LECTURE NOTES ON LIMIT THEOREMS IN PROBABILITY

DAVID STEINSALTZ

1. Review: Laws of Large Numbers

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. (independent, identically distributed) random variables with expectation μ.

Theorem 1.1 (Weak Law of Large Numbers (WLLN))**.**

$$\lim_{n\to\infty} \frac{1}{n}\left(X_1 + \cdots + X_n\right) = \mu \text{ in probability.}$$

That is, for any $\epsilon > 0$*,*

$$\lim_{n\to\infty} \mathbb{P}\left\{\left|\frac{1}{n}\left(X_1 + \cdots + X_n\right) - \mu\right| > \epsilon\right\} = 0.$$

Theorem 1.2 (Strong Law of Large Numbers (SLLN))**.**

$$\lim_{n\to\infty} \frac{1}{n}\left(X_1 + \cdots + X_n\right) = \mu \text{ almost surely.}$$

That is,

$$\mathbb{P}\left\{\lim_{n\to\infty} \frac{1}{n}\left(X_1 + \cdots + X_n\right) = \mu\right\} = 1.$$

Elementary treatments usually assume that the random variables have a finite variance, or even (for the SLLN) a finite fourth moment. This is not required. Is independence required? Clearly we can't dispense with it entirely: If the random variables are all the same, there can't be convergence to a deterministic limit.

One elementary proof of the WLLN goes as follows: Suppose $X_1, X_2, \ldots$ are random variables with expectation μ and variance bounded by S. Let $\bar{X}_n := n^{-1}\left(X_1 + \cdots + X_n\right)$. Then

$$\begin{aligned}\mathrm{Var}(\bar{X}_n) &= n^{-2}\left(\sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{1\le i<j\le n} \mathrm{Cov}\left(X_i, X_j\right)\right) \\ &\le \frac{S}{n} + \frac{2}{n^2}\sum_{1\le i<j\le n} \mathrm{Cov}\left(X_i, X_j\right).\end{aligned}$$

When the random variables are independent, the covariances are all 0, so the variance is bounded by S/n, and by Chebyshev's Inequality

$$\mathbb{P}\left\{\left|\bar{X} - \mu\right| \geq \epsilon\right\} \leq \frac{S}{n\epsilon^2} \xrightarrow{n\to\infty} 0$$

for any $\epsilon > 0$. This is also true, of course, if the random variables are merely uncorrelated. But the covariances don't have to be exactly 0. All this proof requires is that the variance of $\bar{X}_n$ go to 0 as $n \to \infty$. Suppose the random variables approach being uncorrelated as they get further apart in the sequence. That is, suppose there are constants a_k with $\lim_{k\to\infty} a_k = 0$ such that

$$\mathrm{Cov}\left(X_i, X_j\right) \leq a_{|i-j|}.$$

Then

$$\begin{aligned}\mathrm{Var}(\bar{X}_n) &\leq \frac{2}{n^2}\sum_{k=0}^{n-1}(n-k)a_k \\ &\leq \frac{2}{n}\sum_{k=1}^{n} a_k \\ &\xrightarrow{n\to\infty} 0.\end{aligned}$$

The WLLN also has another strengthened form: The Central Limit Theorem (CLT). Dividing the partial sums by n is too much: It crushes out all the randomness. The CLT says that the randomness is actually on the order of $\sqrt{n}$, and that it converges to a Gaussian distribution:

Theorem 1.3.

$$n^{-1/2}\left(X_1 + \cdots + X_n - \mu n\right) \xrightarrow{n\to\infty}_d \mathcal{N}(0, \sigma^2).$$

I have not put explicit conditions on these theorems, because there are a lot of options for how to formulate the conditions. There are three kinds of assumptions that go into these limit theorems: moment conditions, independence conditions, and identical distribution conditions. Since in real applications these conditions might be hard to check, we'd like to know what the minimum required is. The brief answers are:

Moments These can be reduced to the obvious minimum required for the results to make sense. The laws of large numbers are about convergence to a mean, so they require that the random variables have an expectation; in fact, that is all that is required. The CLT is about convergence of distributions to a normal distribution with the same variance, so the X_i must have finite variance. This suffices, though there is a certain tradeoff between moments and identical distribution: If we allow the X_i to have different distributions, we need to at least assume that $\mathbb{E}[|X_i|^{2+\delta}]$ is universally bounded for some $\delta > 0$.

Independence If the random variables are identically distributed in a strong sense, called *stationary*, then we can reduce the independence requirement to a very weak form, called *ergodicity*, meaning that there is no **infinitely long-term** dependence. A description of this **ergodic theory** will be the main topic of this lecture.

Identical distribution The proof of the WLLN doesn't use anything about the distributions except their mean and a bound on their variance. It turns out that we can dispense with identical distribution for the SLLN and the CLT as well, as long as we strengthen the independence to a sort of average independence (still weaker than assuming actual independence) called the *martingale property*.

2. Ergodic theory

2.1. **Some definitions.** These definitions may be found in [11, 3]. Since we are looking at infinite sequences of random variables, a rigorous treatment requires measure-theoretic probability, which we will not use. Instead, we will make some definitions whose consistency will have to be accepted, and proceed from there. An accessible introduction to these concepts may be found in [10].

Let $X_1, X_2, \dots$ be a one-sided infinite discrete-time stochastic process taking values in a state space $\mathcal{X}$. We assume the process is *stationary*, meaning that the distribution of $(X_1, X_2, \dots)$ is the same as the distribution of $(X_n, X_{n+1}, \dots)$ for any n. It will be convenient to think of the sequence as being two-sided: $(\dots, X_{-1}, X_0, X_1, X_2, \dots)$.

Theorem 2.1. *Any one-sided stationary sequence may be embedded in a two-sided stationary sequence.*

Proof. Clearly we may define a stationary sequence $(X_{-n}, X_{-n+1}, \dots)$ for any n. The Kolmogorov Extension Theorem allows this to be extended to an infinite sequence. □

Our sample space is $\Omega = \mathcal{X}^{\mathbb{Z}}$. For $m \le n$ let $\mathcal{B}_m^n$ be the set of bounded $(X_i)_m^n$-*measurable* random variables, by which we mean bounded random variables that can be written as a function of the sequence $(X_m, X_{m+1}, \dots, X_n)$; we write $\mathcal{B}_m$ for $\mathcal{B}_{m,\infty}$ and $\mathcal{B}^n$ for $\mathcal{B}_{-\infty,n}$. We will also say that an event B is in $\mathcal{B}_m^n$ if its indicator is. We define the shift operator $\tau : \Omega \to \Omega$ by shifting every coordinate one to the left. That is, if $X := (\dots, X_{-1}, X_0, X_1, \dots)$ then

$$(\tau X)_i = X_{i+1}.$$

If $Y = f(X_m, \dots, X_n) \in \mathcal{B}_m^n$ then $Y \circ \tau^k = f(X_{m+k}, \dots, X_{n+k}) \in \mathcal{B}_{m+k}^{n+k}$. Note that if $B \subset \Omega$ is any event then

$$\mathbf{1}_{\tau^{-k}B} = \tau^k \mathbf{1}_B.$$

We say the stochastic process is *ergodic*

An important interpretation of *ergodic* is that an ergodic process can't be split into a mixture of processes with different distributions.

Theorem 2.2. *The following are equivalent:*

(i) *If A and B are any events with nonzero probability then $\mathbb{P}(A \cap \tau^{-n} B) > 0$ for some n;*
(ii) *Any shift-invariant event — that is, an event $A \subset \Omega$ such that $\mathbb{P}\left(A \triangle \tau^{-1} A\right) = 0$ — has probability 0 or 1;*
(iii) *Any shift-invariant function of $(X_0, X_1, \dots)$ — that is, a function f such that $f(X_0, X_1, \dots) = f(X_1, X_2, \dots)$ almost surely — is almost surely constant;*
(iv) *There is no way to represent (X_i) by defining two different stationary stochastic processes (Y_i) and (Z_i) with distinct distributions, and an independent Bernoulli(p) random variable ξ (with $0 < p < 1$), such that for all $i = 1, 2, \dots$*

$$X_i = \begin{cases} Y_i & \text{if } \xi = 0, \\ Z_i & \text{if } \xi = 1. \end{cases}$$

Proof. (i)$\Longrightarrow$(iii) Let f be a shift-invariant function. For any x let $A_x := \{f(X_0, X_1, \dots) > x\}$. By shift invariance

$$\tau^{-n} A_x = \left\{X \,:\, f(X_n, X_{n+1}, \dots) > x\right\},$$

so

$$\tau^{-n} A_x \cap A_x^{\mathsf{c}} \subset \left\{f(X_0, X_1, \dots) \neq f(X_n, X_{n+1}, \dots)\right\},$$

which has probability 0 for all n. It follows that A or A^{c} must have probability 0. Since this is true for any x, it follows that f is almost-surely constant.

(iii)$\Longrightarrow$(ii)

(ii)$\Longrightarrow$(i) Consider any sets A and B with nonzero probability such that $\mathbb{P}(A \cap \tau^{-n} B) = 0$ for all n. Then

$$C := \bigcup_{n=0}^{\infty} \tau^{-n} B$$

must have probability strictly between 0 and 1. Also, $\tau^{-1} C \subset C$. Since $\mathbb{P}(\tau^{-1} C) = \mathbb{P}(C)$ (since the process is stationary) it follows that C is invariant. Since $\mathbb{P}(C) \geq \mathbb{P}(A) > 0$, it must be that $\mathbb{P}(C) = 1$. But then

$$0 < \mathbb{P}\left(A \cap C\right) = \mathbb{P}\left(\bigcup_{n=0}^{\infty} A \cap \tau^{-n} B\right).$$

The equivalence to (iv) is left as an exercise. □

We say the process is ergodic if it satisfies these equivalent conditions.

2.2. **Mixing conditions.** The stochastic process is *mixing* (or *strongly mixing*) if for any events A, B

$$\lim_{n\to\infty} \mathbb{P}\left(A \cap \tau^{-n} B\right) = \mathbb{P}(A)\mathbb{P}(B). \tag{1}$$

Finally, the stochastic process is *ϕ-mixing* if for any k, and any

$$\phi(n) := \sup_{A\in\mathcal{B}_{-\infty,m},\, B\in\mathcal{B}_{m+n,\infty}} \left|\mathbb{P}\left(B \mid A\right) - \mathbb{P}\left(B\right)\right| \xrightarrow{n\to\infty} 0. \tag{2}$$

Suppose the process is mixing, and let A be any shift-invariant event. Then

$$\mathbb{P}(A) = \mathbb{P}\left(A \cap \tau^{-n} A\right),$$

so, taking the limit as $n \to \infty$, we have $\mathbb{P}(A) = \mathbb{P}(A)^2$, implying $\mathbb{P}(A)$ is 0 or 1. Thus

$$\phi\text{–mixing} \implies \text{mixing} \implies \text{ergodic}\ .$$

The main result about ergodic processes is that they satisfy the equivalent of the Strong Law of Large Numbers, Birkhoff's Ergodic Theorem, stated in section 2.4. It follows that ergodicity is equivalent to

$$\lim_{n\to\infty} n^{-1} \sum_{i=0}^{n-1} \mathbb{P}\left(A \cap \tau^{-n} B\right) = \mathbb{P}(A)\mathbb{P}(B). \tag{3}$$

for any events A, B.

The definition in terms of the map τ may be hard to understand at first glance. The definition of mixing is essentially a generalisation of the statement that X_0 and X_n are asymptotically independent; that is, for any $A, B \subset \mathcal{X}$

$$\lim_{n\to\infty} \mathbb{P}\left\{X_0 \in A,\, X_n \in B\right\} = \mathbb{P}\left\{X_0 \in A\right\} \cdot \mathbb{P}\left\{X_n \in B\right\}.$$

A thorough discussion of various mixing conditions may be found at `https://www.encyclopediaofmath.org/index.php/Strong_mixing_conditions`

2.3. **Examples.**

2.3.1. *i.i.d. sequences.* Any i.i.d. sequence is mixing, hence ergodic. An immediate consequence is the Kolmogorov zero–one law: We define a *tail event* for an i.i.d. sequence to be an event whose occurrence may be inferred from $X_n, X_{n+1}, \ldots$ for n arbitrarily large. The zero–one law says that all tail events have probability 0 or 1. For example, the event $\{n^{-1}(X_1 + \cdots + X_n)$ converges to a limit$\}$ is a tail event, so must have probability 0 or 1. (The SLLN states that the probability is 1.)

2.3.2. *Rotations of the circle.* Fix $\alpha \in (0,1)$. Let X_0 be uniform on $[0,1)$, and let $X_n = X_0 + \alpha n (\mathrm{mod}\ 1)$. This sequence is ergodic iff α is irrational. Suppose α is irrational, and let f be a shift-invariant function, which may be taken to be $f(X_0)$. We know that there is a unique sequence of Fourier coefficients

$$a_k = \int_0^1 e^{-2\pi \mathrm{i} k x} f(x)\,\mathrm{d}x$$

such that

$$\sum_{k=-K}^{K} a_k e^{2\pi \mathrm{i} k x} \xrightarrow{K\to\infty} f(x).$$

Since $f \circ \tau(x) = f(x+\alpha)$, we have

$$\sum_{k=-K}^{K} \left(a_k e^{2\pi \mathrm{i} k \alpha}\right) e^{2\pi \mathrm{i} k x} \xrightarrow{K\to\infty} f(x+\alpha).$$

Since f is shift-invariant this must be the same as $f(x)$, so we have an alternative Fourier series $a_k e^{2\pi \mathrm{i} k\alpha}$ for the same function. Since the Fourier transform is unique, we must have for all k

$$a_k e^{2\pi \mathrm{i} k\alpha} = a_k.$$

But this happens only if $k\alpha$ is an integer or $a_k = 0$. Thus $a_k = 0$ for all $k \neq 0$, implying that f is constant.

We leave it as an exercise to show the sequence is not ergodic when α is rational, and that it is never mixing.

2.3.3. *Markov chains.* Let $\mathcal{X}$ be finite, and let $X_0, X_1, X_2, \ldots$ be a Markov chain with transition probabilities $P(i,j)$, and X_0 in a stationary distribution π_i satisfying $\pi_j = \sum_{i\in\mathcal{X}} \pi_i P(i,j)$. Then (X_i) is a stationary sequence. It is ergodic if and only if it is irreducible and aperiodic. If $\mathcal{X}$ is countably infinite it must be recurrent.

2.4. **The Birkhoff Pointwise Ergodic Theorem.**

Theorem 2.3 (Birkhoff pointwise ergodic theorem)**.** *Let $X_0, X_1, \ldots$ be an ergodic stationary sequence, and $f : \mathcal{X} \to \mathbb{R}$ any function such that $\mu := \mathbb{E}[f(X_0)]$ is defined. Then*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f(X_i) = \mu \text{ almost surely.} \tag{4}$$

Proof. This proof comes from [6]. Let

$$\begin{aligned} S_n &:= n^{-1} \sum_{i=0}^{n-1} f(X_i), \\ S_N^* &:= \max_{1 \le n \le N} S_n, \\ S^* &:= \sup_N S_N^*, \\ \bar{S} &:= \limsup_{n \to \infty} S_n, \qquad \underline{\mathrm{S}} := \liminf_{n \to \infty} S_n. \end{aligned}$$

Since $\bar{S}$ and $\underline{\mathrm{S}}$ are invariant, they must be constant. Since $\bar{S} \ge \mathbb{E}[X_0] \ge \underline{\mathrm{S}}$, all we need to show is that $\bar{S} = \underline{\mathrm{S}}$.

Suppose f is bounded, and fix a natural number N and some $\epsilon > 0$. Let E_N be the event $\{S_N^* > \bar{S} - \epsilon\}$. Note that if $(X_i) \notin E_N$ then in particular $f(X_0) \le \lambda$, so we always have

$$\big(f(X_0) - \lambda\big)\, \mathbf{1}_{E_N} \ge f(X_0) - \lambda.$$

Note as well that $\mathbf{1}_{E_N}$ are nondecreasing in N, with

$$\lim_{N \to \infty} \mathbf{1}_{E_N} = \mathbf{1}_{S^* \ge \bar{S} - \epsilon} = 1.$$

Since $\mathbb{P}(E_N) > 0$, there is almost surely some k such that $\tau^k(X) = (X_k, X_{k+1}, \dots) \in E_N$. Following that, there is some $n \le N$ such that $\sum_{i=k}^{k+n-1} f(X_i) \ge n\lambda$. Thus, if we look at the sum

$$\sum_{i=0}^{m-1} \big(f(X_i) - \lambda\big)\, \mathbf{1}_{E_N}(X_i, X_{i+1}, \dots)$$

for some large m, it breaks up into sequences of summands that are 0 alternating with sums of $\le N$ terms that total to something positive. The sum can only be negative if it ends in the middle of one of the latter blocks, and then the smallest it can be is $-N(|f|_\infty + \lambda^+)$. Computing the expectation, and using stationarity, we have for all m

$$m\mathbb{E}\left[\big(f(X_0) - \bar{S} + \epsilon\big)\, \mathbf{1}_{E_N}\right] \ge -N\left(|f|_\infty + \bar{S}\right).$$

Dividing by m and sending $m \to \infty$ we then have

$$\mathbb{E}\left[\big(f(X_0) - \bar{S} + \epsilon\big)\, \mathbf{1}_{E_N}\right] \ge 0.$$

We can then send $N \to \infty$ and conclude[1] that

$$\mathbb{E}\left[f(X_0)\right] \ge \bar{S} - \epsilon.$$

[1]The fact that the limit of the expectation is the expectation of the limit in this case is an example of the Dominated Convergence Theorem, which may be found in any text on measure-theoretic probability.

Since this is true for any $\epsilon > 0$, it follows that $\mathbb{E}\left[f(X_0)\right] \geq \bar{S}$. Applying the same result to the function $-f$, since $\limsup(-S_n) = -\liminf S_n$, we have

$$-\mathbb{E}\left[f(X_0)\right] \geq -\underline{\mathrm{S}}.$$

Thus

$$\bar{S} \geq \underline{\mathrm{S}} \geq \mathbb{E}\left[f(X_0)\right] \geq \bar{S},$$

so $\bar{S} = \underline{\mathrm{S}}$.

We have now completed the proof for f bounded. We extend to general f by the usual expedient of truncating at some constant K, which we then let go to ∞. □

2.5. **Recurrence.** If $X = (X_0, X_1, \dots)$ is a stationary ergodic process, then it makes sense to suppose that any state that it has nonzero probability of being in will be returned to infinitely often. This is the content of the Poincaré Recurrence Theorem. The Kac Recurrence Theorem tells us, in addition, that the return will happen at a frequency proportional to the probability.

Theorem 2.4 (Poincaré Recurrence Theorem)**.** *Let $X = (X_0, X_1, \dots)$ be a stationary process. Let $A \subset \Omega$ be any event, then almost every point of A is recurrent. That is,*

$$\mathbb{P}\left\{X \in A, \{n : \tau^n X \in A\} \text{ is finite}\right\} = 0.$$

If the process is ergodic and $\mathbb{P}(A) > 0$ then almost every point is recurrent to A. That is,

$$\mathbb{P}\left\{\{n : \tau^n X \in A\} \text{ is finite}\right\} = 0.$$

The event A may involve the entire future of the process. In the special case where $A = \{\omega : \omega_0 \in E\}$ for some $E \subset \mathcal{X}$, it states that $X_n \in E$ infinitely often with probability 1.

Definition 2.5. *For an event $A \subset \Omega$ we define the first recurrence time*

$$R_A(\omega) := \min\{n > 0 : \tau^n \omega \in A\}$$

The Poincaré Recurrence Theorem tells us that $R_A(\omega) < \infty$ for almost every $\omega \in A$, and for almost every $\omega \in \Omega$ when the process is ergodic.

Theorem 2.6 (Kac Recurrence Theorem)**.** *If (X_n) is ergodic, for any A with $\mathbb{P}(A) > 0$,*

$$\mathbb{E}\left[R_A(X) \;\middle|\; X \in A\right] = \frac{1}{\mathbb{P}(A)}.$$

When $X_0, X_1, \dots$ is a stationary ergodic Markov chain on a finite state space $\mathcal{X}$, we may take $A = \{\omega : \omega_0 = i\}$ for any $i \in \mathcal{X}$. Then $R_A(X) = \min\{n \geq 1 : X_n = i\}$, and the Kac Recurrence Theorem tells us that the expected time to return to i for a process started at i is $1/\pi_i$.

Consider the simple random walk $S_n = X_1 + \cdots + X_n$ with $X_i = \pm 1$ with probability $\frac{1}{2}$. The SLLN tells us that $S_n/n \to 0$, and the CLT tells us that S_n is approximately normal with mean 0 and variance n. But we also know

that S_n is recurrent; that is, S_n returns to 0 infinitely often. We may also want to know how often a random walk returns to 0.

2.6. **The subadditive ergodic theorem.**

Theorem 2.7 (Subadditive ergodic theorem)**.** *Suppose* $(X_{m,n})_{m=0}^{n-1}$, $n = 1, 2, \dots$ *is a triangular array of random variables satisfying*

(i) $X_{0,m} + X_{m,n} \geq X_{0,n}$*;*
(ii) *For each* $k \geq 1$*, the sequence* $Y_n := X_{nk,nk+k}$ *for* $n = 1, 2, dots$ *is stationary and ergodic;*
(iii) *The distribution of* $\left(X_{m,m+k}\right)_{m=1}^{\infty}$ *does not depend on* m*;*
(iv) $\mathbb{E}\left[X_{0,1}^{+}\right] < \infty$ *and* $\mathbb{E}\left[X_{0,n}\right] \geq -\gamma_0 n$ *for some finite* γ_0*.*

Then

$$\lim_{n\to\infty} \frac{X_{0,n}}{n}$$

exists almost surely, and is almost surely equal to

$$\gamma := \inf_m \frac{\mathbb{E}[X_{0,m}]}{m}.$$

Note: This theorem is originally due to Kingman. Following [3] we give the stronger version — that is, with weaker conditions — due to Liggett. In the following diagram, the coloured variables are the starts of stationary sequences described in condition (ii):

$$\begin{array}{cccccc}
X_{0,1} & & & & & \\
X_{0,2} & X_{1,2} & & & & \\
X_{0,3} & X_{1,3} & X_{2,3} & & & \\
X_{0,4} & X_{1,4} & X_{2,4} & X_{3,4} & & \\
X_{0,5} & X_{1,5} & X_{2,5} & X_{3,5} & X_{4,5} & \\
X_{0,6} & X_{1,6} & X_{2,6} & X_{3,6} & X_{4,6} & X_{5,6} \\
\vdots & & & & & \vdots \\
X_{0,n} & \cdots & \cdots & \cdots & \cdots & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{array}.$$

The columns all have the same distribution, as described in condition (iii), but are not necessarily stationary.

To understand why the definitions are this way, consider the case where $\xi_1, \xi_2, \dots$ is a stationary ergodic sequence with finite expectation, and $X_{m,n} := \xi_{m+1} + \cdots + \xi_n$. Then the sequence is additive, rather than merely subadditive as required for (i). The sequences defined in (ii) are stationary because they involve non-overlapping successive partial sums from the stationary sequences.

Example: Range of a random walk: Let $\xi_1, \xi_2, \dots$ be a stationary sequence, and $S_n := \xi_1 + \cdots + \xi_n$. Let $X_{m,n} = \left|\{S_{m+1}, S_{m+2}, \dots, S_n\}\right|$ be the number of points ever hit by the random walk. This satisfies all

the conditions, so we may infer that the range increases linearly at rate γ asymptotically, though this does not immediately tell us anything about how to compute γ.

Products of random matrices: Suppose $M_1, M_2, \ldots$ are a stationary sequence of $k \times k$ random positive matrices satisfying

$$\mathbb{E}\left[\|M_i\|\right] < \infty.$$

Let

$$X_{m,n} := \log \|M_n \cdot \cdots \cdot M_{m+1}\|$$

This satisfies the subadditivity condition $X_{0,m} + X_{m,n} \geq X_{0,n}$. We may conclude that

$$\lambda := \lim_{n\to\infty} n^{-1} \log \|M_n \cdot \cdots \cdot M_1\|$$

exists and is constant almost surely.

In fact, this result holds in significantly greater generality: The matrices do not need to be positive, and the Multiplicative Ergodic Theorem tells us that there are k deterministic Lyapunov exponents $\lambda = \lambda_k \geq lambda_{k-1} \geq \cdots \geq \lambda_1$ such that there is a (random) decomposition of $\mathbb{R}^k$ into $j-dimensional$ subspaces E_j such that $E_j \setminus E_{j-1}$ comprises points x such that

$$\lim_{n\to\infty} n^{-1} \log \|M_n \cdots M_1 x\| = \lambda_j.$$

These are the stochastic analogue of the decomposition of $\mathbb{R}^k$ into eigenspaces of some fixed matrix M.

2.7. **Central Limit Theorem.** Let $X_0, X_1, \ldots$ be a stationary sequence and f a function from $\mathcal{X}$ to $\mathbb{R}$. Let $S_n = X_0 + \cdots + X_{n-1}$ and $\mu = \mathbb{E}[X_i]$. The variance of S_n is given by

$$\sigma_n^2 := \operatorname{Var}(S_n) = n \operatorname{Var}(X_0) + 2\sum_{i=1}^{n-1}(n-i)\operatorname{Cov}(X_0, X_i).$$

The following theorem is due to Ibragimov [5]:

Theorem 2.8. *Suppose the stationary process $X_0, X_1, \ldots$ is ϕ-mixing, $\lim_{n\to\infty} \sigma_n = \infty$, and for some $\delta > 0$*

$$\mathbb{E}\left[|X_i|^{2+\delta}\right] < \infty.$$

Then $(S_n - n\mu)/\sigma_n \xrightarrow{n\to\infty}_d \mathcal{N}(0,1)$.

3. Martingales

An excellent source for all of what is presented here (and much more) is [13].

3.1. **Definitions.** A *martingale* is a mathematical formalisation of the notion of a "fair bet": A process that on average stays constant, regardless of any information about the past behaviour.

A stochastic process $X_0, X_1, X_2, \dots$ is a *martingale difference sequence* if the expectations are finite and for any $m < n$

$$\mathbb{E}\left[X_n \;\middle|\; X_0, \dots, X_m\right] = 0. \tag{5}$$

The partial-sum sequence $M_n := X_0 + \cdots + X_n$ is then called a *martingale*. If the condition '$= 0$' is replaced by '≥ 0' then these are called *submartingale differences* and *submartingale* respectively. If replaced by '≤ 0' then they are *supermartingale differences* and *supermartingale*. Note that a martingale is also a submartingale and a supermartingale.

Examples:

(i) **Symmetric random walk**. Given $X_1, X_2, \dots$ independent with $X_i = \pm 1$ with probability $1/2$, this is a martingale difference sequence, and $S_n = X_1 + \cdots + X_n$ is a martingale.

(ii) **Exponential martingale**. Given $X_1, X_2, \dots$ i.i.d. with moment generating function $e^{m(t)}$ — that is, $\log \mathbb{E}[e^{tX_i}] = m(t)$ — for any t, $M_n := \exp\{tS_n - nm(t)\}$ is a martingale.

(iii) **Polya's Urn**. We have an urn that starts with R_0 red and B_0 blue balls. We pick a ball at random, and then replace it in the urn with k balls of the same colour. Let X_n be the fraction of red balls after the k-th draw. Then $X_0, X_1, \dots$ is a martingale.

Definition 3.1. *A* stopping time *is a random time T such that at time T it can be determined that T has occurred based on $X_0, X_1, \dots, X_T$.*

Definition 3.2. *A stochastic process $X_0, X_1, \dots$ is* uniformly integrable *if*

$$\lim_{K\to\infty} \sup_i \mathbb{E}\left[|X_i| \mathbf{1}_{\{|X_i|>K\}}\right] = 0.$$

Examples: The first entry time $T = \min\{n \geq 0 : X_n \in E\}$ into a set $E \subset \mathcal{X}$ is a stopping time. The last exit time $T = \max\{n \geq 0 : X_n \in E\}$ is not.

3.2. **Stopping and Convergence.** The two key results about martingales are the **Optional Stopping Theorem** and the **Martingale Convergence Theorem**. The Optional Stopping Theorem tells us that a martingale stopped at a random stopping time is still a martingale; in other words, when playing a succession of fair games there is no way to make money on average by choosing some gambling system that tells you when to stop, unless you can look into the future.

Lemma 3.3. *Let X_n be a submartingale and T a stopping time. Suppose that one of these conditions holds:*

(i) $\mathbb{E}[T] < \infty$ *and there is a constant C such that*

$$\mathbb{E}[|X_{n+1} - X_n| \,|\, X_1, \dots, X_n] \leq C \quad \textit{almost surely};$$

(ii) $X_1, X_2, \ldots$ *is uniformly integrable and* T *is almost surely finite;*
(iii) T *is bounded.*

Then the process $X_{n\wedge T}$ *is uniformly integrable.*

Theorem 3.4 (Optional Stopping)**.** *If* (X_n) *is a submartingale then* $(X_{n\wedge T})$ *is a submartingale. If* $X_{n\wedge T}$ *is uniformly integrable then*

$$\mathbb{E}\left[X_T \;\middle|\; X_0\right] \geq X_0.$$

For a supermartingale the same holds with reversed inequalities. And if (X_n) *is a martingale with* $X_{n\wedge T}$ *uniformly integrable then*

$$\mathbb{E}\left[X_T \;\middle|\; X_0\right] = X_0.$$

Note that uniform integrability is definitely required. Consider, for example, the simple random walk started at 0, stopped at $T = \min\{n : S_n = -1\}$. It is a martingale, and T is almost surely finite, but $\mathbb{E}[X_T] = -1 < 0 = X_0$. The reason is that $\{S_n\}$ are not uniformly integrable, and T has infinite expectation.

Theorem 3.5 (Martingale convergence Theorem)**.** *Suppose* $X_0, X_1, \ldots$ *is a submartingale with* $\mathbb{E}[X_n^+] < K$ *for some fixed* K*. Then* $X_\infty := \lim_{n\to\infty} X_n$ *exists almost surely, and* $\mathbb{E}[|X_\infty|] < \infty$*.*

Example: Let $X_1, X_2, \ldots$ be an i.i.d. sequence with $\mathbb{E}[X_i] = 0$ and nonzero variance, with $\mathbb{P}\{X_i \leq K\} = 1$ Let $S_n = X_1 + \cdots + X_n$. Let $T = \min\{n : S_n \geq 0\}$. $S_{T\wedge n}$ is a martingale (hence also submartingale), and $S_{T\wedge n} \leq K$ for all n. Then $S_{T\wedge n}$ converges. Since S_n cannot converge to any finite value (since it is changing by fixed increments at each step), this implies that $S_{n\wedge T}$ is stopped by hitting T. Hence $T < \infty$ almost surely.

3.3. **Martingale Central Limit Theorem.** It is clear that the Martingale CLT cannot be quite as simple as the CLT for identically distributed random variables, since the variance of the successive martingale differences not only could easily be different, but they could be random, depending on past random variables. There are numerous versions, described at length in [4], which is available online at `http://www.stat.yale.edu/~mjk56/MartingaleLimitTheoryAndItsApplication.pdf`

One standard approach is to take a random number of terms to make up a fixed variance.

Theorem 3.6 (Martingale Central Limit Theorem)**.** *Let* $\xi_1, \xi_2, \ldots$ *be a martingale difference sequence with* ξ_i *bounded, and* $X_n = \xi_1 + \cdots + \xi_n$*. Let*

$$\sigma_n^2 := \mathbb{E}\left[\xi_{n+1}^2 \;\middle|\; \xi_1, \ldots, \xi_n\right]$$

be the n*-th conditional variance. For each* $v > 0$ *let*

$$\tau_v := \min\{n : \sum_{i=0}^{n} \sigma_i^2 > v\}.$$

Then if τ_v is almost-surely finite for all v,

$$\frac{X_{\tau_v}}{\sqrt{v}} \xrightarrow{v\to\infty}_d \mathcal{N}(0,1).$$

4. Concentration Inequalities

> In the long run we are all dead. Economists set themselves too easy, too useless a task if in tempestuous seasons they can only tell us that when the storm is past the ocean is flat again. J. M. Keynes

So far, we have examined the long-run behaviour of stochastic sums, in the limit as n goes to ∞. This can provide some assurance that a particular statistical test or error estimate is on the right track. But it is practically meaningless without some information about the *rate* of convergence, ideally with concrete bounds on the error probabilities.

Suppose we have a sequence of random variables $Y_1, Y_2, \ldots, Y_n, \ldots$. These might be partial sums of i.i.d. random variables, or some other function of approximately independent variables. There are fundamentally three different kinds of error measurements we are typically interested in:

Concentration of measure — We want to show that the distribution of Y_n is close — in a quantitative sense — to a limiting distribution.

Tail bounds — The real reason why we usually want a CLT is usually to show that the probability of some extreme value x_{obs} when we expected x_{exp} is very small, like $\mathrm{e}^{-C(x_{\text{obs}}-x_{\text{exp}})^2}$ for some appropriate constant C. General results about convergence in distribution are particularly bad about defining convergence out in the tails, so it's usually better to derive a bound on the tail probability directly.

Strong convergence — If the sequence of random variables Y_i themselves converge to some limit random variable, then we can measure the probability of Y_i still being a certain distance away from its ultimate limit.

Note that the first two approaches are statements about the behaviour of the *distributions* or *laws* of the random variables. The random variables themselves are just convenient placeholders, and there is no assumption that they exist jointly on any probability space. For this reason they describe what is known as *weak convergence*. We will be primarily concerned here with weak convergence.

We note, though, that there are general *representation theorems* that say in great generality that when you have weak convergence, you can create a probability space where all the random variables live together in peace and harmony and strong convergence. This is often a useful tool. For more information see [8, Section IV.3].

4.1. How to measure probabilistic errors.

Recall that a random variable is a function from a sample space Ω to real numbers. A probability measure $\mathbb{P}$ on Ω may be thought of as a map from random variables to $\mathbb{R}$ that

gives you the expected value. We can use the notation $\mathbb{P}f = \int_\Omega f(x)d\mathbb{P}(x)$. Then distributions $\mathbb{P}$ and $\mathbb{Q}$ are similar in the same way that any other mappings are similar: if $|\mathbb{P}f - \mathbb{Q}f|$ is always small. This leads to the definition of *total variation distance*:

$$d_{\mathrm{TV}} = \sup_{|f|\le 1} |\mathbb{P}f - \mathbb{Q}f| . \tag{6}$$

This is a useful definition for discrete state spaces — where we also have $d_{\mathrm{TV}} = \sum_{\omega\in\Omega} \left|\mathbb{P}(\{\omega\}) - \mathbb{Q}(\{\omega\})\right|$ — but for larger state spaces we have the inevitable problem that there are just too many functions. A simple example of where TV distance doesn't behave as you might like is the collection of distributions $\mathbb{P}_n$ on $\Omega = [0,1]$ that are uniformly distributed on the set of points $\{i/n : i = 1,\dots,n-1\}$. It seems like these are approximations to the uniform distribution on $[0,1]$, call it U, but for any distinct primes n and n' we have $d_{\mathrm{TV}}(\mathbb{P}_n, \mathbb{P}_{n'}) = 1$ and for all n $d_{\mathrm{TV}}(\mathbb{P}_n, U) = 1$.

Any set $\mathcal{F}$ of functions $\Omega \to \mathbb{R}$ defines a distance by $\sup_{f\in\mathcal{F}} |\mathbb{P}f - \mathbb{Q}f|$. One commonly used distance when Ω is a metric space is the *Wasserstein distance*, defined by taking $\mathcal{F}$ to be the set of 1-Lipschitz functions bounded by 1. That is,

$$\mathcal{F}_W = \left\{ f : \Omega \to \mathbb{R} : \left|f(x)\right| \le 1 \text{ and } \left|f(x) - f(y)\right| \le d(x,y) \right\}.$$

When $\Omega = \mathbb{R}$ another possibility is the *Kolmogorov–Smirnov distance*, defined by

$$\mathcal{F}_{KS} = \left\{ \mathbf{1}_{(-\infty,x]} : x \in \mathbb{R} \right\}.$$

That is, the Kolmogorov–Smirnov distance is the maximum difference between the cdfs.

Definition 4.1. *Suppose (Ω, d) is a metric space. We say that a sequence of probability measures $\mathbb{P}_n$ converges to the probability* $\mathbb{P}$ weakly *or* in distribution *if for every bounded continuous function $f : \Omega \to \mathbb{R}$,*

$$\lim_{n\to\infty} \mathbb{P}_n f = \mathbb{P}f. \tag{7}$$

It is useful to note that convergence is determined by (7) holding for sufficiently large collections of sets f, which may or may not be contained in the set $\mathcal{F}_{bc}$ of bounded continuous functions. The key criterion is that we need convergence for a sufficiently large class $\mathcal{F}$ of functions such that every function in $\mathcal{F}_{bc}$ can be approximated by linear combinations of functions in $\mathcal{F}$. For example:

- $\mathcal{F}_W$ determines weak convergence;
- $\mathcal{F}_{KS}$ determines weak convergence for distributions on $\mathbb{R}$;
- The set of trigonometric functions $x \mapsto \mathrm{e}^{\mathrm{i}\lambda x}$ for all λ in some real interval. This is the continuity theorem for characteristic functions.
- The set of exponential functions $x \mapsto \mathrm{e}^{\lambda x}$ for all λ in some real interval. This is the continuity theorem for moment generating functions.

Note that these distance measures are quite different, despite producing all the same notion of convergence. Smaller sets of functions produce smaller distances — hence making it easier to bound the distances — but also makes a bound on the distance less useful (since it provides information about fewer test functions).

An excellent source for the results in the rest of this section is [1].

4.2. **Nonasymptotic CLTs.** A basic non-asymptotic version of the Central Limit Theorem is the following:

Theorem 4.2. *Let $X_1, \dots, X_n$ and $Y_1, \dots, Y_n$ be two sequences of independent real-valued random variables, such that $\mathbb{E}[X_i] = \mathbb{E}[Y_i]$ and $\mathbb{E}[X_i^2] = \mathbb{E}[Y_i^2]$ for each i, and there is a finite constant K such that $\mathbb{E}[X_i^3] \le K$ and $\mathbb{E}[Y_i^3] \le K$ for each i. Define*

$$S_X := \sum_{i=1}^{n} \left(X_i - \mathbb{E}[X_i]\right), \quad S_Y := \sum_{i=1}^{n} \left(Y_i - \mathbb{E}[Y_i]\right),$$

$$\sigma := \left(n^{-1} \sum_{i=1}^{n} \mathrm{Var}(X_i)\right)^{1/2}$$

Then for any function f with bounded 3rd derivatives we have

$$\left| \mathbb{E}\left[f\left(\frac{S_X}{\sigma\sqrt{n}}\right)\right] - \mathbb{E}\left[f\left(\frac{S_X}{\sigma\sqrt{n}}\right)\right] \right| \le \frac{\|f'''\|_\infty K}{\sigma^3} n^{-1/2}. \tag{8}$$

Proof. We may assume without loss of generality that $\mathbb{E}[X_i] = \mathbb{E}[Y_i] = 0$. Also, since the assumptions and conclusions make no mention of the joint distribution of X and Y, we may place them on a joint probability space where they are independent. Define, for $1 \le i \le n$, and $x \in \mathbb{R}$,

$$S_i := \sum_{j=1}^{i-1} X_j + \sum_{j=i+1}^{n} Y_j.$$

We have $S_X = S_n + X_n$ and $S_Y = S_1 + Y_1$, and $S_i + X_i = S_{i+1} + Y_{i+1}$ hence

$$\begin{aligned} &\left| \mathbb{E}\left[f\left(\frac{S_X}{\sigma\sqrt{n}}\right)\right] - \mathbb{E}\left[f\left(\frac{S_Y}{\sigma\sqrt{n}}\right)\right] \right| \\ &\qquad \le \sum_{i=1}^{n} \left| \mathbb{E}\left[f\left(\frac{S_i + X_i}{\sigma\sqrt{n}}\right)\right] - \mathbb{E}\left[f\left(\frac{S_i + Y_i}{\sigma\sqrt{n}}\right)\right] \right| \end{aligned} \tag{9}$$

By Taylor's Theorem,

$$f\left(\frac{S_i + x}{\sigma\sqrt{n}}\right) = f\left(\frac{S_i}{\sigma\sqrt{n}}\right) + \frac{x}{\sigma\sqrt{n}} f'\left(\frac{S_i}{\sigma\sqrt{n}}\right) + \frac{x^2}{2\sigma^2 n} f''\left(\frac{S_i}{\sigma\sqrt{n}}\right) + \frac{x^3}{6\sigma^3 n^{3/2}} f'''(\theta),$$

where θ is between $S_i(0)/\sigma\sqrt{n}$ and $S_i(x)/\sigma\sqrt{n}$. Thus

$$\begin{aligned}\mathbb{E}\left[f\left(\frac{S_i+X_i}{\sigma\sqrt{n}}\right)\right] - \mathbb{E}\left[f\left(\frac{S_i+Y_i}{\sigma\sqrt{n}}\right)\right] \\ = \mathbb{E}\left[\frac{X_i-Y_i}{\sigma\sqrt{n}}\sigma\sqrt{n}f'\left(\frac{S_i}{\sigma\sqrt{n}}\right)\right] + \mathbb{E}\left[\frac{X_i^2-Y_i^2}{2\sigma^2 n}f''\left(\frac{S_i}{\sigma\sqrt{n}}\right)\right] \\ + \mathbb{E}\left[\frac{X_i^3}{6\sigma^3 n^{3/2}}f'''\left(\theta_x\right)\right] - \mathbb{E}\left[\frac{Y_i^3}{6\sigma^3 n^{3/2}}f'''\left(\theta_y\right)\right].\end{aligned}$$

Since X_i and Y_i are both independent of S_i, and since $\mathbb{E}[X_i - Y_i] = \mathbb{E}[X_i^2 - Y_i^2] = 0$ we get

$$\begin{aligned}\mathbb{E}\left[f\left(\frac{S_i+X_i}{\sigma\sqrt{n}}\right)\right] - \mathbb{E}\left[f\left(\frac{S_i+Y_i}{\sigma\sqrt{n}}\right)\right] \\ = \mathbb{E}\left[\frac{X_i^3}{6\sigma^3 n^{3/2}}f'''\left(\theta_x\right)\right] - \mathbb{E}\left[\frac{Y_i^3}{6\sigma^3 n^{3/2}}f'''\left(\theta_y\right)\right].\end{aligned}$$

Thus

$$\left|\mathbb{E}\left[f\left(\frac{S_i+X_i}{\sigma\sqrt{n}}\right)\right] - \mathbb{E}\left[f\left(\frac{S_i+Y_i}{\sigma\sqrt{n}}\right)\right]\right| \leq \frac{\|f'''\|_\infty}{6\sigma^3 n^{3/2}}\left(\mathbb{E}\left[|X_i|^3\right] + \mathbb{E}\left[|Y_i|^3\right]\right).$$

Plugging this into (9) completes the proof. □

It may not be immediately obvious why this is a CLT. This theorem says that if convolve independent random variables with each other, as long as they have the same mean and variance, the distributions get closer together. Convolution is a **contraction** on the space of probability distributions. In any such situation, repeated application of a contraction leads to a unique fixed point. We already know that the Gaussian distribution is a fixed point, so that's where the process leads.

To put it differently, if Y_i are themselves Gaussian with mean μ and variance σ^2, then $S_Y/\sigma\sqrt{n}$ is standard Gaussian. This tells us that the average of n independent random variables with mean μ and variance σ^2 converges in distribution to this standard Gaussian, with a distance bounded by the right-hand side of (8).

The undergraduate version of the CLT and convergence in distribution is expressed in terms of cumulative distribution functions: Letting $\Phi(x) = \int_{-\infty}^{x}(2\pi)^{-1/2}\mathrm{e}^{-z^2/2}\,\mathrm{d}z$ be the standard Gaussian cdf, and F_n the cdf of $(X_1 + \cdots + X_n)/\sigma\sqrt{n}$. Then for each x,

$$\lim_{n\to\infty} F_n(x) = \Phi(x).$$

The cdf is the expected value of the indicator function $\mathbf{1}_{(-\infty,x)}$, which is not a differentiable function. We can derive the CLT from Theorem 4.2

by approximating the indicator with smooth functions. In order to derive bounds on the rate of convergence we need more sophisticated tools. For details see chapter 2 of [12]. One version of such a theorem is called the *Berry–Esseen Theorem*.

Theorem 4.3. *Given the conditions of Theorem 4.2, with F_n the cdf of S_X. Then*

$$\int_{-\infty}^{\infty} \left|F_n(x) - \Phi(x)\right| \, \mathrm{d}x \leq \frac{6\tau}{\sqrt{n}}. \tag{10}$$

Also, for each x,

$$\left|F_n(x) - \Phi(x)\right| \leq \frac{10\tau}{\sqrt{n}}. \tag{11}$$

Another sort of generalisation abandons the assumption that we are looking at sums. Theorem 4.2 states that for any independent random variables $X_1, \ldots, X_n$ with bounded third moments, and any alternative random variables $Y_1, \ldots, Y_n$ with the same first and second moments and bounded third moments, the random variables

$$g(X_1, \ldots, X_n) \text{ and } g(Y_1, \ldots, Y_n)$$

have approximately the same distribution, with an error of order $n^{-1/2}$, where g is the function $g(\mathbf{x}) = n^{-1/2} \sum (x_i - \mu_i)$, for $\mu_i := \mathbb{E}[X_i]$. Essentially the same proof may be generalised to any function g whose dependence (as measured by partial derivatives up to third order) on any individual component is small. One version of this statement is Theorem 1.2 of [2], of which we give a simplifed form here.

Theorem 4.4. *Let $g : \mathbb{R}^n \to \mathbb{R}$ be thrice differentiable, and let all k-th mixed partials be bounded by $C_k n^{-k/2}$, for $k = 1, 2, 3$. Let $f : \mathbb{R} \to \mathbb{R}$ be thrice differentiable with $|f^{(k)}|_\infty \leq B_k$, for $k = 1, 2, 3$. Let $L_2 := C_2 B_1^2 + C_1 B_2$ and $L_3 := C_3 B_1^3 + C_1 B_3 + 3 C_2 B_2^2$. Finally, let $X_1, \ldots, X_n$ be i.i.d. random variables with third and fourth central moments m_3 and m_4 respectively, and define $U := g(X_1, \ldots, X_n)$ and $V = g(Y_1, \ldots, Y_n)$. Then*

$$\left| \mathbb{E}\left[f(U)\right] - \mathbb{E}\left[f(V)\right] \right| \leq \left(9.5 m_4^{1/2} L_2 + 13 m_3 L_3 \right) n^{-1/2}.$$

In the special case of the classical CLT we have $C_1 = 1$ and $C_2 = C_3 = 0$, since the function g is linear.

Theorems of this sort are called *invariance principles*, since they tell us that a certain function of random variables is (approximately) invariant under changes in the detailed distribution of the components.

4.3. **Tail bounds.** Chatterjee's result tells us that for a large class of functions $g : \mathbb{R}^n \to \mathbb{R}$, and a large class of distributions, if we have i.i.d. random variables $X_1, \ldots, X_n$, then for $Y = g(X_1, \ldots, X_n)$

$$Y \approx \mathbb{E}[Y] + Z\sqrt{\mathrm{Var}(Y)},$$

where Z is a standard normal random variable. (Those g for which the dominant contribution is linear.) That means that if $\mu = \mathbb{E}[Y]$ and $\sigma^2 = \mathrm{Var}(Y)$,

$$\mathbb{P}\{Y > \mu + y\} \approx 1 - \Phi(y/\sigma).$$

Since large z we have

$$1 - \Phi(z) \approx \frac{1}{\sqrt{2\pi}z} \mathrm{e}^{-z^2/2},$$

so we may suppose that for y/σ large

$$\mathbb{P}\{Y > \mu + y\} \approx \frac{\sigma}{\sqrt{2\pi}y} \mathrm{e}^{-y^2/2\sigma^2}.$$

But this depends on what we mean by "large". The errors in the CLT — or in Chatterjee's theorem — are small when n is large for each fixed value of y.

Suppose we let $Y_n = n^{-1/2}(X_1 + \cdots + X_n)$, where X_i are i.i.d. with mean 0 and variance σ^2. Since Y_n is approximately normal with mean 0 and variance σ^2 for large n we might expect a bound like

$$\mathbb{P}\{Y_n > zn^\alpha\} \leq C\mathrm{e}^{-n^{2\alpha}z^2/2\sigma^2}.$$

It turns out that this is true, but only under further assumptions on the tails of X_i.

The simplest tail bounds come from applications of Markov's inequality. Let Y be any real-valued random variable. If you have a bound on $\mathbb{P}\phi(Y)$, where ϕ is any positive increasing function, then you automatically have a bound on the tail probabilities of Y. The faster ϕ grows, the faster the tail probabilities fall, since we have

$$\mathbb{P}\{Y \geq y\} = \mathbb{P}\left\{\phi(Y) \geq \phi(y)\right\} \leq \frac{\mathbb{P}\phi(Y)}{\phi(y)}.$$

Thus, if we have a bound on the moment generating function $M_Y(\lambda) = \mathbb{E}[\mathrm{e}^{\lambda Y}]$, we have the exponential tail bound known as *Chernoff's bound*.

$$\mathbb{P}\{Y \geq y\} \leq M_Y(\lambda)\mathrm{e}^{-\lambda y}.$$

Note that this is a whole family of inequality, so we can choose whichever value of λ makes this bound the smallest. We write $\psi_Y(\lambda) := -\log M_Y(\lambda)$, yielding

$$\log \mathbb{P}\{Y \geq y\} \leq -\sup_{\lambda > 0} \left(\lambda y - \psi_Y(\lambda)\right). \tag{12}$$

Definition 4.5. *We say a random variable Y is* sub-gaussian *with* variance factor ν *if for all $\lambda \geq 0$*

$$\psi_Y(\lambda) \leq \frac{\lambda^2 \nu}{2}$$

If Y is sub-gaussian with variance factor ν,

$$\log \mathbb{P}\{Y \geq y\} \leq -\sup_{\lambda>0}\left(\lambda y - \frac{\lambda^2\nu}{2}\right) = -\frac{y^2}{2\nu}. \tag{13}$$

Lemma 4.6. *If $X_1, \ldots, X_n$ are independent sub-gaussian random variables with variance factors $\nu_1, \ldots, \nu_n$, then $Y = k(X_1 + \cdots + X_n)$ is sub-gaussian with variance factor $k^2(\nu_1 + \cdots + \nu_n)$.*

Lemma 4.7. *If $a \leq X \leq b$ almost surely, and $\mathbb{E}[X] = 0$, then X is sub-gaussian with variance factor $(b-a)^2/4$.*

Proof. **From** [1]**:** For each $\lambda \geq 0$ define X_λ to be a random variable with density $\mathrm{e}^{-\psi_X(\lambda)+\lambda x}$ with respect to X. We have

$$\psi_X'(\lambda) = \mathrm{e}^{-\psi_X(\lambda)}\mathbb{E}\left[X\mathrm{e}^{\lambda X}\right],$$

so

$$\begin{aligned}
\psi_X''(\lambda) &= -\psi_X'(\lambda)\mathrm{e}^{-\psi_X(\lambda)}\mathbb{E}\left[X\mathrm{e}^{\lambda X}\right] + \mathrm{e}^{-\psi_X(\lambda)}\mathbb{E}\left[X^2\mathrm{e}^{\lambda X}\right] \\
&= -\mathrm{e}^{-2\psi_X(\lambda)}\mathbb{E}\left[X\mathrm{e}^{\lambda X}\right]^2 + \mathrm{e}^{-\psi_X(\lambda)}\mathbb{E}\left[X^2\mathrm{e}^{\lambda X}\right] \\
&= \mathrm{Var}(X_\lambda) \\
&\leq \frac{(b-a)^2}{4}
\end{aligned}$$

since $a \leq X_\lambda \leq b$. By Taylor's Theorem, for some $\theta \in [0, \lambda]$,

$$\psi_X(\lambda) = \psi_X(0) + \lambda\psi_X'(0) + \frac{\lambda^2}{2}\psi_X''(\theta) \leq \frac{\lambda^2(b-a)^2}{8}.$$

□

Combining this with the bound (13) yields immediately a sub-gaussian tail inequality for sums of bounded random variables.

Theorem 4.8 (Hoeffding's Inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables for which there are deterministic constants a_i, b_i such that $a_i \leq X_i \leq b_i$. Let $S = \sum_{i=1}^n (X_i - \mathbb{E}[X_i])$. Then for every $t > 0$,*

$$\mathbb{P}\{S \geq t\} \leq \exp\left\{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right\}.$$

In this proof we are taking the worst possible case for the variance, under the assumption of boundedness. It turns out, we can do better, even dispensing with the assumption of boundedness. We present **Bernstein's Inequality** without proof:

Theorem 4.9 (Bernstein's Inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables, and let c be a constant such that for all integers $q \geq 3$*

$$\mathbb{E}\left[(X_i)_+^q\right] \leq \frac{q!}{2}c^{q-2}\,\mathrm{Var}(X_i).$$

Let $S = \sum_{i=1}^n (X_i - \mathbb{E}[X_i])$. Then for every $t > 0$,

$$\mathbb{P}\left\{S \geq t\right\} \leq \exp\left\{-\frac{t^2}{2(\nu + ct)}\right\}.$$

If there is a bound b such that $X_i \leq b$ almost surely, then we may take $c = b/3$

4.4. **Maximal inequalities.** There are many situations where we are interested in bounding the tails of the maximum difference between two random functions. A paradigm example is the problem of convergence of *empirical processes*. Consider the problem of estimating a cdf F, given n i.i.d. observations $X_1, \ldots, X_n$ from that distribution. Clearly we would estimate $F(t)$ for any fixed t by

$$F_n(t) := n^{-1} \# \left\{i \,:\, X_i \leq t\right\}.$$

Since $F_n(t)$ is simply a binomial random variable with parameters $n, F(t)$, we know it has expected value $F(t)$ (so it is an unbiased estimator for $F(t)$), satisfies the Law of Large Numbers, and the Central Limit Theorem, so

$$\widetilde{F}_n(t) := \frac{F_n(t) - F(t)}{\sqrt{n}}$$

is approximately Gaussian with mean 0 and variance $F(t)\big(1 - F(t)\big)$. We have the bounds in distribution and tail bounds from sections 4.3 and 4.2. But it is possible, in principle, for each $F_n(t)$ to converge to $F(t)$, but for there always to be random exceptional points where $F_n(t) - F(t)$ is still large. We would like to have bounds on $\|F_n - F\| := \sup_{t \in \mathbb{R}} |F_n(t) - F(t)|$, and some kind of limit theorem for the entire function $\widetilde{F}_n(t)$. (Note for future reference that the supremum norm $\|\cdot\|$ is a convex function.)

Since $\|F_n - F\|$ is a maximum over infinitely many random variables, it is not obvious how we can get any control at all over its distribution. There are two basic general techniques: *symmetrisation* and *chaining*. Chaining depends on having control over the pointwise errors, as well as a bound on the difference in errors between points that are close together. We then find a sequence of *skeletons*, finite collections of points that cover the whole space to within 2^{-k}. The total error at any point t is bounded by the sums of differences in errors between the closest approximation in the k-th skeleton, for $k = 1$ to ∞. In this way, if we can control the number of points in the skeleton — which depends on the geometry of the space we are working in — we can get excellent bounds on the distribution of the maximum. The details may be found in [9] or [7].

In one dimension we can do quite a lot with symmetrisation alone, and we will focus on that here. We express $F_n(t)$ as a sum of indicator functions

$$n^{-1} \sum_{i=1}^n f_i(t), \quad \text{where } f_i(t) := \mathbf{1}\{X_i \leq t\}.$$

Given any positive random variable Y, we have the general formula

$$\mathbb{E}\left[\mathrm{e}^{\lambda Y}\right] = 1 + \int_0^\infty \lambda \mathrm{e}^{\lambda y} \mathbb{P}\left\{Y > y\right\} \mathrm{d}y.$$

Thus, if we have another random variable Y' and a constant $c \geq 1$ such that $\mathbb{P}\{Y' > y\} \leq c\mathbb{P}\{Y > y\}$ for all y,

$$\mathbb{E}\left[\mathrm{e}^{\lambda Y'}\right] \leq c\mathbb{E}\left[\mathrm{e}^{\lambda Y}\right].$$

In particular, this implies that if Y is sub-gaussian with variance factor ν, then Y' is sub-gaussian with variance factor $c\nu$.

We apply this to empirical processes by means of *Lévy's Reflection Principle*.

Lemma 4.10 (Reflection Principle)**.** *Let $G(t)$ be a random function on an interval $[a, b]$ (where a or b may be infinite) that is* martingale symmetric[2]*, in the sense that for any stopping time T, the distribution of $G\big|_{[T,b]} - G(T)$ conditioned on $G\big|_{[a,T]}$ is symmetric; that is, it is the same as the distribution of $G(T) - G\big|_{[T,b]}$. Then for any $y > 0$,*

$$\mathbb{P}\left\{\|G\| \geq y\right\} \leq 2\mathbb{P}\left\{\sup G(t) \geq y\right\} \leq 4\mathbb{P}\left\{G(b) \geq y\right\}.$$

Proof. Let T be the smallest t such that $G(t) \geq y$, or ∞ if no such t exists. Then for any $t < \infty$,

$$\mathbb{P}\left\{G(b) \geq y \,|\, T = t\right\} \geq \mathbb{P}\left\{G(b) > G(t) \,|\, T = t\right\} = \frac{1}{2}.$$

Thus

$$\mathbb{P}\left\{G(b) \geq y\right\} \geq \mathbb{P}\left\{G(b) \geq y \text{ and } T < \infty\right\} \geq \frac{1}{2}\mathbb{P}\left\{T < \infty\right\} = \frac{1}{2}\mathbb{P}\left\{\sup G(t) \geq y\right\}.$$

□

The classic application is to symmetric random walk, or Brownian motion. The empirical processes are not symmetric, so we need to symmetrise them. We define a sequence of *Rademacher variables* ρ_i, i.i.d. taking on the values ± 1 with probability $\frac{1}{2}$, and assumed independent of the X_i. Let

$$\widetilde{G}_n(t) := n^{-1/2} \sum_{i=1}^n \rho_i f_i(t).$$

Then $\widetilde{G}_n$ is a function to which we can apply the Reflection Principle. Since $f_i(\infty) = 1$ for all i, by Lemma 4.7,

$$\mathbb{E}_\rho\left[\exp\left\{\lambda\left\|\widetilde{G}_n\right\|\right\}\right] \leq \mathbb{E}_\rho\left[\exp\left\{\lambda n^{-1/2}\sum_{i=1}^n \rho_i\right\}\right] \leq \mathrm{e}^{\lambda^2/2} \tag{14}$$

[2]This is not standard terminology.

for any choice of the step functions f_i.

Theorem 4.11. *For all $z > 0$,*

$$\mathbb{P}\left\{\sup_{t\in\mathbb{R}} \left|F_n(t) - F(t)\right| \geq zn^{-1/2}\right\} \leq 2\mathrm{e}^{-z^2/8} \tag{15}$$

Proof. Let $X_1', \ldots, X_n'$ be an independent copy of $X_1, \ldots, X_n$, and let f_i' be the corresponding indicators. We also write $\mathbb{E}'$ for the expected value with respect to the X' variables. Then

$$\begin{aligned}
\mathbb{E}\left[\mathrm{e}^{\lambda\|\widetilde{F}_n\|}\right] &= \mathbb{E}\left[\exp\left\{\lambda n^{-1/2}\Big\|\sum_{i=1}^n \big(f_i(t) - F(t)\big)\Big\|\right\}\right] \\
&\leq \mathbb{E}\left[\exp\left\{\lambda n^{-1/2}\Big\|\mathbb{E}'\sum_{i=1}^n \big(f_i(t) - f_i'(t)\big)\Big\|\right\}\right] \\
&\leq \mathbb{E}\left[\exp\left\{\lambda n^{-1/2}\Big\|\mathbb{E}'\mathbb{E}_\rho\sum_{i=1}^n \rho_i\big(f_i(t) - f_i'(t)\big)\Big\|\right\}\right] \\
&\leq \mathbb{E}\mathbb{E}'\mathbb{E}_\rho\left[\exp\left\{\lambda n^{-1/2}\Big\|\sum_{i=1}^n \rho_i\big(f_i(t) - f_i'(t)\big)\Big\|\right\}\right]
\end{aligned}$$

by Jensen's Inequality, since the exponential and the supremum are both convex functions. Then

$$\begin{aligned}
\mathbb{E}\left[\mathrm{e}^{\lambda\|\widetilde{F}_n\|}\right] &\leq \mathbb{E}\mathbb{E}'\mathbb{E}_\rho\left[\exp\left\{\lambda n^{-1/2}\Big\|\sum_{i=1}^n \rho_i f_i(t)\Big\| + \lambda n^{-1/2}\Big\|\sum_{i=1}^n -\rho_i f_i'(t)\Big\|\right\}\right] \\
&\leq 2\mathbb{E}\left[\mathbb{E}_\rho\left[\exp\left\{2\lambda\Big\|\widetilde{G}_n\Big\|\right\}\right]^{1/2}\right]^2 \text{ by Cauchy–Schwartz} \\
&\leq 2\mathrm{e}^{-2\lambda^2}.
\end{aligned}$$

(15) follows then by the same calculation as in (13). □

References

[1] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[2] Sourav Chatterjee. A generalization of the Lindeberg principle. *Annals of Probability*, 34(6):2061–2076, 2006.

[3] Rick Durrett. *Probability: theory and examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, fourth edition, 2010.

[4] Peter Hall and Christopher C. Heyde. *Martingale Limit Theory and its Application*. Academic Press, New York, London, 1980.
[5] I. A. Ibragimov. Some limit theorems for stationary processes. *Theory of Probability and its Applications*, 7:349–82, 1962.
[6] Michael Keane and Karl Petersen. Easy and nearly simultaneous proofs of the ergodic theorem and maximal ergodic theorem. *Lecture Notes–Monograph Series: Dynamics & Stochastics*, 48:248–51, 2006.
[7] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces*. Springer-Verlag, New York, Heidelberg, Berlin, 1991.
[8] David Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, New York, Heidelberg, Berlin, 1984.
[9] David Pollard. *Empirical Processes: Theory and Applications*, volume 2 of *CBMS-NSF Regional Conference Series in Probability and Statistics*. Institute of Mathematical, Hayward, California, 1990.
[10] Jeffrey S. Rosenthal. *A First Look at Rigorous Probability Theory*. World Scientific, second edition, 2006.
[11] Gennady Samorodnitsky. *Stochastic processes and long range dependence*. SV, 2016.
[12] Daniel W. Stroock. *Probability theory, an analytic view*. Cambridge University Press, Cambridge, 1993.
[13] David Williams. *Probability with Martingales*. Cambridge University Press, 1991.