Appendix B

Probability Theory

B.1 Measurability and Measures

B.1.1 Information Structure

The *information structure* of probability models provides the basic building block of the theory. It formalizes the notion of "event spaces" or "states of nature" usually denoted by Ω, and the notion of "events", being those objects that we want to describe within the model. Events must be included in the information structure so that we may assign likelihoods or probabilities to them. If the model admits an event, it must also admit its complement. This characteristic is best described with the notion of a σ-field, usually denoted by the symbol $\mathfrak{F}$. We now summarize the basic definitions for a (general) measurable space and random variables.

Definition B.1 *Let $S \neq \emptyset$ be a set. A* σ-field *$\mathcal{S}$ on S is a collection of subsets of S with the following properties:*

(a) $S \in \mathcal{S}$,

(b) if $A \in \mathcal{S}$, *then* $A^c \in \mathcal{S}$, *where* $A^c = \{s \in S : s \notin A\}$, *and*

(c) if $A_i \in \mathcal{S}$, *for* $i \in \mathbb{N}$, *then* $\bigcup_{i \in \mathbb{N}} A_i \in \mathcal{S}$.

Elements of a σ-field are often referred to as "events" in probability theory.

Definition B.2 *Let $\mathcal{A}$ denote a collection of subsets of S. The σ-field* generated *by $\mathcal{A}$ is the smallest σ-field that contains $\mathcal{A}$.*

Definition B.3 *Let $(S, \mathcal{T})$ be a topological space. The* Borel field *of S, denoted by $\mathcal{B}$, is the σ-field generated by the collection of open sets $\mathcal{T}$, in formula: $\mathcal{B} = \sigma(\mathcal{T})$.*

Definition B.4 *Given a set $S \neq \emptyset$ and a σ-field $\mathcal{S}$ on S the pair $(S, \mathcal{S})$ is called a* measurable space.

Definition B.5 *Let $(S, \mathcal{S})$ and $(R, \mathcal{R})$ be two measurable spaces. A mapping $g : S \to R$ is said to be* measurable *(w.r.t. $\mathcal{S}$) if for any $A \in \mathcal{R}$ it holds true that $\{s \in S : g(s) \in A\} \in \mathcal{S}$. A measurable mapping is also called a* random variable.

Definition B.6 *The* σ-field generated *by a random variable* X *on a measurable space* $(\Omega, \mathfrak{F})$*, denoted* $\sigma(X)$ *is the smallest* σ*-field with respect to which* X *is measurable. Similarly, if* $X_1, \ldots, X_n$ *are random variables defined on a common space* $(\Omega, \mathfrak{F})$*, then* $\sigma(X_1, \ldots, X_n)$ *is the smallest* σ*-field w.r.t which every* $X_i; i = 1, \ldots, n$ *is measurable.*

Result: Suppose that X is a random variable on $(\Omega, \mathfrak{F})$. If $\sigma(X) \subset \mathcal{G} \subset \mathfrak{F}$, then X is also $\mathcal{G}$-measurable.

The notion of time dynamics is captured with the appropriate information structure that defines the "arrow of time". The history of the process is represented by having more refined σ-fields as time increases. Knowledge increases only in one direction, that of increasing time. That is, all of the events that are described at any given time are also described in future times.

Definition B.7 *Given a set* $T \subset \mathbb{R}$ *(or* $T \subset \mathbb{N}$*) and a measurable space* $(\Omega, \mathfrak{F})$ *a* filtration $\mathbb{F}$ *is a sequence of increasing* σ*-fields* $\mathbb{F} = \{\mathfrak{F}_t; t \in T\}$ *satisfying* $\mathfrak{F}_s \subset \mathfrak{F}_t \subset \mathfrak{F}$*, for every* $s < t; s, t \in T\}$*.*

Definition B.8 *Given a measurable space* $(\Omega, \mathfrak{F})$ *and a filtration* $\mathbb{F} = \{\mathfrak{F}_t; t \in T\}$ *on it, a* stochastic process $\{X_t; t \in T\}$ *is a collection of random variables on the common measurable space, where for each* $t \in T$*,* X_t *is* $\mathfrak{F}_t$*-measurable. The particular case where* $\mathfrak{F}_t = \sigma(X_t)$ *is called the* natural filtration *of the process.*

Definition B.9 *A* random stopping time τ *adapted to the filtration* $\{\mathfrak{F}_t; t \in T\}$ *on* $(\Omega, \mathfrak{F})$ *is a random variable on* $(\Omega, \mathfrak{F})$ *that satisfies:*

$$\{\omega : \tau(\omega) \leq t\} \in \mathfrak{F}_t; \quad \forall t \in T.$$

The σ*-field* $\mathfrak{F}_\tau$ *contains all events* A *such that* $A \cup \{\tau(\omega) \leq t\} \in \mathfrak{F}_t$*.*

REMARK. Notice that all the concepts related to measurability are related to the information structure given to the model, and they are independent of the concept of "measure", to which we come now. In other words, random variables and stochastic processes, event spaces and filtrations are defined without need to introduce any probability measure.

B.1.2 Measures

Definition B.10 *A* measure μ *on a measurable space* $(S, \mathcal{S})$ *is a mapping* $\mu : \mathcal{S} \to \mathbb{R} \cup \{-\infty, \infty\}$ *such that for any sequence* $\{A_n\}$ *of mutually disjoint elements of* $\mathcal{S}$ *it holds that*

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu\Big(A_n\Big) .$$

The measure m on $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ denotes the Borel field on $\mathbb{R}$, assigning $m((a, b]) = b - a$ to an interval $(a, b]$ is called *Lebesgue measure*. It generalizes the notion of length in geometry and is the case closest to everyday intuition.

Definition B.11 *The collection* $(S, \mathcal{S}, \mu)$ *is called* measure space*.*

Definition B.12 *A measure* μ *is called* signed *if* $\mu(A) < 0$ *for some* $A \in \mathcal{S}$ *and otherwise it is called* non-negative*. Furthermore, a measure* μ *is called* finite *if* $\mu(A) \in \mathbb{R}$ *for any* $A \in \mathcal{S}$*. We denote the set of signed measures on* $(S, \mathcal{S})$ *by* $\mathcal{M}$*. A non-negative measure* μ *is called* σ-finite *if there exist countably many sets* A_i *in* $\mathcal{S}$ *such that* $\mu(A_i) < \infty$ *and* $\bigcup_i A_i = S$*.*

Let $\mu \in \mathcal{M}$ be non-negative. Then for any measurable mapping $g : S \to \mathbb{R}$ the *μ-integral* of g, denoted by

$$< g, \mu >= \int_S g(s)\, \mu(ds)\,,$$

is defined although it may take values in $\{-\infty, \infty\}$. In particular, for any $A \in \mathcal{S}$,

$$< 1_A, \mu >= \mu(A)\,,$$

where $1_A : S \to \mathbb{R}$ is defined by $1_A(s) = 1$ for $s \in A$ and $1_A(s) = 0$ otherwise.

For any signed measure $\mu \in \mathcal{M}$ a measurable set S_μ^+ exists such that, for any $A \in \mathcal{S}$, it holds that $\mu(A \cap S_\mu^+) \geq 0$, whereas $\mu(A \cap (S \setminus S_\mu^+)) \leq 0$, see, for example, Proposition IV.1.1 in **[?]** for a proof. The positive part of μ is defined by

$$[\mu]^+(A) \;=\; \mu(A \cap S_\mu^+)\,, \quad A \in \mathcal{S}$$

and the negative part by

$$[\mu]^-(A) \;=\; -\mu(A \cap (S \setminus S_\mu^+))\,, \quad A \in \mathcal{S}\,.$$

The pair $([\mu]^+, [\mu]^-)$ is called *Hahn-Jordan decomposition*. The absolute measure $|\mu|$ is defined by $|\mu| = [\mu]^+ + [\mu]^-$. Integration with respect to a signed measure is defined by

$$< g, \mu >=< g, [\mu]^+ > \; - < g, [\mu]^- >$$

and integration with respect to an absolute measure is defined by

$$< g, |\mu| >=< g, [\mu]^+ > \; + < g, [\mu]^- >\,, \tag{B.1}$$

provided that the terms on the right-hand side of the above formulas are finite. The Hahn-Jordan decomposition is unique in the sense that if $\hat{G}$ is another set, such that $\mu(A \cap \hat{G}) \geq 0$ and $\mu(A \cap \hat{G}^c) \leq 0$ for any $A \in \mathcal{S}$, then $\mu(A \cap G) \;=\; \mu(A \cap S_\mu^+)$ for any $A \in \mathcal{S}$. A signed measure $\mu \in \mathcal{M}$ is finite if $[\mu]^+(S)$ and $[\mu]^-(S)$ are finite.

Definition B.13 *A* probability measure *μ is a non-negative measure such that $\mu(S) = 1$ (which already implies that $\mu(\emptyset) = 0$). If μ is a probability measure on $(S, \mathcal{S})$, then the collection $(S, \mathcal{S}, \mu)$ is called* probability space.

Theorem B.1 (Continuity Theorem for decreasing (increasing) events) *Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a probability space. If $\{A_n\}$ is a sequence of decreasing (inreasing)events in $\mathfrak{F}$ (that is, $A_{n+1} \subset A_n$ in case of decreasing, and $A_n \subset A_{n+1}$ in case of increasing), then $\mathbb{P}(\lim_n A_n) = \lim_n \mathbb{P}(A_n)$.*

Definition B.14 *Let μ and ν be σ-finite measures on a measurable space $(S, \mathcal{S})$. The measure μ is said to be* absolutely continuous *with respect to ν (denoted $\mu << \nu$) if $\nu(A) = 0$, for $A \in \mathcal{S}$, implies $\mu(A) = 0$.*

Theorem B.2 (Radon-Nikodym) *Two σ-finite measures on a measurable space $(\Omega, \mathfrak{F})$ satisfy $\mu << \nu$ if, and only if, there exists a non-negative measurable mapping $d\mu/d\nu : S \to \mathbb{R}$ such that*

$$\mu(A) \;=\; \int_A \frac{d\mu}{d\nu}(s)\, \nu(ds)\,, \quad \textit{for every } A \in \mathfrak{F}\,.$$

*The random variable $d\mu/d\nu$ is called ν-*density of *μ, or* Radon-Nikodym derivative.

REMARK. In the particular case that $\nu = m$ is the Lebesgue measure on $\mathbb{R}$ and μ is a probability measure, the Radon-Nikodym derivative is the usual probability density function (p.d.f.)

Let $(S, \mathcal{S}, \mu)$ and $(T, \mathcal{T}, \nu)$ be probability spaces. The product of μ and ν on $S \times T$, denoted by $\mu \times \nu$, is a measure such that

$$\forall A \in \mathcal{S}, B \in \mathcal{T} : \quad (\mu \times \nu)(A \times B) = \mu(A)\,\nu(B)$$

and *Fubini's theorem* states that

$$\begin{aligned} \int_{S\times T} f(s,t)\,(\mu\times\nu)(ds,dt) &= \int_T \left(\int_S f(s,t)\,\mu(ds)\right)\nu(dt) \\ &= \int_S \left(\int_T f(s,t)\,\nu(dt)\right)\mu(ds)\,, \end{aligned}$$

for any measurable mapping $f : S \times T \to \mathbb{R}$.

Definition B.15 *Let $(S, \mathcal{S}, \mu)$ be a measure space and $g : S \to \mathbb{R}$ a measurable mapping from $(S, \mathcal{S})$ to $(\mathbb{R}, \mathcal{R})$, where $\mathcal{R}$ is a σ-field over $\mathbb{R}$. The* induced measure *of g, denoted by μ^g, is defined as follows*

$$\mu^g(A) = \mu\Big(\{s \in S : g(s) \in A)\}\Big)\,, \quad A \in \mathcal{R}\,.$$

Definition B.16 *The cumulative distribution function (c.d.f.) of a real-valued random variable X defined on a probability space $(S, \mathcal{S}, \mu)$ is the function $F : [-\infty, \infty] \to [0, 1]$, where*

$$F(x) = \mu^X(\,(-\infty, x]\,)\,, \quad -\infty \le x \le \infty\,.$$

We take the domain $[-\infty, \infty]$ since it is natural to assign the values 0 and 1 to $F(-\infty)$ and $F(\infty)$. A c.d.f. has the decomposition $F(x) = F'(x)+F''(x)$, where $F'(x)$ is positive only on a set of Lebesgue measure zero, and $F''(x)$ is absolutely continuous with respect to the Lebesgue measure. The Radon-Nikodym derivative of F'' with respect to the Lebesgue measure exists and is called probability density function (p.d.f.). If f is the p.d.f. of the c.d.f. F, then it holds that $f(x) = dF(x)/dx$ except for a set of Lebesgue measure zero.

Definition B.17 *Two random variables X, Y on a common probability space $(\Omega, \mathfrak{F})$ said to be are equal* almost surely *(a.s., or w.p.1) if $\mathbb{P}(\omega\colon X(\omega) = Y(\omega)) = 1$. Two random variables X and Y not necessarily defined on a common probability space are said to be equal em in distribution (denoted $X \stackrel{\mathcal{L}}{=} Y$) if $\mathbb{P}(X \le x) = \mathbb{P}(Y \le x)$ for every $x \in \mathbb{R}$.*

Theorem B.3 (Skorokhod Representation: random variables) *Given a real-valued random variable X on a probability space $(\Omega, \mathfrak{F}, \mathbb{P})$ there exists a real-valued random variable $\tilde{X}$ on the* canonical *probability space $([0, 1], \mathcal{B}([0, 1], m)$ (where m is the Lebesgue measure) such that $X \stackrel{\mathcal{L}}{=} \tilde{X}$.*

Let μ be a finite measure on $(S, \mathcal{S})$, where S is a locally compact Hausdorff space (see any book on functional analysis for definitions) and $\mathcal{S}$ contains the Borel field on S. The measure μ is called *regular* if

$$\mu(A) = \inf\{\mu(U) : U \text{ open in } S, A \subset U\}\,, \quad A \in \mathcal{S}\,,$$

and for any open set $U \subset S$ it holds

$$\mu(U) = \sup\{\mu(F) : F \text{ is compact in } S, F \subset U\}\,.$$

B.2 Expectations and Conditioning

Definition B.18 *Let $\mathbb{P}$ be a probability measure and let X be a random variable on $(\Omega, \mathfrak{F})$. The* expectation *of X is defined as the integral:*

$$\mathbb{E}(X) = \int_\Omega X(\omega)\,\mathbb{P}(d\omega) = \int_{\mathbb{R}} xF(dx),$$

where F is the c.d.f. of X.

Definition B.19 *Let X be a random variable on $(\Omega, \mathfrak{F}, \mathbb{P})$. The* variance *of X is defined as:*

$$\mathsf{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2.$$

Theorem B.4 (Markov Inequality) *Let X be a non-negative random variable on the probability space $(\Omega, \mathfrak{F}, \mathbb{P})$. Then for any positive real number $a \in \mathbb{R}^+$*

$$\mathbb{P}(X > a) \leq \frac{\mathbb{E}(X)}{a}.$$

A corollary to this theorem is known as Chebyshev's inequality:

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq a) \leq \frac{\mathsf{Var}(X)}{a^2}.$$

Theorem B.5 (Wald's Equality) *Let $\{X(n)\}$ be an i.i.d. sequence such that $\mathbb{E}[X(1)]$ is finite. Furthermore, let η be a non-negative integer-valued random variable with finite mean. If for all $m \geq 0$ the event $\{\eta = m\}$ is independent of $\{X(m+n) : n \geq 1\}$, then*

$$\mathbb{E}\left[\sum_{i=1}^{\eta} X(i)\right] \;=\; \mathbb{E}[\eta]\,\mathbb{E}[X(1)] \;<\infty\,.$$

Definition B.20 *Let X be a random variable on the probability space $(\Omega, \mathfrak{F}, \mathbb{P})$ such that $\mathbb{E}(|X|) < \infty$, and let $\mathcal{G} \subset \mathfrak{F}$ be another σ-field on Ω. The* conditional expectation *of X given $\mathcal{G}$, is a random variable Z on satisfying:*

(a) Z is $\mathcal{G}$-measurable,

(b) For all $B \in \mathcal{G}$, $\mathbb{E}(Z\mathbf{1}_{\{B\}}) = \mathbb{E}(X\mathbf{1}_{\{B\}})$.

We often denote Z by $\mathbb{E}(X \,|\, \mathcal{G})$.

Definition B.21 *Given two random variables on $(\Omega, \mathfrak{F}, \mathbb{P})$ the* conditional expectation *of X given Y is defined as $\mathbb{E}(X \,|\, Y) = \mathbb{E}(X \,|\, \sigma(Y))$.*

Theorem B.6 *[Summary of Properties of Conditional Expectations] Consider a probability space $(\Omega, \mathfrak{F}, \mathbb{P})$ and sub σ-algebras $\mathcal{G}_i \subset \mathfrak{F}$.*

(a) If X is $\mathcal{G}$-mbl then $\mathbb{E}(X \,|\, \mathcal{G}) = X$ a.s.

(b) If $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \mathfrak{F}$ then $\mathbb{E}(\mathbb{E}(X \,|\, \mathcal{G}_1) \,|\, \mathcal{G}_2) = \mathbb{E}(\mathbb{E}(X \,|\, \mathcal{G}_2) \,|\, \mathcal{G}_1) = \mathbb{E}(X \,|\, \mathcal{G}_1)$.

(c) $\mathbb{E}(X \,|\, \{\Omega, \emptyset\}) = \mathbb{E}(X)$.

(d) For any real valued function h such that $\mathbb{E}(h(X)) < \infty$, $\mathbb{E}(\mathbb{E}(h(X) \,|\, \mathcal{G})) = \mathbb{E}(h(X))$.

(e) If Y is $\mathcal{G}$-mbl and both $\mathbb{E}(X)$ and $\mathbb{E}(Y)$ exist, then $\mathbb{E}(XY \,|\, \mathcal{G}) = Y\,\mathbb{E}(X \,|\, \mathcal{G})$.

(f) **Jensen's inequality:** *if h is a convex function and $\mathbb{E}|h(X)| < \infty$ then $h\,(\mathbb{E}(X)) \leq \mathbb{E}(h(X)) \quad a.s.$*

B.3 Polish Spaces

I think this section should go to the appendix 1.

Let S be a nonempty set with zero element 0_S. A *norm* is a mapping $\|\cdot\| : S \to [0,\infty)$ having the properties (i) $0 < \|x\| < \infty$ for $x \neq 0_S$ and $\|0_S\| = 0$, (ii) $\|\alpha x\| = |\alpha|\,\|x\|$ for $\alpha \in \mathbb{R}$ and (iii) $\|x+y\| \leq \|x\| + \|y\|$ (triangle inequality), for any $x, y \in S$.

A *metric* is a mapping $d : S \times S \to [0,\infty)$ having the properties (i) $d(x,y) = d(y,x)$, (ii) $d(x,y) = 0 \Leftrightarrow x = y$, and (iii) $d(x,z) \leq d(x,y) + d(y,z)$ (triangle inequality), for any $x, y, z \in S$. If (i) and (iii) hold but $d(x,y) = 0$ is possible when $x \neq y$, we call d a *pseudo metric*. A *metric space* (S,d) is a set S paired with metric d.

An *open set* of (S,d) is a set $A \subset S$ such that, for each $s \in A$, $\delta > 0$ exists such that $\{x \in S : d(x,s) < \delta\} \subset A$. The collection of open subsets of S is denoted by $\mathcal{T}(d)$. Hence, $(S, \mathcal{T}(d))$ is a *topological space*. The Borel field on a metric space (S,d) is the σ-field generated by $\mathcal{T}(d)$.

A metric is said to be *complete* if the metric space (S,d) is complete, that is, if the limiting point of any Cauchy sequence in S lies in S. If there is a countable collection of open subsets of $\mathcal{T}(d)$ such that any open subset of S can be written as union of these sets, then $\mathcal{T}(d)$ is said to have a *countable basis*. A topological space $(S,\mathcal{T})$ is called a *Polish space* if (i) its topology is defined by a complete metric (that is, there exists a metric d such that $\mathcal{T} = \mathcal{T}(d)$) and (ii) $\mathcal{T}$ has a countable basis.

B.4 Convergence of random sequences

Let X, X_n, $n \geq 0$, be real-valued random variables defined on a common probability space $(\Omega, \mathfrak{F}, \mathbb{P})$ with state space $S \subset \mathbb{R}$ and let S be equipped with the Borel field $\mathcal{B}(S)$.

B.4.1 Types of Convergence

When we talk about "convergence" there must be an implicit norm defined on an appropriate space to study "how close" the sequence gets to the limit. For random variables, there are several ways to define the notion of being "close". Recall that two random variables may be equal in distribution yet they may be defined on entirely different measurable spaces. The following are the various most common concepts of convergence for random sequences.

Definition B.22 *The sequence* $\{X_n\}$ *converges* almost surely *(or w.p.1) to* X *as* n *tends to* ∞ *if for any* $\delta > 0$

$$\mathbb{P}\Big(\lim_{n\to\infty} \sup_{m\geq n} |X_m - X| > \delta \Big) = 0$$

and yet another equivalent condition is that the event $\Big\{ \lim_{n\to\infty} X_n = X \Big\}$ *has probability one. This is denoted* $X_n \to X$ *a.s.*

Using the continuity theorem for decreasing events, if $X_n \to X$ a.s., then it holds that $\lim_{n\to\infty} \mathbb{P}\Big(\sup_{m\geq n} |X_m - X| > \delta \Big) = 0$.

Definition B.23 *The sequence* $\{X_n\}$ *converges* in probability *to* X *as* n *tends to* ∞ *if for any* $\delta > 0$

$$\lim_{n\to\infty} \mathbb{P}(\,|X_n - X| \geq \delta\,) = 0\,,$$

or, equivalently,

$$\lim_{n\to\infty} \mathbb{P}\Big(|X_n - X| > \delta \Big) \;=\; 0\,.$$

This is denoted $X_n \xrightarrow{P} X$.

Result: Almost sure convergence of $\{X_n\}$ to X implies convergence in probability of $\{X_n\}$ to X. On the other hand, convergence in probability of $\{X_n\}$ to X implies a.s. convergence of a subsequence of $\{X_n\}$ to X.

Definition B.24 *The sequence* $\{X_n\}$ *converges* in the r-th mean *to* X *as* n *tends to* ∞ *if*

$$\mathbb{E}(|X_n - X|^r) \to 0.$$

This is denoted $X_n \xrightarrow{r} X$.

Theorem B.7 *If* $\lim_{m,n\to\infty} X_n - X_n = 0$ *a.s., or in the* r*-th mean, or in probability, then there exists a random variable* X *such that* $X_n \to X$ *in the same sense. That is, Cauchy convergence implies convergence.*

Definition B.25 *Let* $C_b(\mathbb{R})$ *denote the set of bounded continuous mapping from* S *onto* $\mathbb{R}$*. A sequence* $\{\mu_n\}$ *of measures on* S *is said to* converge weakly *to a distribution* μ *if*

$$\lim_{n\to\infty} \int_S f\, d\mu_n \;=\; \int_S f\, d\mu\,, \quad \text{for any } f \in C_b(\mathbb{R})\,.$$

Definition B.26 *Let* F_n *denote the distribution of* X_n *and* F *the distribution of* X*. If* $\{F_n\}$ *converges weakly to* F *as* n *tends to* ∞*, then we say that* $\{X_n\}$ converges in distribution *to* X*. Equivalently,*

$$F_n(x) \to F(x);\ \textit{for every point of continuity of } F.$$

This is denoted $X_n \xRightarrow{\mathcal{L}} X$.

Result: Convergence in probability implies convergence in distribution but the converse is not true.

Definition B.27 *The* total variation norm *of a (signed) measure* μ *on* S *is defined by*

$$||\mu||_{tv} \;=\; \sup_{\substack{f\in C_b(\mathbb{R}) \\ |f|\le 1}} \left| \int_S f\, d\mu \right|\,.$$

Definition B.28 *Let* μ_n *denote the distribution of* X_n *and* μ *the distribution of* X*. If* $\{\mu_n\}$ *converges in total variation to* μ *as* n *tends to* ∞ *(that is, if* $\lim_{n\to\infty} ||\mu_n - \mu||_{tv} = 0$*) then we say that* $\{X_n\}$ converges in total variation *to* X*. The convergence in total variation of* $\{X_n\}$ *to* X *can be expressed equivalently by*

$$\lim_{n\to\infty} \sup_{A\in\mathcal{S}} |\, \mathbb{P}(X_n \in A) \;-\; \mathbb{P}(X \in A)\,| \;=\; 0\,.$$

Result: Convergence in total variation implies convergence in distribution (or, weak convergence) but the converse is not true.

Theorem B.8 (Continuous Mapping Theorem) *Let $h : \mathbb{R} \to \mathbb{R}$ be measurable with discontinuity points confined to a set D_h, where $\mu(D_h) = 0$. If μ_n converges weakly towards μ as n tends to ∞, then μ_n^h tends to μ^h as n tends to ∞, or, equivalently,*

$$\lim_{n\to\infty} \int f(h(x))\, \mu_n(dx) \;=\; \int f(h(x))\, \mu(dx)\,, \quad f \in C_b(\mathbb{R})\,.$$

Hence, if $\{X_n\}$ converges weakly and h is continuous, then $\{h(X_n)\}$ converges weakly.

A random sequence may converge a.s. and yet not in expectation, as the following example illustrates. Consider the canonical probability space $([0,1], \mathcal{B}([0,1]), m)$ and let

$$X_n(\omega) = \begin{cases} n & \omega \in [0, 1/n) \\ 0 & \text{otherwise.} \end{cases}$$

and define $X(\omega) \equiv 0$. Clearly, for every $\omega > 0$ there is an integer n such that $X_n(\omega) = 0$. Because $m(0) = 0$, then the event $\{\omega\colon \lim_{n\to\infty} X_n(\omega) = 0\}$ has measure 1. Thus, $X_n \to X$ a.s. However, notice that for each n,

$$\mathbb{E}(X_n) = \int_0^1 X_n(\omega) n(d\omega) = \int_1^{1/n} nm(d\omega) \equiv 1,$$

so that $\mathbb{E}(X_n) \not\to \mathbb{E}(X) = 0$. The problem with the example stems from the fact that on very small sets the random variables X_n can become large without bound. The following theorem states the condition under which a.s. convergence ensures also convergence in expectation.

Add MEAN VALUE THEORM

Theorem B.9 (Dominated Convergence) *Let $(S, \mathcal{S}, \mu)$ be a probability space. Let $f_n : S \to \mathbb{R}$, for $n \in \mathbb{N}$, be measurable and assume that $f, g : S \to \mathbb{R}$ are measurable mappings such that, for any $n \in \mathbb{N}$, the set of points $s \in S$ with $|\, f_n(s)\,| \le g(s)$ and $\lim_{n\to\infty} f_n(s) = f(s)$ has μ-measure one. If*

$$\int_S |\, g(s)\,|\, \mu(ds) \;<\; \infty\,,$$

then

$$\lim_{n\to\infty} \int_S f_n(s)\, \mu(ds) \;=\; \int_S f(s)\, \mu(ds)\,.$$

Equivalently, if $X_n \to X$ a.s. and there is a random variable Z on $(\Omega, \mathfrak{F}, \mathbb{P})$ with $\mathbb{E}(Z) < \infty$ and $|X_n| \le Z$ for every n, then $\mathbb{E}(X_n) \to \mathbb{E}(X)$.

Theorem B.10 (Monotone Convergence) *Let $\{X_n\}$ be a sequence of random variables on $(\Omega, \mathfrak{F}, \mathbb{P})$ increasing a.s. to a limit, that is, $X_{n+1} \ge X_n$ a.s. and $X = \lim_{n\to\infty} X_n$ exists a.s. Then we can interchange the limit and expectation: $\lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(X)$.*

Recall (Chapter 2) that a set A in a normed space is called (sequentially) compact if any infinite sequence $\{a_n\}$ possesses at least one subsequence that converges to an element in the set, and all accumulation points belong to A. Also, Cauchy convergence and compactness imply convergence. These "compactness" arguments are very useful when trying to establish convergence of sequences

(notably our stochastic algorithms). Unfortunately the set of distribution functions is not compact. To see this, consider the distribution functions:

$$F_n(x) = \begin{cases} 0 & x < n \\ 1 & x \geq n, \end{cases}$$

so that $F_n(x) \to 0$ for every $x \in \mathbb{R}$. The limit function (zero) is not a probability distribution and no subsequence converges to a probability distribution. The problem here is that the "mass" where the probability is concentrated floats away to infinity. The following material establishes the conditions under which we can characterize the compact sets of distributions.

Definition B.29 *Let μ_n denote the measure on $\mathcal{B}(S)$ induced by the random variable X_n. The sequence of measures $\{\mu_n\}$ is said to be* tight *if*

$$\lim_{K\to\infty} \sup_n \mu_n([-K,K]^c) = \lim_{K\to\infty} \sup_n \mathbb{P}(|X_n| \geq K) = 0.$$

Sometimes we say that the sequence of distributions is mass preserving*, or that the sequence of random variables $\{X_n\}$ is* uniformly bounded in probability*. Equivalently, for each n and $\alpha > 0$ there are N_α, K_α such that*

$$\mathbb{P}(|X_n| \geq K_\alpha) \leq \alpha \quad \textit{for every } n \geq N_\alpha.$$

Theorem B.11 *If the sequence of probability measures $\{\mu_n\}$ is tight then it is relatively compact; that is, it contains at least one weakly convergent subsequence and all accumulation points are probability measures.*

Theorem B.12 (Skorokhod representation: convergence) *Consider the space $(\Omega, \mathfrak{F}, \mathbb{P})$ and let $X_n; X : \Omega \to S$ be S-valued random variables, where S is a complete separable metric space with metric $d(\cdot,\cdot)$. Suppose that $X_n \stackrel{\mathcal{L}}{\Longrightarrow} X$. Then there is a probability space $(\tilde\Omega, \tilde{\mathfrak{F}}, \tilde{\mathbb{P}})$ with associated random variables $\tilde X_n; \tilde X$ such that $X_n \stackrel{\mathcal{L}}{=} \tilde X_n$ for every n, $\tilde X \stackrel{\mathcal{L}}{=} X$ and*

$$\lim_{n\to\infty} d(\tilde X_n, X) = 0 \quad \tilde{\mathbb{P}} - a.s.$$

that is, $\tilde X_n \to \tilde X$ with $(\tilde{\mathbb{P}})$-probability one.

Definition B.30 *A sequence of random variables $\{X_n\}$ is* uniformly integrable *if*

$$\sup_n \lim_{K\to\infty} \mathbb{E}(|X_n| \mathbf{1}_{\{|X_n|>K\}}) = 0.$$

Definition B.31 *A continuous time stochastic process $\{\vartheta(t)\}$ is called a* cadlag *process if it is right-continuous with left limits at every point t. The space $D[0,\infty]$ is the space of piecewise constant cadlag processes with the Skorokhod topology (a modified sup norm, please see Billingsley for details).*

Definition B.32 *Let $X(t)$ be a continuous time stochastic process on $(\Omega, \mathfrak{F}, \mathbb{P})$ with natural filtration $\mathbb{F}$. we say that X is* locally Lipschitz continuous with probability 1 *if for each $T > 0$ there is a $\mathfrak{F}_T$-mbl random variable $K(T) > 0$ with finite expectation and such that*

$$|X(t+s) - X(t)| \leq K(T)s; \quad \textit{for all } t \leq t+s \leq T.$$

Theorem B.13 *Let $\{Y_n^\epsilon; n \geq 1, \epsilon > 0\}$ be a family of random variables on $(\Omega, \mathfrak{F}, \mathbb{P})$ which is uniformly integrable and define the piecewise constant function:*

$$\vartheta^\epsilon(t) = \vartheta^\epsilon(0) + \epsilon \sum_{n=1}^{\lfloor \epsilon/t \rfloor} Y_n^\epsilon.$$

Then the sequence $\{\vartheta^\epsilon\}$ of processes is tight in $D[0,\infty]$ and all its accumulation points (as $\epsilon \to 0$) ϑ^* are Lipschitz continuous with probability 1.

B.5 Weak Convergence and Norm Convergence

Let (S, d) be a separable metric space and denote the set of continuous real-valued mappings on S by $C(S)$. Let $v : S \to \mathbb{R}$ be a measurable mapping such that

$$\inf_{s \in S} v(s) \geq 1.$$

The set of mappings from S to $\mathbb{R}$ can be equipped with the so-called v-*norm* introduced presently. For $g : S \to \mathbb{R}$, the v-norm of g, denoted by $||g||_v$, is defined by

$$\|g\|_v \stackrel{\text{def}}{=} \sup_{s \in S} \frac{|g(s)|}{v(s)},$$

see, for example, **[?]** for the use of the v-norm in the theory of measure-valued differentiation of Markov chains. If g has finite v-norm, then $|g(s)| \leq c\, v(s)$ for any $s \in S$ and some finite constant c. For example, the set of real, continuous v-dominated functions, defined by

$$\mathcal{D}_v(S) \stackrel{\text{def}}{=} \{g \in C(S) \,|\exists c > 0 : |g(s)| \leq c v(s), \forall s \in S\}, \tag{B.2}$$

can be characterized as the set of all continuous mappings $g : S \to \mathbb{R}$ having finite v-norm. Note that $C^b(S)$ is a particular $\mathcal{D}_v(S)$-space, obtained for $v = const$. Moreover, the condition that $\inf_{s \in S} v(s) \geq 1$ implies that $C^b(S) \subset \mathcal{D}_v(S)$ for any choice of v.

The v-norm of a measure μ on $(S, \mathcal{S})$, with $\mathcal{S}$ the Borel-field with respect to the metric d, is defined through

$$||\mu||_v \stackrel{\text{def}}{=} \sup_{||g||_v \leq 1} \left| \int_S g(s)\, \mu(ds) \right|,$$

or, more explicitly,

$$||\mu||_v = \sup_{|g| \leq v} \left| \int_S g(s)\, \mu(ds) \right|.$$

In particular, it holds that

$$||\mu||_v = \int v(s)\, |\mu|(ds), \tag{B.3}$$

see (B.1). Let $\{\mu_n\}$ be a sequence of measures on $(S, \mathcal{S})$ and let μ be a measure on $(S, \mathcal{S})$. We say that μ_n converges in v-norm towards μ if

$$\lim_{n \to \infty} ||\mu_n - \mu||_v = 0.$$

It can be shown that the set $\mathcal{D}_v(S)$ endowed with the v-norm is a Banach space. This last remark indicates the following fact: For each measure μ with $\int v(s)\mu(ds)$ finite, the mapping $T_\mu : \mathcal{D}_v(S) \to \mathbb{R}$ defined through

$$T_\mu(g) \stackrel{\text{def}}{=} \int g d\mu,$$

is a continuous linear functional on the Banach space $\mathcal{D}_v(S)$ and the operator norm of T_μ satisfies $\|T_\mu\| = \|\mu\|_v$. The Cauchy-Schwartz inequality thus holds for v-norms, i.e.,

$$\left| \int g(s)\mu(ds) \right| \leq \|g\|_v \cdot \|\mu\|_v,$$

for all $g \in \mathcal{D}_v(S)$ and μ such that v is μ-integrable.

Let $\{\mu_n\}$ be a sequence of measures on $(S, \mathcal{S})$ and let μ be a measure on $(S, \mathcal{S})$. We say that μ_n converges weakly in $\mathcal{D}_v(S)$-sense towards μ if

$$\lim_{n\to\infty} \int_S g(s)\, \mu_n(ds) = \int_S g(s)\, \mu(ds),$$

for all $g \in \mathcal{D}_v(S)$; in symbols $\mu_n \stackrel{\mathcal{D}_v(S)}{\Longrightarrow} \mu$.

REMARK. Note that v-norm convergence implies $\mathcal{D}_v(S)$-convergence. This can be seen as follows. According to Cauchy-Schwartz Inequality, for each $g \in \mathcal{D}_v(S)$ it holds that:

$$\left| \int g(s)\mu_n(ds) - \int g(s)\mu(ds) \right| = \left| \int g(s)(\mu_n - \mu)(ds) \right| \leq \|g\|_v \cdot \|\mu_n - \mu\|_v.$$

Hence, $\|\mu_n - \mu\|_v \to 0$ implies that the left-hand side in the above relation converges to 0 as $n \to \infty$.

B.6 Martingale processes

Let $(\Omega, \mathfrak{F}, \mathbb{P})$ be a probability space and $\mathbb{F} = \{\mathfrak{F}_t; t \in T\}$ a filtration on it. Let $\{M_t; t \in T\}$ be a stochastic process adapted to $\mathbb{F}$ (we may without loss of generality consider the natural filtration of the process). When $T = \mathbb{R}^+$ it is a continuous time process, and when $T = Z^+$ it is a discrete time process.

Definition B.33 *If $\mathbb{E}(|M_t|) < \infty$ for all t, then we say that:*

- $\{M_t\}$ *is a* $\mathbb{F}$-martingale *if* $M_t = \mathbb{E}(M_{t+s} \,|\, \mathfrak{F}_t)$ *a.s. for all* $s \geq 0$.
- $\{M_t\}$ *is a* $\mathbb{F}$-submartingale *if* $M_t \leq \mathbb{E}(M_{t+s} \,|\, \mathfrak{F}_t)$ *a.s. for all* $s \geq 0$.
- $\{M_t\}$ *is a* $\mathbb{F}$-supermartingale *if* $M_t \geq \mathbb{E}(M_{t+s} \,|\, \mathfrak{F}_t)$ *a.s. for all* $s \geq 0$.

Vector-valued martingales are processes where each component is a martingale.

One of the simple examples of martingales is the "gambling" process where the chances for winning at each play are equal to the chances of losing. M_n denotes the current wealth of the player, who always bets the same amount (until $M_n = 0$ at which point the game is over). Unfair games are modeled with sub or supermartingales. More interesting to our discourse are the algorithms that seek a given direction. Consider for example a stochastic approximation driven by an unbiased gradient estimator:

$$\theta_{n+1} = \theta_n - \epsilon Y_n,$$

where $\mathbb{E}[Y_n \,|\, \mathfrak{F}_n] = \nabla J(\theta_n)$. Then using Taylor's expansion and calling $M_n = J(\theta_n)$ we have

$$M_{n+1} = M_n - \epsilon \|\nabla J(\theta_n)\|^2 + \epsilon^2 \, \xi_n,$$

where ξ_n is a random variable (measurable w.r.t. $\mathfrak{F}_n$). If we know that $\{\xi_n\}$ are uniformly bounded, then there is ϵ_0 such that for all $\epsilon \leq \epsilon_0$ the process M_n is a non-negative supermartingale.

Definition B.34 *Consider a martingale process $\{M_n\}$. The process $\{\delta M_n\}$, where $\delta M_n = M_{n+1} - M_n$ is called the* martingale difference process.

Proposition B.1 *If $\mathbb{E}[|M_n|^2] < \infty$ then the martingale differences are uncorrelated: $\mathbb{E}[\delta M_n \, \delta M_m] = 0$ for all $m \neq n$.*

Let Y be a $\mathfrak{F}$-measurable random variable with $\mathbb{E}|Y| < \infty$. Then the random process $M_n = \mathbb{E}(Y \,|\, \mathfrak{F}_n)$ form a martingale, because $\mathbb{E}(M_{n+1} \,|\, \mathfrak{F}_n) = \mathbb{E}(\mathbb{E}(Y \,|\, \mathfrak{F}_n)) = \mathbb{E}(Y \,|\, \mathfrak{F}_n) = M_n$. This is called *Doob's martingale* and it models the case where information about Y is "accumulated" as the number of observations grow.

Example: Let $\{Y_n\}$ be a sequence of random variables and construct the natural filtration with $\mathfrak{F}_n = \sigma(Y_1, \ldots, Y_n)$. Express;

$$Y_n = Y_n - \mathbb{E}[Y_n \,|\, \mathfrak{F}_{n-1}] + \mathbb{E}[Y_n \,|\, \mathfrak{F}_{n-1}] = \delta M_n + \mathbb{E}[Y_n \,|\, \mathfrak{F}_{n-1}].$$

The "unpredictable" term in Y_n is a martingale difference with corresponding martngale process given by:

$$M_n = \sum_{j=1}^{n} (Y_j - \mathbb{E}[Y_j \,|\, \mathfrak{F}_{j-1})].$$

The following proposition is a corollary to Kolmogorov's inequality for submartinagles.

Proposition B.2 *Let M_n be a square-integrable martingale with $c = \mathbb{E}(M_0)$. For any constant $\Delta > 0$ and any finite time N*

$$\mathbb{P}\left(\sup_{m \leq N} |M_m - c| \geq \Delta \right) \leq \frac{\mathsf{Var}(M_N)}{\Delta^2}.$$

Proposition B.3 *Let $\{M_n\}$ be a $\mathbb{F}$-martingale and $q(\cdot)$ a non-decreasing non-negative convex function (commonly, we use $q(x) = |x|, x^2$, or $e^{ax}, a > 0$.). Then for all $n < N$ and $\lambda > 0$*

$$\mathbb{P}\left(\sup_{n\leq m\leq N}|M_m|\geq\lambda\,|\,\mathfrak{F}_n\right)\leq\frac{\mathbb{E}(q(M_N\,|\,\mathfrak{F}_n)}{q(\lambda)}. \tag{B.4}$$

$$\mathbb{E}\left(\sup_{n\leq m\leq N}|M_m|^2\,|\,\mathfrak{F}_n\right)\leq 4\,\mathbb{E}\left(|M_n|^2\,|\,\mathfrak{F}_n\right) \tag{B.5}$$

$$\mathbb{P}\left(\sup_{n\leq m\leq N}M_m\geq\lambda\,|\,\mathfrak{F}_n\right)\leq\frac{M_n}{\lambda}. \tag{B.6}$$

The first expression is a generalization of Proposition B.2. The third expression is a generalization of the Markov inequality; it is known as Kolmogorov submartingale inequality and it is also true if the process is a submartingale.

Definition B.35 *Let $\{M_n\}$ be a $\mathbb{F}$-martingale and let τ be a $\mathbb{F}$-adapted stopping time. The* stopped martingale *$\{M_{\tau\wedge n}\}$ is defined by:*

$$M_{\tau\wedge n}=\begin{cases}M_n & n<\tau\\ M_\tau & n\geq\tau.\end{cases}$$

Proposition B.4 *Let τ be a $\mathbb{F}$-adapted stopping time, then*

- *If $\{M_n\}$ is a martingale then $\{M_{\tau\wedge n}\}$ is a martingale.*
- *If $\{M_n\}$ is a submartingale then $\{M_{\tau\wedge n}\}$ is a submartingale.*
- *If $\{M_n\}$ is a supermartingale then $\{M_{\tau\wedge n}\}$ is a supermartingale.*

In general, the final value of a stopped process depends mostly on the stopping criterion, for example if a stochastic process X_t is stopped at $\tau=\min(t\colon X_t=10)$ and we know that $\tau<N$ a.s. for some value of N, then $X_\tau=10$. So, in general, it is not true that $\mathbb{E}(X_\tau)=\mathbb{E}(X_0)$ for a martingale process. The following result establishes the conditions under which this is true.

Theorem B.14 (Optional sampling theorem) *Let $\{M_n\}$ be a discrete-time martingale and τ a stopping time with values in $\mathbb{N}\cup\{\infty\}$, both with respect to the filtration $\mathbb{F}$. Assume that one of the following three conditions holds:*

(a) The stopping time τ is almost surely bounded, i.e., there exists a constant $c\in\mathbb{N}$ such that $\tau\leq c$ a.s.

(b) The stopping time τ has finite expectation and the conditional expectations of the absolute value of the martingale increments are almost surely bounded, more precisely, $\mathbb{E}[\tau]<\infty$ and there exists a constant c such that $\mathbb{E}\big[|M_{t+1}-M_t|\,\big|\,\mathcal{F}_t\big]\leq c$ almost surely on the event $\{\tau>t\}$ for all $t\in\mathbb{N}$.

(c) There exists a constant c such that $|M_{\tau\wedge n}|\leq c$ a.s. for all $n\in\mathbb{N}$.

Then M_τ is an almost surely well defined random variable and $\mathbb{E}[M_\tau]=\mathbb{E}[M_0]$.

Similarly, if the stochastic process M is a submartingale or a supermartingale and one of the above conditions holds, then $\mathbb{E}[M_\tau]\geq\mathbb{E}[M_0]$, for a submartingale, and $\mathbb{E}[M_\tau]\leq\mathbb{E}[M_0]$, for a supermartingale.

Theorem B.15 *Let $\{M_n\}$ be a real-valued submartingale with $\sup_n \mathbb{E}[|M_n|] < \infty$. Then $\{M_n\}$ converges w.p.1 as $n \to \infty$. If $\{M_n\}$ is a real-valued supermartingale with $\mathbb{E}[\max(0, -M_n)] < \infty$ then supermartingale converges w.p.1.*

Theorem B.16 *A continuous time martingale $\{M_t\}$ whose paths are locally Lipschitz continuous w.p.1 on each bounded time interval is a constant w.p.1.*

In order to use any of the martingale convergence theorems, one must verify that the process is a martingale. The following is a useful characterization of a martingale.

Theorem B.17 (Martingale characterization) *Consider a process $\{\vartheta(t); t \geq 0\}$ on $(\Omega, \mathfrak{F}, \mathbb{P})$ adapted to a filtration $\mathbb{F} = \{\mathfrak{F}_t\}$. Let $\{M(t)\}$ be another $\mathbb{F}$-adapted process and suppose that for each $t, r \geq 0$ and for each $p \in \mathbb{N}, s_i \leq t, i = 1, \ldots, p$ and any bounded and continuous real valued function $h(\cdot)$ the following holds:*

$$\mathbb{E}\Big(h(\vartheta(s_i); i = 1, \ldots, p)\,(M(t+r) - M(t))\Big) = 0.$$

Then $\{M(t)\}$ is a martingale.

B.7 Regenerative Processes

Let $\{X(n)\}$ denote a stochastic process with state space $(S, \mathcal{S})$. A random time τ_k is called *stopping time* if the occurrence or non-occurrence of τ_k at time t is known from $\{X(n) : n \leq t\}$, that is, the event $\{\tau_k = t\}$ lies in $\sigma(\{X(n) : n \leq t\})$, for any t. The process $\{X(n)\}$ is called *classical regenerative*, or, *regenerative* if there exists a sequence of stopping times $\{\tau_k\}$ also called *regeneration times* such that

- $\{\tau_{k+1} - \tau_k\,,\ k \geq 0\}$ is an i.i.d. sequence;
- for every sequence of times $0 < t_1 < t_2 < \cdots < t_n$ and every $k \geq 0$, the random vectors $(X(t_1), X(t_2), \ldots, X(t_n))$ and $(X(\tau_k{+}t_1), X(\tau_k{+}t_2), \ldots, X(\tau_k{+}t_n))$ have the same distributions, and the processes $\{X(n) : n \leq \tau_k\}$ and $\{X(n) : n > \tau_k\}$ are independent.

Thus, in a regenerative process, the regeneration points $\{\tau_k : k \geq 0\}$ cut the process into independent and identically distributed *cycles* of the form $\{X(n) : \tau_k \leq n < \tau_{k+1}\}$. A distribution function is called *lattice* if it assigns probability one to a set of the form $\{0, \delta, 2\delta, \ldots\}$, for some $\delta > 0$, and it is called *non-lattice* otherwise.

Result: Let $\{X(n)\}$ be a regenerative process such that the distribution of $\tau_{k+1} - \tau_k$ is non-lattice. If, for a measurable mapping $f : S \to \mathbb{R}$, $\mathbb{E}[\sum_{n=\tau_1}^{\tau_2 - 1} f(X(n))]$ is finite, then

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(X(n)) \;=\; \frac{\mathbb{E}\left[\sum_{n=\tau_1}^{\tau_2-1} f(X(n))\right]}{\mathbb{E}[\tau_2 - \tau_1]} \quad \text{a.s.}$$