REVIEW: MATH 546–PROBABILITY

SHANSHAN DING

Fall 2009

Contents

1. Introduction

Definition 1.1. *A* **σ-field** *on a space Ω is a collection $\mathcal{F}$ of subsets of Ω such that*

- $\emptyset, \Omega \in \mathcal{F}$,
- *If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$, and*
- *If a countable sequence $\{A_n\}$ is in $\mathcal{F}$, then so is $\cup_{n=1}^{\infty} A_n$.*

Remark. If $\Omega = \mathbb{R}$, then we may take $\mathcal{F}$ to be the collection $\mathcal{B}$ of Borel sets on $\mathbb{R}$, which is the σ-field generated by open subsets of $\mathbb{R}$.

Definition 1.2. *A* **measure** *on $(\Omega, \mathcal{F})$ is a function μ on $\mathcal{F}$ such that*

- *For all $A \in \mathcal{F}$, $\mu(A) \geq 0 = \mu(\emptyset)$ (non-negativity and null empty set), and*
- *For any sequence $\{A_i : i \geq 1\}$ of disjoint sets in $\mathcal{F}$, $\mu(\cup_i A_i) = \sum_i \mu(A_i)$ (countable additivity).*

If, in addition, $\mu(\Omega) = 1$, then μ is a **probability measure**.

Definition 1.3. *The pair $(\Omega, \mathcal{F})$ is called a* **measurable space**, *the triple $(\Omega, \mathcal{F}, \mu)$ is called a* **measure space**, *and members of $\mathcal{F}$ are called* **measurable sets**. *A function $f : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$ is* **measurable** *if $f^{-1}(A) \in \mathcal{F}$ for every $A \in \mathcal{S}$.*

Definition 1.4. *A* **probability space** *is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where*

- *Ω is a space (a set), sometimes called the universe,*
- *$\mathcal{F}$ is a σ-field on Ω, meaning a collection of events, and*
- *$\mathbb{P} : \mathcal{F} \to [0, 1]$ is a probability measure on $\mathcal{F}$.*

Definition 1.5. *An* **event** *is a measurable subset of Ω, that is, an element of $\mathcal{F}$. The* **indicator function $\mathbf{1}_A$** *of an event A is the function defined on Ω as $\mathbf{1}_A(\omega) = 1$ if $\omega \in A$ and $\mathbf{1}_A(\omega) = 0$ otherwise.*

Definition 1.6. *A (real-valued)* **random variable** *X is a measurable map from $(\Omega, \mathcal{F}, \mathbb{P})$ to $(\mathbb{R}, \mathcal{B})$. The* **distribution** *or* **law** *of X is the measure on $\mathbb{R}$ which is the push-forward of $\mathbb{P}$ by X. This means that the law of X is the measure μ given by $\mu(A) = \mathbb{P}(X^{-1}(A))$.*

The law of X completely determines its **distribution function** $F(x) = \mathbb{P}(X \leq x)$ (we use the shorthand $\mathbb{P}(X \in A)$ for $\mathbb{P}(\omega \in \Omega : X(\omega) \in A)$. F is non-decreasing, has limits 0 and 1 at $-\infty$ and $+\infty$ respectively, and is right-continuous. If F is continuous, then we say that the corresponding law is **non-atomic**. If F is absolutely continuous, then F is differentiable almost everywhere, in which case it makes sense to speak of $F'(x)$.

Definition 1.7. *If F is absolutely continuous and $F'(x) = g(x)$ a.e., then we say that μ has* **density** *g. Here*

$$\mu((a, b]) = F(b) - F(a) = \int_a^b g(x)\, dx, \tag{1.1}$$

and more generally

$$\mu(A) = \int_A g(x)\, dx = \int \mathbf{1}_A g(x)\, dx. \tag{1.2}$$

Remark. A distribution F has a density function if and only if F is absolutely continuous, so in particular discrete distributions do not have density functions.

Definition 1.8. *If X is a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, then its* **expected value** *is defined to be $\mathbb{E}X = \int X \, d\mathbb{P}$.*

That this integral is well-defined follows from the standard construction of the Lebesgue integral. We say that $\mathbb{E}X = \mathbb{E}X^+ - \mathbb{E}X^-$ *exists* if either $\mathbb{E}X^+$ or $\mathbb{E}X^-$ is finite. The following is a series of classic real analytic results, stated in the language of expectations:

Theorem 1.9 (Fatou's lemma). *If $X_n \geq 0$, then $\mathbb{E}(\liminf_{n\to\infty} X_n) \leq \liminf_{n\to\infty} \mathbb{E}(X_n)$.*

Theorem 1.10 (Monotone convergence). *If $0 \leq X_n \uparrow X$, then $\mathbb{E}X_n \uparrow \mathbb{E}X$.*

Theorem 1.11 (Lebesgue dominated convergence). *If $X_n \to X$ a.s., $|X_n| \leq Y$ for all n, and $\mathbb{E}Y < \infty$, then $\mathbb{E}X_n \to \mathbb{E}X$.*

Remark. If Y is constant, this is sometimes known as the *bounded convergence theorem.*

Theorem 1.12 (Fubini). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the n-fold product of $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$, and suppose $f \leq 0$ or $\int |f| \, d\mathbb{P} < \infty$, then*

$$\int_\Omega f \, d\mathbb{P} = \int_{\Omega_1} \cdots \left(\int_{\Omega_n} f \, d\mathbb{P}_n \right) \cdots d\mathbb{P}_1, \tag{1.3}$$

and in fact we can iterate in any order.

Theorem 1.13 (Markov's inequality). *Let $A \in \mathbb{R}$. For any non-negative function f,*

$$\mathbb{P}(X \in A) \leq \frac{\mathbb{E}f(X)}{\inf_{x\in A} f(x)}. \tag{1.4}$$

Proof. Since $f \geq 0$, we have that $\mathbb{E}f(X) \geq \mathbb{E}[f(X)\mathbf{1}_A(X)] \geq \mathbb{E}[\inf_{x\in A} f(x)\mathbf{1}_A(X)]$. As $\mathbb{P}(X \in A) = \mathbb{E}\mathbf{1}_A(X)$, the result follows if we divide both ends by $\inf_{x\in A} f(x)$. □

Theorem 1.14 (Chebyshev's inequality). *As a special case of Markov's inequality with $A = \{x : |x - b| \geq a\}$ and $f(x) = (x - b)^2$,*

$$\mathbb{P}(|X - b| \geq a) \leq \frac{\mathbb{E}(X - b)^2}{a^2}. \tag{1.5}$$

Theorem 1.15 (Jensen's inequality). *Let f be a convex function, then $f(\mathbb{E}X) \leq \mathbb{E}f(X)$, provided that both expectations exist.*

Remark. A useful case is when $f(x) = |x|$, which gives that $|\mathbb{E}X| \leq \mathbb{E}|X|$.

Proof. Let l be the tangent line to f at $\mathbb{E}X$, then by the convexity of f, we see that $\mathbb{E}f(X) \geq \mathbb{E}l(X) = l(\mathbb{E}X) = f(\mathbb{E}X)$. □

Theorem 1.16 (Cauchy-Schwarz). $\mathbb{E}|XY| \leq (\mathbb{E}X^2\mathbb{E}Y^2)^{1/2}$*; in particular, if $Y \equiv 1$, then* $\mathbb{E}|X| \leq (\mathbb{E}X^2)^{1/2}$.

Remark. Combining the latter with Jensen's inequality yields the more familiar-looking $(\mathbb{E}X)^2 \leq \mathbb{E}X^2$.

Definition 1.17. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space.*

- *Events A and B are* **independent** *if $\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$. More generally, a collection of events is* **mutually independent** *if, for any finite subset $A_1, \ldots, A_n$ of the collection, $\mathbb{P}(\cap_{i=1}^n A_i) = \prod_{i=1}^n \mathbb{P}(A_i)$.*

- *Random variables X and Y are* **independent** *if, for all measurable (i.e. Borel) subsets C, D of $\mathbb{R}$, the events $\{X \in C\}$ and $\{Y \in D\}$ are independent.*
- *Let $\mathcal{A}$ and $\mathcal{B}$ be sub-σ-fields of $\mathcal{F}$. Then $\mathcal{A}$ and $\mathcal{B}$ are* **independent** *if, whenever $A \in \mathcal{A}$ and $B \in \mathcal{B}$, the events A and B are independent.*

Mutually independent random variables and σ-fields are defined analogously.

Remark. Pairwise independence is weaker than mutual independence: consider

$$(X, Y, Z) = \begin{cases} (0,0,0) & \text{with probability } 1/4, \\ (0,1,1) & \text{with probability } 1/4, \\ (1,0,1) & \text{with probbaility } 1/4, \\ (1,1,0) & \text{with probability } 1/4. \end{cases}$$

Remark. If $\mathbb{E}(XY) = (\mathbb{E}X)(\mathbb{E}Y)$, we say that X and Y are **uncorrelated**. This is a weaker condition than independence: consider $X = \chi$ and $Y = X^2$. Uncorrelatedness is by definition a pairwise property.

2. Weak Law of Large Numbers

Definition 2.1. *X_n* **converges** *to X* **in probability** *if $\mathbb{P}(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$ for all $\epsilon > 0$. X_n* **converges** *to X* **almost surely** *if $\mathbb{P}(X_n \to X \text{ as } n \to \infty) = 1$.*

Remark. Almost sure convergence implies convergence in probability, which in turn implies convergence in distribution (to be defined later).

Definition 2.2. *X_n* **converges** *to X* **in** L^p *if $\mathbb{E}|X_n - X|^p \to 0$ as $n \to \infty$.*

Definition 2.3. *If $\mathbb{E}X^2 < \infty$, then the* **variance** *of X is defined to be*

$$\operatorname{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}X^2 - (\mathbb{E}X)^2. \tag{2.1}$$

It is easy to see that $\operatorname{Var}(aX + b) = a^2 \operatorname{Var}(X)$. Furthermore, if $X_1, \dots, X_n$ are uncorrelated random variables such that $\mathbb{E}(X_i^2) < \infty$, then $\operatorname{Var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \operatorname{Var}(X_i)$.

Lemma 2.4. *If $p > 0$ and $\mathbb{E}|Z_n|^p \to 0$, then $Z_n \to 0$ in probability.*

Proof. Follows from Markov's inequality by setting $f(x) = |x|^p$ and $A = \{x : |x| \geq \epsilon\}$. □

Theorem 2.5 (L^2 weak law)**.** *Let X_1, X_2, … be uncorrelated random variables with $\mathbb{E}X_i = \mu$ and $\operatorname{Var}(X_i) \leq C < \infty$. If $S_n = \sum_{i=1}^n X_i$, then $S_n/n \to \mu$ in L^2 and in probability.*

Proof. To see the convergence in L^2, observe that $\mathbb{E}(S_n/n) = \mu$, so that

$$\mathbb{E}(S_n/n - \mu)^2 = \operatorname{Var}(S_n/n) = \frac{1}{n^2}(\operatorname{Var}(X_1)) + \cdots + \operatorname{Var}(X_n)) \leq \frac{Cn}{n^2} \to 0. \tag{2.2}$$

To see the convergence in probability, apply Lemma 2.4 to $Z_n = S_n/n - \mu$. □

Theorem 2.6 (Weak law for triangular arrays)**.** *For each n, let $X_{n,k}, 1 \leq k \leq n$, be independent. Let $b_n > 0$ with $b_n \to \infty$, and let $\bar{X}_{n,k} = X_{n,k}\mathbf{1}_{|X_{n,k}| \leq b_n}$. Suppose that as $n \to \infty$,*

$$\sum_{k=1}^n \mathbb{P}(|X_{n,k}| > b_n) \to 0 \ \text{ and } \ b_n^{-2} \sum_{k=1}^n \mathbb{E}\bar{X}_{n,k}^2 \to 0. \tag{2.3}$$

If we let $S_n = \sum_{k=1}^n X_{n,k}$ *and put* $a_n = \sum_{k=1}^n \mathbb{E}\bar{X}_{n,k}$*, then* $(S_n - a_n)/b_n \to 0$ *in probability.*

Theorem 2.7. *Let* $X_1, X_2, \dots$ *be IID with* $x\mathbb{P}(|X_i| > x) \to 0$ *as* $x \to \infty$*, and let* $\mu_n = \mathbb{E}(X_1 \mathbf{1}_{|X_1| \le n})$*. Then* $S_n/n - \mu_n \to 0$ *in probability.*

Remark. The proof of the theorem involves showing that the conditions of (2.3) are satisfied (with $X_{n,k} = X_k$ and $b_n = n$).

Theorem 2.8 (Weak law of large numbers)**.** *Let* $X_1, X_2, \dots,$ *be IID random variables with* $\mathbb{E}|X_i| < \infty$ *and* $\mathbb{E}X_i = \mu$*. Then* $S_n/n \to \mu$ *in probability.*

Proof. Observe that

$$x\mathbb{P}(|X_1| > x) = x\mathbb{E}(\mathbf{1}_{|X_1|>x}) \le \mathbb{E}(|X_1|\mathbf{1}_{|X_1|>x}), \tag{2.4}$$

which by Lebesgue dominated convergence goes to zero as $x \to \infty$. By Theorem 2.7, $\mathbb{P}(|S_n/n - \mu_n| > \epsilon/2) \to 0$. Lebesgue dominated convergence also shows that

$$\mu_n = \mathbb{E}(X_1\mathbf{1}_{|X_1|\le n}) \to \mathbb{E}X_1 = \mu \tag{2.5}$$

as $n \to \infty$, from which it follows that $\mathbb{P}(|S_n/n - \mu| > \epsilon) \to 0$. □

3. Borel-Cantelli Lemmas and Strong Law of Large Numbers

If A_n is a sequence of subsets of Ω, we define

$$\limsup A_n = \lim_{m\to\infty} \cup_{n=m}^\infty A_n = \{\omega : \omega \in A_n \text{ i.o.}\}. \tag{3.1}$$

For instance, $X_n \to 0$ a.s. if and only if for all $\epsilon > 0$, $\mathbb{P}(|X_n| > \epsilon \text{ i.o.}) = 0$.

Theorem 3.1 (Borel-Cantelli I)**.** *If* $\sum \mathbb{P}(A_n) < \infty$*, then* $\mathbb{P}(A_n \text{ i.o.}) = 0$.

Proof. Let $N = \sum \mathbf{1}_{A_n}$, then

$$\mathbb{E}(N) = \mathbb{E}(\sum \mathbf{1}_{A_n}) = \sum \mathbb{E}\mathbf{1}_{A_n} = \sum \mathbb{P}(A_n) < \infty, \tag{3.2}$$

so that $\mathbb{P}(N = \infty) = 0$. □

Theorem 3.2 (Borel-Cantelli II)**.** *If* $\{A_n\}$ *are independent, then* $\sum \mathbb{P}(A_n) = \infty$ *implies that* $\mathbb{P}(A_n \text{ i.o}) = 1$.

Proof. By independence and the inequality $1 - x \le e^{-x}$, for every M,

$$\begin{aligned}\mathbb{P}(\cap_{n=M}^N A_n^c) = \prod_{n=M}^N (1 - \mathbb{P}(A_n)) &\le \prod_{n=M}^N \exp(-\mathbb{P}(A_n)) \\ &= \exp\left(-\sum_{n=M}^N \mathbb{P}(A_n)\right) \to 0\end{aligned} \tag{3.3}$$

as $N \to \infty$. Thus $\mathbb{P}(\cup_{n=M}^\infty A_n) = 1$ for all M. Since $\cup_{n=M}^\infty A_n \downarrow \limsup A_n = \{A_n \text{ i.o.}\}$, we see that $\mathbb{P}(A_n \text{ i.o.}) = 1$. □

Example. Let $X_1, X_2, \dots$ be independent random variables with $\mathbb{P}(X_n = 0) = 1 - \frac{1}{n}$ and $\mathbb{P}(X_n = 1) = \frac{1}{n}$. Then X_n converges in probability to zero. However, by Borel-Cantelli II, since $\sum \mathbb{P}(X_n = 1) = \sum \frac{1}{n}$ diverges, $\mathbb{P}(X_n = 1 \text{ i.o.}) = 1$, and therefore X_n does not converge a.s. to zero.

Theorem 3.3 (Borel-Cantelli II, extended version)**.** *If $\{A_n\}$ are pairwise independent and $\sum \mathbb{P}(A_n) = \infty$, then*

$$\frac{\sum_{k=1}^n \mathbf{1}_{A_k}}{\sum_{k=1}^n \mathbb{P}(A_k)} \to 1 \tag{3.4}$$

almost surely as $n \to \infty$.

Proof. Let $N_n = \sum_{k=1}^n \mathbf{1}_{A_k}$ count how many of the first n events occur. By pairwise independence,

$$\mathrm{Var}(N_n) = \sum_{k=1}^n \mathrm{Var}(\mathbf{1}_{A_k}) \le \sum_{k=1}^n \mathbb{E}\mathbf{1}_{A_k}^2 = \sum_{k=1}^n \mathbb{E}\mathbf{1}_{A_k} = \mathbb{E}N_n. \tag{3.5}$$

Chebyshev's inequality shows that for any $\delta > 0$,

$$\mathbb{P}(|N_n - \mathbb{E}N_n| > \delta\mathbb{E}N_n) \le \frac{\mathrm{Var}(N_n)}{(\delta\mathbb{E}N_n)^2} \le \delta^{-2}\frac{\mathbb{E}N_n}{(\mathbb{E}N_n)^2}, \tag{3.6}$$

which tends to zero as $n \to \infty$ because $\sum \mathbb{P}(A_n) = \infty$ implies $\mathbb{E}N_n \to \infty$.

The last computation shows that $N_n/\mathbb{E}N_n \to 1$ in probability. To get almost sure convergence, we have to take subsequences. Let $n_j = \inf\{n : \mathbb{E}N_n \ge j^2\}$, and let $T_j = N_{n_j}$. Observe that since $\mathbb{E}\mathbf{1}_{A_k} \le 1$, $j^2 \le \mathbb{E}T_j \le j^2 + 1$, so by (3.6),

$$\mathbb{P}(|T_j - \mathbb{E}T_j| > \delta\mathbb{E}T_j) \le 1/(\delta^2 j^2), \tag{3.7}$$

and hence

$$\sum_{j=0}^\infty \mathbb{P}(|T_j - \mathbb{E}T_j| > \delta\mathbb{E}T_j) < \infty. \tag{3.8}$$

Borel-Cantelli I then shows that $\mathbb{P}(|T_j - \mathbb{E}T_j| > \delta\mathbb{E}T_j \text{ i.o}) = 0$, and since δ is arbitrary, we see that $T_j/\mathbb{E}T_j \to 1$ almost surely. To show that $N_n/\mathbb{E}N_n \to 1$ a.s., note that if $n_j \le n \le n_{j+1}$, then by monotonicity of N_n and $\mathbb{E}N_n$,

$$\frac{T_j}{\mathbb{E}T_{j+1}} \le \frac{N_n}{\mathbb{E}(N_n)} \le \frac{T_{j+1}}{\mathbb{E}T_j}. \tag{3.9}$$

Now write

$$\frac{\mathbb{E}T_{j+1}}{\mathbb{E}T_j} \cdot \frac{T_j}{\mathbb{E}T_{j+1}} \le \frac{N_n}{\mathbb{E}(N_n)} \le \frac{T_{j+1}}{\mathbb{E}T_j} \cdot \frac{\mathbb{E}T_j}{\mathbb{E}T_{j+1}}, \tag{3.10}$$

and since $j^2 \le \mathbb{E}T_j \le \mathbb{E}T_{j+1} \le (j+1)^2 + 1$, $\mathbb{E}T_j/\mathbb{E}T_{j+1} \to 1$, which forces $N_n/\mathbb{E}(N_n) \to 1$ whenever $T_j/\mathbb{E}T_j \to 1$, i.e. almost surely. $\square$

Theorem 3.4 (Strong law of large numbers)**.** *Let $X_1, X_2, \ldots,$ be IID random variables with $\mathbb{E}|X_i| < \infty$ and $\mathbb{E}X_i = \mu$. Then $S_n/n \to \mu$ almost surely.*

Remark. The SLLN holds even if we replace mutual independence of the X_i by pairwise independence.

Theorem 3.5 (Glivenko-Cantelli)**.** *Let $\{X_i\}$ be IID with distribution F, and let*

$$F_n(x) = \frac{1}{n}\sum_{k=1}^n \mathbf{1}_{X_k \le x} \tag{3.11}$$

be the empirical distribution functions. *Then* $\sup_x|F_n(x) - F(x)| \to 0$ *a.s. as $n \to \infty$.*

Proof. Fix x, and let $N_n(x)$ count the number of $X_1, \ldots, X_n$ that are in $(-\infty, x]$. Then $F_n = N_n/n$, so by the SLLN $F_n(x) \to \mathbb{E}\mathbf{1}_{X_1 \le x} = \mathbb{P}(X_1 \le x) = F(x)$ a.s. Now if F is continuous, then the theorem follows from the fact that a sequence of non-decreasing functions converging pointwise to a bounded continuous limit converges uniformly. Otherwise, let $F(x-) = \mathbb{P}(X_1 < x)$, and let

$$F_n(x-) = \frac{1}{n}\sum_{k=1}^{n} \mathbf{1}_{X_k < x}. \tag{3.12}$$

Then $F_n(x-) \to F(x-)$ pointwise, which, along with the monotonicity of F_n and F , implies uniform convergence. □

Let $\mathcal{F}_n' = \sigma(X_n, X_{n+1}, \ldots)$, the smallest σ-field with respect to which all X_m such that $m \ge n$ are measurable, i.e. the future after time n. Define the remote future, or **tail σ-field**, to be $\mathcal{T} = \cap_n \mathcal{F}_n'$. Intuitively, $A \in \mathcal{T}$ if and only if changing a finite number of values does not affect the occurrence of the event. For instance, Let $S_n = X_1 + \ldots + X_n$ as usual, then $\{\lim_{n\to\infty} S_n \text{ exists}\} \in \mathcal{T}$ whereas $\{\limsup_{n\to\infty} S_n > 0\} \notin \mathcal{T}$.

Theorem 3.6 (Kolmogorov's 0-1 law)**.** *If $X_1, X_2, \ldots$ are independent and $A \in \mathcal{T}$, then $\mathbb{P}(A) = 0$ or 1.*

Remark. Since $\{A_n \text{ i.o}\} \in \mathcal{T}$, if $A_1, A_2, \ldots$ are independent, then $\mathbf{1}_{A_1}, \mathbf{1}_{A_2}, \ldots$ are independent, so that $\mathbb{P}(A_n \text{ i.o}) = 0$ or 1.

Theorem 3.7 (Kolmogorov's maximal inequality)**.** *Suppose $X_1, \ldots, X_n$ are independent with $\mathbb{E}X_i = 0$ and $\mathrm{Var}(X_i) < \infty$, then $\mathbb{P}(\max_{1\le k\le n}|S_k| \ge x) \le x^{-2}\mathrm{Var}(S_n)$.*

Remark. In comparison, under the same hypotheses, Chebyshev's inequality gives only $\mathbb{P}(|S_n| \ge x) \le x^{-2}\mathrm{Var}(S_n)$.

Theorem 3.8 (Kolmogorov's three-series)**.** *Let $X_1, X_2, \ldots$ be independent. Let $A > 0$ and let $Y_i = X_i \mathbf{1}_{|X_i| \le A}$. Then $\sum_{n=1}^{\infty} X_n$ converges a.s. if and only if*

(1) $\sum_{n=1}^{\infty} \mathbb{P}(|X_n| > A) < \infty$,
(2) $\sum_{n=1}^{\infty} \mathbb{E}Y_n$ converges, and
(3) $\sum_{n=1}^{\infty} \mathrm{Var}(Y_n) < \infty$.

4. Large Deviations

The SLLN tells us that $\mathbb{P}(S_n > an) \to 0$ for any $a > \mathbb{E}X_1$. Large deviations theory concerns the rate at which this goes to zero.

For any fixed λ, Markov's inequality with $A = \{x : x > an\}$ and $f(x) = e^{\lambda x}$ gives us that $\mathbb{P}(S_n > an) \le e^{-\lambda an}\mathbb{E}e^{\lambda S_n}$. Taking logs and dividing by n, we get that

$$\frac{1}{n}\log \mathbb{P}(S_n > an) \le -\lambda a + \psi(\lambda), \tag{4.1}$$

where $\psi(\lambda) = \log \phi(\lambda) = \log \mathbb{E}e^{\lambda X_1}$.

Suppose that we have found the optimal λ which minimizes $\psi(\lambda) - a\lambda$; call it $\lambda_0(a)$. Define the rate function $\beta \le 0$ by $\beta(a) = \psi(\lambda_0(a)) - a\lambda_0(a) = \inf_\lambda(\psi(\lambda) - a\lambda)$.

Theorem 4.1 (Chernoff)**.** *The function ψ is convex. On the interior of the interval J on which ψ is finite, ψ' exists and the image under ψ' of the interior of J is some (possibily infinite) open interval U. If a is in the interior of U, then there is a unique λ_0 with $\psi'(\lambda_0) = a$, and this is also the unique λ_0 minimizing $\psi(\lambda) - \lambda a$. For this value of λ_0, we have the bounds*

$$\frac{1}{n} \log \mathbb{P}(S_n > an) \leq \beta(a) \text{ and} \tag{4.2}$$

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n > an) = \beta(a). \tag{4.3}$$

We essentially have already demonstrated the upper bound (4.2) using Markov's inequality. After a brief discussion of Radon-Nikodym derivatives, we will prove the lower bound (4.3) via the tilted distribution.

Theorem 4.2 (Radon-Nikodym)**.** *Let π and ν be σ-finite measures on $(\Omega, \mathcal{F})$. If ν is absolutely continuous with respect to π, then there is a function g such that $\nu(A) = \int_A g \, d\pi$ for all $A \in \mathcal{F}$. We say that ν has* **Radon-Nikodym derivative** *g with respect to π on $(\Omega, \mathcal{F})$ and write $d\nu/d\pi = g$.*

Remark. If π is supported on a finite set $\{x_1, \ldots, x_n\}$, then so is ν and the Radon-Nikodym derivative is just the likelihood ratio $\frac{d\nu}{d\pi}(x_i) = \frac{\nu(x_i)}{\pi(x_i)}$.

When $\Omega = \mathbb{R}$, the cumulative distribution functions F and G of π and ν are related by $G(x) = \int_{-\infty}^x g(t) \, dF(t)$.

Proposition 4.3. *If $g > 0$ and we construct a measure ν to have $d\nu/d\pi = g$ as above, then we also have that $d\pi/d\nu = 1/g$.*

If we try to set $g(x) = e^{\lambda x}$ and let π be the the law of X_1, then after normalizing, $G(x) = \frac{1}{\phi(\lambda)} \int_{-\infty}^x e^{\lambda t} \, dF(t)$–where $\phi(\lambda) = \int_{-\infty}^\infty e^{\lambda t} \, dF(t)$–defines a probability CDF. This is the so-called **tilted distribution**, which we denote as ν_λ. Moreover, we denote the likilihood ratio $e^{\lambda x}/\phi(\lambda)$ as $g_\lambda(x)$. It can be shown that

$$\frac{d\nu^n}{d\pi^n} = \prod_{i=1}^n g_\lambda(x_i) = \frac{e^{\lambda g(x_1, \ldots, x_n)}}{(\phi(\lambda))^n}. \tag{4.4}$$

Proof of upper bound. We need to show that for all $\epsilon > 0$ and $a \in U$,

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n > an) \geq \beta(a) - \epsilon. \tag{4.5}$$

Suppose $a' \in U$ with $a < a' < a + \epsilon/4$. Let $\lambda' = \lambda_0(a')$, and apply the WLLN to $\nu_{\lambda'}$ to see that $\nu_{\lambda'}^n(\{an < S_n < (a+\epsilon/2)n\}) \to 1$ as $n \to \infty$. Now $d\pi^n/d\nu_{\lambda'}^n = (\phi(\lambda'))^n e^{-\lambda' S_n}$, the inf of which on the event $\{an < S_n < (a+\epsilon/2)n\}$ is at least $\exp[n(\psi(\lambda') - (a+\epsilon/2)\lambda')]$, so that

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n > an) \geq \psi(\lambda') - (a+\epsilon/2)\lambda'. \tag{4.6}$$

As $a' \downarrow a$, the RHS converges to $\beta(a) - \epsilon$. $\square$

Theorem 4.4. *Assume that* $\phi(\lambda) = \mathbb{E}e^{\lambda X_1} < \infty$ *for some* $\lambda > 0$*, that the distribution of* X_1 *does not have a point mass at* $\mathbb{E}X_1$*, and that there is a* $\lambda_0(a) \in (0, \sup\{\lambda : \phi(\lambda) < \infty\})$ *such that* $a = \phi'(\lambda_0(a))/\phi(\lambda_0(a))$*. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(S_n > an) = \psi(\lambda_0(a)) - a\lambda_0(a). \tag{4.7}$$

5. Convergence in Distribution

Let Ω be a topological vector space with Borel σ-field $\mathcal{F}$. Let $BC(\Omega)$ be the space of bounded continuous functions $f : \Omega \to \mathbb{R}$. Denote the space of finite measures on $(\Omega, \mathcal{F})$ by $\mathcal{M}(\Omega, \mathcal{F})$ and the subspace of probability measures by $\mathcal{P}(\Omega, \mathcal{F})$. View $f \in BC(\Omega)$ as an element of the dual space $\mathcal{M}(\Omega, \mathcal{F})^*$ via $f(\mu) = \int f\, d\mu$ for $\mu \in \mathcal{M}(\Omega, \mathcal{F})$. The **weak topology** induced by $BC(\Omega)$ on $\mathcal{M}(\Omega, \mathcal{F})$ is the smallest topology (the topology with fewest open sets) such that every $f \in BC(\Omega)$ is continuous on $\mathcal{M}(\Omega, \mathcal{F})$.

Definition 5.1. *We say that* μ_n **converges** *to* μ **in distribution** *if* $\mu_n \to \mu$ *in the weak topology, that is, if* $f(\mu_n) \to f(\mu)$ *for every* $f \in BC(\Omega)$*.*

Theorem 5.2. *Let* $\{\mu_n\}$ *be probability measures on* $(\mathbb{R}, \mathcal{B})$*, and let* F_n *denote the CDF of* μ_n*. Then* $\mu_n \to \mu$ *in distribution if and only if* $F_n(x) \to F(x)$ *for every* x *at which* F*, the CDF of* μ*, is continuous.*

Definition 5.3. *A family* $\{\mu_\alpha, \alpha \in A\}$ *of probability measures on* Ω *is* **tight** *if, for every* $\epsilon > 0$*, there is a compact set* K *such that* $\mu_\alpha(K^c) < \epsilon$ *simultaneously for every* $\alpha \in A$*. If* $\Omega = \mathbb{R}$*, this is equivalent to the condition that* $\sup_\alpha \mu_\alpha([-M, M])^c) \to 0$ *as* $M \to \infty$*. In the language of CDFs, the condition is that* $\sup_\alpha[1 - F_\alpha(M) + F_\alpha(-M)] \to 0$ *as* $M \to \infty$*.*

Example. If the family $\{X_n\}$ satisfies $\mathbb{E}|X_n| \le C$ for all n, then it is tight.

Theorem 5.4. *A family of measures is tight if and only if every sequence of measures has a subsequential limit in distribution.*

6. Characteristic Functions

Definition 6.1. *If* X *is a random variable, we define its* **characteristic function** *by* $\varphi(t) = \mathbb{E}e^{itX}$*.*

Example. If X is uniform on $[a, b]$, then its characteristic function is given by

$$\mathbb{E}e^{itX} = \int_a^b \frac{e^{itx}}{b-a}\, dx = \left[\frac{e^{itx}}{it(b-a)}\right]_a^b = \frac{e^{itb} - e^{ita}}{it(b-a)}. \tag{6.1}$$

Example. If X is the standard normal whose density is given by $\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, then using the change of variable $y = x - it$, we compute its characteristic function to be

$$\begin{aligned} \int_{-\infty}^{\infty} \frac{e^{itx}e^{-x^2/2}}{\sqrt{2\pi}}\, dx &= e^{-t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(x-it)^2/2}\, dx \\ &= e^{-t^2/2} \int_{-\infty-it}^{\infty-it} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy = e^{-t^2/2}. \end{aligned} \tag{6.2}$$

Example. If X is an exponential of mean $1/\lambda$ whose density is given by $\lambda e^{-\lambda x}$, then its characteristic function is

$$\int_0^\infty e^{itx}\lambda e^{-\lambda x}\,dx = \int_0^\infty \lambda e^{x(it-\lambda)}\,dx = \lim_{a\to\infty}\left[\frac{\lambda e^{x(it-\lambda)}}{it-\lambda}\right]_0^a = \frac{\lambda}{\lambda - it}. \tag{6.3}$$

Example. If X is a Poisson of mean λ, where the probability of exactly k occurences is $\frac{\lambda^k e^{-\lambda}}{k!}$, then its characteristic function is given by

$$\sum_{k=0}^\infty e^{itk}\frac{\lambda^k e^{-\lambda}}{k!} = \sum_{k=0}^\infty \frac{(\lambda e^{it})^k}{k!}e^{-\lambda} = \exp(\lambda e^{it})e^{-\lambda} = \exp[\lambda(e^{it}-1)]. \tag{6.4}$$

Theorem 6.2 (Inversion formula). *Let $\varphi(t) = \int e^{itx}\mu(dx)$, where μ is a probability measure. For $a < b$,*

$$\mu((a,b)) + \frac{1}{2}\mu(\{a,b\}) = \frac{1}{2\pi}\lim_{T\to\infty}\int_{-T}^{T}\frac{e^{-ita}-e^{-itb}}{it}\varphi(t)\,dt. \tag{6.5}$$

Moreover, if $\int|\varphi(t)|dt < \infty$, then μ has bounded continuous density given by

$$f(y) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-ity}\varphi(t)\,dt. \tag{6.6}$$

Theorem 6.3 (Continuity lemma). *Let $\{\mu_n\}$ be probability measures with characteristic functions $\{\varphi_n\}$.*

- *If $\mu_n \to \mu$ in distribution for some μ with characteristic function φ, then $\varphi_n \to \varphi$ pointwise, and*
- *Conversely, if $\varphi_n \to \varphi$ pointwise for some φ that is continuous at zero, then $\mu_n \to \mu$ in distribution, where μ is a measure whose characteristic function is φ.*

Theorem 6.4. *Let $\{\mu_\alpha\}$ be a family of probability measures with corresponding characteristic functions $\{\varphi_\alpha\}$. Then $\{\mu_\alpha\}$ is tight if and only if $\{\varphi_\alpha\}$ is equicontinuous at zero (i.e. if for every $\epsilon > 0$ there exists a $\delta > 0$ such that simultaneously for all α, $|t| < \delta$ implies $|\varphi_\alpha(t) - 1| < \epsilon$).*

Proposition 6.5. *If $\mathbb{E}|X|^n < \infty$, then the characteristic function φ of X is n times differentiable, and $\varphi^{(n)}(t) = \mathbb{E}(iX)^n e^{itX}$.*

7. Central Limit Theorem

Theorem 7.1 (Central limit theorem). *Let $X_1, X_2, \dots$ be IID with mean μ and variance $\sigma^2 \in (0,\infty)$. Then $(S_n - n\mu)/\sigma\sqrt{n} \to \chi$ in distribution, where χ is the standard normal.*

Proof. It suffices to prove the result for $\mu = 0$. By Proposition 6.5, the characteristic function φ of X_1 is twice-differentiable with derivatives

$$\varphi'(0) = \mathbb{E}(iX_1) = 0 \text{ and } \varphi''(0) = -\mathbb{E}X_1^2 = -\sigma^2. \tag{7.1}$$

Thus Taylor's estimate gives

$$\varphi(t) = 1 - \sigma^2 t^2\frac{1+g(t)}{2}, \tag{7.2}$$

where $g(t) \to 0$ as $t \to 0$. The characteristic function $\varphi_{S_n/\sigma\sqrt{n}}$ of $S_n/\sigma\sqrt{n}$ is

$$\begin{aligned}\mathbb{E}\exp\left(\frac{itS_n}{\sigma\sqrt{n}}\right) &= \mathbb{E}\left(\exp\left(iX_1\frac{t}{\sigma\sqrt{n}}\right)\right)^n \\ &= \left(\varphi\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n = \left(1 - t^2\frac{1+g(t/\sigma\sqrt{n})}{2n}\right)^n.\end{aligned} \tag{7.3}$$

For fixed t, let $c_n = -\frac{1}{2}t^2[1+g(t/\sigma\sqrt{n})]$. Then $c_n \to -t^2/2$, so that

$$\varphi_{S_n/\sigma\sqrt{n}} = (1+c_n/n)^n \to e^{-t^2/2}, \tag{7.4}$$

which by the continuity lemma shows that $S_n/\sigma\sqrt{n} \to \chi$. □

Theorem 7.2 (Lindeberg-Feller CLT). *Let $\{X_{n,k} : 1 \leq k \leq n < \infty\}$ be a triangular array, with independence assumed within each row. Assume all variables have mean zero,*

$$\lim_{n\to\infty}\sum_{k=1}^{n}\mathbb{E}X_{n,k}^2 = \sigma^2 > 0, \tag{7.5}$$

and that for all $\epsilon > 0$,

$$\lim_{n\to\infty}\sum_{k=1}^{n}\mathbb{E}X_{n,k}^2\mathbf{1}_{|X_{n,k}|>\epsilon} = 0. \tag{7.6}$$

Then $\sum_{k=1}^n X_{n,k} \to \sigma\chi$ in distribution as $n \to \infty$, where $\sigma\chi$ is a normal with mean zero and variance σ^2.

Recall that if X and Y are random variables with means μ and ν respectively, then their **covariance** is defined as $\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mu)(Y-\nu)]$. A **covariance matrix** is the matrix of covariances between elements of a random vector.

Definition 7.3 (Multivariate Gaussian). *Let Γ be a $d \times d$ non-negative definite, real symmetric matrix, and let B be a $d \times d$ positive definite, real symmetric matrix.*

1. *The Gaussian distribution with mean zero and covariance matrix Γ is the distribution with characteristic function $\varphi(\theta) = \exp(-\frac{1}{2}\theta^T\Gamma\theta)$, where $\theta \in (\mathbb{R}^d)^*$.*
2. *The Gaussian distribution with mean zero and covariance matrix Γ is the law of the linear image $A(\mathbf{Y})$ of a vector $\mathbf{Y}$ of IID standard normals under a linear map A such that the covariance matrix of $A(\mathbf{Y})$ is Γ; there exists one and only one such law.*
3. *The Gaussian distribution with mean zero and covariance matrix B^{-1} is the distribution on $\mathbb{R}^d$ with density $(2\pi)^{-d/2}|B|^{1/2}\exp(-\frac{1}{2}\mathbf{x}^TB\mathbf{x})$.*

Theorem 7.4 (Multivariate CLT). *Let $\{\mathbf{X}_n\}$ be IID with mean zero and finite covariances $\mathbb{E}X_iX_j = \Gamma_{ij}$. If $\mathbf{S}_n = \sum_{k=1}^n \mathbf{X}_k$, then $\mathbf{S}_n/\sqrt{n} \to \mathcal{N}(0,\Gamma)$ in distribution as $n \to \infty$.*

8. Poisson Limits and Poisson Processes

Definition 8.1. *Let μ and ν be two probability measures on a countable set S. We define their* **total variation distance** *to be $||\mu-\nu||_{TV} = \sum_{x\in S}|\mu(x)-\nu(x)|$.*

Proposition 8.2. *With μ and ν as above, $||\mu-\nu||_{TV} = 2\sup_{A\subset S}|\mu(A)-\nu(A)|$.*

Proof. For any $A \subset S$,

$$\sum_{x \in S} |\mu(x) - \nu(x)| \geq |\mu(A) - \nu(A)| + |\mu(A^c) - \nu(A^c)| = 2|\mu(A) - \nu(A)|, \tag{8.1}$$

and equality holds when $A = \{x \in S : \mu(x) \geq \nu(x)\}$. □

Lemma 8.3. *Let* $(\mu_1 \times \mu_2)(x, y) = \mu_1(x)\mu_2(y)$ *be the product measure on* $\mathbb{Z} \times \mathbb{Z}$*. Then*

$$||\mu_1 \times \mu_2 - \nu_1 \times \nu_2||_{TV} \leq ||\mu_1 - \nu_1||_{TV} + ||\mu_2 - \nu_2||_{TV}. \tag{8.2}$$

Lemma 8.4. *Let* $(\mu_1 * \mu_2)(x) = \sum_y \mu_1(x-y)\mu_2(y)$ *be the convolution of* μ_1 *and* μ_2*. Then*

$$||\mu_1 * \mu_2 - \nu_1 * \nu_2||_{TV} \leq ||\mu_1 \times \mu_2 - \nu_1 \times \nu_2||_{TV}. \tag{8.3}$$

Lemma 8.5. *Let* μ *and* ν *be the respective laws of a Bernoulli and a Poisson with the same parameter* p*. Then* $||\mu - \nu||_{TV} \leq 2p^2$*.*

Proof. Since $1 - p < e^{-p}$, only 1 has higher probability under the Bernoulli. Therefore

$$||\mu - \nu||_{TV} = 2(p - pe^{-p}) = 2p(1 - e^{-p}) \leq 2p^2 \tag{8.4}$$

by Proposition 8.2 and the symmetry of total variation. □

Theorem 8.6 (Poisson convergence)**.** *Let* $\{X_{n,k} : 1 \leq k \leq n < \infty\}$ *be a triangular array of random variables taking* 0 *or* 1*. Let* $p_{n,k} = \mathbb{E}X_{n,k} = \mathbb{P}(X_{n,k} = 1)$ *and* $S_n = \sum_{k=1}^n X_{n,k}$*. Suppose that as* $n \to \infty$*,*

$$\sum_{k=1}^{n} p_{n,k} \to \lambda \in (0, \infty) \text{ and } \max_{1 \leq k \leq n} p_{n,k} \to 0. \tag{8.5}$$

Then $S_n \to \mathcal{P}(\lambda)$ *in distribution as* $n \to \infty$*.*

Proof. Let $\mu_{n,k}$ be the law of $X_{n,k}$ and let μ_n be the law of S_n. Moreover let $\nu_{n,k}$, ν_n, and ν be the laws of Poissons with means $p_{n,k}$, $\lambda_n = \sum_{k \leq n} p_{n,k}$, and λ respectively. Since (see Theorem 9.1) $\mu_n = \mu_{n,1} * \cdots * \mu_{n,n}$ and $\nu_n = \nu_{n,1} * \cdots * \nu_{n,n}$,

$$\sup_A |\mu_n(A) - \nu_n(A)| = \frac{1}{2}||\mu_n - \nu_n||_{TV} \leq \sum_{k=1}^{n} p_{n,k}^2 \tag{8.6}$$

by the previous four results. The RHS is bounded by $\max_k p_{n,k} \sum_{k=1}^n p_{n,k}$, which by hypothesis converges to $0 \cdot \lambda = 0$. Since $\nu_n \to \nu$ in distribution, the result follows. □

Let I be an at most countably infinite index set. Let $\{X_\alpha : \alpha \in I\}$ be random variables, each taking 0 or 1. For each $\alpha \in I$ we choose a set B_α, which loosely speaking is a set of indices β such that X_α and X_β are highly dependent. Let $p_\alpha = \mathbb{E}X_\alpha$, $p_{\alpha,\beta} = \mathbb{E}X_\alpha X_\beta$,

$$\begin{aligned} b_1 &= \sum_{\alpha \in I} \sum_{\beta \in B_\alpha} p_\alpha p_\beta, \\ b_2 &= \sum_{\alpha \in I} \sum_{\alpha \neq \beta \in B_\alpha} p_{\alpha,\beta}, \text{ and} \\ b_3 &= \sum_{\alpha \in I} \mathbb{E}|\mathbb{E}(X_\alpha \mid X_\beta : \beta \notin B_\alpha) - p_\alpha|. \end{aligned} \tag{8.7}$$

Intuitively, b_1 measures the total size of the dependence neighborhoods, b_2 measures how many dependent pairs are likely to arise, and b_3 measures how far X_α is from being independent of $\{X_\beta : \beta \notin B_\alpha\}$.

Theorem 8.7 (Arratia-Goldstein-Gordon)**.** *Let $W = \sum_{\alpha \in I} X_\alpha$, and let Z be a Poisson with mean $\lambda = \mathbb{E}W = \sum_{\alpha \in I} p_\alpha$. Then the total variation distance between W and Z is at most $b_1 + b_2 + b_3$.*

Say we want to have a set of points springing up in an arbitrary space $(S, \mathcal{S})$. Let μ be a non-atomic σ-finite measure on $(S, \mathcal{S})$. We would like each small patch of measure $d\mu$ to have probability $d\mu$ of containing one of the points (and probability essentially zero of containing more than one point), and we want the numbers of points springing up in the sets $A_1, \dots, A_n$ to be independent if these sets are disjoint. What kind of formal object must we be dealing with?

Definition 8.8. *Let μ be a non-atomic σ-finite measure on $(S, \mathcal{S})$. A* **Poisson process with intensity** *μ is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ together with a map $N : \Omega \times \mathcal{S} \to \mathbb{Z}^+$ such that*

(1) For each $\omega \in \Omega$, the map $A \to N(\omega, A)$ is a measure on $(S, \mathcal{S})$, and
(2) If $A_1, \dots, A_k \in \mathcal{S}$ are disjoint sets of finite measure, then $N(\cdot, A_1), \dots, N(\cdot, A_k)$ are independent Poissons with means $\mu(A_1), \dots, \mu(A_k)$.

Proposition 8.9. *Let μ be a non-atomic σ-finite measure on $(S, \mathcal{S})$. Suppose N is a random counting measure such that $\mathbb{E}N(A) = \mu(A)$ and $\mathbb{P}(N(A) \geq 2) = O(\mu(A)^2)$. Suppose also that the $N(A_k)$ are independent if the A_k are disjoint. Then N is a Poisson process with intensity μ.*

Explicitly, when $\|\mu\| = \lambda < \infty$, the construction of a Poisson process is as follows: let $\mathcal{P}$ be a Poisson with mean λ, then place $\mathcal{P}$ points with ordinates X_i that are IID $\sim \nu = \mu/\lambda$ and independent of $\mathcal{P}$. Let φ_ν denote the characteristic function of X_i. The random variable $Z = \sum_{i=1}^{\mathcal{P}} X_i$ is a compound poisson with characteristic function

$$\begin{aligned}
\mathbb{E}e^{itZ} &= \mathbb{E}\left[\exp\left(it\sum_{j=1}^{\mathcal{P}} X_j\right)\right] = \mathbb{E}\left(\prod_{j=1}^{\mathcal{P}} e^{itX_j}\right) = \mathbb{E}(e^{itX_1})^{\mathcal{P}} \\
&= \mathbb{E}(\varphi_\nu(t))^{\mathcal{P}} = \sum_{k=0}^{\infty} \mathbb{P}(\mathcal{P} = k)(\varphi_\nu(t))^k = \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!}(\varphi_\nu(t))^k \qquad (8.8) \\
&= e^{-\lambda}\sum_{k=0}^{\infty} \frac{(\lambda\varphi_\nu(t))^k}{k!} = e^{-\lambda}e^{\lambda\varphi_\nu(t)} = \exp[\lambda(\varphi_\nu(t) - 1)].
\end{aligned}$$

When μ is the Lebesgue measure on $\mathbb{R}^+$, the random measure $N(\omega)$ has distribution function $X_t(\omega) = F(\omega, t) = N(\omega, (-\infty, t])$. The process $\{X_t : t \geq 0\}$ is sometimes known as the **standard Poisson process**. The set of discontinuities of F is the random set W. Let $T_1 < T_2 < \dots$ enumerate W, set $T_0 = 0$, and denote $\xi_j = T_j - T_{j-1}$.

Proposition 8.10. *The ξ_j are IID exponentials of mean 1. Conversely, given ξ_j IID exponentials of mean 1, setting $T_n = \sum_{j=1}^{n} \xi_j$ and $X_t = \sup\{n : T_n \leq t\}$, there is a unique measure $N(\omega)$ for which $X_t(\omega) = N(\omega, (-\infty, t])$ for all t, and this N is a Poisson process with intensity dt on $(0, \infty)$.*

The usual interpretation for the Poisson process on $\mathbb{R}^+$ is that a timer goes off at times corresponding to the points of the process. The amount of time between successive rings is exponential in distribution. The chance to go off in time dt does not depend on the past, and this is inherited from the memoryless property of the exponential. In elementary texts, the Poisson process on $\mathbb{R}^+$ is axiomatized by

(1) For $s < t$, the law of $N(s,t] = N_t - N_s$ is Poisson with mean $t - s$, and
(2) For a finite collection $\{I_1, \ldots, I_k\}$ of finite disjoint intervals, $\{N(I_1), \ldots, N(I_k)\}$ are independent.

If the intensity of a (possibly non-homogeneous) Poisson process is given by a rate function $\lambda(t)$, we see that the expected number of events between time a and time b is $\lambda_{a,b} = \int_a^b \lambda(t)\,dt$. The number of arrivals in the time interval $(a,b]$ follows a Poisson distribution with associated parameter $\lambda_{a,b}$, that is,

$$\mathbb{P}(N_b - N_a = k) = \frac{e^{-\lambda_{a,b}}(\lambda_{a,b})^k}{k!}. \tag{8.9}$$

9. Infinite Divisibility and Stable Laws

Theorem 9.1. *Suppose X and Y are independent random variables with distributions $F(x) = \mathbb{P}(X \leq x)$ and $G(y) = \mathbb{P}(Y \leq y)$, then $\mathbb{P}(X + Y \leq z) = \int F(z-y)\,dG(y)$. This integral is called the* **convolution** *of F and G and is denoted as $(F * G)(z)$.*

Definition 9.2. *A random variable Z has an* **infinitely divisible distribution** *if it is a limit of sums of n IID random variables as $n \to \infty$. Equivalently, Z is equal in distribution to a sum of n IID random variables for any n.*

Definition 9.3. *A random variable Z is* **stable** *if, given two of its independent copies Z_1 and Z_2 and any constants a and b, $aZ_1 + bZ_2$ has the same distribution as $cZ + d$ for some constants c and d.*

Remark. Clearly stable laws are infinitely divisible, and as such, they arise from summing the points of Poisson processes.

10. Random Walks

Definition 10.1. *Let $\{X_n\}$ be IID elements of $\mathbb{R}^d$, with partial sums S_n. The sequence $\{S_n\}$ is known as a* **random walk**.

The standard space here is $(\Omega, \mathcal{F}, \mathbb{P}) = (S, \mathcal{S}, F)^\infty$, where F is some probability distribution on $(S, \mathcal{S})$, usually either $(\mathbb{R}^d, \mathcal{B})$ or $(\mathbb{Z}^d, 2^{\mathbb{Z}^d})$.

Definition 10.2. *Let $\{X_\alpha : \alpha \in A\}$ be a family of random variables. The* **σ-field generated by** *$\{X_\alpha : \alpha \in A\}$, denoted as $\sigma(X_\alpha : \alpha \in A)$, is the smallest σ-field with respect to which each X_α is measurable.*

Remark. The interpretation of this σ-field is that it is the information we know by looking at the value of each X_α. In particular, an event is in $\mathcal{F}_n = \sigma(X_1, \ldots, X_n)$ if we can always tell whether it has happened by looking at the values of $X_1, \ldots, X_n$.

Definition 10.3. *A* **stopping time** *is a random variable τ taking values in $\mathbb{Z}^+ \cup \{+\infty\}$ such that for each n, the event $\{\tau \leq n\}$ is in $\mathcal{F}_n$.*

Proposition 10.4. *If τ, ρ are stopping times, then* $\max(\tau, \rho)$, $\min(\tau, \rho)$, *and* $\tau + \rho$ *are stopping times.*

Proof. Since τ, ρ are stopping times, $\{\tau \leq n\}$ and $\{\rho \leq n\}$ are in $\mathcal{F}_n$ for each n, hence $\{\tau \leq n\} \cap \{\rho \leq n\} = \{\max(\tau, \rho) \leq n\} \in \mathcal{F}_n$ for each n. This shows that $\max(\tau, \rho)$ (and similarly $\min(\tau, \rho)$) is a stopping time. Now observe that

$$\{\tau + \rho \leq n\} = \bigcup_{n_1 + n_2 \leq n} (\{\tau \leq n_1\} \cup \{\rho \leq n_2\}) \in \mathcal{F}_n \tag{10.1}$$

for each n, so $\tau + \rho$ is a stopping time. □

Proposition 10.5. *The random time τ is a stopping time if and only if the event $\{\tau = n\}$ is in $\mathcal{F}_n$ for each n.*

Definition 10.6. *Let τ be a stopping time. Define the σ-field $\mathcal{F}_\tau$ to be the set of events A such that for all finite n, $A \cap \{\tau = n\}$ is in $\mathcal{F}_n$.*

It is not difficult to see that τ is $\mathcal{F}_\tau$-measurable. Indeed, consider $\{\tau = m\} \cap \{\tau = n\}$: if $m \neq n$, then this is the empty set, which is in any σ-field, and if $m = n$, then note that $\{\tau = n\} \in \mathcal{F}_n$ for every n. Furthermore, define X_τ as the function[1] taking the value $X_n(\omega)$ when $\tau(\omega) = n$, then for any Borel subset B of $\mathbb{R}$,

$$\{X_\tau \in B\} \cap \{\tau = n\} = \{X_n \in B\} \cap \{\tau = n\}. \tag{10.2}$$

Since $\{X_n \in B\} \in \mathcal{F}_n$ for every n, it follows that X_τ (and similarly S_τ) is $\mathcal{F}_\tau$-measurable.

Theorem 10.7 (Wald's identity)**.** *Let $\{X_i\}$ be IID with $\mathbb{E}|X_1| < \infty$, and let τ be a stopping time with $\mathbb{E}\tau < \infty$. Then $\mathbb{E}S_\tau = (\mathbb{E}\tau)(\mathbb{E}X_1)$.*

Proof. First suppose that the X_i are all non-negative. Then

$$\begin{aligned} \mathbb{E}(S_\tau) &= \sum_{n=1}^{\infty} \int S_n \mathbf{1}_{\tau=n} \, d\mathbb{P} \\ &= \sum_{n=1}^{\infty} \sum_{k=1}^{n} \int X_k \mathbf{1}_{\tau=n} \, d\mathbb{P} = \sum_{k=1}^{\infty} \sum_{n=k}^{\infty} \int X_k \mathbf{1}_{\tau=n} \, d\mathbb{P}, \end{aligned} \tag{10.3}$$

where the switch in order of summation is justified because all summands are non-negative. Collapsing the inner sum yields $\mathbb{E}S_\tau = \sum_{k=1}^{\infty} \int X_k \mathbf{1}_{\tau \geq k} \, d\mathbb{P}$. Since the event $\{\tau \geq k\}$ is the complement of the event $\{\tau \leq k-1\}$, it is in $\mathcal{F}_{k-1}$, and therefore independent of X_k. Hence

$$\mathbb{E}S_\tau = \sum_{k=1}^{\infty} \int X_k \mathbf{1}_{\tau \geq k} \, d\mathbb{P} = \sum_{k=1}^{\infty} \mathbb{P}(\tau \geq k)(\mathbb{E}X_k) = (\mathbb{E}\tau)(\mathbb{E}X_1). \tag{10.4}$$

For the general case, use the previous argument on $\{|X_n|\}$ to see that

$$\sum_{k=1}^{\infty} \sum_{n=k}^{\infty} \int |X_k| \mathbf{1}_{\tau=n} \, d\mathbb{P} \tag{10.5}$$

is absolutely convergent. Now we can repeat the above argument, justifying the switch in order of summation by absolute convergence, and obtain $\mathbb{E}S_\tau = (\mathbb{E}\tau)(\mathbb{E}X_1)$ as before. □

[1] Usually events in terms of X_τ are intersected with $\{\tau < \infty\}$ in order to be well-defined.

Suppose $\{S_n\}$ is a random walk on $\mathbb{Z}^d$ with steps $\{X_n\}$ that are IID, and define $\tau = \inf\{n \geq 1 : S_n = \mathbf{0}\}$ to be the time of the first return to the origin.

Definition 10.8. *If every x in the support of $\{S_n\}$ is almost surely achieved infinitely often, we call the random walk* **recurrent**. *If almost surely no x is achieved infinitely often, we call the random walk* **transient**.

Theorem 10.9. *For any random walk, the following are equivalent:*

(1) $\mathbb{P}(\tau < \infty) = 1$.
(2) $\mathbb{P}(S_n = \mathbf{0} \text{ i.o.}) = 1$.
(3) $\sum_{n=1}^{\infty} \mathbb{P}(S_n = \mathbf{0}) = \infty$.

If a random walk satisfies any of the above, it is recurrent. Otherwise, $\mathbb{P}(\tau < \infty) = \alpha < 1$, in which case $\mathbb{E}(|n : S_n = \mathbf{0}|) = 1/(1-\alpha)$, and the random walk is transient.

Theorem 10.10. *A simple random walk on $\mathbb{Z}^d$ is recurrent if and only if $d \leq 2$.*

Proof. If $d = 1$, then Stirling's formula[2] gives

$$\mathbb{P}(S_{2n} = 0) = 2^{-2n}\tbinom{2n}{n} \sim \frac{1}{2^{2n}} \frac{(2n)^{2n}}{n^n n^n} \frac{\sqrt{2\pi(2n)}}{(\sqrt{2\pi n})^2} = \frac{1}{\sqrt{\pi n}}, \tag{10.6}$$

which is not summable, hence the walk is recurrent. In $d = 2$, start with a product of two independent one-dimensional simple random walks (i.e. a two-dimensional random walk whose steps are uniform on the four values $(\pm 1, \pm 1)$), then rotate and shrink by $\sqrt{2}$ to see that this is isomorphic to a two-dimensional simple random walk. It follows from (10.6) that $\mathbb{P}(S_{2n} = \mathbf{0}) \sim 1/\pi n$, which again is not summable, so the walk is recurrent.

When $d = 3$, to return to $\mathbf{0}$, the number of steps in each positive axis direction must be equal to the number of steps in the corresponding negative axis direction. Therefore

$$\begin{aligned}
\mathbb{P}(S_{2n} = \mathbf{0}) &= 6^{-2n} \sum_{j,k} \frac{(2n)!}{(j!k!(n-j-k)!)^2} \\
&= 2^{-2n}\binom{2n}{n}\left(3^{-n}\frac{n!}{j!k!(n-j-k)}\right)^2 \\
&\leq 2^{-2n}\binom{2n}{n} \max_{j,k} 3^{-n} \frac{n!}{j!k!(n-j-k)!},
\end{aligned} \tag{10.7}$$

where the inequality used the fact that if $a_{j,k} \geq 0$ sum to 1, then $\sum_{j,k} a_{j,k}^2 \leq \max_{j,k} a_{j,k}$. It can be shown that the maximum occurs when j, k, and $n - j - k$ are as close as possible to $n/3$, and Stirling's formula then demonstrates that the maximum is $O(n^{-1})$. Thus $\mathbb{P}(S_{2n} = \mathbf{0}) = O(n^{-3/2})$, and the random walk is transient.

When $d \geq 3$, let $T_n = (S_n^1, S_n^2, S_n^3)$, i.e. the first three coordinates of S_n. Let $N(0) = 0$ and $N(n) = \inf\{m > N(n-1) : T_m \neq T_{N(n-1)}\}$, so that $T_{N(n)}$ is a three-dimensional simple random walk. Then $T_{N(n)}$ returns infinitely often to $\mathbf{0}$ with probability 0, and since the first three coordinates of S_n are constant in between the indices $N(n)$, we see that $S_n = \mathbf{0}$ finitely often a.s., thus S_n is transient. □

In the non-lattice case, a general random walk is said to be **recurrent** if, for every $\epsilon > 0$, $\mathbb{P}(|S_n| < \epsilon \text{ i.o.}) = 1$.

[2]The version most useful to us is that $n! \sim n^n e^{-n}\sqrt{2\pi n}$ as $n \to \infty$.

Theorem 10.11 (Chung-Fuchs). *In one dimension, if the weak law holds ($S_n/n \to 0$ in probability), then the walk is recurrent.*

For a fixed n, let $\mathbf{x}$ denote a vector $(x_1, \dots, x_n)$ of real numbers, and let s_k denote their partial sums ($s_0 = 0$). Set $M(\mathbf{x}) = \max_{0\le k\le n} s_k$ and $\sigma_n = \inf\{j \le n : s_j = M(\mathbf{x})\}$, so that σ_n is the smallest j that $S_j = \max_{0\le k\le n} S_k$.

Theorem 10.12 (Arcsine law). *If $\{X_i\}$ are IID, symmmetric (so that $-X_i = X_i$ in distribution), and non-atomic, then $n^{-1}\sigma_n \to Z$ in distribution, where*

$$\mathbb{P}(Z \in [a, b]) = \int_a^b \frac{1}{\pi\sqrt{x(1-x)}}\, dx. \tag{10.8}$$

An equivalent expression is $[\arcsin(2b-1) - \arcsin(2a-1)]/\pi$.

Consider the sequence $\{S_n\}$ as being represented by a polygonal line with segments $(k-1, S_{k-1}) \to (k, S_k)$. A **path** is a polygonal line that is a possible outcome of a SRW.

Theorem 10.13 (Reflection principle). *If $x, y > 0$, then the number of paths from $(0, x)$ to (n, y) that are 0 at some time is equal to the number of paths from $(0, -x)$ to (n, y).*

Definition 10.14. *An event A is* **permutable** *if $\pi^{-1}A = \{\omega : \pi\omega \in A\}$ is equal to A for any finite permutation π on $\mathbb{N}$. In other words, A is permutable if its occurence is not affected by rearranging the random variables.*

Definition 10.15. *The collection of pemutable events is a σ-field called the* **exchangeable σ-field** *and denoted by $\mathcal{E}$.*

Theorem 10.16 (Hewiit-Savage 0-1 law). *If $\{X_i\}$ are IID and $A \in \mathcal{E}$, then $\mathbb{P}(A)$ is either 0 or 1.*

Appendix A. Common Probability Distributions

	Support	PDF/PMF	Mean	Var	Char Fn
Uniform	$[a, b]$	$\frac{1}{b-a}$ for $x \in [a, b]$, 0 otherwise	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{itb}-e^{ita}}{it(b-a)}$
Normal	$\mathbb{R}$	$\frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$	μ	σ^2	$\exp(i\mu t - \frac{\sigma^2 t^2}{2})$
Exponential	$[0, \infty)$	$\lambda e^{-\lambda x}$	$1/\lambda$	$1/\lambda^2$	$\frac{\lambda}{\lambda - it}$
Bernoulli	$\{0, 1\}$	$\mathbb{P}(X = 1) = p = 1 - q$	p	pq	$p + qe^{it}$
Geometric	$\mathbb{N}_1$	$\mathbb{P}(X = k) = (1-p)^{k-1}p$	$1/p$	$\frac{1-p}{p^2}$	$\frac{pe^{it}}{1-(1-p)e^{it}}$
Poisson	$\mathbb{N}_0$	$\mathbb{P}(Z = k) = \lambda^k e^{-\lambda}/k!$	λ	λ	$\exp[\lambda(e^{it} - 1)]$

Remark. The exponential distribution occurs naturally when describing the lengths of the inter-arrival times in a homogeneous Poisson process. It is **memoryless**, i.e. if X is exponentially distributed, then $\mathbb{P}(X > s + t \mid X > s) = \mathbb{P}(X > t)$ for all $s, t \ge 0$.

Remark. The geometric distribution is the probability distribution of the number X of Bernoulli trials needed to get one success.

Remark. If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Y \sim \mathcal{N}(\nu, \tau^2)$, then $X + Y \sim \mathcal{N}(\mu + \nu, \sigma^2 + \tau^2)$. If $X \sim \mathcal{P}(\lambda_1)$ and $Y \sim \mathcal{P}(\lambda_2)$, then $X + Y \sim \mathcal{P}(\lambda_1 + \lambda_2)$.

E-mail address: shanshand@math.upenn.edu