Supplement to Graduate Program Applications

Kole Butterer

March 26, 2025

What follows is meant to demonstrate my mathematical abilities and interests. The material stems from my advanced studies with professors and graduate students, and a graduate measure theoretic probability class at Northwestern University. Jump to whatever section interests you most!

Contents

1 Von Neumann ergodic theorem

The standard Birkhoff Ergodic Theorem says that the time average of a function f under some transformation T is equal to the expectation of the function when the transformation T is measure preserving. Von Neumann's ergodic theorem says that the time average is equal to a *part* of f that is unaffected by—or, invariant under—the transformation T. We prove Von Neumann's version below.

Von Neumann Ergodic Theorem: *Let $T : X \to X$ be a measure preserving transformation with respect to measure μ. And let $P : L^2(X,\mu) \to L^2(X,\mu)$ be the orthogonal projection onto the subspace $L^2(X,\mu,I)$ of T-invariant functions. Then for any $f \in L^2(X,\mu)$, we have in L^2* that

$$\frac{1}{N}\sum_{n=0}^{N-1} f \circ T^n \overset{N\longrightarrow\infty}{\longrightarrow} Pf.$$

Proof. To start, let f be T-invariant i.e. $f \in L^2(X,\mu,I)$. Then

$$\frac{1}{N}\sum_{n=0}^{N-1} f \circ T^n = \frac{1}{N}\sum_{n=0}^{N-1} f \longrightarrow f = Pf$$

in L^2 trivially. Now suppose f is not necessarily T-invariant, but that $f = g - (g \circ T)$ for some $g \in L^2(X,\mu)$. Then we have

$$\begin{aligned}
\frac{1}{N}\sum_{n=0}^{N-1} f \circ T^n &= \frac{1}{N}\sum_{n=0}^{N-1} (g - g\circ T)\circ T^n \\
&= \frac{1}{N}\sum_{n=0}^{N-1} g\circ T^n - \frac{1}{N}\sum_{n=0}^{N-1} g \circ T^{n+1} \\
&= \frac{1}{N} g \circ T^{n+1}. && \text{(notice the telescoping terms)}
\end{aligned}$$

We know that this function tends to 0 in L^2 because

$$\begin{aligned}
\left\| \frac{1}{N}(g - g\circ T^{n+1}) \right\|_2 2 &\le \left\| \frac{1}{N} g \right\|_2 2 + \left\| \frac{1}{N} g \circ T^{n+1} \right\|_2 2 && \text{(Triangle Inequality)} \\
&= \frac{1}{N}(\|g\|_2 2 + \|g\|_2 2) && (T \text{ preserves norms}) \\
&= \frac{2}{N}\|g\|_2 2 \overset{N\longrightarrow\infty}{\longrightarrow} 0.
\end{aligned}$$

Now suppose f is the limit of a sequence of functions of the same form i.e. $f_k \longrightarrow f$ pointwise and $f_k = g_k - (g_k \circ T)$ for $g_k \in L^2(X,\mu)$. We claim that the sequence of averages of $f \circ T^n$ tends to 0 as more terms are included in each average. We choose k such that $\|f - f_k\|_2 2 \le \epsilon$. We have

$$\begin{aligned}
\left\| \frac{1}{N}\sum_{n=0}^{N-1} f\circ T^n - 0 \right\|_2 2 &= \left\| \frac{1}{N}\sum_{n=0}^{N-1} f_k \circ T^n - \frac{1}{N}\sum_{n=0}^{N-1}(f - f_k)\circ T^n \right\|_2 2 \\
&\le \left\| \frac{1}{N}\sum_{n=0}^{N-1} g - g_k \circ T^n \right\|_2 2 + \frac{1}{N}\sum_{n=0}^{N-1} \|f - f_k\|_2 2 && \text{(Triangle Inequality and } T \text{ preserves norms)} \\
&\le \left\| \frac{1}{N}\sum_{n=0}^{N-1} g - g_k \circ T^n \right\|_2 2 + \epsilon \overset{N\longrightarrow\infty}{\longrightarrow} \epsilon. && \text{(chose } k \text{ accordingly, and first term goes to 0)}
\end{aligned}$$

Hence, $\left\| \frac{1}{N}\sum_{n=0}^{N-1} f\circ T^n \right\|_2 2 \overset{N\longrightarrow\infty}{\longrightarrow} 0$ when f is the limit of some sequence of functions of that particular form.

We have just shown if $f \in \overline{\{g - (g\circ T) \mid g \in L^2(X,\mu)\}}$ then its sequence of averages tends to 0 as we include more terms. Now we claim that $L^2(X,\mu,I)$, the set of T-invariant functions, is the orthogonal

complement to $\{g - (g \circ T) \mid g \in L^2(X, \mu)\}$. If this is true (we assume it is, then show it is after), then $f = Pf + f_\perp$ for some $f_\perp \in \{g - (g \circ T) \mid g \in L^2(X, \mu)\}$. Thus, we have

$$\begin{aligned}\frac{1}{N}\sum_{n=0}^{N-1} f \circ T^n &= \frac{1}{N}\sum_{n=0}^{N-1} (Pf + f_\perp) \circ T^n \\ &= \frac{1}{N}\sum_{n=0}^{N-1} Pf \circ T^n + \frac{1}{N}\sum_{n=0}^{N-1} f_\perp \circ T^n \\ &\xrightarrow{N\longrightarrow\infty} Pf + 0 = Pf\end{aligned}$$

where the second term in the second line tends to 0 because what we have shown above.

Now we must prove that the set of T-invariant functions is orthogonal to the set of functions who take the form of $g - (g \circ T)$, i.e. $L^2(X, \mu, I) = \{g - (g \circ T) \mid g \in L^2(X, \mu)\}^\perp$. For the forwards inclusion, let $f \in L^2(X, \mu, I)$ and $g \in L^2(X, \mu)$. Then

$$\begin{aligned}\langle f, g - (g \circ T)\rangle &= \langle f, g\rangle - \langle f, g \circ T\rangle \\ &= \langle f, g\rangle - \langle f \circ T, g \circ T\rangle && \text{(since } f \text{ is } T\text{-invariant)} \\ &= \langle f, g\rangle - \langle f, g\rangle && \text{(T is measure preserving and so preserves inner products)} \\ &= 0.\end{aligned}$$

Now, for the opposite inclusion, suppose $\langle f, g - (g \circ T)\rangle = 0$ for $f, g \in L^2(X, \mu)$. Pick f such that $f = g$. Then we have

$$0 = \langle f, g - (g \circ T)\rangle = \langle f, f\rangle - \langle f, f \circ T\rangle$$

which implies $\langle f, f\rangle = \langle f, f \circ T\rangle$. Consider the norm of the difference between f and $f \circ T$. We have

$$\begin{aligned}\|f - (f \circ T)\|_2 2^2 &= \langle f - (f \circ T), f - (f \circ T)\rangle \\ &= \langle f, f\rangle - 2\langle f, f \circ T\rangle + \langle f \circ T, f \circ T\rangle \\ &= 2\langle f, f\rangle - 2\langle f, f \circ T\rangle && \text{(because } T \text{ preserves inner products)} \\ &= 2(\langle f, f\rangle - \langle f, f \circ T\rangle) \\ &= 0.\end{aligned}$$

Thus, $f = f \circ T$, so $f \in L^2(X, \mu, I)$, and we are done. □

2 Limiting proportion of time spent in state

We're often interested in the proportion of time a Markov chain would spend in a given state if the chain ran forever. We can use the following lemma to prove equivalencies of ergodicity in Markov chains and irreducibility later on.

Lemma: The limit

$$q_{ij} = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} p_{ij}^{(k)},$$

exists and is well defined for each $i, j \in \{0, 1, ..., N-1\}$ in the state space, where the quantity $p_{ij}^{(k)}$ is the k-step transition probability from state i to state j.

Proof. Let T be the left-shift transformation. It is measure preserving. Let x be a sequence of random variables generated by a Markov chain. By the ergodic theorem, we have that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{1}_{\{x:x_k=j\}}(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{1}_{\{x:x_0=j\}}(T^k x) = f^*(x),$$

such that f^* is integrable. Using the above equality, and the fact that $\frac{1}{n} \sum_{k=0}^{n-1} \mathbb{1}_{\{x:x_0=j\}}(T^k x) \leq 1$ for all n, we can use the dominated convergence theorem to rearrange the equality for q_{ij}. We have

$$\begin{aligned}
q_{ij} &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} p_{ij}^{(k)} \\
&= \frac{1}{\pi_i} \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mu(\{x \in X : x_0 = i, x_k = j\}) \\
&= \frac{1}{\pi_i} \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \int_X \mathbb{1}_{\{x:x_0=i,x_k=j\}} \, d\mu \\
&= \frac{1}{\pi_i} \int_X \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \mathbb{1}_{\{x:x_0=i,x_k=j\}} \, d\mu \qquad \text{(by DCT)} \\
&= \frac{1}{\pi_i} \int_X f^*(x) \mathbb{1}_{\{x:x_0=i\}} \, d\mu \\
&= \frac{1}{\pi_i} \int_{\{x:x_0=i\}} f^*(x) \, d\mu.
\end{aligned}$$

Since f^* is integrable, we know q_{ij} exists and is well-defined. □

3 Spectral bounds for mixing times

When working with finite-state, irreducible, aperiodic, reversible Markov chains, we can use the eigenvalues of the transition matrix to bound mixing times. These are some notes that I took and presented during an independent study with Ursula Porod.

3.1 Spectral decomposition

First we note that the stationary distribution π is strictly positive because the Markov chain is irreducible. Therefore, we define

$$\mathbf{D} := \text{diag}(\pi(1), ..., \pi(n))$$

and

$$\mathbf{P}^* := \mathbf{D}^{\frac{1}{2}}\mathbf{P}\mathbf{D}^{-\frac{1}{2}}.$$

Then

$$P^*_{ij} = \frac{\sqrt{\pi(i)}}{\sqrt{\pi(j)}}P_{ij}.$$

This follows directly from the computation of $\mathbf{P}^*$. Since $\mathbf{P}^*$ is assumed to be reversible, we know $P^*_{ij} = P^*_{ji}$. Hence $\mathbf{P}^*$ is symmetric, and its eigenvectors are orthonormal by the real spectral theorem. It shares the same eigenvalues with $\mathbf{P}$ since they are similar matrices. Moreover, since our Markov chain is assumed to be irreducible and aperiodic, we know $\mathbf{P}^* s_1 = s_1$ where s_1 is the stationary distribution corresponding to $\mathbf{P}^*$, and the rest of the eigenvalues $\lambda_2, ..., \lambda_k$ are all in the range $(-1, 1)$. Let $\mathbf{S}$ be the matrix of eigenvectors of $\mathbf{P}^*$ i.e. each column is an eigenvector. Defining $\mathbf{\Lambda}$ as the diagonal matrix of eigenvalues, we have

$$\mathbf{P}^* = \mathbf{S}\mathbf{\Lambda}\mathbf{S}^T$$

since the inverse of $\mathbf{S}$ is its transpose (it is orthogonal). It follows that

$$\mathbf{P}^*_{ij} = \sum_{k=1}^{n} S_{ik}\lambda_k S_{jk},$$

and then

$$\mathbf{P}_{ij} = \frac{\sqrt{\pi(j)}}{\sqrt{\pi(i)}}\sum_{k=1}^{n} S_{ik}\lambda_k S_{jk}.$$

Now, since

$$\mathbf{P} = \mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{\Lambda}\mathbf{S}^T\mathbf{D}^{\frac{1}{2}},$$

we have

$$\mathbf{P}^m = \mathbf{D}^{-\frac{1}{2}}\mathbf{S}\mathbf{\Lambda}^{\mathbf{m}}\mathbf{S}^T\mathbf{D}^{\frac{1}{2}},$$

and equivalently

$$P^m_{ij} = \frac{\sqrt{\pi(j)}}{\sqrt{\pi(i)}}\sum_{k=1}^{n} S_{ik}\lambda_k^m S_{jk}.$$

If we take the first term out of the sum since $\lambda_1 = 1$, we have

$$P^m_{ij} = \frac{\sqrt{\pi(j)}}{\sqrt{\pi(i)}}S_{i1}S_{j1} + \sum_{k=2}^{n} S_{ik}\lambda_k^m S_{jk}.$$

Note that the since $P^m_{ij} \xrightarrow{m\longrightarrow\infty} \pi(j)$ and $\lim_{m\to\infty}\lambda_k^m = 0$, it follows that

$$P^m_{ij} = \pi(j) + \sum_{k=2}^{n} S_{ik}\lambda_k^m S_{jk}.$$

This equation tells us that mixing times depend on the non-trivial eigenvalues of the transition matrix. Specifically, there exists constants for each $i, j \leq n$ such that

$$|P^m_{ij} - \pi(j)| \leq C_{ij}\lambda_*^m$$

where $\lambda_* := \max\{\lambda_2, ..., \lambda_n\}$. Taking $C = \max_{i,j\leq n} C_{ij}$, we can conclude that for any initial distribution μ_0, and $\mu_m := \mu_0\mathbf{P}^m$, we have

$$||\mu_m - \pi||_{TV} \leq C\lambda_*^m.$$

3.2 Proving real spectral theorem

We now prove the real spectral theorem. The following are useful lemmas in the proof, and are proved afterwards.

Lemma 1: *Let T be a self-adjoint operator, and let $b, c \in \mathbb{R}$ such that $b^2 < 4c$. Then the operator $T^2 + bT + cI$ is invertible.*

Lemma 2: *If T is a self-adjoint operator on a non-zero real vector space V, then there exist $\lambda_1, ..., \lambda_m \in \mathbb{R}$ such that its minimal polynomial factors completely into the form*

$$p(x) = (x - \lambda_1) \cdot ... \cdot (x - \lambda_m).$$

Lemma 3: *If the minimial polynomial of an operator T on a finite dimensional vector space V factors completely into the form*

$$p(x) = (x - \lambda_1) \cdot ... \cdot (x - \lambda_m)$$

for some $\lambda_1, ..., \lambda_m \in \mathbb{R}$ or $\mathbb{C}$, then there exits some basis of V such that the matrix of T is upper-triangular with respect to that basis.

Real spectral theorem: *Let T be an operator on a real, finite dimensional inner product space V. Then the following are equivalent.*

(i) *T is self-adjoint.*

(ii) *There exists an orthonormal basis of V such that the matrix of T is diagonal with respect to that basis.*

(iii) *There exists an orthonormal basis of V consisting of eigenvectors of T.*

Proof. If $\dim V = 0$, the theorem is trivial. Let $\dim V > 0$.

(i) $\implies$ (ii). Let T be self-adjoint. Let p be the minimal polynomial of T. From Lemma 1.2, there exist $\lambda_1, ..., \lambda_m \in \mathbb{R}$ such that for each $x \in \mathbb{R}$

$$p(x) = (x - \lambda_1) \cdot ... \cdot (x - \lambda_n).$$

Then there exists a basis of V for which the matrix of T is upper triangular by Lemma 1.3. And since T is self-adjoint, $T = T^*$, and the matrices of T and T^* are the same (the adjoint of a real valued matrix is its transpose). So the matrix of T denoted $M(T)$ must be diagonal.

(ii) $\implies$ (i). Assume there exists an orthonormal basis of V for which the matrix of T is diagonal with respect to that basis. Denote this matrix $M(T)$. Since $M(T)$ is diagonal, $M(T) = M(T)^*$. Thus $T = T^*$, and T is self-adjoint.

(ii) $\iff$ (iii). Let $e_1, ..., e_n$ be an orthonormal basis of V. Then for $k = 1, ..., n$, we have

$$T(e_k) = \lambda_k e_k$$

if and only if $\lambda_1, ..., \lambda_n$ are on the diagonal of the matrix of T. This is equivalent to λ_k being an eigenvalue for eigenvector e_k. Thus, there exists an orthonormal basis of V for which the matrix of T is diagonal if and only if there exists an orthonormal basis of V consisting of eigenvectors of T.
We have shown all three statements are equivalent. □

3.2.1 Proof of Lemma 1

Let T be a self-adjoint operator, and let $b, c \in \mathbb{R}$ such that $b^2 < 4c$. Then the operator $T^2 + bT + cI$ is invertible.

Proof. Assume $b, c \in \mathbb{R}$ and $b^2 < 4c$. Let $v \in V$ and $v \neq 0$. We show that the operator $T^2 + bT + cI$ is injective i.e. $\text{null}(T^2 + bT + cI) = \{0\}$ by showing the following inner product is never equal to 0. We have

$$\begin{aligned}
\langle (T^2 + bT + cI)(v), v\rangle &= \langle (T^2)(v), v\rangle + b\langle T(v), v\rangle + c\langle v, v\rangle \\
&= \langle T(v), T(v)\rangle + b\langle T(v), v\rangle + c\langle v, v\rangle && \text{(since T is self-adjoint)} \\
&\geq \|T(v)\|_2^2 - |b||\langle T(v), v\rangle| + c\|v\|_2^2 \\
&\geq \|T(v)\|_2^2 - |b|\,\|T(v)\|_2\,\|v\|_2 + c\|v\|_2^2 && \text{(by Cauchy-Schwarz)} \\
&= \|T(v)\|_2^2 - |b|\,\|T(v)\|_2\,\|v\|_2 - \frac{b^2}{4}\|v\|_2^2 + \frac{b^2}{4}\|v\|_2^2 + c\|v\|_2^2 && \text{(completing the square)} \\
&= (\|T(v)\|_2 - \frac{|b|}{2}\|v\|_2)^2 + \|v\|_2^2\,(c - \frac{b^2}{4}) \\
&> 0
\end{aligned}$$

where the last line follows because the first term is squared and therefore non-negative, and the second term is positive by assumption. Hence, the inner product between the operator $T^2 + bT + cI$ applied to v and v itself is always positive. Therefore, the operator $T^2 + bT + cI \neq 0$ and $\text{null}(T^2 + bT + cI) = \{0\}$, and so it is injective, and equivalently invertible. □

3.2.2 Proof of Lemma 2

If T is a self-adjoint operator on a non-zero real vector space V, then there exists $\lambda_1, ..., \lambda_m$ such that its minimal polynomial factors completely into the form $p(x) = (x - \lambda_1) \cdot ... \cdot (x - \lambda_m)$.

Proof. Let $\lambda_1, ..., \lambda_m$ be the eigenvalues of T. They are real-valued because T is self-adjoint, and are the zeros of the minimal polynomial p of T. Assume for sake of contradiction that

$$p(x) \neq (x - \lambda_1) \cdot ... \cdot (x - \lambda_m).$$

If p is constant, then V is the zero vector space, but we assumed V is non-zero, so p can't be constant. Then, (by another lemma not included) it must be the case that

$$p(x) = q(x)(x^2 + bx + c)$$

where $b, c \in \mathbb{R}$ and $b^2 < 4c$ and $\deg q < \deg p$. Since p is the minimal polynomial of T, we have

$$p(T) = q(T)(T^2 + bT + cI) = 0.$$

And the operator $T^2 + bT + cI$ is invertible (i.e. $\text{null}(T^2 + bT + cI) = \{0\}$) by Lemma 1.1, so $q(T) = 0$. But we assumed $\deg q < \deg p$, so we have a contradiction. Thus, $p(x) = (x - \lambda_1) \cdot ... \cdot (x - \lambda_m)$ for real-valued $\lambda_1, ..., \lambda_m$. □

3.2.3 Proof of Lemma 3

Lemma 1.3: *If the minimial polynomial of an operator T on a finite-dimensional vector space V factors completely into the form*

$$p(x) = (x - \lambda_1) \cdot ... \cdot (x - \lambda_m)$$

for some $\lambda_1, ..., \lambda_m \in \mathbb{R}$ or $\mathbb{C}$, then there exists some basis of V such that the matrix of T is upper-triangular with respect to that basis.

Proof. Let $p(x) = (x - \lambda_1) \cdot ... \cdot (x - \lambda_m)$ be the minimal polynomial of the operator T. We use induction on m. Let $m = 1$. Then $p(T) = (T - \lambda_1 I) = 0$ by assumption, which implies $T = \lambda_1 I$ which is of course upper-triangular with respect to any basis of V. Now let $m > 1$, and let the desired result be true for all positive integers less than m. Define $\mathcal{U} := \text{range}(T - \lambda_m I)$. We know $\mathcal{U}$ is invariant under T (null spaces and ranges of polynomials applied to operators are invariant under that operator). So $T|_{\mathcal{U}}$ is an operator on $\mathcal{U}$. If $u \in \mathcal{U}$, then $u = (T - \lambda_m I)(v)$ for some $v \in V$. Now define q to be the minimal polynomial of $T|_{\mathcal{U}}$, and define $r(T) = (T - \lambda_1 I) \cdot ... \cdot (T - \lambda_{m-1} I)$. Then

$$r(T)(u) = (T - \lambda_1 I) \cdot ... \cdot (T - \lambda_m I)(v) = p(T)(v) = 0$$

by our assumption about the form of the minimal polynomial p of T. Hence, $r(T)$ is a polynomial multiple of $q(T)$. Therefore, there exist $\alpha_1, ..., \alpha_n \in \{\lambda_1, ..., \lambda_{m-1}\}$ for $n \leq m-1$ such that $q(T) = (T - \alpha_1 I) \cdot ... \cdot (T - \alpha_n I) = 0$. By our induction hypothesis, there exits a basis $u_1, ..., u_M$ of $\mathcal{U}$ such that $T|_{\mathcal{U}}$ is upper-triangular with respect to that basis. Extend this basis of $\mathcal{U}$ to a basis of V so that $u_1, ..., u_M, v_1, ..., v_N$ is a basis of V. Now we show an equivalent condition for upper-triangulability, which is that $T(a_k) \in \text{span}(a_1, ..., a_k)$ for each $k = 1, ..., \dim V$ where $a_1, ..., a_k$ is a basis for V. In our case, if we take any u_k for $k = 1, ..., M$ we have

$$T(u_k) = T|_{\mathcal{U}}(u_k) \in \text{span}(u_1, ..., u_k)$$

since $T|_{\mathcal{U}}$ is upper-triangular. Now take any v_k for $k = M+1, ..., N$. We make a basic manipulation to the expression $T(v_k)$ so that we have

$$T(v_k) = (T - \lambda_m I)(v_k) + \lambda_m v_k.$$

We see that $(T - \lambda_m I)(v_k) \in \mathcal{U}$ by definition, and therefore $(T - \lambda_m I)(v_k) \in \text{span}(u_1, ..., u_M)$. Consequently,

$$(T - \lambda_m I)(v_k) + \lambda_m v_k \in \text{span}(u_1, ..., u_M, v_k) \subset \text{span}(u_1, ..., u_M, v_1, ..., v_k).$$

This is equivalent to T being upper-triangular with respect to the extended basis. We are done. $\square$

4 Strong law of large numbers

We prove the strong law of large numbers in 4 steps. We state the theorem below.

Strong Law of Large Numbers. *Let $X_1, X_2, ...$ be iid random variables and assume $\mathbb{E}[|X_1|] < \infty$. Let $S_n := X_1 + ... + X_n$. Then*

$$\frac{S_n}{n} \xrightarrow{a.s.} \mathbb{E}[X_1].$$

4.1 Step 1 (Kolmogorov's inequality).

Theorem. Let $X_1, ..., X_n$ be mutually independent random variables. Suppose $\mathbb{E}[X_i] = 0$ and $\sigma_i^2 := \mathbb{E}[X_i^2] < \infty$. Let $\lambda > 0$. Then

$$\mathbb{P}\left(\max_{1\leq i\leq n}|X_1 + ... + X_i| \geq \lambda\right) \leq \frac{\sum_{i=1}^n \sigma_i^2}{\lambda^2}.$$

Proof. Note that this is a stronger version of Chebyshev's inequality. It bounds the probability of the absolute value of the maximum partial sum exceeding λ. Let $S_k := \sum_{i=1}^k X_i$. And $A := \{\max_{1\leq k\leq n}|S_k| \geq \lambda\} = \bigcup_{k=1}^n \{|S_k| \geq \lambda\}$ i.e. the event that the maximum partial sum excedes λ. Now let $A_k := \{\max_{1\leq i\leq k-1}|S_i| < \lambda, |S_k| \geq \lambda\}$ i.e. the event that the largest partial sum over indices less than or equal to k exceeds λ no sooner than the kth partial. Then $A = \bigcup_{k=1}^n A_k$. Also note that $A_k \cap A_\ell = \emptyset$ for $k \neq \ell$. So

$$\mathbb{P}(A) = \sum_{k=1}^n \mathbb{P}(A_k) = \mathbb{E}[\mathbb{1}_{\{\omega\in A_k\}}] \leq \sum_{k=1}^n \mathbb{E}\left[\frac{S_k}{\lambda^2}\cdot \mathbb{1}_{\{\omega\in A_k\}}\right] = \frac{1}{\lambda^2}\sum_{k=1}^n \mathbb{E}[S_k^2 \cdot \mathbb{1}_{\{\omega\in A_k\}}].$$

Now just consider the summation. We have

$$\begin{aligned}
\sum_{k=1}^n \mathbb{E}[S_k^2 \cdot \mathbb{1}_{\{\omega\in A_k\}}] &\leq \sum_{k=1}^n \left[\mathbb{E}[S_k^2 \cdot \mathbb{1}_{\{\omega\in A_k\}}] + \mathbb{E}[(S_n - S_k)^2 \cdot \mathbb{1}_{\{\omega\in A_k\}}]\right] && \text{(second term is non-negative)}\\
&= \sum_{k=1}^n \mathbb{E}[S_n^2] \cdot \mathbb{1}_{\{\omega\in A_k\}} && (**)\\
&= \mathbb{E}[S_n^2 \sum_{i=1}^n \mathbb{1}_{\{\omega\in A_k\}}]\\
&= \mathbb{E}[S_n^2 \cdot \mathbb{1}_{\{\omega\in A\}}]\\
&\leq \mathbb{E}[S_n^2] && \text{(monotonicity of integral)}\\
&= \sum_{i=1}^n \sigma_i^2 && \text{(independence and mean zero).}
\end{aligned}$$

Now we explain (**) since it's not immediately obvious. We focus on the two terms in the sum. We have

$$\begin{aligned}
\mathbb{E}[S_k^2 \cdot \mathbb{1}_{\{\omega\in A_k\}}] + \mathbb{E}[(S_n - S_k)^2 \cdot \mathbb{1}_{\{\omega\in A_k\}}] &= \mathbb{E}\left[[S_k^2 + (S_n - S_k)^2] \cdot \mathbb{1}_{\{\omega\in A_k\}}\right] && \text{(linearity of expectation)}\\
&= \mathbb{E}[S_n^2 + 2S_k(S_k - S_n)] && \text{(expanding everything)}\\
&= \mathbb{E}[S_n^2] + \mathbb{E}[2S_k(S_k - S_n)] && \text{(linearity of expectation)}\\
&= \mathbb{E}[S_n^2] + \mathbb{E}[2S_k]\mathbb{E}[S_k - S_n] && (S_k \text{ and } (S_k - S_n) \text{ independent})\\
&= \mathbb{E}[S_n^2] && \text{(each partial has expectation 0).}
\end{aligned}$$

Putting everything together, we have

$$\mathbb{P}\left(\max_{1\leq i\leq n}|X_1 + ... + X_i| \geq \lambda\right) \leq \frac{\sum_{i=1}^n \sigma_i^2}{\lambda^2}.$$

□

4.2 Step 2 (Infinite sum of finite, zero-mean r.vs. exists and is finite).

Theorem. Let $X_1, X_2, ...$ be mutually independent random variables, and assume $\mathbb{E}[X_i] = 0$, and $\sigma_i^2 := \mathbb{E}[X_i^2] < \infty$, and $\sum_{i=1}^n \sigma_i^2 < \infty$. Then $\lim_{n\to\infty} \sum_{i=1}^n X_i$ exists and is finite almost surely.

Proof. Let $S_n := X_1 + ... + X_n$. We will prove that the sequence $\{S_n\}_{n\geq 1}$ is a Cauchy sequence with probability 1. Define the event $A_{N,r} := \left\{\exists i, j \geq N : |S_i - S_j| \geq \frac{1}{r}\right\}$ Then event that $\{S_n\}_{n\geq 1}$ is not Cauchy is

$$\bigcup_{r=1}^{\infty} \bigcap_{N=1}^{\infty} A_{N,r}.$$

Note that $A_{N,r}$ is increasing in r and decreasing in N, so we can use sequential continuity of measure. We have

$$\begin{aligned}\mathbb{P}\left(\{S_n\}_{n\geq 1} \text{ is not Cauchy}\right) &= \mathbb{P}\left(\bigcup_{r=1}^{\infty} \bigcap_{N=1}^{\infty} A_{N,r}\right) \\ &= \lim_{r\to\infty} \lim_{N\to\infty} A_{N,r}.\end{aligned}$$

Now we show for every $r \geq 1$ we have $\lim_{N\to\infty} A_{N,r} = 0$. Fix $r \geq 1$, and define $B_{N,r} := \left\{\exists i \geq N : |S_i - S_N| \geq \frac{1}{2r}\right\}$. Then

$$\begin{aligned}\mathbb{P}\left(\left\{\exists i \geq N : |S_i - S_N| \geq \frac{1}{2r}\right\}\right) &= \mathbb{P}\left(\bigcup_{N=N'}^{\infty} \left\{\exists N \leq i \leq N' : |S_i - S_N| \geq \frac{1}{2r}\right\}\right) \\ &= \lim_{N'\to\infty} \mathbb{P}\left(\max_{N\leq i\leq N'} |X_{N+1} + X_{N+2} + ... + X_{N'}| \geq \frac{1}{2r}\right) \\ &\leq \lim_{N'\to\infty} 4r^2 \sum_{i=N}^{N'} \sigma_i^2 \\ &= 4r^2 \sum_{i=N}^{\infty} \sigma_i^2.\end{aligned}$$

The second line follows from Kolmogorov's inequality shown in Step 1. Hence

$$\lim_{N\to\infty} \mathbb{P}(B_{N,r}) \leq \lim_{N\to\infty} 4r^2 \sum_{i=N}^{\infty} \sigma_i^2 = 0$$

because the sum of the variances converges by assumption, and therefore its tail sums go to 0. To conclude, note that $A_{N,r} \subseteq B_{N,r}$ by the triangle inequality. By monotonicity, we have $\lim_{N\to\infty} \mathbb{P}(A_{N,r}) = 0$. Equivalently,

$$\mathbb{P}\left(\{S_n\}_{n\geq 1} \text{ is Cauchy}\right) = 1.$$

□

4.3 Step 3 (SLLN with special condition)

First we state a useful lemma that will help us prove the strong law of large numbers when we have summable variances. Then we will use this step to prove a version of the strong law of large numbers in general.

Lemma. Suppose $\{a_n\}_{n\geq 1}$ is a sequence such that $\sum_{k=1}^{\infty} \frac{a_k}{k} < \infty$. Then $\frac{1}{n} \sum_{i=1}^{\infty} a_k \xrightarrow{n\to\infty} 0$.

SLLN with summable variances. Let $X_1, X_2, ...$ be mutually independent random variables, and define $\mu_i := \mathbb{E}[X_i], \sigma_i^2 := \text{Var}(X_i) < \infty$, and $\sum_{i=1}^{\infty} \frac{\sigma_i^2}{i^2} < \infty$. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{\infty} (X_i - \mu_i) \xrightarrow{\text{a.s.}} 0.$$

Proof. Define $Y_i := \frac{X_i - \mu_i}{i}$. Then $\mathbb{E}[Y_i] = 0$, and $\text{Var}(Y_i) = \frac{\sigma_i^2}{i^2}$. By assumption, we have $\sum_{i=1}^{\infty} \text{Var}(Y_i) = \sum_{i=1}^{\infty} \frac{\sigma_i^2}{i^2} < \infty$. By Step 2, $\sum_{i=1}^{\infty} Y_i < \infty$ almost surely. Thus $\sum_{i=1}^{\infty} Y_i = \sum_{i=1}^{\infty} \frac{X_i - \mu_i}{i} < \infty$. Hence, by the lemma, it follows that

$$\frac{1}{n} \sum_{i=1}^{\infty} (X_i - \mu_i) \xrightarrow{\text{a.s.}} 0.$$

□

4.4 Step 4 (General case of SLLN)

Now we can prove SLLN in the more general case where we do not require the variances to be summable as they are in Step 3.

Strong Law of Large Numbers without variance condition. Assume $X_1, X_2, ...$ are iid random variables with $\mathbb{E}[|X_1|] < \infty$ and $\mathbb{E}[X_1] = 0$. Let $S_n := X_1 + ... + X_n$. Then

$$\frac{S_n}{n} \xrightarrow{\text{a.s.}} \mathbb{E}[X_1].$$

Proof. We use truncation to prove the statement. Define $Y_k := X_k \cdot \mathbb{1}_{\{|X_k| \leq k\}}$ and $Z_k := X_k \cdot \mathbb{1}_{\{|X_k| > k\}}$. Then $X_k = Y_k + Z_k$. So

$$\frac{1}{n} \sum_{i=1}^{n} X_i = \frac{1}{n} \sum_{i=1}^{n} Y_i + \frac{1}{n} \sum_{i=1}^{n} Z_i.$$

We want to show that both sums on the right tend to zero as n tends to ∞. We start with the term summing Z_i. Our goal is to show that only finitely many Z_k are non-zero. This way, the entire term tends to 0. We use Borel-Cantelli. Define $A_k := \{Z_k \neq 0\}$. Then

$$\begin{aligned}
\sum_{k=1}^{\infty} \mathbb{P}(A_k) &= \sum_{k=1}^{\infty} \mathbb{P}(X_k > k) + \mathbb{P}(X_k < -k) \\
&= \sum_{k=1}^{\infty} F(-k) + (1 - F(k)) \\
&\leq \sum_{k=1}^{\infty} \left[\int_{-k}^{-(k-1)} F(y)\, dy + \int_{k}^{k+1} (1 - F(y))\, dy \right] && \text{(monotonicity of integral)} \\
&= \int_{-\infty}^{0} F(y)\, dy + \int_{0}^{\infty} (1 - F(y))\, dy \\
&= -\int_{-\infty}^{0} x dF(x) + \int_{0}^{\infty} x dF(x) && \text{(integration by parts)} \\
&= \mathbb{E}[|X_1|] < \infty && \text{(by assumption).}
\end{aligned}$$

Hence, by Borel-Cantelli, finitely many Z_k are non-zero almost surely, and thus $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Z_i = 0$ almost surely. Now we can show that $\frac{1}{n} \sum_{i=1}^{n} Y_i \xrightarrow{\text{a.s.}} 0$. This is equivalent to showing

$$\left[\frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbb{E}[Y_i]) + \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[Y_i] \right] \xrightarrow{\text{a.s.}} 0.$$

We want to show that $\sum_{i=1}^{\infty} \frac{Y_i}{i^2} < \infty$ so that by Step 3 the first term in the sum above goes to 0. Define

$$a_n := \int_{n-1}^{n} x dF(x) - \int_{-n}^{-n+1} x dF(x).$$

As shown above, $\sum_{n=1}^{\infty} a_n = \mathbb{E}[|X_1|] < \infty$. We have

$$\int_{n-1}^{n} x^2 dF(x) + \int_{-n}^{-n+1} x^2 dF(x) \leq \int_{n-1}^{n} n x dF(x) - \int_{-n}^{-n+1} n x dF(x) = n a_n.$$

It follows that

$$\begin{aligned}
\sum_{k=1}^{\infty} \frac{\mathrm{Var}(Y_k)}{k^2} &\leq \sum_{k=1}^{\infty} \frac{\mathbb{E}[Y_k^2]}{k^2} \\
&= \sum_{k=1}^{\infty} \frac{1}{k^2} \int_{-k}^{k} x^2 dF(x) \\
&= \sum_{k=1}^{\infty} \frac{1}{k^2} \cdot \sum_{\ell=1}^{k} \left[\int_{\ell-1}^{\ell=1} x^2 dF(x) + \int_{-\ell}^{-\ell+1} x^2 dF(x) \right] \\
&\leq \sum_{k=1}^{\infty} \frac{1}{k^2} \cdot \sum_{\ell=1}^{k} \ell a_\ell && \text{(how we've defined } a_n\text{)} \\
&= \sum_{\ell=1}^{\infty} \ell a_\ell \sum_{k=\ell}^{\infty} \frac{1}{k^2} && \text{(interchanging sums)} \\
&\leq \sum_{\ell=1}^{\infty} \ell a_\ell \frac{c}{\ell} && \text{(can be shown each tail sum proportional to } \frac{1}{\ell}\text{)} \\
&= c \sum_{\ell=1}^{\infty} a_\ell \\
&= c\, \mathbb{E}[|X_1|] \\
&< \infty && \text{(by assumption).}
\end{aligned}$$

Above, c is some constant bounding the tail sums from above. Thus, $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbb{E}[Y_i]) \xrightarrow{\text{a.s.}} 0$ as desired. Now we need to take care of the $\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i]$ term. This average should converge to 0 almost surely because Y_i tends to X_1 as $i \to \infty$ and $\mathbb{E}[X_1] = 0$. Formally, we have

$$\begin{aligned}
\lim_{k\to\infty} \mathbb{E}[Y_k] &= \lim_{k\to\infty} \int_{-k}^{k} x \; dF(x) \\
&= \lim_{k\to\infty} \int_{-\infty}^{\infty} x \; \mathbb{1}_{\{|X_k| \leq k\}} \; dF(x) \\
&= \int_{-\infty}^{\infty} \lim_{k\to\infty} x \; \mathbb{1}_{\{|X_k| \leq k\}} \; dF(x) && \text{(Dominated Convergence Theorem)} \\
&= \int_{-\infty}^{\infty} x \; dF(x) \\
&= \mathbb{E}[X_1] = 0.
\end{aligned}$$

It can be shown that for a sequence $\{a_n\}_{n\geq 1}$, if $a_n \longrightarrow 0$, then $\lim_{n\to\infty} \frac{1}{n}\sum_{n=1}^{\infty} a_n = 0$. It follows that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i] \xrightarrow{n\to\infty} 0 \text{ almost surely.}$$

□

5 Central limit theorem

We prove the central limit theorem in two different ways. The first using Lindenberg's swapping trick, and the other using characteristic functions.

Central Limit Theorem. *Let $X_1, X_2, ...$ be iid random variables with $\mathbb{E}[X_1^2] < \infty$. Let $S_n := X_1 + ... + X_n$. Then*

$$\frac{S_n - \mathbb{E}[S_n]}{\sqrt{Var(S_n)}} \xrightarrow{d} \mathcal{N}(0,1).$$

5.1 Lindenberg's swapping trick

Proof. For this proof, we do not assume the random variables are iid, just mutually independent. Assume that $\mathbb{E}[X_i] = 0$ and $\mathrm{Var}(X_i) = 1$ (otherwise take $Y_i := \frac{X_i - \mathbb{E}[X_i]}{\sqrt{\mathrm{Var}(X_i)}}$ and proceed as follows). Let $Z \sim \mathcal{N}(0,1)$. Assuming a finite third absolute moment i.e. $\mathbb{E}[|X_i|^3] < \infty$, it is enough to show that for every continuous, bounded, three-times differentiable function g, we have

$$\mathbb{E}\left[g\left(\frac{S_n}{\sqrt{n}}\right)\right] \xrightarrow{n\to\infty} \mathbb{E}[g(Z)] \quad (**).$$

Observation 1. If $Z_1, ..., Z_n$ are iid standard normal, then $\frac{Z_1+...+Z_n}{\sqrt{n}} \sim \mathcal{N}(0,1)$. Now define $T_n := Z_1 + ... + Z_n$ and $S_n := X_1 + ... + X_n$.
Observation 2. We note that (**) is equivalent to showing

$$\mathbb{E}\left[g\left(\frac{S_n}{\sqrt{n}}\right)\right] - \mathbb{E}\left[g\left(\frac{T_n}{\sqrt{n}}\right)\right] \xrightarrow{n\to\infty} 0.$$

Now, the idea is to swap the X_i's to Z_i's one by one. And when n is large, the difference in expectations for the given sums tends to 0. We define auxillary random variables

$$\begin{aligned}
S_n^{(0)} &:= X_1 + X_2 + ... + X_n \\
S_n^{(1)} &:= Z_1 + X_2 + ... + X_n \\
S_n^{(2)} &:= Z_1 + Z_2 + X_3 + ... + X_n \\
&\vdots \\
S_n^{(j)} &:= Z_1 + ... + Z_j + X_{j+1} + ... + X_n \\
&\vdots \\
S_n^{(n)} &:= Z_1 + ... + Z_n.
\end{aligned}$$

Note that $S_n^{(0)} = S_n$ and $S_n^{(n)} = T_n$. Now we can rewrite the expression that we want to tend towards 0 as a telescoping sum. We have

$$\begin{aligned}
\mathbb{E}\left[g\left(\frac{S_n}{\sqrt{n}}\right)\right] - \mathbb{E}\left[g\left(\frac{T_n}{\sqrt{n}}\right)\right] &= \sum_{j=1}^{n}\left(\mathbb{E}\left[g\left(\frac{S_n^{(j-1)}}{\sqrt{n}}\right)\right] - \mathbb{E}\left[g\left(\frac{S_n^{(j)}}{\sqrt{n}}\right)\right]\right) \\
&= \sum_{j=1}^{n}\mathbb{E}\left[g\left(\frac{S_n^{(j-1)}}{\sqrt{n}}\right) - g\left(\frac{S_n^{(j)}}{\sqrt{n}}\right)\right].
\end{aligned}$$

If we let $R_j := Z_1 + ... + Z_{j-1} + X_{j+1} + ... + X_n$ i.e. leaving out the jth random variable, then $S_n^{(j-1)} = R_j + X_j$, and $S_n^{(j)} = R_j + Z_j$. Since g is bounded and three-times differentiable, we can use its Taylor expansion centered around the point r. By Taylor's Theorem, we have

$$g(r+x) = g(r) + xg'(r) + \frac{x^2}{2}g''(r) + \frac{x^3}{6}g'''(r')$$

for some $r' \in (r, r+x)$. If we take the expectation of the Taylor approximation and let r and x be independent, we get

$$\begin{aligned}\mathbb{E}[g(r+x)] &= \mathbb{E}[g(r) + xg'(r) + \frac{x^2}{2}g''(r) + \frac{x^3}{6}g'''(r')] \\ &= \mathbb{E}[g(r)] + \mathbb{E}[x]\mathbb{E}[g'(r)] + \frac{\mathbb{E}[x^2]}{2}\mathbb{E}[g''(r)] + \mathbb{E}\left[\frac{x^3}{6}g'''(r')\right].\end{aligned}$$

Note that we can't split up the expectation in the cubic term because x and r' depend on each other. Now, letting $r = \frac{R_j}{\sqrt{n}}$ and $x = \frac{X_j}{\sqrt{n}}$ which are independent, we have

$$\begin{aligned}\mathbb{E}\left[g\left(S_n^{(j-1)}\right)\right] &= \mathbb{E}\left[g\left(\frac{R_j}{\sqrt{n}}\right)\right] + \mathbb{E}[X_j]\mathbb{E}\left[g'\left(\frac{R_j}{\sqrt{n}}\right)\right] + \frac{1}{2}\mathbb{E}[X_j^2]\mathbb{E}\left[g''\left(\frac{R_j}{\sqrt{n}}\right)\right] + \frac{1}{6}\mathbb{E}\left[X_j^3 g'''\left(\frac{R_j'}{\sqrt{n}}\right)\right] \\ &= \mathbb{E}\left[g\left(\frac{R_j}{\sqrt{n}}\right)\right] + \frac{1}{2n}\mathbb{E}\left[g''\left(\frac{R_j}{\sqrt{n}}\right)\right] + \frac{1}{6n^{3/2}}\mathbb{E}\left[X_j^3 g'''\left(\frac{R_j'}{\sqrt{n}}\right)\right].\end{aligned}$$

Note the first order term disappears because X_j has mean zero. Applying the same technique to $S_n^{(j)}$, we get

$$\mathbb{E}\left[g\left(\frac{S_n^{(j)}}{\sqrt{n}}\right)\right] = \mathbb{E}\left[g\left(\frac{R_j}{\sqrt{n}}\right)\right] + \frac{1}{2n}\mathbb{E}\left[g''\left(\frac{R_j}{\sqrt{n}}\right)\right] + \frac{1}{6n^{3/2}}\mathbb{E}\left[Z_j^3 g'''\left(\frac{\tilde{R}_j'}{\sqrt{n}}\right)\right].$$

The zero, first, and second order terms are all the same, so they cancel when we consider the difference in the telescoping sum above i.e.

$$\mathbb{E}\left[g\left(\frac{S_n^{(j-1)}}{\sqrt{n}}\right)\right] - \mathbb{E}\left[g\left(\frac{S_n^{(j)}}{\sqrt{n}}\right)\right] = \frac{1}{6n^{3/2}}\left(\mathbb{E}\left[X_j^3 g'''\left(\frac{R_j'}{\sqrt{n}}\right)\right] - \mathbb{E}\left[Z_j^3 g'''\left(\frac{\tilde{R}_j'}{\sqrt{n}}\right)\right]\right).$$

Recall that $\mathbb{E}[|X_j^3|] < C_1$ and $|g'''| \leq C_2$ for some $C_1, C_2 \in \mathbb{R}^+$ by assumption, and it's also true that $\mathbb{E}[|Z_j^3|] \leq C_3$ for some $C_3 \in \mathbb{R}^+$. Hence,

$$\mathbb{E}\left[g\left(\frac{S_n^{(j-1)}}{\sqrt{n}}\right)\right] - \mathbb{E}\left[g\left(\frac{S_n^{(j)}}{\sqrt{n}}\right)\right] \leq \frac{1}{6n^{3/2}}[C_2(C_1 + C_3)].$$

Thus,

$$\left|\sum_{j=1}^{n}\left(\mathbb{E}\left[g\left(\frac{S_n^{(j-1)}}{\sqrt{n}}\right)\right] - \mathbb{E}\left[g\left(\frac{S_n^{(j)}}{\sqrt{n}}\right)\right]\right)\right| \leq \sum_{j=1}^{n}\frac{c'}{n^{3/2}} = \frac{c'}{\sqrt{n}}$$

for some $c' \in \mathbb{R}$, and

$$\frac{c'}{\sqrt{n}} \xrightarrow{n\longrightarrow\infty} 0.$$

We have proved the central limit theorem. □

5.2 Using characteristic functions

We show that the characteristic function for the scaled sum converges to the characteristic function for a standard normal random variable as $n \longrightarrow \infty$. It is known that when a characteristic function converges to another characteristic function, we have corresponding weak convergence.

Proof. Let $X_1, ..., X_n$ be iid random variables with $\mathbb{E}[X_1] = 0$ and $\mathrm{Var}(X_1) = 1$. Let $S_n := X_1 + ... + X_n$. We'll take the Taylor expansion of φ_{X_1} centered around $t = 0$ up to order m assuming it's m-times differentiable (X has finite absolute mth moment). We have

$$\varphi_{X_1}(t) = \sum_{k=0}^{m}\frac{\varphi_{X_1}^{(k)}(0)}{k!}t^k + \mathcal{O}(t^m) = \sum_{k=0}^{m}\frac{\mathbb{E}[X_1^k]}{k!}(it)^k + \mathcal{O}(t^m).$$

Now fix $t \in \mathbb{R}$. Then

$$\begin{aligned}
\varphi_{\frac{S_n}{\sqrt{n}}}(t) &= \varphi_{S_n}(\frac{t}{\sqrt{n}}) \\
&= \prod_{i=1}^{n} \varphi_{X_i}(\frac{t}{\sqrt{n}}) \\
&= (\varphi_{X_1}(\frac{t}{\sqrt{n}}))^n \\
&= \left(1 + \frac{1}{\sqrt{n}}\mathbb{E}[X_1]\ it - \frac{\mathbb{E}[X_1^2]}{2n}\ t^2 + \mathcal{O}(\frac{t^2}{n})\right)^n \\
&= \left(1 - \frac{t^2}{2n} + \mathcal{O}(\frac{t^2}{n})\right)^n \qquad (X_1 \text{ has mean zero and unit variance}).
\end{aligned}$$

As $n \longrightarrow \infty$, we have

$$\lim_{n\to\infty} \left(1 - \frac{t^2}{2n} + \mathcal{O}(\frac{t^2}{n})\right)^n = e^{-t^2/2}$$

which is the characteristic function for a standard normal random variable. □

6 Graduate measure theoretic probability homework

6.1 Coupon collector's problem

Let $U_1, U_2, ...$ be i.i.d. random variables, each distributed uniformly on $\{1, 2, ..., n\}$. And let $|\{U_1, ..., U_k\}|$ be the number of distinct elements among the first k variables, and let $T_n := \inf\{k : |\{U_1, ..., U_k\}| = n\}$. Show that $\frac{T_n}{n \log n} \xrightarrow{\mathbb{P}} 1$.

Proof. We defined T_n to be the number of draws until we've seen all coupons. We can make this problem easier if we define some auxillary random variables t_i for $i \geq 1$ where each t_i is the number of draws it takes to see the i'th unique coupon once we've seen the $(i-1)$'th unique coupon. Then we have

$$T_n = t_1 + ... + t_n.$$

Let $\epsilon > 0$. By Chebyshev, we have

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{T_n}{n \log n} - \mathbb{E}\left[\frac{T_n}{n \log n}\right]\right| > \epsilon\right) &\leq \frac{\operatorname{Var}\left(\frac{T_n}{n \log n}\right)}{\epsilon^2} \\
&= \frac{\operatorname{Var}(T_n)}{(n \log n\ \epsilon)^2} \\
&= \frac{\sum_{i=1}^n \operatorname{Var}(t_i)}{(n \log n\ \epsilon)^2} \qquad \text{(the } t_i \text{ are pairwise independent).}
\end{aligned}$$

Also note that each t_i is a geometric random variable with success probability $p_i := \frac{n-i+1}{n}$. So we have a closed form for its variance: $\operatorname{Var}(t_i) = \frac{1-p_i}{p_i^2}$. So for our term in the numerator in the last line of the inequality above, we have

$$\begin{aligned}
\sum_{i=1}^n \operatorname{Var}(t_i) &= \sum_{i=1}^n \frac{1-p_i}{p_i^2} \\
&= \sum_{i=1}^n \frac{n(i-1)}{(n-i+1)^2} \\
&= n\left(0 + \frac{1}{(n-1)^2} + \frac{2}{(n-2)^2} + ... + \frac{n-1}{1^2}\right) \\
&\leq n\left(\frac{1}{n^2} + \frac{1}{(n-1)^2} + \frac{2}{(n-2)^2} + ... + \frac{n-1}{1^2}\right) \quad \text{(adding } \frac{1}{n^2} \text{ to the sum)} \\
&\leq n^2\left(\frac{1}{n^2} + \frac{1}{(n-1)^2} + \frac{1}{(n-2)^2} + ... + 1\right) \qquad \text{(making numerators all equal to } n\text{, then pulling it out)} \\
&= n^2 \sum_{i=1}^n \frac{1}{i^2} \\
&< n^2 \frac{\pi^2}{6} \qquad \text{(Basel problem).}
\end{aligned}$$

Returning to our original inequality, letting $c > \frac{\pi^2}{6}$ be a constant, we have

$$\mathbb{P}\left(\left|\frac{T_n}{n \log n} - \mathbb{E}\left[\frac{T_n}{n \log n}\right]\right| > \epsilon\right) \leq \frac{n^2 c}{n^2 (\log n\ \epsilon)^2}$$

for every $\epsilon > 0$. Taking limits as $n \longrightarrow \infty$, we have

$$\lim_{n \to \infty} \frac{c}{(\log n\ \epsilon)^2} = 0$$

and thus

$$\frac{T_n}{n \log n} \xrightarrow{\mathbb{P}} \mathbb{E}\left[\frac{T_n}{n \log n}\right].$$

So all we need to show is

$$\lim_{n \to \infty} \mathbb{E}\left[\frac{T_n}{n \log n}\right] = 1.$$

To do so, note that

$$\mathbb{E}[T_n] = \mathbb{E}[t_1 + ... + t_n] = \sum_{i=1}^{n} \frac{1}{p_i} = n \sum_{i=1}^{n} \frac{1}{i}.$$

We use the handy inequality that

$$\log(n+1) \leq \sum_{i=1}^{n} \frac{1}{i} \leq \log n + 1,$$

which gives us

$$\frac{n \log(n+1)}{n \log n} \leq \mathbb{E}\left[\frac{T_n}{n \log n}\right] \leq \frac{n \log n + 1}{n \log n}$$

and clearly both outer expressions tend to 1 as $n \longrightarrow \infty$. □

6.2 Almost law of iterated algorithm

Let $X_1, X_2, ...$ be standard normal i.i.d random variables and let $S_n := X_1 + ... + X_n$.

(i) Show that

$$\frac{1}{2\pi}\left[\frac{1}{x} - \frac{1}{x^3}\right] e^{-x^2/2} \leq \mathbb{P}(X_1 \geq x) \leq \frac{1}{2\pi}\frac{1}{x} e^{-x^2/2}.$$

Proof. Since X_1 is standard normal, we know

$$\mathbb{P}(X_1 \geq x) = \int_x^\infty \frac{1}{2\pi} e^{-y^2/2}\, dy.$$

We can simplify the problem if we make a change of variables corresponding to a shift of x. Once we change variables, we shift the point x to the origin. We let $z = y - x$, and therefore the integral becomes

$$\begin{aligned} &= \int_0^\infty \frac{1}{2\pi} e^{-(z+x)^2/2}\, dz \\ &= \frac{1}{2\pi} e^{-x^2/2} \left(\int_0^\infty e^{-(z^2+2zx)/2}\, dz \right). \end{aligned}$$

Now we deal with the integral on the right. Note that $z^2 + 2zx \geq 2zx$, and therefore $e^{-(z^2+2zx)} \leq e^{-2zx}$, and therefore

$$\int_0^\infty e^{-(z^2+2zx)/2}\, dz \leq \int_0^\infty e^{-zx}\, dz = \frac{1}{x}.$$

So we've shown the right side of the inequality. Now we show the left side. First split up the exponentials, and then make use of the fact that $1 - x \leq e^{-x}$, so we have

$$\begin{aligned} \int_0^\infty e^{-(z^2+2zx)/2}\, dz &= \int_0^\infty e^{-z^2/2} e^{-zx}\, dz \\ &\leq \int_0^\infty \left(1 - \frac{z^2}{2}\right) e^{-zx}\, dz. \end{aligned}$$

Integrating by parts yields the desired inequality

$$\frac{1}{2\pi}\left[\frac{1}{x} - \frac{1}{x^3}\right] e^{-x^2/2} \leq \mathbb{P}(X_1 \geq x) \leq \frac{1}{2\pi}\frac{1}{x} e^{-x^2/2}.$$

□

(ii) Show that $\limsup_{n\to\infty} \frac{X_n}{\sqrt{2 \log n}} = 1$ almost surely.

Proof. We use Borel-Cantelli and the inequality we've shown above to squeeze the lim sup to 1. Specifically, for every $\epsilon > 0$, we claim

$$\mathbb{P}\left(1 - \epsilon \leq \limsup_{n\to\infty} \frac{X_n}{\sqrt{2 \log n}} \leq 1 + \epsilon\right) = 1.$$

Let $\epsilon > 0$. Define the event $A_n^\epsilon := \{\frac{X_n}{\sqrt{2\log n}} \geq 1+\epsilon\}$. We have

$$\begin{aligned}\mathbb{P}\left(\frac{X_n}{\sqrt{2\log n}} \geq 1+\epsilon\right) &= \mathbb{P}(X_n \geq \sqrt{2\log n}(1+\epsilon)) \\ &\leq \frac{1}{2\pi}\frac{1}{\sqrt{2\log n}(1+\epsilon)} e^{-(\sqrt{2\log n}(1+\epsilon))^2/2} \\ &= \frac{1}{2\pi}\frac{1}{\sqrt{2\log n}(1+\epsilon)}\frac{1}{n^{(1+\epsilon)^2}}.\end{aligned}$$

So

$$\sum_{n=1}^{\infty} \mathbb{P}(A_n^\epsilon) < \infty$$

by the p-series test, and by Borel-Cantelli I we have,

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{X_n}{\sqrt{2\log n}} \geq 1+\epsilon\right) = 0.$$

We follow the same approach to show the other bound. Let $B_n^\epsilon := \left\{\frac{X_n}{\sqrt{2\log n}} \geq 1-\epsilon\right\}$. So

$$\begin{aligned}\mathbb{P}\left(\limsup_{n\to\infty} \frac{X_n}{\sqrt{2\log n}} \geq 1-\epsilon\right) &\geq \frac{1}{2\pi}\left[\frac{1}{\sqrt{2\log n}(1-\epsilon)} - \frac{1}{(\sqrt{2\log n}(1-\epsilon))^3}\right] e^{(\sqrt{2\log n}(1-\epsilon))^2/2} \\ &= \frac{1}{2\pi}\left[\frac{1}{\sqrt{2\log n}(1-\epsilon)} - \frac{1}{(\sqrt{2\log n}(1-\epsilon))^3}\right]\frac{1}{n^{(1-\epsilon)^2}}\end{aligned}$$

which is not summable i.e. $\sum_{n=1}^{\infty}\mathbb{P}(B_n^\epsilon) = \infty$. Thus, by Borel-Cantelli II,

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{X_n}{\sqrt{2\log n}} \geq 1-\epsilon\right) = 1.$$

Putting everything together, we've shown

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{X_n}{\sqrt{2\log n}} = 1\right) = 1.$$

□

(iii) Let C be a constant such that $C > \sqrt{2}$. Show that $\limsup_{n\to\infty}\left(\frac{S_n}{\sqrt{2\log n}}\right) < C$ almost surely.

Proof. We use the same approach as before. Let $\epsilon > 0$, and let $C = \sqrt{2}+\epsilon$. Note that $\frac{S_n}{\sqrt{n}}$ is standard normal. Let $C_n^\epsilon := \left\{\frac{S_n}{\sqrt{n}} \geq (\sqrt{2}+\epsilon)\sqrt{\log n}\right\}$. Then for every n, we have

$$\begin{aligned}\mathbb{P}\left(\frac{S_n}{\sqrt{n}} > (\sqrt{2}+\epsilon)\sqrt{\log n}\right) &\leq \frac{1}{2\pi}\frac{1}{(\sqrt{2}+\epsilon)\sqrt{\log n}} e^{-((\sqrt{2}+\epsilon)\sqrt{\log n})^2/2} \\ &= \frac{1}{2\pi}\frac{1}{(\sqrt{2}+\epsilon)\sqrt{\log n}}\frac{1}{n^{(\sqrt{2}+\epsilon)^2/2}} \\ &\leq \frac{1}{n^\alpha}\end{aligned}$$

for some $\alpha > 1$. Thus, $\sum_{n=1}^{\infty}\mathbb{P}(C_n^\epsilon) < \infty$ and

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{S_n}{\sqrt{n}} \geq (\sqrt{2}+\epsilon)\sqrt{\log n}\right) = 0.$$

Equivalently,

$$\mathbb{P}\left(\limsup_{n\to\infty} \frac{S_n}{\sqrt{n\log n}} < C\right) = 1.$$

□

6.3 Poisson approximation to the binomial distribution

Let $\{p_n\}_{n\geq 1}$ be a positive sequence such that $\lim_{n\to\infty} p_n = 0$ and $\lim_{n\to\infty} np_n = \lambda$ where $\lambda \in (0,\infty)$. Show that $\text{Bin}(n,p_n)$ converges in distribution to $Poi(\lambda)$ as $n \longrightarrow \infty$.

Proof. Let $\lambda_n := np_n$ i.e. the expected value of $X_n \sim \text{Bin}(n,p_n)$. The distribution for X_n is as follows:

$$\begin{aligned}
\mathbb{P}(X_n = k) &= \binom{n}{k} p_n^k (1-p_n)^{n-k} \\
&= \frac{n!}{(n-k)!k!}\left(\frac{\lambda_n}{n}\right)^k \left(1-\frac{\lambda_n}{n}\right)^{n-k} \\
&= \frac{(n)(n-1)(n-2)\cdots(n-k+1)}{k!}\left(\frac{\lambda_n}{n}\right)^k \left(1-\frac{\lambda_n}{n}\right)^{n-k} \\
&= \left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)\cdots\left(\frac{n-k+1}{n}\right)\left(\frac{\lambda_n^k}{k!}\right)\left(1-\frac{\lambda_n}{n}\right)^{n-k}
\end{aligned}$$

where all we did going from the third to fourth line was swap the position of $k!$ and n^k, and expanding n^k into k number of terms. Taking the limits as $n \longrightarrow \infty$, we see that

$$\begin{aligned}
\lim_{n\to\infty} \frac{n}{n}\frac{n-1}{n}\frac{n-2}{n}\cdots\frac{n-k+1}{n}\frac{\lambda_n^k}{k!}\left(1-\frac{\lambda_n}{n}\right)^{n-k} &= 1\cdot\left[\lim_{n\to\infty}\frac{\lambda_n^k}{k!}\right]\cdot\left[\lim_{n\to\infty}\left(1-\frac{\lambda_n}{k!}\right)^{n-k}\right] \\
&= \left[\lim_{n\to\infty}\frac{\lambda_n^k}{k!}\right]\cdot\left[\lim_{n\to\infty}\left(1-\frac{\lambda_n}{k!}\right)^{n-k}\right] \\
&= \frac{\lambda^k}{k!}\cdot\left[\lim_{n\to\infty}\left(1+\frac{\lambda_n}{n}\right)^{n}\right]\cdot\left[\lim_{n\to\infty}\left(1+\frac{\lambda_n}{n}\right)^{-k}\right] \\
&= \frac{\lambda^k}{k!}\cdot\lim_{n\to\infty} e^{-\lambda_n} \\
&= \frac{\lambda^k}{k!}e^{-\lambda}
\end{aligned}$$

which is the Poisson probability mass at k. All we did in the first line was realize that $\lim_{n\to\infty}\frac{n}{n}\frac{n-1}{n}\frac{n-2}{n}\cdots\frac{n-k+1}{n} = 1$. □

6.4 Exponential approximation to geometric distribution

Let X_p be a random variable with geometric distribution with parameter $p \in (0,1)$. Show that pX_p converges in distribution to Z where $Z \sim \text{Exp}(1)$.
First we prove a useful lemma.
Lemma. If $c_n \longrightarrow 0$ and $a_n \longrightarrow \infty$ but $a_n c_n \longrightarrow \lambda$, then $(1+c_n)^{a_n} \longrightarrow e^{\lambda}$.

Proof. We prove that $\lim_{n\to\infty} a_n \log(1+c_n) = \lambda$, so that in the end we have

$$\lim_{n\to\infty}(1+c_n)^{a_n} = \lim_{n\to\infty} e^{\log[(1+c_n)^{a_n}]} = \lim_{n\to\infty} e^{a_n\log(1+c_n)} = e^{\lambda}.$$

First we derive the Taylor expansion of $\log(1+x)$. We use the integral definition of log (and make a simple substitution), to get

$$\log(1+x) = \int_0^x \frac{1}{1+t}\,dt.$$

Now we realize we can write the integrand as a geometric sum, yielding

$$\frac{1}{1+t} = \frac{1}{1-(-t)} = \sum_{n=0}^{\infty}(-t)^n = 1 - t + t^2 - t^3 + ...$$

Thus we can integrate the infinite series on the right term by term which yields

$$\begin{aligned}
\log(1+x) = \int_0^x 1\,dt - \int_0^x t\,dt + \int_0^x t^2\,dt - ... &= x - \frac{x^2}{2} + \frac{x^3}{3} - ... \\
&= \sum_{n=1}^{\infty}\frac{(-1)^{n+1}x^n}{n}.
\end{aligned}$$

Now we can determine that $a_n \log(1+c_n) = \lambda$ using the Taylor expansion above. We have

$$\begin{aligned}
\lim_{n\to\infty} a_n \log(1+c_n) &= \lim_{n\to\infty} \left[a_n \left(c_n - \frac{c_n^2}{2} + \mathcal{O}(c_n^3) \right) \right] \\
&= \lim_{n\to\infty} \left[a_n c_n - \frac{a_n c_n^2}{2} + a_n \mathcal{O}(c_n^3) \right] \\
&= \lambda - \lim_{n\to\infty} \left[a_n c_n \frac{c_n}{2} + a_n c_n \mathcal{O}(c_n^2) \right] \\
&= \lambda \qquad\qquad \text{(since } c_n \longrightarrow 0\text{)}.
\end{aligned}$$

We've proven the lemma. Now, showing weak convergence is straight forward. We have

$$\mathbb{P}(pX_p > x) = \mathbb{P}(X_p > \frac{x}{p}) = (1-p)^{\lfloor \frac{x}{p} \rfloor}.$$

Note that for small p, the quantity $(1-p)^{\lfloor \frac{x}{p} \rfloor} \approx (1-p)^{\frac{x}{p}}$. Hence we have

$$\lim_{p\to 0} (1-p)^{\lfloor \frac{x}{p} \rfloor} = \lim_{p\to 0} \left((1-p)^{\frac{1}{p}} \right)^x = e^{-x}.$$

Thus

$$\lim_{p\to 0} \mathbb{P}(pX_p > x) = e^{-x} = \mathbb{P}(Z > x).$$

□

6.5 Weak LLN for weakly correlated random variables

Let $r : \mathbb{N} \to \mathbb{R}$ be a bounded function such that $r(k) \to 0$ as $k \to \infty$. Let $X_1, X_2, ...$ be identical but not necessarily independent random variables with mean zero and finite variance. Suppose that the covariances of the random variables satisfy $\text{Cov}(X_i, X_j) \leq r(|i-j|)$ for every $i, j \geq 1$. Let $S_n := X_1 + ... + X_n$. Show that $\frac{S_n}{n} \xrightarrow{\mathbb{P}} 0$.

Proof. Fix $\epsilon > 0$. By Chebyshev, we have that

$$\begin{aligned}
\mathbb{P}\left(\left| \frac{S_n}{n} - 0 \right| > \epsilon \right) &\leq \frac{\text{Var}\left(\frac{S_n}{n}\right)}{\epsilon^2} \\
&= \frac{\sum_{i=1}^n \text{Var}(X_i) + 2\sum_{i=1}^n \sum_{j=i+1}^n \text{Cov}(X_i, X_j)}{n^2 \epsilon^2}.
\end{aligned}$$

Since $r(|i-j|) \longrightarrow 0$ as $|i-j| \longrightarrow \infty$ there exists some $K \in \mathbb{N}$ such that $\text{Cov}(X_i, X_j) \leq r(|i-j|) \leq \delta$ for all i, j such that $|i-j| \geq K$. So we split the sum of the covariances into two sums. The first sum is over the indices where the distance between i and j is less than N, and the second sum is over i and j where the distance between them is greater than or equal to K. For the term in the numerator, we have that

$$\begin{aligned}
\sum_{i=1}^n \text{Var}(X_i) + 2\sum_{i=1}^n \sum_{j=i+1}^n \text{Cov}(X_i, X_j) &= \sum_{i=1}^n \text{Var}(X_i) + 2\sum_{i=1}^n \sum_{j=i+1}^{i+K} \text{Cov}(X_i, X_j) + 2\sum_{i=1}^n \sum_{j=i+K+1}^n \text{Cov}(X_i, X_j) \\
&\leq \sum_{i=1}^n r(0) + 2\sum_{i=1}^n \sum_{j=i+1}^{i+K} r(|i-j|) + 2\sum_{i=1}^n \sum_{j=i+K+1}^n r(|i-j|).
\end{aligned}$$

Now letting $M := \max\{r(k) : k \leq K\}$, we have

$$\begin{aligned}
&\leq nr(0) + 2\sum_{k=1}^{K-1} (n-k) r(k) + 2\sum_{k=K}^{n-K} (n-k) r(k) \\
&\leq nM + 2KMn + 2\delta n^2.
\end{aligned}$$

Plugging this back into the original expression, we have

$$\begin{aligned}\mathbb{P}\left(\left|\frac{S_n}{n}-0\right|>\epsilon\right) &\leq \frac{\mathrm{Var}\left(\frac{S_n}{n}\right)}{\epsilon^2} \\ &\leq \frac{nM+2KMn+2\delta n^2}{n^2\epsilon^2} \\ &= \frac{M}{n\epsilon^2}+\frac{2KM}{n\epsilon^2}+\frac{2\delta}{\epsilon^2}.\end{aligned}$$

Taking the limit as $n \to \infty$ we have

$$\lim_{n\to\infty} \frac{M}{n\epsilon^2}+\frac{2KM}{n\epsilon^2}+\frac{2\delta}{\epsilon^2} = \frac{2\delta}{\epsilon^2},$$

and since δ was arbitrary, we can take it to 0. Hence for every $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{S_n}{n}-0\right|>\epsilon\right) = 0,$$

and equivalently,

$$\frac{S_n}{n} \xrightarrow{\mathbb{P}} 0.$$

□