \infty)$, show that if $\sup_n \mathbb{E}g(|X_n|) < \infty$ then $\{X_n\}$ is UI. Take $g(x) = x^{1+\varepsilon}$ for some $\varepsilon > 0$, what conclusion do we get?*

Proof.

$$\mathbb{E}|X_n|\mathbb{I}_{|X_n|\geq\lambda} = \mathbb{E}\frac{|X_n|}{g(|X_n|)}g(|X_n|)\mathbb{I}_{|X_n|\geq\lambda} \tag{204}$$

since $\frac{x}{g(x)} \to 0$ $(x \to \infty)$,

$$\forall \varepsilon > 0, \exists K, \forall x > K, \frac{x}{g(x)} < \varepsilon \tag{205}$$

when λ is large enough, $|X_n| \geq \lambda$ implies $|X_n| > K$ so $\frac{|X_n|}{g(|X_n|)} < \varepsilon$,

$$\sup_n \mathbb{E}|X_n|\mathbb{I}_{|X_n|\geq\lambda} \leq \varepsilon \sup_n \mathbb{E}g(|X_n|) \tag{206}$$

since $\sup_n \mathbb{E}g(|X_n|) < \infty$, this proves UI.

Take $g(x) = x^{1+\varepsilon}$ for $\varepsilon > 0$ so it satisfies the condition of this conclusion, if $\exists \varepsilon > 0, \sup_n \mathbb{E}|X_n|^{1+\varepsilon} < \infty$, then $\{X_n\}$ is UI. □

Remark. *Uniform boundedness in $1+\varepsilon$-th moment implies UI but uniform boundedness in the first moment does not imply UI! Think about if the converse of the conclusion above is true, i.e. if $\{X_n\}$ UI, can we guarantee that $\exists \varepsilon > 0, \sup_n \mathbb{E}|X_n|^{1+\varepsilon} < \infty$?*

Lemma 27 (Dominated condition implies UI)**.** *If $\mathbb{E}\sup_n |X_n| < \infty$, then $\{X_n\}$ is UI.*

Proof.

$$\mathbb{E}|X_n|\mathbb{I}_{|X_n|\geq\lambda} \leq \mathbb{E}\sup_n |X_n|\mathbb{I}_{\sup_n |X_n|\geq\lambda} \tag{207}$$

take sup on both sides w.r.t. n

$$\sup_n \mathbb{E}|X_n|\mathbb{I}_{|X_n|\geq\lambda} \leq \mathbb{E}\sup_n |X_n|\mathbb{I}_{\sup_n |X_n|\geq\lambda} \tag{208}$$

since $\sup_n |X_n| \in L^1$,from Cauchy principle, RHS converges to zero as $\lambda \to \infty$ concludes the proof. □

At this point, one is welcome to go back to the final problem in the sample final to check why UI holds there. Although Vitali's convergence theorem provides a strong connection between convergence in probability and L^p convergence, we remark that it tells nothing about almost sure convergence, even if when all the r.v. has special structures.

Lemma 28. *Consider $M_n = X_n Y_n$ where*

$$X_n = \begin{cases} 1 & w.p.\ \frac{1}{n} \\ 0 & w.p.\ 1-\frac{1}{n} \end{cases} \tag{209}$$

$\{X_n\}$ is a sequence of independent r.v.,

$$Y_n = \prod_{i=1}^{n} \xi_i \tag{210}$$

where

$$\xi_n = \begin{cases} n^2 & w.p.\ \frac{1}{n^2} \\ 0 & w.p.\ 1 - \frac{1}{n^2} \end{cases} \tag{211}$$

where $\{\xi_n\}$ *is a sequence of independent r.v. and* $\{\xi_n\}$ *is independent of* $\{X_n\}$.

Set $\mathscr{F} = \sigma(X_1, X_2, ...)$, *prove that* M_n *converges almost surely and in* L^1 *to some limit* M, $\{M_n\}$ *is UI, and* $\mathbb{E}(M_n|\mathscr{F})$ *converges in* L^1 *to* $\mathbb{E}(M|\mathscr{F})$ *but not almost surely.*

Proof. By Borel-Cantelli, almost surely $\xi_n = 0$ eventually and $\mathbb{E}\xi_n = 1$ so $Y_n = 0$ eventually. This results in $M_n \overset{a.s.}{\to} 0$ $(n \to \infty)$. Now check

$$\mathbb{E}M_n = \mathbb{E}X_n \mathbb{E}Y_n = \frac{1}{n} \to 0\ (n \to \infty) \tag{212}$$

so by Vitali's convergence, $M_n \overset{L^1}{\to} 0$ $(n \to \infty)$, and $\{M_n\}$ is UI.

When it comes to the conditional expectation,

$$\mathbb{E}|\mathbb{E}(M_n|\mathscr{F}) - 0| \leq \mathbb{E}[\mathbb{E}(|M_n||\mathscr{F})] = \mathbb{E}|M_n| = \mathbb{E}|X_n| \cdot \mathbb{E}|Y_n| = \frac{1}{n} \to 0 \tag{213}$$

so $\mathbb{E}(M_n|\mathscr{F}) \overset{L^1}{\to} 0$ $(n \to \infty)$, however,

$$\mathbb{E}(M_n|\mathscr{F}) = X_n \mathbb{E}(Y_n) = X_n \tag{214}$$

and by Borel-Cantelli, $\mathbb{P}(X_n = 1\ i.o.) = 1$ so almost sure convergence of $\mathbb{E}(M_n|\mathscr{F})$ fails.

$\square$

Diagram for Convergence Mode

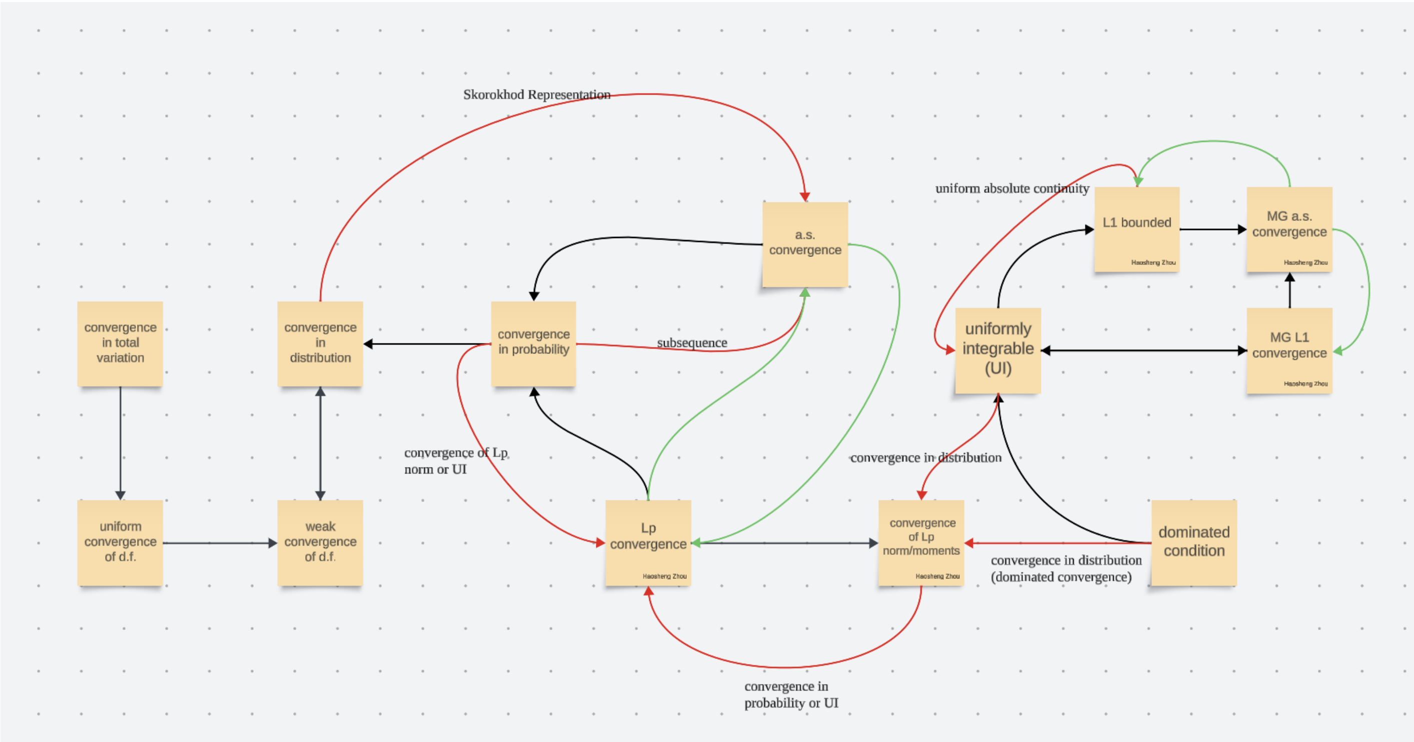

Figure 1: Diagram for Convergence Mode

Black edges as implication, red edges as implication under extra conditions, green edges as important counterexamples to keep in mind.

Weakly Stationary Process and Spectral Density

A weakly stationary process $\{X_t\}$ has its mean function $m(t) = \mathbb{E}X_t$ to be constant and covariance kernel $C(s,t) = cov(X_s, X_t)$ to be only a function on the time lag $t-s$, i.e. $C(s,t) = c(t-s)$. Interestingly, as long as $c(0) > 0$ is strictly positive and $c(t)$ is continuous at $t = 0$, the normalization

$$\rho(t) = \frac{c(t)}{c(0)} \tag{215}$$

gives a function ρ such that $\rho(0) = 1$ and is actually a c.f. of some underlying distribution, guaranteed by Bochner's theorem. This is the spectral theorem, the underlying distribution is denoted F and the spectral density $F' = f$ exists if F is a continuous probability distribution.

In this case, the Fourier transform of f gives ρ so an inverse Fourier transform (Levy's inversion formula) on ρ gives spectral density f, just to recall the result we have proved, if spectral density exists, it can be calculated through

$$f(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-itx}\rho(t)\,dt \tag{216}$$

if ρ is integrable. Let's prove below that for any distribution F we can always build up a weakly stationary process such that it has spectral distribution F.

Lemma 29. *$U \sim U(-\pi,\pi)$ and V is independent of U, $V \sim F$. Show that*

$$X_n = e^{i(U-Vn)} \tag{217}$$

defines a weakly stationary process with spectral distribution F.

Proof. Check

$$\mathbb{E}X_n = \mathbb{E}e^{iU}\mathbb{E}e^{-inV} \tag{218}$$

where $\mathbb{E}e^{iU} = 0$ so the mean function does not depend on n. Check covariance kernel, be careful here that since we are in the complex field, the covariance contains a complex conjugate

$$\forall s,t, C(s,t) = cov(X_s,X_t) = \mathbb{E}e^{i(t-s)V} = \phi_V(t-s) \tag{219}$$

it a function of the time lag $t-s$, so $c(t) = \phi_V(t)$ and $\rho(t) = c(t) = \phi_V(t)$. Obviously, the spectral distribution is just the distribution of V, which is F.

$\square$

Week 8

Spectral Theorem

Weakly stationary process X with zero mean, unit variance, continuous autocorrelation function and spectral distribution F always has the following representation

$$X_t = \int_{\mathbb{R}} e^{it\lambda}\, dS_\lambda \tag{220}$$

for some complex-valued process S. Moreover, such S is guaranteed to have orthogonal increments and $\mathbb{E}|S_v - S_u|^2 = F(v) - F(u)$ for $u \leq v$. If we are in the discrete time setting, the integral is on $(-\pi, \pi]$.

Some problems arise with this notation: how to define the integral w.r.t. a stochastic process? If the trajectory of a process gives a d.f., it's clear that this integration is well-defined having the structure of expectation. If the trajectory of a process gives an increasing function, it induces a measure on $\mathbb{R}$ (not necessarily finite) so this integration is well-defined as Lebesgue-Stieljes integral. More generally, if the trajectory of a process has finite total variation, any function with finite total variation can be written as the difference of two increasing functions. It induces a signed measure on $\mathbb{R}$ so this integration is well-defined through the extension Lebesgue-Stieljes integral preserving linearity. However, it's always the case that the trajectory of a process has very frequent change in its value so the total variation might not be finite. This is exactly the difficulty stochastic integration has to face.

After defining the stochastic integral is defined, the spectral theorem is proved through the white noise construction $S_\lambda = \mu(\mathbb{I}_\lambda)$ with μ to be the isometry between two Hilbert spaces. Actually it's fine if we don't understand all the details of the proof at this moment since we will come back to this again when we learn Brownian motion and martingale representation theorem. Let's just learn how to apply the spectral theorem.

Lemma 30. *If the spectral process S exists for X, then increments of S has zero mean.*

Proof. Clearly for $\forall u \leq v$, X_t is the L^2 limit of

$$\sum_i e^{it\lambda_{i-1}}(S_{\lambda_i} - S_{\lambda_{i-1}}) \tag{221}$$

since $\forall t, \mathbb{E}X_t = 0$,

$$\forall t, \sum_i e^{it\lambda_{i-1}}\mathbb{E}(S_{\lambda_i} - S_{\lambda_{i-1}}) = 0 \tag{222}$$

proves $\mathbb{E}(S_{\lambda_i} - S_{\lambda_{i-1}}) = 0$. □

Lemma 31. *X is discrete-time weakly stationary with strictly positive spectral density f, zero mean and spectral process S. Let*

$$Y_n = \int_{(-\pi,\pi]} \frac{e^{in\lambda}}{\sqrt{2\pi f(\lambda)}}\, dS_\lambda \tag{223}$$

show that $\{Y_n\}$ is a sequence of uncorrelated r.v. with zero mean and unit variance.

Proof. Since S has increments of zero mean, $\mathbb{E}Y_n = 0$. Calculate the covariance kernel of Y

$$\forall s \le t, cov(Y_s, Y_t) = \mathbb{E}Y_t\overline{Y_s} = \mathbb{E}\int_{(-\pi,\pi]} \frac{e^{-isu}}{\sqrt{2\pi f(u)}}\, d\overline{S}_u \int_{(-\pi,\pi]} \frac{e^{itu}}{\sqrt{2\pi f(u)}}\, dS_u \tag{224}$$

$$= \int_{(-\pi,\pi]} \frac{e^{i(t-s)u}}{2\pi f(u)}\, d\mathbb{E}[\overline{S}_u S_u] \tag{225}$$

since $\mathbb{E}[\overline{S}_u S_u] = F(u)$, it's clear that

$$\int_{(-\pi,\pi]} \frac{e^{i(t-s)u}}{2\pi f(u)}\, d\mathbb{E}[\overline{S}_u S_u] = \int_{(-\pi,\pi]} \frac{e^{i(t-s)u}}{2\pi f(u)} f(u)\, du \tag{226}$$

$$= \frac{1}{2\pi}\int_{(-\pi,\pi]} e^{i(t-s)u}\, du = \delta_{t=s} \tag{227}$$

we proved that $\{Y_n\}$ is a sequence of uncorrelated r.v. with unit variance. $\square$

Remark. *If one can find Fourier coefficients $\{a_j\}$ such that the Fourier series $\sum_{j\in\mathbb{Z}} a_j e^{-iju} = \sqrt{2\pi f(u)}$ for $\forall u \in (-\pi,\pi]$, then we have actually proved that the moving average decomposition*

$$X_n = \sum_{j\in\mathbb{Z}} a_j Y_{n-j} \tag{228}$$

holds.

Strong Stationarity and Ergodic Theorem

Consider strongly stationary sequence $\{X_n\}$ embedded into the probability space $(\mathbb{R}^T, \mathscr{B}^T, \mathbb{Q})$ through Kolmogorov's extension theorem, i.e. each sample point is a real sequence $x = (x_1, x_2, ..)$. The shift operator τ is defined mapping $(x_1, x_2, ...)$ to $(x_2, x_3, ...)$, i.e. shifting one time unit. In this case, τ is measure preserving, i.e. $\forall A \in \mathscr{B}^T, \mathbb{Q}(A) = \mathbb{Q}(\tau^{-1}A)$ for strongly stationary process $\{X_n\}$.

Ergodicity is formally defined through the invariant sigma field $\mathscr{I}$. The definition is simple: $A \in \mathscr{I}$ iff $A = \tau^{-1}A$. Let's check that this is indeed a valid sigma field.

Lemma 32. *Prove that $\mathscr{I}$ is actually a sigma field.*

Proof. First check that it contains $\emptyset$ and Ω. This is trivially correct.

Then check that it's closed under complement. If $A \in \mathscr{I}$, then $\forall x \in A^c, x \notin A$ so $x \notin \tau^{-1}A$ meaning $\tau(x) \notin A$, $\tau(x) \in A^c$ so $x \in \tau^{-1}A^c$. Conversely one can also check that it's true so $A^c \in \mathscr{I}$.

At last, check that it's closed under countable union. If $A_1, A_2, ... \in \mathscr{I}$, then $\forall x \in \bigcup_n A_n$, $\exists j, x \in A_j = \tau^{-1}A_j$ so $x \in \tau^{-1}\bigcup_n A_n$ proves $\bigcup_n A_n = \tau^{-1}\bigcup_n A_n, \bigcup_n A_n \in \mathscr{I}$. $\square$

A process is called ergodic if its invariant sigma field is trivial, i.e. all events within has probability zero or one happening. The following Birkhoff's ergodic theorem is fundamental for many applications in probability.

Theorem 2 (Birkhoff)**.** *If the process $\{X_n\}$ is generated by a single r.v. X on the probability space $(\mathbb{R}^T, \mathscr{B}^T, \mathbb{Q})$ and the shift operator τ, i.e.*

$$\forall x \in \mathbb{R}^T, X_n(x) = X(\tau^{n-1}(x)) \tag{229}$$

where $X_n(x_1, x_2, ...) = x_n$ is the projection onto the n-th component. Then

$$\frac{1}{n}\sum_{i=1}^n X_i \overset{a.s.,L^1}{\to} \mathbb{E}(X|\mathscr{I}) \ (n \to \infty) \tag{230}$$

Some examples of ergodic process includes a sequence of *i.i.d.* r.v. and an ergodic Markov chain (check the definition!) In this sense, SLLN and the SLLN for Markov chain both become the special case of this Birkhoff's ergodic theorem.

Week 9

Example of Ergodic and Non-ergodic Process

Last week we talk about the definition of ergodic process and the implications it might have. Let's have an intuitive understanding on what kind of process is ergodic. Consider strongly stationary process $\{X_n\}$, for which the invariant sigma field w.r.t. the shift operator τ is well-defined.

One important example is that an ergodic Markov chain (irreducible, positive recurrent, aperiodic) starting from stationary distribution is ergodic. Firstly, if a Markov chain is irreducible and positive recurrent, the stationary distribution must exist and is unique, denoted π. Consider the Markov chain $\{X_n\}$ with $X_0 \sim \pi$, then $\forall n, X_n \sim \pi$ so it's obviously strongly stationary due to Markov property. To argue that it's ergodic, we have to prove $\forall A \in \mathscr{I}$, $\mathbb{P}(A)$ is either zero or one. Notice that $\forall A \in \mathscr{I}, \forall n, A = \tau^{-n}A$ by the definition of invariant sigma field so why don't we combine this with the Markov property?

$$\mathbb{E}_{X_0\sim\pi}\mathbb{I}_A = \mathbb{E}(\mathbb{I}_{\tau^{-n}A}|\mathscr{F}_n) = \mathbb{E}(\mathbb{I}_A|\mathscr{F}_n) \overset{a.s.,L^1}{\to} \mathbb{E}(\mathbb{I}_A|\mathscr{F}_\infty) = \mathbb{I}_A \ (n\to\infty) \tag{231}$$

the first inequality is from Markov property and the convergence is from Levy's zero-one law where $\mathscr{F}_\infty = \sigma(\bigcup_{n=1}^\infty \mathscr{F}_n)$ so $A \in \mathscr{F}_\infty$ (one is welcome to prove this result on ones' own from the definition of conditional expectation). As a result, $\mathbb{P}_{X_0\sim\pi}(A) = \mathbb{I}_A \ a.s.$ is either zero or one, proves the ergodicity.

As a result, the proof of the SLLN-type theorem for ergodic Markov chain can be naturally carried out through the Birkhoff's ergodic theorem. For ergodic Markov chain $\{X_n\}$ with any initial distribution, we have proved that $X_n \overset{d}{\to} \pi \ (n\to\infty)$ so after a long enough time we expect to see this Markov chain reach the stationary distribution, which is equivalent to saying the asymptotic behavior of this Markov chain is the same as the asymptotic behavior of Markov chain $\{Y_n\}$, which has the same transition dynamics as $\{X_n\}$ but $Y_0 \sim \pi$. Since this $\{Y_n\}$ is an ergodic process, Birkhoff's ergodic theorem finishes the proof.

Some other important examples can be constructed based on the process $\{X_n\}$ as a sequence of Bernoulli random variables (not necessarily independent of course). Consider the distribution of $\{X_n\}$ given by

$$\mathbb{P}((X_0, X_1, ...) = (1,0,1,0,...)) = \frac{1}{2}, \mathbb{P}((X_0, X_1, ...) = (0,1,0,1,...)) = \frac{1}{2} \tag{232}$$

then it's strongly stationary. It's also ergodic since $(1,0,1,0,...)$ and $(0,1,0,1,...)$ has exactly the same tail behavior switching between zero and one.

On the other hand, consider the distribution of $\{X_n\}$ given by

$$\mathbb{P}((X_0, X_1, ...) = (0,0,0,0,...)) = \frac{1}{2}, \mathbb{P}((X_0, X_1, ...) = (1,1,1,1,...)) = \frac{1}{2} \tag{233}$$

then it's strongly stationary. It's not ergodic since $A = \{\lim_{n\to\infty} X_n = 0\} \in \mathscr{I}$ but $\mathbb{P}(A) = \frac{1}{2}$.

Remark. *The comparison of those two examples explain the interpretation of ergodicity that ergodic process traverse through all possible values it can take in an irregular enough way, i.e. it's chaotic enough. In the example above,* *the process takes value in $\{0,1\}$ at each time step, the first example is chaotic enough in that it always alternates between zero and one while the second example is not since the sequence is either always zero or always one. As a result, Birhoff's ergodic theorem is saying that if a strongly stationary sequence is chaotic enough, then it owns similar property to i.i.d. sequence that SLLN-type theorem holds.*

Interestingly, if one consider the distribution of $\{X_n\}$ given by $X_0 = X_1 = ...$ where $X_0 \sim B(1,p)$ and $p \sim U(0,1)$ independent of the whole sequence, it's strongly stationary and $A = \{\lim_{n\to\infty} X_n = 0\} \in \mathscr{I}$ now has $\mathbb{P}(A|p) = 1-p$ so $\mathbb{P}(A) = \mathbb{E}(1-p) = \frac{1}{2}$ so the sequence is still not ergodic, meaning that introducing a uniformly random p does not introduce enough chaos into the system. One can also check this statement through Birkhoff's ergodic theorem.

The ergodicity also has some connection with number theory. Consider $\Omega = [0,1)$ with $\mathbb{P} = \lambda$ as the Lebesgue measure, fix $\theta \in (0,1)$ and consider $X_n(\omega) = (\omega + n\theta)\ mod\ 1$ as a strongly stationary process. This process is ergodic iff $\theta \in \mathbb{R} - \mathbb{Q}$, i.e. it's an irrational number (try to prove this through an analysis argument). This example is called the rotation on the circle.

Application of Ergodic Theorem

Let's apply ergodic theorem for the rotation on the circle, for $\forall \theta \in \mathbb{R} - \mathbb{Q}$, $X_n(\omega) = (\omega + n\theta)\ mod\ 1$ is ergodic. Consider fixing any Borel set $A \subset \Omega$, then $\{\mathbb{I}_A(X_n(\omega))\}$ is also ergodic so

$$\forall \omega \in [0,1), \frac{\sum_{m=0}^{n-1} \mathbb{I}_A((\omega + m\theta)\ mod\ 1)}{n} \overset{a.s.,L^1}{\rightarrow} \mathbb{E}\mathbb{I}_A = \lambda(A)\ (n \to \infty) \tag{234}$$

set $\omega = 0$ to get the Weyl's equidistribution theorem

$$\forall \theta \in \mathbb{R} - \mathbb{Q}, \forall A \in \mathscr{B}_{[0,1)}, \frac{\sum_{m=0}^{n-1} \mathbb{I}_A((m\theta)\ mod\ 1)}{n} \overset{a.s.,L^1}{\rightarrow} \lambda(A)\ (n \to \infty) \tag{235}$$

with the limit as the Lebesgue measure of A. This theorem has an interesting corollary, known as the Benford's law. Consider the set of numbers $\{2^0, 2^1, ..., 2^{n-1}\}$, we want to know the asymptotic frequency of the first digit of the numbers appearing in this set. One might guess that digit $1, ..., 9$ appear with equal probability but that's not the case! The first digit of a number 2^m is $k \in \{1,2,...,9\}$ iff there exists some integer l such that $k10^l \leq 2^m < (k+1)10^l$ iff $\log_{10} k + l \leq m \log_{10} 2 < \log_{10}(k+1) + l$, i.e. $m \log_{10} 2\ mod\ 1 \in [\log_{10} k, \log_{10}(k+1))$. As a result, we set

$$\theta = \log_{10} 2 \in \mathbb{R} - \mathbb{Q}, A_k = [\log_{10} k, \log_{10}(k+1)) \tag{236}$$

and apply the Weyl's equidistribution theorem to conclude that

$$\forall k \in \{1,2,...,9\}, \frac{\sum_{m=0}^{n-1} \mathbb{I}_{A_k}((m\theta)\ mod\ 1)}{n} \overset{a.s.,L^1}{\rightarrow} \log_{10} \frac{k+1}{k}\ (n \to \infty) \tag{237}$$

notice that $\sum_{m=0}^{n-1} \mathbb{I}_{A_k}((m\theta)\ mod\ 1)$ is the count of k appearing as the first digit in the set $\{2^0, 2^1, ..., 2^{n-1}\}$ so LHS is just the frequency we want to calculate! Finally, we get the famous Benford's law that for the collection of all powers of 2, $\left\{2^0, 2^1, ...\right\}$, the asymptotic frequency of the first digit of the numbers in this set is

$$\forall k \in \{1, 2, ..., 9\}, p_k = \log_{10} \frac{k+1}{k} \tag{238}$$

amazingly, this number theory problem is solvable from a probabilistic point of view and 1 is appearing the most often, which is counterintuitive at the first glance. This example should be enough to show the power of the ergodic theorem.

Gaussian Process

Lemma 33. *$\{X_n\}$ is GP, prove it's strongly stationary iff it's weakly stationary.*

Proof. Obviously, any strongly stationary process is weakly stationary.

If GP is weakly stationary, we have $\forall k, \mathbb{E}X_k = \mu, C(s,t) = cov(X_{s+k}, X_{t+k})$. By definition, we need to prove that $\forall t_1, ..., t_n, \forall k, (X_{t_1}, ..., X_{t_n}) \overset{d}{=} (X_{t_1+k}, ..., X_{t_n+k})$. Since it's GP, $(X_{t_1}, ..., X_{t_n}), (X_{t_1+k}, ..., X_{t_n+k})$ are both jointly Gaussian, we only need to prove they have the same mean vector and covariance matrix. Obviously, both have mean vector $\mu\vec{1}$ from the mean function. On the other hand, both have covariance matrix

$$\Sigma = \begin{bmatrix} C(t_1,t_1) & C(t_1,t_2) & ... & C(t_1,t_n) \\ C(t_2,t_1) & C(t_2,t_2) & ... & C(t_2,t_n) \\ ... & & & \\ C(t_n,t_1) & C(t_n,t_2) & ... & C(t_n,t_n) \end{bmatrix} \tag{239}$$

from the covariance kernel, this concludes the proof. Basically, this result is just from the fact that mean and variance are enough to identify a Gaussian distribution. □

Lemma 34. *$\{X_n\}$ is stationary GP, with zero mean and unit variance, and it's also Markov. Find the spectral density in terms of $\rho = cov(X_1, X_2)$.*

Proof. The covariance kernel $C(s,t) = c(t-s)$ with $c(0) = 1$ and $c(1) = \rho$. Notice that

$$c(t+s) = \mathbb{E}(X_0 X_{t+s}) = \mathbb{E}[X_0 \mathbb{E}(X_{t+s}|\mathscr{F}_s)] = \mathbb{E}[X_0 \mathbb{E}(X_{t+s}|X_s)] \tag{240}$$

$$= \mathbb{E}[X_0 c(t) X_s] = c(t)c(s) \tag{241}$$

it's easily seen that $c(t) = \rho^{|t|}$. Here due to Markov property $\mathbb{E}(X_{t+s}|X_s = x) = \mathbb{E}(X_t|X_0 = x)$. Since (X_0, X_t) is jointly Gaussian, $\mathbb{E}(X_t|X_0 = x) = c(t)x$ comes from the standard calculation for Gaussian distribution.

As a result, normalize c to get $\rho(t) = \frac{c(t)}{c(0)} = \rho^{|t|}$ as the c.f. of the spectral distribution. Now $\rho^t = \int e^{itx} f(x)\, dx$ so Levy's inversion formula (discrete inverse Fourier transform) tells us

$$f(x) = \frac{1}{2\pi} \sum_{t\in\mathbb{Z}} \rho^{|t|} e^{-itx} = \frac{1-\rho^2}{2\pi|1-\rho e^{ix}|^2}, x \in [-\pi, \pi] \tag{242}$$

which is exactly the Poisson kernel for the unit disk.

$\square$