ORF 526: Summary of Important Results

Compiled by Vu T. Chau

1 Lebesgue Integration and Convergence Theorem

Theorem 1.1 (Monotone Convergence Theorem). *Let $g, f_1, f_2, ...$ be measurable functions such that $g \in L^1$, $g \leq f_1$ a.e. and $f_n \nearrow f$ a.e. Then:*

$$\int f_n d\mu \to \int f d\mu$$

Lemma 1.2 (Fatou's Lemma). *Let $g, f_1, f_2, ...$ be measurable functions such that $g \in L^1$, $f_n \geq g$ a.e. for $\forall n$. Then:*

$$\liminf_n \int f_n d\mu \leq \int \liminf_n f_n d\mu$$

Theorem 1.3 (Lebesgue's Dominated Convergence Theorem). *Let $g, f_1, f_2, ...$ be measurable functions such that $g \in L^1$, $|f_n| \leq |g_n|$ a.e. $\forall n$, and $f_n \to f$ a.e. Then, we have*

$$\lim_{n\to\infty} \int f_n d\mu = \int f d\mu$$

and

$$\lim_{n\to\infty} \int |f - f_n| = 0$$

If μ is a probability measure, then μ_f is a probability measure too, and called the distribution of f.

Proposition 1.4 (Push-forward measure[1]). *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space, and $(\Omega', \mathcal{F}')$ a measurable space. $f : (\Omega, \mathcal{F}) \mapsto (\Omega', \mathcal{F}')$ a measurable function. Then: $\mu_f(B) = \mu[f^{-1}(B)]$ defines a measure on $(\Omega', \mathcal{F}')$, and*

$$\int_\Omega g \circ f d\mu = \int_{\Omega'} g d\mu_f$$

Definition (Expectation, Variance, and Covariance). *We define:*
$\mathbb{E}X = \int_\Omega X dP$ *for* $X \in L^1$.
$Var(X) = \int_\Omega (X - \mathbb{E}X)^2 dP = \mathbb{E}[X^2] - (\mathbb{E}X)^2$
$Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] = \mathbb{E}[XY] - (\mathbb{E}X)(\mathbb{E}Y)$

So far, we are saying that given a measure space and a random variable, we can characterize the distribution of that variable by the cdf $F_X(x) = (P \circ X^{-1})[X \leq x]$. Now we define the inverse function of F_X:

Definition (Right-quantile function). *Define $q_X : (0, 1) \mapsto \mathbb{R}$ by $q_X(u) = sup\{x : F_X(x) \leq u\}$. We note that $q_X =^d X$.*

So far we have linked the relationship between pointwise convergence and convergence in L^1 using the two convergence theorems. The following section explores more types of convergence, and the notion of "uniform integrable."

2 Types of Convergence and Uniform Integrability

Definition. *Let $X_1, X_2, ...$ be randome variables on a probability space $(\Omega, \mathcal{F}, P)$.*

- $X_n \to X$ *a.s.* $\Leftrightarrow \exists N \in \mathcal{F}$ *such that* $P[N] = 0$ *and* $\lim_{n\to\infty} X_n(\Omega) = X(\Omega)$ *for* $\forall \omega \in \Omega \setminus N$.[2]
- $X_n \xrightarrow{L^p} X \Leftrightarrow \|X_n - X\|_p \to 0$ *as* $n \to \infty$.[3].
- $X_n \xrightarrow{P} X \Leftrightarrow P[|X_n - X|] > \epsilon \to 0$ *as* $n \to \infty$ *for all* $\epsilon > 0$.

[1] Take $g(x) = x$, we end up with $\mathbb{E}X = \int_\Omega X dP = \int_\mathbb{R} x d(P \circ X^{-1})$
[2] Almost surely convergence is pointwise convergence at all points except for a set of measure zero.
[3] Recall that $\|f\|_p = (\mathbb{E}[|f|^p])^{1/p}$. Note that when $p = 1$ we get convergence in L^1, which we discussed extensively earlier.

- $X_n \xrightarrow{D} X \Leftrightarrow \mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ *as* $n \to \infty$ *for all bounded continuous functions* $f : \mathbb{R} \mapsto \mathbb{R}$.

Definition (Uniform Integrable). *A family of random variables* $(X_i)_{i\in I}$ *on a probability space is uniformly integrable if* $\sup_{i\in I} \int_{|X_i|>c} |X_i| dP \to 0$ *as* $c \to \infty$.

It turns out there is a very simple condition for uniform integrability: if the sequence of random variables is bounded by an integrable RV X. This is stated clearly below:

Lemma 2.1 (Conditions for Uniform Integrability). *If* $(X_i)_{i\in I}$ *is a family of random variables on a probability space* $(\Omega, \mathcal{F}, P)$ *such that* $|X_i| \leq |X|$ *for some* $X \in L^1(\Omega, \mathcal{F}, P)$ *and* $\forall\, i \in I$, *then* $(X_i)_{i\in I}$ *is uniformly integrable.*

It follows that finitely many random variables $X_1, X_2, ..., X_n \in L^1(\Omega, \mathcal{F}, P)$ *are uniformly integrable.*

Now we turn to the necessary and sufficient condition for uniform integrability. In simple words, a family of RVs is uniformly integrable if (1) each RV in the family is L^1-bounded, and (2) the L^1-norm of each *RV* vanishes on any small enough set. Particularly:

Theorem 2.2 (Necessary and Sufficient Conditions for Uniform Integrability). *Let* $(X_i)_{i\in I}$ *be a family of RVs. This family is uniformly integrable if:*

- *(L^1-boundedness)* $\sup_{i\in I} \|X_i\|_{L^1} < \infty$*; and,*
- *(vanishes with small domain) For any* $\epsilon > 0$, $\exists \delta > 0$ *s.t.* $\mathbb{E}[|X_i| \cdot 1_A] < \epsilon$ *for all* i *and set* $A \in \mathcal{F}$ *with* $P[A] < \delta$.

Below is a major result:

Theorem 2.3 (Uniform Integrability + Convergence in Probability = Convergence in L^1). *Let* $X_1, X_2, ...$ *be RVs on* $(\Omega, \mathcal{F}, P)$, $X_n \xrightarrow{P} X$, *and* $\{X_n\}_{n\geq 1}$ *uniformly integrable. Then* X *is integrable (*$\mathbb{E}X < \infty$*),* $\|X_n - X\|_{L^1} \to 0$, *and, as a result,* $\mathbb{E}X_n \to \mathbb{E}X$.

Recall that we have a sufficient condition for uniform integrability before, which is (1) (X_i) has to be L^1-bounded, and (2) a really long and complicated condition. The following lemma and its corollary helps simplify the second condition:

Lemma 2.4 (de la Vallee - Poussin). *A family of RVs* $(X_i)_{i\in I}$ *on* $(\Omega, \mathcal{F}, P)$ *is uniformly integrable if and only if there exists* $\varphi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ *such that* $\lim_{x\to\infty} \frac{\varphi(x)}{x} = \infty$ *and* $\sup_{i\in I} \mathbb{E}[\varphi(|X_i|)] < \infty$.

φ *can always be chosen to be convex and non-decreasing.*

Corollary 2.5 (L^p-bounded ($p > 1$) implies uniform integrability). *A family of random variables* $(X_i)_{i\in I}$ *defined on* $(\Omega, \mathcal{F}, P)$ *is uniformly integrable if* $\sup_{i\in I} \|X_i\|_{L^p} < \infty$ *for some* $p \in (1, \infty]$.

Using Jensen's inequality[4], we can prove the following result:

Lemma 2.6. *Let* X *be a RV. Then* $\|X\|_{L^p} \leq \|X\|_{L^q}$ *for all* $1 \leq p \leq q \leq \infty$.

3 Weak Law of Large Numbers

The law of large number states that the sample mean of n uncorrelated variables with the same mean and finite second moment will converge to the population mean in L^2 and in probability.

Theorem 3.1 (Weak Law of Large Number). *Let* $X_1, X_2, ..., X_n$ *be uncorrelated RVs in* L^2 *s.t.* $\mathbb{E}[X_n] = m$ $\forall\, n$, *and* $\sup_n \mathbb{E}[X_n^2] < \infty$. *Then* $\frac{1}{n} \sum_{i=1}^{n} X_i \to m$ *in* L^2 *(and hence, in probability).*

[4]Let X be a RV and $\varphi : \mathbb{R} \mapsto \mathbb{R}$ a convex function such that $X, \varphi \in L^1$. Then: $\mathbb{E}[\varphi(X)] \geq \varphi(\mathbb{E}[X])$

4 Distributions and Densities

4.1 Discrete distributions

Definition (Dirac measure)**.** *Let Ω be a set, $\omega \in \Omega$. The Dirac measure $\delta_\omega : 2^\Omega \mapsto \{0,1\}$ is given by:*

$$\delta_\omega \triangleq \begin{cases} 1 & \text{if } \omega \in A, \\ 0 & \text{else.} \end{cases}$$

Let μ be a measure on $\mathbb{R}^d$.

Definition. *μ is called **discrete** if $\mu = \sum_{n\geq 1} p_n \delta_{x_n}$ for $\{x_n\}_{n\geq 1}$ and $p_1, p_2, ... \geq 0$.*

Some **examples** of discrete distributions on $\mathbb{R}$:

(1) Discrete uniform distribution: $\mu(n) = \frac{1}{N}$, $n = 1, 2, ..., N$.

(2) Bernoulli: $\mu(0) = 1 - p$, $\mu(1) = p$ for $p \in [0,1]$.

(3) Binomial: $\mu(n) = \binom{N}{n} p^n (1-p)^{N-n}$ for $n = 0, 1, ..., N$.

(4) Poisson: $\mu(n) = e^{-\lambda} \frac{\lambda^n}{n!}$, $n = 0, 1, ...$

(5) Geometric: $\mu(n) = (1-p)^n p$, $n = 1, 2, ..$

4.2 Continuous Distributions

Definition. *μ is called **continuous** if $\mu(A)$ is continuous in $\mathbb{R}^d$, for $A = (a_1, b_1] \times (a2, b2] \times ... \times (a_d, b_d]$.*

Definition. *μ is called **absolutely continuous** if there exists a Borel measurable function $f : \mathbb{R}^d \mapsto \mathbb{R}_+$ such that:*

$$\mu\left((a_1, b_1] \times ... \times (a_d, b_d]\right) = \int_{a_1}^{b_1} ... \int_{a_d}^{b_d} f(y_1, y_2, ..., y_d) dy_d ... dy_1$$

f is called a "density" and $\mu[A] = \int_A f(\mathbf{x}) d\mathbf{x}\ \forall\ A \in \mathcal{B}(\mathbb{R})^{\otimes d}$. Also, for any g measurable, we have

$$\int_{\mathbb{R}^d} g d\mu(\mathbf{x}) = \int_{\mathbb{R}^d} g(x) f(x) dx$$

Examples:

(1) Normal distribution (or, *Gaussian distribution*):

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$$

(2) Uniform on $[a, b]$: $f(x) = \frac{1}{b-a}$ for $x \in [a, b]$ and zero outside.

(3) Exponential: $f(x) = \lambda e^{-\lambda x}$, for $x \geq 0, \lambda > 0$.[5] (memoryless)

(4) Bilateral exponential: $f(x) = \frac{1}{2} \lambda e^{-\lambda |x|}$ for $x \in \mathbb{R}, \lambda > 0$.

(5) Cauchy distribution: It is the distribution of a ratio of two standard normal variables. The density is:

$$f(x) = \frac{\theta}{\pi(x^2 + \theta^2)}$$

This distribution is 'pathological' in the sense that it has no moment (everything, mean, variance, etc. is undefined).

[5] This is the continuous analogue of geometric distribution. It describes the time between events in a Poisson process.

5 Characteristic Function

Definition (Characteristic Function)**.** *Let X be a $d-$dimensional RV with $F(x_1, ..., x_d) = \mathbb{P}(X_1 \leq x_1, ..., X_d \leq x_d)$. The* ***characteristic function*** *$\psi_X : \mathbb{R}^d \mapsto \mathbb{C}$ of X is given by:*

$$\psi_X(\mathbf{u}) = \mathbb{E}\left[e^{i\mathbf{u}^T\mathbf{X}}\right] = \int_{\mathbb{R}^d} e^{i\mathbf{u}^T\mathbf{X}} dF(\mathbf{x})$$

Some properties of characteristic function:

Lemma 5.1. *Let X be a $d-$dimensional RV. Then:*

(1) $|\psi_X(u)| \leq 1$

(2) $\psi_X(-u) = \psi_{-X}(u) = \overline{\psi_X(u)}$

(3) If ψ_X is real-valued, then X and $-X$ has the same distribution.

(4) $\psi_X(u)$ is uniformly continuous in $\mathbb{R}^d$.[6]

(5) If $Y = aX + b$, then the characteristic function of Y is $\psi_Y = e^{iu^T b}\psi_X(au)$.[7]

(6) The characteristic function of $Z = X + Y$ where X, Y are independent RVs is $\psi_Z(u) = \psi_X(u)\psi_Y(u)$.

5.1 Common Characteristic Function

(1) Dirac delta measure $\mathbb{P}[X = x] = 1$: $\psi_X(u) = \exp(iu^T x)$.

(2) Bernoulli: $\mathbb{P}[X = 0] = 1 - p$, $\mathbb{P}[X = 1] = p$. Then: $\psi_X(u) = \mathbb{E}[e^{iuX}] = (1-p)e^{iu \cdot 0} + pe^{iu \cdot 1} = (1-p) + pe^{iu}$.

(3) Poisson distribution ($f(n) = e^{-\lambda}\frac{\lambda^n}{n!}$):[8] $\psi_X(u) = \sum_{n=0}^{\infty} f(n)e^{iun} = e^{-\lambda}\sum_{n=0}^{\infty}\frac{[\lambda e^{iu}]^n}{n!} = e^{-\lambda(1-e^{iu})}$

(4) Gaussian distribution $X \sim \mathcal{N}(\mu, \sigma^2)$: $\psi_X(u) = \exp(iu\mu - \frac{1}{2}\sigma^2 u^2)$.
Standard normal ($\mu = 0, \sigma = 1$), then $\mu_X(u) = e^{-\frac{1}{2}u^2}$.

5.2 Inversion Formula

One way to check if a probability measure admits a density is to see if the characteristic function is integrable. If that's the case, the theorem below helps uncover the original density:

Theorem 5.2 (Inversion Formula)**.** *Let $\psi_X(u) = \int e^{iux}d\mu$ such that $\int_{\mathbb{R}} |\psi_X(u)|du < \infty$, then a density $f(x)$ exists and satisfies*

$$f(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-iux}\psi(u)du$$

5.3 Levy's Continuity Theorem

We have that the distribution of a random variable X is uniquely determined by the characteristic function ψ_x due to the inversion vormula. This applies to a vector of RVs as well.

Theorem 5.3 (Levy's Continuity Theorem)**.** *Let X_n be RVs defined on probability spaces $(\Omega_n, \mathcal{F}_n, P_n)$ with characteristic function $\psi_n(u)$, $n \geq 1$. Then convergence in distribution of $(X_n)_{n\geq 1}$ implies pointwise convergence in the characteristic functions, and vice versa. More rigorously:*

(1) If $X_n \to^d X$ for some RV X, then $\psi_n(u) \to \psi(u)$ for all $u \in \mathbb{R}$.

(2) If $\psi_n(u) \to \psi(u)$ for all $u \in \mathbb{R}$, and $\psi(u)$ is continuous at $u = 0$, then ψ is the characteristic function of some RV X, and $X_n \to^d X$.

---

[6] Quick proof: Fix $\epsilon > 0$. $|\psi_X(u+t) - \psi_X(u)| = \mathbb{E}[\exp(iu^T X)(\exp(it^T X) - 1)] \leq |\mathbb{E}[\exp(it^T X) - 1] \to 0$ as $t \to 0$

[7] $\psi_Y(u) = \mathbb{E}[\exp(iu^T Y)] = \mathbb{E}[[\exp(i(au)^T X)]] = \exp(iu^T b)\psi_X(au)$

[8] Recall that $e^x = \sum_{n=0}^{\infty}\frac{x^n}{n!}$

5.4 Central Limit Theorem

Theorem 5.4 (CLT). *Let $X_1, X_2, ..., X_n$ be RVs on $(\Omega, \mathcal{F}, P)$ such that $\mathbb{E}[X^2] < \infty$ and $\sigma_n = \sqrt{Var(X_n)} > 0$ for $\forall\, n$. Then, $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mathbb{E}[X_i]}{\sigma(X_i)} \to^d Z$ where $Z \sim \mathcal{N}(0,1)$.*

6 Dynamical System

Definition. *A random vector $X \in \mathbb{R}^d$ is **normal** (or **Gaussian**) if and only $\psi_X(u) = \exp\left(iu^T\mu - \frac{1}{2}u^TCu\right)$ for some $\mu \in \mathbb{R}^d$ and C symmetric positive semidefinite. Notation: $X \sim \mathcal{N}_d(\mu, C)$.*

Note that when $d = 1$ we get the characteristic function $\psi_X(u) = \exp(iu\mu - \frac{1}{2}\sigma^2u^2)$ of a normal distribution mean μ and variance σ^2. **Importantly, if C is not invertible, then the density doesn't exist.**

Proposition 6.1. *Let $\mu \in \mathbb{R}^d$ and $C \in \mathbb{R}_{d\times d}$ be a symmetric positive semidedinite matrix. Then:*

(1) There exists A such that $A^2 = C$.[9] [10] *If $Z \sim \mathcal{N}_d(0, I_d)$, then $X = \mu + AZ \sim \mathcal{N}_d(\mu, C)$.*

(2) The components of X are independent if and only if they are uncorrelated.[11]

(3) If C is invertible, then X has a density:

$$f(x) = \frac{1}{(2\pi)^{d/2}\sqrt{det(C)}} \exp\left\{\frac{-1}{2}(X-\mu)^TC^{-1}(X-\mu)\right\}$$

If C is not-invertible, then A is not invertible, and $A\mathbb{R}^d$ is a strict subspace of $\mathbb{R}^d$, and X cannot have a density.

(4) For all $V \in \mathbb{R}^k$ and $M \in \mathbb{R}_{k\times d}$, $Y = V + MX \sim \mathcal{N}_d(V + M\mu, MCM^T)$

Proposition 6.2. *A $d-$dimensional random vector X is normal if and only if v^TX is normal for all $v \in \mathbb{R}^d$.*

6.1 Gaussian Process

A Gaussian process is a process in which any finite set samples in time (or space) is normally distributed. Formally:

Definition (Gaussian Process). *Let I be a non-empty set. A family $(X_i)_{i\in I}$ of RVs on a probability space is called a **Gaussian process** if $(X_{i_1}, X_{i_2}, ..., X_{i_d})$ is a finite dimensional Gaussian for all d and $i_1, i_2, ..., i_d \in I$.*

As we move to a (possibly) continuous index set I (for example, time $[0, T]$), we need to extend the notion of covariance matrix C from the last session.

Definition (Symmetric, positive semidefiniteness of function). *A function $C : I^2 \mapsto \mathbb{R}$ is **symmetric** if $C(i,j) = C(j,i)$ for all $i, j \in I$. It is **positive semidefinite** if $[C(i_k, i_l)]_{k,l=1}^d$ is positive semidefinite for all d and $(i_1, ..., i_d) \in I^d$.*

For a Gaussian process $(X_i)_{i\in I}$, denote $\mu(i) = \mathbb{E}[X_i]$ and $C^X(i,j) = cov(X_i, X_j)$ for $i, j \in I$. Then, C^X is symmetric and positive semidefinite by definition (since any finite sample follows multivariate normal, its covariance matrix must be positive semidefinite). However, provided a mean function $\mu^X(i)$ and covariance function $C^X(i,j)$, a Gaussian process with that mean and covariance function is guaranteed to exist:

[9] Linear algebra note: C is symmetric, then the followings are equivalent: (1) C is positive semidefinite; (2) C has all nonnegative eigenvalues; (3) all principal minors of C are nonnegative; and (4) There exists A such that $A^TA = C$.

[10] Connection between definiteness and principal minors: Let D_k and Δ_k denotes **leading principal minor** and **principal minor** of order k respectively. Positive definiteness = $D_k > 0$ for all D_k; negative definiteness = $(-1)^kD_k > 0$ for all D_k; positive semidefiniteness = $Delta_k \geq 0$ for all Δ_k; negative semidefiniteness = $(-1)^k\Delta_k \geq 0$ for all Δ_k.

[11] Generally, zero correlation does not imply independence. However, if two variables are jointly normal, then correlation **does** imply independence. Note however that two variables which are separately normal may still not have this property, since there is no guarantee that the joint distribution is normal.

Theorem 6.3 (Existence of Gaussian process). *Let I be a non-empty set, $\mu : I \mapsto \mathbb{R}$ an arbitrary function, and $C : I^2 \mapsto \mathbb{R}$ an arbitrary symmetric and positive semidefinite function. Then there exists a Gaussian process $(X_i)_{i\in I}$ with $\mu^X = \mu$ and $C^X = C$.*

Examples of Gaussian processes:

(1) White noise: Gaussian process with $\mu^X(i) = 0$ for $\forall\, i \in I$ and $C^X(i,j) = 1_{i=j}$ for $\forall\, i, j \in I$. [12]

(2) Brownian motion: $\mu^X(i) = 0$ for $\forall i$, and $C^X(i,j) = (i \wedge j)$ for $\forall\, i, j \in I$. Properties of a Brownian motion:

- $E[X_0^2] = 0 \wedge 0 = 0 \Rightarrow X_0 = 0$ a.s.
- Increments are normally distributed: For $t > s$, $\begin{pmatrix} X_s \\ X_t \end{pmatrix} \sim \mathcal{N}_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} s & s \\ s & t \end{pmatrix}\right)$, so $X_t - X_s = \begin{pmatrix} -1 & 1 \end{pmatrix} \begin{pmatrix} X_s \\ X_t \end{pmatrix} \sim \mathcal{N}(0, t-s)$.
- Increments are independent: For $s < t < u < v$:$cov(X_v - X_u, X_t - X_s) = cov(X_v, X_t) + cov(X_u, X_s) - cov(X_v, X_s) - cov(X_u, X_t) = t + s - s - t = 0$. Since increments are distributed jointly Gaussian, this also implies they are independent.

7 Martingales

7.1 Conditional Expectation

Definition (Absolute Continuous Measure). *Let μ_1, μ_2 be two measures on $(\Omega, \mathcal{F})$. μ_2 is said to be absolutely continuous wrt μ_1 if $\mu_1[A] = 0$ implies $\mu_2[A] = 0$ for all $A \in \mathcal{F}$.*[13] *Notation: $\mu_1 >> \mu_2$.*

Note that $\mu_2 = \int_A f d\mu_1 = \mathbb{E}[1_A f]$ defines a measure that is absolutely continuous wrt μ_1. This means that given a function, then we can generate a corresponding absolutely continuous measure. The other direction is guaranteed by the Radon-Nikodyn theorem, which posits that given a measure, we can find such a function f:

Theorem 7.1 (Radon-Nikodym). *Let μ_1, μ_2 be 2 measures on $(\Omega, \mathcal{F})$ such that $\mu_2 << \mu_1$ and μ_1 is $\sigma-$finite. Then, there exists $f : \Omega \mapsto \mathbb{R}_+ \cup \{\infty\}$ such that $\mu_2[A] = \int_A f d\mu_1 = \mathbb{E}[1_A f]$ for all $A \in \mathcal{F}$.*

We use the remark above and the R-N theorem to guarantee the existence of a conditional expectation which is defined below:

Definition (Conditional Expectation). *Let $(\Omega, \mathcal{F}, P)$ be a probability space, and $\mathcal{G}$ a sub-algebra of $\mathcal{F}$. The conditional expectation of $X \in L^1(\Omega, \mathcal{F}, P)$ **given** $\mathcal{G}$ is a RV Y which satisfies:*

- *Y is $\mathcal{G}-$measurable, $Y \in L^1(\Omega, \mathcal{G}, P)$.*
- *$\mathbb{E}[1_A Y] = \mathbb{E}[1_A X]$ for all $A \in \mathcal{G}$.*

Notation: $Y = \mathbb{E}[X|\mathcal{G}]$. We also define expectation wrt another RV: $\mathbb{E}[X|Z] = \mathbb{E}[X|\sigma(Z)]$

Proposition 7.2. *Such a conditional expectation always exists uniquely (up to a.e. equal).*

The idea is that: (1) given $\mathcal{F}$, P, and X, we can generate another measure $P_2 = \mathbb{E}[1_A f] << P$ for $\forall\, A \in \mathcal{G}$ (and $P_2[A] = 0$ for $A \notin \mathcal{G}$). This measure, therefore, is limited onto $\mathcal{G}$ only. Then, applies Radon-Nikodym on $(\Omega, \mathcal{G})$ guarantees the existence of $Y \in L^1(\Omega, \mathcal{G})$ such that $P_2[A] = \mathbb{E}[1_A Y]$ for all $A \in \mathcal{G}$. It follows that $\mathbb{E}[1_A Y] = \mathbb{E}[1_A X]$ for all $A \in \mathcal{G}$.

Properties of conditional expectation:

[12] This means that the variance of the noise at each point in time is $var(X_i) = 1$, and the noises are independent: $cov(X_i, X_j) = 0$ (note: they are distributed normally, so zero correlation implies independence).

[13] We can understand that μ_2 is a more "narrowly defined" measure, and the expectation on mu_2 will be the conditional expectation viewed from μ_1.

Proposition 7.3 (Properties of Conditional Expectation)**.** *We have:*

(1) *(monotonicity)* $\mathbb{E}[X|\mathcal{G}] \geq \mathbb{E}[Y|\mathcal{G}]$ *if* $X \geq Y$ *a.e.*

(2) *(Pulling out what's known)* $\mathbb{E}[XY|\mathcal{G}] = Y\mathbb{E}[X|\mathcal{G}]$ *if* Y *is* $\mathcal{G}-$*measurable.*

(3) *(Irrelevant information)* $\mathbb{E}[X|\mathcal{G}] = \mathbb{E}[X]$ *if* X *is independent of* $\mathcal{G}$.

(4) *(Jensen's Inequality)* $\mathbb{E}[\varphi(X)|\mathcal{G}] \geq \varphi(\mathbb{E}[X|\mathcal{G}])$ *for* $\varphi : \mathbb{R} \mapsto \mathbb{R}$, $\varphi \in L^1(\Omega, \mathcal{F}, P)$ *convex.*

(5) *(Conditional MCT) If* $Y \leq X_1 \leq X_2 \leq ...$ *for* $Y, X_1, X_2, ... \in L^1(\Omega, \mathcal{F}, P)$ *then* $\lim_{n\to\infty} \mathbb{E}[X_n|\mathcal{G}] = \mathbb{E}[\lim_{n\to\infty} X_n|\mathcal{G}]$ *a.s.*

(6) *(Conditional Fatou) If* $X_n \geq Y$ *for some* $Y \in L^1(\Omega, \mathcal{F}, P)$*, then* $\mathbb{E}[\liminf_n X_n|\mathcal{G}] \leq \liminf_n \mathbb{E}[X_n|\mathcal{G}]$.

(7) *(Conditional DCT) If* $|X_n| \leq Y$ *for some* $Y \in L^1(\Omega, \mathcal{F}, P)$*, then* $\lim_{n\to\infty} \mathbb{E}[X_n|\mathcal{G}] = \mathbb{E}[\lim_{n\to\infty} X_n|\mathcal{G}]$.

7.2 Martingales

Definition. $(X_n)_{n\geq 0}$ *is a martingale wrt* $(\mathcal{F}_n)_{n\geq 0}$ *if* $X_n \in L^1(\Omega, \mathcal{F}_n, P)$ *and* $\mathbb{E}[X_{n+1}|\mathcal{F}_n] = X_n$.

Definition. $(X_n)_{n\geq 0}$ *is a supermartingale wrt* $(\mathcal{F}_n)_{n\geq 0}$ *if* $X_n \in L^1(\Omega, \mathcal{F}_n, P)$ *and* $\mathbb{E}[X_{n+1}|\mathcal{F}_n] \leq X_n$.

Definition. $(X_n)_{n\geq 0}$ *is a submartingale wrt* $(\mathcal{F}_n)_{n\geq 0}$ *if* $X_n \in L^1(\Omega, \mathcal{F}_n, P)$ *and* $\mathbb{E}[X_{n+1}|\mathcal{F}_n] \geq X_n$.

Proposition 7.4. $X_n = \mathbb{E}[X|\mathcal{F}_n]$ *defines a uniformly integrable martingale.*

Definition (Martingale Transform)**.** *Let* $(X_n)_{n\geq 0}$ *and* $(V_n)_{n\geq 0}$ *be 2 stochastic processes on* $(\Omega, \mathcal{F}, P)$*. Define* $\Delta X_n = X_n - X_{n-1}$ *for* $n \geq 1$*. Define a new process* $(V \cdot X)_n = \begin{cases} \sum_{i=1}^n V_i \Delta X_i & \text{if } n \geq 1 \\ 0 & \text{if } n = 0. \end{cases}$

A process $(V \cdot X)_n$ such as above is called a Martingale transform. If (X_n) is a martingale, and (V_n) is predictable, then the transform is also a martingale. More rigorously:

Theorem 7.5. *Let* $(X_n)_{n\geq 0}$ *is a martingale wrt* $(F_n)_{n\geq 0}$*,* $(V_n)_{n\geq 0}$ *is predictable wrt* $(F_n)_{n\geq 0}$*. If* $(V \cdot X)_n$ *is* L^1-*bounded (i.e.* $\sup_n \mathbb{E}|(V \cdot X)_n| < \infty$*), then* $(V \cdot X_n)_{n\geq 0}$ *is a* $(\mathcal{F}_n)-$*martingale.*

Definition (Stopping Time[14])**.** *A* ***stopping time*** *wrt* $(\mathcal{F}_n)_{n\geq 0}$ *on* $(\Omega, \mathcal{F}, P)$ *is a RV* $\tau : \Omega \mapsto \mathbb{N} \cup \{+\infty\}$ *such that*

$$\{\tau = n\} = \{\omega \in \Omega : \tau(\omega) = n\} \in \mathcal{F}_n$$

Definition (Stopping Algebra)**.** *Given a stopping time* τ*, the algebra of that stopping time is the collection of events such that for each event, the decision to stop for each state of that event depends solely on the information available up to then:* $(\mathcal{F}_\tau) = \{A \in \mathcal{F} : A \cap \{\tau \leq n\} \in \mathcal{F}_n\}$*. It is easy to verify that this is a* $\sigma-$*algebra.*

Lemma 7.6. *If* τ, σ *are stopping times, then so are* $\tau + \sigma$*,* $\tau \wedge \sigma$*,* $\tau \vee \sigma$*.*
Also, $\mathcal{F}_{\tau\wedge\sigma} = \mathcal{F}_\tau \cap \mathcal{F}_\sigma$*. If* $\tau \leq \sigma$*, then* $\mathcal{F}_\tau \subset \mathcal{F}_\sigma$*.*

Below we define X_n^τ as a **stopped process**: it takes value of X_n as normal until the stopping time, then stay constant thereafter. This process will be shown to be a martingale also.

Lemma 7.7. $(X_n)_{n\geq 0}$ *a stochastic process adapted to* $(\mathcal{F}_n)_{n\geq 0}$*, and* τ *is a* $(\mathcal{F}_n)-$*stopping time. Then:*

- *the random variable* $X_\tau 1_{\{\tau<\infty\}}$ *is* $\mathcal{F}_\tau-$*measurable.*
- $X_n^\tau = X_{n\wedge\tau}$ *is* $(\mathcal{F}_n)-$*adapted.*

Corollary 7.8. *Let* X_n *be* $(\mathcal{F}_n)$*-martingale,* τ *a* $(\mathcal{F}_n)-$*stopping time. Then* $(X_n^\tau)_{n\geq 0}$ *is a martingale wrt* $(\mathcal{F}_n)_{n\geq 0}$*. In particular, if* τ *is bounded, then* $\mathbb{E}[X_\tau] = \mathbb{E}[X_0]$*.*

[14] We can think of $\tau(\omega)$ as a stopping decision: the decision to stop at any time τ is based solely on the information up to τ. This definition is equivalent to $\{\tau \leq n\} \in \mathcal{F}_n$ for all n.

This implies that the expected value of any stopped process under a bounded stopping time is the same as the expected value of the process at time 0. All this result boils down to a major theorem: **Doob's Optional Stopping theorem**. This theorem posits that given a martingale process, a bounded stopping time in which the decision to stop in every state depends only the information up to that time. Then, that stopping time is **optional** in the sense that it will not alter the expected value at stopping time. More formally:

Theorem 7.9 (Doob's Optional Stopping). *Let $(X_n)_{n\geq 0}$ be a $(\mathcal{F}_n)-$martingale, and $\sigma < \tau \leq N$ two bounded stopping times wrt $(\mathcal{F}_n)_{n\geq 0}$. Then $\mathbb{E}[X_\tau|\mathcal{F}_\sigma] = X_\sigma$.*
Taking $\sigma(\omega) = 0$ for $\forall\, \omega$, we have $\mathbb{E}[X_\tau] = X_0$.

Theorem 7.10 (Doob's Decomposition Theorem). *Let $(X_n)_{n\geq 0}$ be a sub-martingale wrt $(\mathcal{F}_n)_{n\geq 0}$. Then there exists a unique decomposition $X_n = M_n + A_n$ in which $(M_n)_{n\geq 0}$ is a martingale and $(A_n)_{n\geq 0}$ is a non-decreasing predictable process.*

Before, we have that transforming a martingale using a predictable process yields another martingale. We now transform submartingale:

Corollary 7.11 (Submartingale Transformation). *Let $(X_n)_{n\geq 0}$ be a submartingale, $(V_n)_{n\geq 0}$ be a non-negative predictable process wrt $(\mathcal{F}_n)_{n\geq 0}$, and $\mathbb{E}[|(V \cdot X)_n|] < \infty$ for $\forall\, n$. Then $(V \cdot X_n)_{n\geq 0}$ is a submartingale.*

Proposition 7.12. *Let $(X_n)_{n\geq 0}$ be a submartingale wrt $(\mathcal{F}_n)_{n\geq 0}$, and $\varphi : \mathbb{R} \mapsto \mathbb{R}$ convex, $\varphi(X_n) \in L^1$ for all n. If X_n is a martingale or φ is non-decreasing, then $\varphi(X_n)$ is a martingale.*

Let $(X_n)_{n\geq 0}$ be a submartingale wrt $(\mathcal{F}_n)_{n\geq 0}$ and an interval $[a, b]$ Let $\beta_n(a, b)$ denotes the number of times $(X_n)_n$ crosses b from below before n. It is easy to show that the time in which $(X_n)_n$ crosses from below is a stopping time. We also have an upper bound on the $\beta_n(a, b)$:

Theorem 7.13 (Upcrossing Inequality). *If $(X_n)_{n\geq 0}$ is a submartingale, then $\mathbb{E}[\beta_n(a, b)] \leq \frac{\mathbb{E}[(X_n - a)^+]}{b-a}$.*

This upcrossing inequality is used to prove the following convergence theorem of sub-martingale:

Theorem 7.14 (Sub/super-martingale Convergence). *$(X_n)_{n\geq 0}$ is L^1-bounded submartingale. Then there exists X_∞ such that $X_n \to X_\infty$ a.s., and $\|X_\infty\|_{L^1} \leq sup_n \|X_n\|_{L^1}$*
Similarly, an L^1-bounded supermartingale converges almost surely.

Corollary 7.15. *If $(X_n)_{n\geq 0}$ is a uniformly integrable submartingale, then $X_n \to X_\infty$ a.s. and in L^1. More over, $\mathbb{E}[X_\infty|\mathcal{F}_n] \geq X_n$ for all n.*
Similarly, a uniformly integrable supermartingale converges a.s. and in L^1.

Corollary 7.16. *A submartingale $(X_n)_{n\geq 0}$ that is bounded above converges a.s. Similarly, a supermartingale that is bounded from below converges a.s.*

Corollary 7.17. *Similarly, an L^p-bounded super/sub-martingale $(p > 1)$ converges a.s. and in L^1.*[15]

Corollary 7.18 (Convergence of Conditional Expectation Martingale). *Let $X \in L^1(\Omega, \mathcal{F}, P)$ and $(\mathcal{F}_n)_{n\geq 0}$ a filtration. Denote $X_n = \mathbb{E}[X|\mathcal{F}_n]$, and $X\infty = \mathbb{E}[X|\mathcal{F}_\infty]$ where $\mathcal{F}_\infty = \bigcup_n \mathcal{F}_n = \sigma(\bigcup_n \mathcal{F}_n)$. Then $X_n \to X_\infty$ a.s. and in L^1.*

Theorem 7.19 (Reverse Martingale Convergence). *Let $\mathcal{G}_0 \supset \mathcal{G}_1 \supset ...$ be a decreasing sequence of $\sigma-$algebras on a probability space $(\Omega, \mathcal{F}, P)$ and $X \in L^1(\Omega, \mathcal{F}, P)$. Then $\mathbb{E}[X|\mathcal{G}_n] \to \mathbb{E}[X|\mathcal{G}]$ a.s. and in L^1, where $\mathcal{G} = \bigcap_{n\geq 0} \mathcal{G}_n$.*

The Reverse Martingale Convergence theorem is useful to prove the strong law of large number.

Lemma 7.20. *Let $X_1, X_2, ...$ be i.i.d. RV in L^1 and define $S_n \sum_{i=1}^n X_i$. Then $\mathbb{E}[X_i|S_n] = \frac{S_n}{n}$.*

Theorem 7.21 (STRONG LAW OR LARGE NUMBER). *Let $X_1, X_2, ...$ be i.i.d. RV in $L^1(\Omega, \mathcal{F}, P)$. Then $\frac{1}{n}\sum_{i=1}^n X_i \to \mathbb{E}[X_i]$ a.s. and in L^1.*

[15] This is straightforward, because L^p-bounded for $p > 1$ implies uniform integrability.

Theorem 7.22 (Doob's Maximal Inequality for Probabilities)**.** *Let $(X_n)_{n\geq 0}$ be a sub-martingale and $\lambda > 0$. Then:*

$$\lambda \mathbb{P}[\max_{k\leq n} X_k \geq \lambda] \leq \mathbb{E}[X_n 1_{\{\max_{k\leq n} X_k \geq \lambda\}}] \leq \mathbb{E}[X_n^+]$$

Corollary 7.23. *If $(X_n)_{n\geq 0}$ is a martingale, then $\mathbb{P}[\max_{k\leq n} |X_k| \geq \lambda] \leq \frac{\mathbb{E}|X_n|^p}{\lambda^p}$.*

Theorem 7.24 (Doob's L^p maximal inequality)**.** *Let $(X_n)_{n\geq 0}$ be a non-negative submartingale. Then: $\| \max_{k\leq n} X_k\|_{L^p} \leq \frac{p}{p-1}\|X_n\|_{L^p}$ and $\| \max_{k\leq n} X_k\|_{L^1} \leq \frac{e}{e-1}(1 + \|X_n \log^+ X_n\|_{L^1})$.*

8 Stochastic Kernel and Regular Conditional Probabilities

Definition. *Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be two measurable spaces. A mapping $K : \Omega \times \Omega' \mapsto \mathbb{R}_+ \cup \{\pm\infty\}$ is a stochastic kernel from $(\Omega, \mathcal{F})$ to $(\Omega', \mathcal{F}')$ if:*

- *$\omega \to K(\omega, A)$ is $\mathcal{F}-$measurable for $\forall\, A \in \mathcal{F}'$.*
- *$A \to K(\omega, A$ is a measure on $(\Omega', \mathcal{F}')$ for all $\omega \in \Omega$.*

9 Markov Chains

Let I be a countable state space. A vector $(\lambda_i)_{i\in I}$ is a distribution if $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$. A matrix $(P_{ij})_{i,j\in I}$ is called stochastic if all its rows are distributions.

Definition. *A stochastic process $(X_n)_{n\geq 0}$ on a probability space $(\Omega, \mathcal{F}, P)$ is a Markov chain with initial distribution λ and transition matrix p if:*

(i) $\mathbb{P}[X_0 = i] = \lambda_i$

(ii) $\mathbb{P}[X_{n+1} = i_{n+1} | X_n = i_n, ..., X_0 = i_0] = p_{i_n i_{n+1}}$

Notation: $(X_n)_{n\geq 0}$ is Markov(λ, p).

Theorem 9.1 (Equivalent definition of Markov Chain)**.** *A discrete time stochastic process $(X_n)_{n\geq 0}$ is Markov(λ, p) if and only if $\mathbb{P}[X_{n+1} = i_{n+1} | X_n = i_n, ..., X_0 = i_0] = \lambda_{i_0} p_{i_0 i_1} ... p_{i_{n-1} i_n}$ for all $n \geq 0$ and $i_0, i_1, ..., i_n \in I$.*

Theorem 9.2 (Markov Property)**.** *Let $(X_n)_{n\geq 0}$ be the Markov(λ, p). If $\mathbb{P}[X_m = i] > 0$, then conditional on $X_m = i$, $(X_{m+n})_{n\geq 0}$ is Markov(δ_i, p) and is independent of $(X_0, ..., X_m)$*

Note that the state space I decomposes into **communicating classes**. A communicating class $C \subset I$ is called **closed** if $i \to j$ for $i \in C$ implies $j \in C$. $i \in I$ is called absorbing if $\{i\}$ is a closed communicating class. A Markov chain or transition matrix is called **irreducible** if I is a communicating class.

Definition. *Let $A \subset I$. Let $H^A = \inf\{n : X_n \in A\}$ be the hitting time of the subset A. $h^A = \mathbb{P}_i[H^A < \infty]$ the probability that A will be reached in finite time, and $k^A = \mathbb{E}_i[H^A]$ the expected hitting time given that we start at i.*

Theorem 9.3. *The vector of hitting probabilities $(h_i^A)_{i\in I}$ is the minimal non-negative solution to the system*

$$\begin{cases} 1 & \text{if } i \in A \\ \sum_{j\in I} p_{ij} h_j^A & \text{if } i \notin A \end{cases}$$

Theorem 9.4. *The vector of hitting time expectation $(k_i^A)_{i\in I}$ is the minimal non-negative solution to the system of linear equations:*

$$\begin{cases} 1 & \text{if } 0 \in A \\ 1 + \sum_{j\in I} p_{ij} k_j^A & \text{if } i \notin A \end{cases}$$

Theorem 9.5 (Strong Markov Property). *Let $(X_n)_{n\geq 0}$ be Markov(λ, p) and a stopping time τ wrt $(\mathcal{F}_n^X)_{n\geq 0}$. Let $i \in I$ be such that $\mathbb{P}[\tau < \infty$ and $X_\tau = i] > 0$. Then, conditional on $\tau < \infty$ and $X_\tau = i$, $(X_{\tau+n})$ is Markov(δ_i, p) and independent of $\mathcal{F}_\tau^X$.*

Definition (Recurrent and Transient State)**.** *A state is **recurrent** if $\mathbb{P}_i[X_n = i$ for infinitely many $n] = 1$. A state is **transient** if $\mathbb{P}_i[X_n = i$ for infinitely many $n = 0$.*

Define the hitting times $\tau_i^0 = 0$, $\tau_i^1 = \tau_i = \inf\{n \geq 1 : X_n = i\}$, $\tau_i^2 = \inf\{n > \tau_i^1 : X_n = i\}$, etc. Also, define the time difference between two hittings $S_i^k = \tau_i^k - \tau_i^{k-1}$ if $\tau_i^{k-1} < \infty$ and 0 otherwise.

Lemma 9.6. *For $k = 2, 3, ...$, conditional on $\tau_i^{k-1} < \infty$, S_i^k is independent of $\mathcal{F}^X_{\tau_i^{k-1}}$ and $\mathbb{P}[S_i^k = n | \tau_i^{k-1} < \infty] = \mathbb{P}_i[\tau_i = n]$.*

This lemma is a direct consequence of the Strong Markov Property.
Define $V_i = \sum_{n\geq 0} 1_{X_n = i}$ (number of visits to i), and $f_i = \mathbb{P}_i[\tau_i < \infty]$.

Lemma 9.7. $\mathbb{E}_i[V_i] = \sum_{n\geq 0} p_{ii}^{(n)}$. *Also, $\mathbb{P}_i[V_i > k] = f_i^k$, for $k = 1, 2, ...$*

Theorem 9.8 (Determining Recurrent or Transient). *If $\mathbb{P}_i[\tau_i < \infty] = 1$, then i is recurrent and $\sum_{n\geq 0} p_{ii}^n = \infty$. If $\mathbb{P}_i[\tau_i < \infty] < 1$, then i is transient and $\sum_{n\geq 0} p_{ii}^n < \infty$. In particular, every state is either transient or recurrent.*

Theorem 9.9. *The states of a given communicating class are either all recurrent or transient.*

Therefore, one can speak of a recurrent/transient class.

Theorem 9.10. *Every recurrent class is closed.*

Theorem 9.11. *Every finite closed class is recurrent.*

Theorem 9.12. *If p is an irreducible and recurrent. Then, for every initial distribution λ, $\mathbb{P}[\tau_j < \infty] = 1$ for all $j \in I$.*

9.1 Equilibrium Distribution

Definition. *A measure λ on I is said to be **invariant** wrt a transition matrix p if $\lambda p = \lambda$.*

Theorem 9.13. *If $(X_n)_{n\geq 0}$ is Markov(λ, p) and λ is invariant wrt p, then for every $m \geq 1$, $(X_{m+n})_{n\geq 0}$ is again Markov(λ, p).*

Theorem 9.14. *Let I be finite. Assume there exists $i \in I$ such that $p_{ij}^n \to \Pi_j$ as $n \to \infty$ for all j. Then $(\Pi_j)_{j\in I}$ is an invariant distribution.*

Definition. *For $i, k \in I$, set $\gamma_i^k \triangleq \mathbb{E}_k[\sum_{n=0}^{\tau_k - 1} 1_{X_n = i}]$ (starting at k, number of visits to i before visiting k again).*

Theorem 9.15. *Let p be an irreducible recurrent transition matrix. Then:*

(1) $\gamma_k^k = 1$.

(2) $(\gamma_i^k)_{i\in I}$ *solves* $\gamma^k p = \gamma^k$

(3) $0 < \gamma_i^k < \infty$ *for all* $i \in I$.

Theorem 9.16. *Let p be an irreducible transition matrix and λ an invariant measure such that $\lambda_k = 1$. Then $\lambda \geq \gamma^k$. In addition, if p is recurrent, then $\lambda = \gamma^k$.*