1. LaTeX Examples

2. Real Analysis

For purposes of our discussion of measure theory, we often make little use of the structure of the reals. In many cases it is with little effort that we can state results much more generally. Sometimes the results will be true of arbitrary sets but in other cases we need the most basic notions of metric spaces.

Definition 2.1. A metric space is a set S together with a function $d : SxS \to \mathbb{R}$ satisfying

(i) $d(x, y) = 0$ if and only if $x = y$.
(ii) For all $x, y \in S$, $d(x, y) = d(y, x)$.
(iii) For all $x, y, z \in S$, $d(x, z) \leq d(x, y) + d(y, z)$.

Lemma 2.2. *Given a metric space (S, d), we have $d(x, y) \geq 0$ for all $x, y \in S$.*

Proof. Let $x, y \in S$ and observe

$$\begin{aligned} d(x, y) &= \frac{1}{2}(d(x, y) + d(y, x)) \text{ by symmetry} \\ &\geq \frac{1}{2}d(x, x) \text{ by triangle inequality} \\ &= 0 \end{aligned}$$

□

It's pretty easy to see that standard notions of limits and continuity extend to the case of metric spaces.

Definition 2.3. A sequence of elements $x_n \in S$ converges to $x \in S$ if for every $\epsilon > 0$, there exists $N > 0$ such that $d(x_n, x) < \epsilon$ for all $n > N$.

Definition 2.4. A function between metric spaces $f : (S, d) \to (S', d')$ is continuous at $x \in S$ if for every $\epsilon > 0$, there exists $\delta > 0$ such that for $y \in S$ such that $d(x, y) < \delta$ we have $d'(f(x), f(y)) < \epsilon$. A function f that is continuous at all points $x \in S$ is said to be continuous.

Lemma 2.5. *$f : (S, d) \to (S', d')$ is continuous at $x \in S$ if and only if for every $x_n \to x$ we have $f(x_n) \to f(x)$.*

Proof. Suppose f is continuous and let $\epsilon > 0$ be given. By continuity, we can pick $\delta > 0$ such that for all $y \in S$ with $d(x, y) < \delta$ we have $d'(f(x), f(y)) < \epsilon$. Now by convergence of the sequence x_n, we can find N such that for all $n > N$, we have $d(x_n, x) < \delta$. Hence for all $n > N$, we have $d'(f(x), f(x_n)) < \epsilon$.

Now suppose that for every $x_n \to x$ we have $f(x_n) \to f(x)$. We argue by contradiction. Suppose f is not continuous at x. There exists $\epsilon > 0$ such that we can find $x_n \in S$ such that $d(x, x_n) < 2^{-n}$ and $d'(f(x_n), f(x)) \geq \epsilon$. Note that the sequence $x_n \to x$ but $f(x_n)$ doesn't converge to $f(x)$. □

Definition 2.6. For $x \in S$ and $r \geq 0$, the open ball at x or radius r is the set

$$B(x; r) = \{y \in S | d(x, y) < r\}$$

Definition 2.7. A set $U \subset S$ is open if for every $x \in U$ there exists $r > 0$ such that $B(x; r) \subset U$. The complement of an open set is called a closed set.

Lemma 2.8. *A set $A \subset S$ is closed if and only if for every $x_n \to x$ with $x_n \in A$, we have $x \in A$.*

Proof. Suppose A is closed. Then A^c is open. Let $x_n \in A$ converge to x. If $x \notin A$, then $x \in A^c$ and we can find an open ball $B(x;\epsilon) \subset A^c$. Pick $N > 0$ such that $d(x_n, x) < \epsilon$ for all $n > N$. Then $x_n \notin A$ for all $n > N$ which is a contradiction.

Now suppose A contains all of its limit points. We show that A^c is open. Let $x \in A^c$ and suppose the balls $B(x; 2^{-n}) \bigcap A \neq \emptyset$. Then we can construct a sequence $x_n \in A$ such that $x_n - > x$. This is a contradiction, hence for some n, we have $B(x; 2^{-n}) \bigcap A = \emptyset$ and therefore A^c is open. □

As it turns out continuity of a function can be expressed entirely in terms of open sets.

Lemma 2.9. *A function between metric spaces $f : (S, d) - > (T, d')$ is continuous if and only if for every open subset $U \subset T$, we have $f^{-1}(U)$ is an open subset of S.*

Proof. For the only if direction, let $U \subset T$ be an open set and pick $x \in f^{-1}(U)$. Now, $f(x) \in U$ and by openness of U we can find $\epsilon > 0$ such that $B(f(x);\epsilon) \subset U$. By continutity of f we can find a $\delta > 0$ such that for all $y \in S$ with $d(x, y) < \delta$ we have $d'(f(x), f(y)) < \epsilon$. This is just another way of saying $B(x;\delta) \subset f^{-1}(U)$ which shows that $f^{-1}(U)$ is open.

For the if direction, pick $x \in S$ and suppose we are given $\epsilon > 0$. The ball $B(f(x);\epsilon)$ is an open set in T. By assumption we know that $f^{-1}(B(f(x);\epsilon))$ is an open set in S containing x. By definition of openness, we can pick a $\delta > 0$, such that $B(x;\delta) \subset f^{-1}(B(f(x);\epsilon))$. Unwinding this statement shows that for all $y \in S$ with $d(x, y) < \delta$, we have $d'(f(x), f(y)) < \epsilon$ and we have show that f is continuous at x. Since $x \in S$ was arbitrary we have shown f is continuous on all of S. □

Definition 2.10. A sequence of elements $x_n \in S$ is said to be a *Cauchy sequence* if for every $\epsilon > 0$, there exists $N > 0$ such that $d(x_n, x_m) < \epsilon$ for all $n, m > N$.

Note that any convergent sequence is Cauchy.

Lemma 2.11. *If a sequence of elements $x_n \in S$ converges to $x \in S$ then it is a Cauchy sequence.*

Proof. Pick $\epsilon > 0$ and then pick $N > 0$ so that $d(x_n, x) < \frac{\epsilon}{2}$ for all $n > N$. Then by the triangle inequality, $d(x_n, x_m) \leq d(x_n, x) + d(x, x_m) < \epsilon$ for $n, m > N$. □

It is also easy to construct examples of Cauchy sequences that do not converge by looking at spaces with *holes*.

Example 2.12. Consider the sequence $\frac{1}{n}$ on $\mathbb{R} \setminus \{0\}$. It is Cauchy but does not converge.

The existence of non-convergent Cauchy sequences is in some sense the definition of what it means for a general metric space to have holes. This motivates the following definition.

Definition 2.13. A metric space (S, d) is said to be *complete* if every Cauchy sequence is convergent.

Definition 2.14. The real line $\mathbb{R}$ is complete.

Proof. Suppose we are given a Cauchy sequence x_n. Let $a = \liminf_{n\to\infty} x_n$ and $b = \limsup_{n\to\infty} x_n$. We proceed by contradiction and suppose that $a < b$ (note that the *completeness axiom* of the reals is used in the definition of $\liminf$ and $\limsup$). Let $M = b - a$ then for any $0 < \epsilon < M$, $N > 0$ we can find $k, m > N$ such that $|a - x_k| < \frac{M-\epsilon}{2}$ and $|b - x_m| < \frac{M-\epsilon}{2}$ thus showing $|x_k - x_m| \geq \epsilon$ and contradicting the assumption that x_n was a Cauchy sequence. □

The following is a simple fact about $\mathbb{R}$.

Lemma 2.15. *Let x_n be a nondecreasing sequence in $\mathbb{R}$. Suppose there is an infinite subsequence x_{n_k} such that $\lim_{k\to\infty} x_{n_k} = x$, then $\lim_{n\to\infty} x_n = x$.*

Proof. TODO: This is actually pretty much obvious. □

In our treatment of measure theory we'll want to have a detailed understanding of the structure of the topology of the real line. It can be described quite simply.

Lemma 2.16. *The open sets in $\mathbb{R}$ are precisely the countable unions of disjoint open intervals.*

Proof. Pick an open set $U \subset \mathbb{R}$. Define an equivalence relation on U such that $a \equiv b$ if and only if $[a, b] \subset U$ or $[b, a] \subset U$. It is easy to see this is an equivalence relation. Reflexivity and symmetry are entirely obvious. Transitivity follows from taking a union of intervals (carefully taking order into consideration).

Now, consider the equivalence classes of the relation. As equivalence classes these sets are disjoint and their union is U. Call the family of equivalence classes U_α.

We have to show that the equivalence classes are open intervals. Consider $x \in U_\alpha \subset U$. Openness of U_α follows from using openness of U to find a small ball (open interval) around $x \in U$ and noting that every point of the ball is $\equiv$-related to x. Therefore the same open ball demonstrates the openness of U_α.

To see that equivalence classes are intervals, pick an equivalence class U_α and consider the open interval $(\inf U_\alpha, \sup U_\alpha)$. Since U_α is nonempty and open, $\inf U_\alpha \neq \sup U_\alpha$ and this interval is non-empty. By definition of inf and sup and the openness of U_α we can see that $U_\alpha \subset (\inf U_\alpha, \sup U_\alpha)$ (otherwise we could find an element of U_α bigger than sup or less than inf). On the other hand, suppose we are given $x \in (\inf U_\alpha, \sup U_\alpha)$. We can find elements $y, z \in U_\alpha$ such that $\inf U_\alpha < y < x < z < \sup U_\alpha$. By definition of the equivalence relation, this shows $[y, z] \subset U_\alpha$ and therefore $x \in U_\alpha$. Therefore we have shown that $U_\alpha = (\inf U_\alpha, \sup U_\alpha)$ is an open interval.

The fact that there are at most countably many equivalence classes follows from the density and countability of $\mathbb{Q}$. □

Lemma 2.17. *Let $A \subset \mathbb{R}$ be a countable set. Then A^c is dense in $\mathbb{R}$.*

Proof. Pick an $x \in \mathbb{R}$ and consider an interval $I_n = (x - \frac{1}{n}, x + \frac{1}{n})$ for $n > 0$. Then if $A^c \cap I_n = \emptyset$ we have $I_n \subset A$ which implies that I_n is countable. This is clearly false (since otherwise we could write the reals as a countable union of countable sets which would imply the reals themselves are countable). □

Just as an aside at this point, we note that notions of open and closed set are really all that is needed to make sense of the notions of convergence and continuity.

Definition 2.18. A topological space is a set S together with a collection of subsets τ satisfying

(i) τ contains $\emptyset$ and S.
(ii) τ is closed under arbitrary union.
(iii) τ is closed under finite intersection.

The collection τ is called a topology on S. The elements of τ are called the open sets of S and the complement of the open sets are called closed sets. As we have shown above, if one defines continuity of a function between topological spaces as inverse images of open sets being open we have a definition that is a compatible generalization of the ϵ/δ definition of calculus.

Theorem 2.19 (Taylor's Formula)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be a function which is m-times continuously differentiable. Then for all $0 \leq n < m$,*

$$f(b) = \sum_{k=0}^{n} \frac{(b-a)^k}{k!} f^{(k)}(a) + R_n(b)$$

where the remainder term *is of the form*

$$R_n(b) = \int_a^b \frac{(b-x)^n}{n!} f^{(n+1)}(x)\, dx$$

Proof. We proceed by induction. Note that for $n = 1$, then Taylor's Formula simply says $f(b) = f(a) + \int_a^b f'(x)\, dx$ which is just the Fundamental Theorem of Calculus. For the induction step, we integrate the remainder term by parts. Consider the integral $\int_a^b \frac{(b-x)^{(n-1)}}{(n-1)!} f^{(n)}(x)\, dx$ and let $u = f^{(n)}(x)$ and $dv = \frac{(b-x)^{(n-1)}}{(n-1)!} dx$. Then $du = f^{(n+1)}(x)dx$ and $v = -\frac{(b-x)^n}{n!}$, so

$$\begin{aligned}\int_a^b \frac{(b-x)^{(n-1)}}{(n-1)!} f^{(n)}(x)\, dx &= -\frac{(b-x)^n}{n!} f^{(n)}(x) \mid_a^b + \int_a^b \frac{(b-x)^n}{n!} f^{(n+1)}(x)\, dx \\ &= \frac{(b-a)^n}{n!} f^{(n)}(a) + \int_a^b \frac{(b-x)^n}{n!} f^{(n+1)}(x)\, dx\end{aligned}$$

which proves the result. □

The version of Taylor's Formula above expresses the "integral form" of the remainder term. It is often useful to transform the remainder term in Taylor's Formula into the *Lagrange form*.

Lemma 2.20. *There is a number $c \in [a, b]$ such that $\int_a^b R_n(b) = f^{(n+1)}(c) \frac{(b-a)^{n+1}}{(n+1)!}$*

Proof. By continuity of $f^{(n+1)}(x)$ and compactness of $[a, b]$ we know that there exist $m, M \in \mathbb{R}$ such that $m = \min_{x\in[a,b]} f^{(n+1)}(x)$ and $M = \max_{x\in[a,b]} f^{(n+1)}(x)$. Therefore we have the bounds

$$\begin{aligned}m\frac{(b-a)^{(n+1)}}{(n+1)!} &= m \int_a^b \frac{(b-x)^n}{n!}\, dx \\ &\leq \int_a^b \frac{(b-x)^n}{n!} f^{(n+1)}(x)\, dx \\ &\leq M \int_a^b \frac{(b-x)^n}{n!}\, dx = M\frac{(b-a)^{(n+1)}}{(n+1)!}\end{aligned}$$

hence

$$m \leq \frac{(n+1)!}{(b-a)^{(n+1)}} \int_a^b \frac{(b-x)^n}{n!} f^{(n+1)}(x)\,dx \leq M$$

By continuity of $f^{(n+1)}(x)$ and the Intermediate Value Theorem, we know that $f^{(n+1)}(x)$ takes every value in $[m, M]$ and therefore there exists $c \in [a, b]$ such that $f^{(n+1)}(c) = \frac{(n+1)!}{(b-a)^{(n+1)}} \int_a^b \frac{(b-x)^n}{n!} f^{(n+1)}(x)\,dx$. □

Lemma 2.21. *Let X be a real normed vector space with a subspace Y of codimension 1. Then any bounded linear functional λ on Y extends to a bounded linear functional on X with the same operator norm.*

Proof. We first assume that λ has operator norm 1. Let v be any vector that is not in Y. Then every element of X is of the form $y + tv$, hence by linearity all we really have to choose is the value of $\lambda(v)$ so that the operator norm doesn't increase. First, note that it suffices to show $|\lambda(y + v)| \leq \|y + v\|$ for all y. For it that if that is true then

$$\begin{aligned} |\lambda(y + tv)| &= |t\lambda(y/t + v)| \\ &\leq |t|\|y/t + v\| \\ &= \|y + tv\| \end{aligned}$$

We rewrite the constraint $|\lambda(y + v)| \leq \|y + v\|$ for all y as

$$-\lambda(y) - \|y + v\| \leq \lambda(v) \leq \|y + v\| - \lambda(y)$$

To see that it is possible to satisfy the constraint derived above, we use the triangle inequality (subadditivity) of the operator norm. For all $y_1, y_2 \in Y$,

$$\begin{aligned} \lambda(y_1) - \lambda(y_2) &\leq |\lambda(y_1 - y_2)| \\ &\leq \|y_1 - y_2\| \\ &= \|y_1 + v - v - y_2\| \\ &\leq \|y_1 + v\| + \|y_2 + v\| \end{aligned}$$

From which we conclude by rearranging terms

$$\sup_{y_2 \in Y} -\lambda(y_2) - \|y_2 + v\| \leq \inf_{y_1 \in Y} \|y_1 + v\| - \lambda(y_1)$$

Picking any value between the two terms of the above inequality results in a valid extension. To handle the case of operator norm not equal to 1, notice that the extension is trivial if the operator norm is 0 (i.e. $\lambda = 0$), otherwise define the extension by $\|\lambda\|$ times the extension of $\lambda/\|\lambda\|$. □

Theorem 2.22 (Hahn-Banach Theorem (Real case))**.** *Let X be a real normed vector space with a subspace Y. Then any bounded linear functional λ on Y extends to a bounded linear functional on X with the same operator norm.*

Proof. We proceed by using the codimension 1 case proved above and then applying Zorn's Lemma. We define a partial extension of λ to be a pair (Y', λ') such that $Y \subset Y' \subset X$ and λ' is an extension of λ with the same operator norm. Put a partial order on the set of extensions by declaring $(Y', \lambda') \leq (Y'', \lambda'')$ if and only if $Y' \subset Y''$ and $\lambda'' \,|_{Y'} = \lambda'$.

To apply Zorn's Lemma, we need to show that every chain has an upper bound. If we are given a chain $(Y_\alpha, \lambda_\alpha)$ then we define $Z = \cup_\alpha Y_\alpha$ and for any $z \in Z$ we

define $\tilde{\lambda}(z) = \lambda_\alpha(z)$ for any α such that $z \in Y_\alpha$. It is immediate that this well defined. It is easy to show linearity and to show that $\|\tilde{\lambda}\| = \|\lambda\|$ (TODO: do this).

Now we can apply Zorn's Lemma to conclude that there is a maximal element (Y', λ'). The codimension one case show us that $Y' = X$ for otherwise we can construct an extension that shows (Y', λ') is not maximal. □

Note that the use of Zorn's Lemma here is not accidental; the Hahn Banach Theorem cannot be proven in set theory without the Axiom of Choice (though according to Tao it can be proven without the full power of the Axiom of Choice using what is know as the Ultrafilter Lemma).

2.1. **Compactness.**

Definition 2.23. Let (S, d) be a metric space, then we say $K \subset S$ is *sequentially compact* if and only if for every sequence $x_1, x_2, \cdots \in K$ there exists a convergent subsequence x_{n_j} such that $\lim_{j\to\infty} x_{n_j} \in K$.

Definition 2.24. Let (S, d) be a metric space, then we say S is *compact* if and only if for every collection U_α of open sets such that $\bigcup_\alpha U_\alpha \supset S$ there exists a finite subcollection $U_1, \ldots, U_n$ such that $\bigcup_{j=1}^n U_j \supset S$.

Definition 2.25. Let (S, d) be a metric space, then we say S is *totally bounded* if and only if for every $\epsilon > 0$ there exists a finite set of points $F \subset S$ such that for every $x \in S$ there is a $y \in F$ such that $d(x, y) < \epsilon$.

Definition 2.26. Let (S, d) be a metric space, then we say $x \in S$ is *limit point* of a set $A \subset S$ if and only if for every open set U containing x, $A \cap (U \setminus \{x\}) \neq \emptyset$.

Theorem 2.27. *In a metric space (S, d) the following are equivalent*

(i) *S is compact*
(ii) *S is complete and totally bounded*
(iii) *Every infinite subset of S has a limit point*
(iv) *S is sequentially compact*

Proof. First we show that (i) implies (ii). Given $\epsilon > 0$ note that we have a covering by open balls $\cup_{x\in S} B(x, \epsilon)$. By compactness we have a finite set $x_1, \ldots, x_m$ such that $\cup_{i=1}^m B(x_i, \epsilon) = S$. Thus given $y \in S$, we know there is an x_j such that $y \in B(x_j, \epsilon)$ and we have shown total boundedness. To show completeness, let $x_1, x_2, \ldots$ be a Cauchy sequence in S. For every $m > 0$ we know there exists N_m such that $d(x_{N_m}, x_n) < \frac{1}{m}$ for every $n > N_m$. Now define $U_m = \{x \in S \mid d(x_{N_m}, x) > \frac{1}{m}\}$ and note that U_m is open. Furthermore we know that $x_n \notin U_m$ for all $n > N_m$. By virtue of this latter fact we can see that there is no finite subset of U_m that covers S; for given $U_1, \ldots, U_m$ then $x_n \notin \cup_{k=1}^m U_k$ for any $n > \max(N_1, \ldots, N_m)$. By compactness of S we know that the U_m do not cover S and therefore there is an $x \in S \setminus \cup_{m=1}^\infty U_m$. For such an x, by definition of U_m we know that $d(x_{N_m}, x) \leq \frac{1}{m}$ for all $m > 0$. By the triangle inequality we then get that $d(x_n, x) \leq \frac{2}{m}$ for all $n > N_m$ and $m > 0$ which shows that x_n converges to x. Thus S is complete.

Next we show that (ii) implies (iii). Suppose $A \subset S$ is an infinite set. By the assumption of total boundedness, for each $n > 0$, we can find a finite set F_n such that for every $y \in S$ there exists $x \in F_n$ such that $d(x, y) < \frac{1}{n}$. Since the finite sets $B(y, 1)$ for $y \in F_1$ cover S there is an $y_1 \in F_1$ such that $A \cap B(y_1, 1)$ is infinite. Then arguing inductively we construct for every $n > 0$ a $y_n \in F_n$ such that

$A \cap B(y_1, 1) \cap \cdots \cap B(y_n, \frac{1}{n})$ is infinite. Note that for $n > m > 0$, by the triangle inequality using any of the infinite number of elements in $B(y_n, \frac{1}{n}) \cap B(y_m, \frac{1}{m})$, we have $d(y_n, y_m) < \frac{1}{m} + \frac{1}{n} < \frac{2}{m}$. This shows that y_n is a Cauchy sequence and by assumption we know that this converges to some $y \in S$ and by the above estimate on $d(y_n, y_m)$, we know that for every $m > 0$, $d(y, y_m) < \frac{2}{m}$. Therefore we have the inclusion $B(y_m, \frac{1}{m}) \subset B(y, \frac{3}{m})$ and therefore $A \cap B(y, \frac{3}{m})$ is also infinite which shows y is a limit point of A.

Next we show that (iii) implies (iv). Let $x_1, x_2, \ldots$ be an infinite sequence with an infinite range and by (iii) we can get a limit point $x \in S$. Thus we can find a subsequence $x_{n_1}, x_{n_2}, \ldots$ such that $x_{n_k} \in B(x, \frac{1}{k})$ which shows that the subsequence converges. If the sequence has a finite range then it is eventually constant and converges.

Lastly let's show that (iv) implies (i). Pick an open cover $\mathcal{U}_\alpha$ of S. Our first subtask is to show that there exists a radius $r > 0$ such that for every $x \in S$, the ball $B(x, r)$ is contained in some element of $\mathcal{U}_\alpha$. To that end, for every $x \in S$ let

$$f(x) = \sup\{r \mid B(x, r) \subset U_\alpha \text{ for some } \alpha\}$$

We claim that $\inf\{f(x) \mid x \in S\} > 0$. To verify the claim, we argue by contradiction and assume we can find a sequence x_n with $f(x_n) < \frac{1}{n}$ (i.e. the ball $B(x_n, \frac{1}{n})$ is not contained in any U_α). By sequential compactness we have a convergent subsequence x_{n_k} that converges to $x \in S$. Because $\mathcal{U}_\alpha$ is a open cover there we can find an $r > 0$ and U_α such that $B(x, r) \subset U_\alpha$. Pick $N_1 > \frac{2}{r}$. By convergence of x_{n_k} we can find $N_2 > 0$ such that for $n_k > N_2$ we have $d(x, x_{n_k}) < \frac{r}{2}$. For $n_k > \max(N_1, N_2)$, by the triangle inequality we have $B(x_{n_k}, \frac{1}{n_k}) \subset B(x, r) \subset U_\alpha$, so we have a contradiction.

With the claim verified we return to the problem of proving compactness. Pick an arbitrary $x_1 \in S$ and let $c = 2 \wedge \inf_{x \in S} f(x)$. We define x_n inductively by the following algorithm: if there is exists x_n such that $d(x_n, x_j) > \frac{c}{2}$ for all $j = 1, \ldots, n-1$ then pick it otherwise stop. We claim that the algorithm terminates after a finite number of steps. If it didn't then we'd have constructed an infinite sequence x_n such that for all $m, n > 0$ we have $d(x_n, x_m) > \frac{c}{2}$ which implies there is no Cauchy subsequence hence has no convergent subsequence contradicting sequential compactness. Therefore there is an $n > 0$ such that $S = \cup_{k=1}^{n} B(x_k, \frac{c}{2})$; however by construction we know that for every x_k there is a U_k such that $B(x_k, \frac{c}{2}) \subset U_k$. Then $U_1, \ldots, U_n$ is a finite subcover of S and we are done. □

It is worth noting that the equivalence of the finite subcover property and sequential compactness does not hold in general topological spaces. In general sequential compactness is equivalent to the weaker property that *countable* open covers have finite subcovers (sometime this property is refered to as countable compactness). It turns out that in these circumstances that the full power of the finite subcover property is generally needed.

Corollary 2.28. *Every closed subset of a compact set is compact.*

Proof. Let B be a compact set and let $A \subset B$ be closed. Then by the previous result and the compactness of B, any infinite sequence in A has a subsequence that converges to a point in B. Because A is closed the limit of the subsequence is in fact in A. □

Theorem 2.29. *Let $f : (S, d) \to (S', d')$ be continuous. If S is compact then $f(S)$ is compact.*

Proof. Let U_α be an open cover of $f(S)$. By continuity of f, $f^{-1}(U_\alpha)$ is an open cover of S and therefore has a finite subcover $f^{-1}(U_1), \ldots, f^{-1}(U_n)$. It is easy to see that $U_1, \ldots, U_n$ is a finite subcover of $f(S)$: if $y \in f(S)$, we can write $y = f(x)$ for $x \in S$; picking i so that $x \in f^{-1}(U_i)$, we see that $y \in U_i$. □

The following is a characterization of compact sets in $\mathbb{R}^n$.

Theorem 2.30. *[Heine-Borel Theorem]A subset $A \subset \mathbb{R}^n$ is closed and bounded if and only if it is compact.*

TODO: I don't think it is worth doing the proof from scratch; this is a simple corollary of the result.

Proof. By Lemma 2.27 it suffices to show that a closed and bounded set in $\mathbb{R}^n$ is complete and totally bounded. Completeness is simple as any Cauchy sequence in A converges in $\mathbb{R}^n$ by completeness of $\mathbb{R}^n$ but then the limit is in A because A is closed. To see total boundedness, pick an $\epsilon > 0$ and then pick $N > \frac{\sqrt{n}}{\epsilon}$. Since A is bounded, there exists $M > 0$ such that $A \subset [-M, M] \times \cdots \times [-M, M]$. It suffices to show that the latter set is totally bounded. Pick the finite set of points $\{(x_1/N, \ldots, x_n/N) \mid -MN \leq x_j \leq MN\}$ and note that

$$[-M, M] \times \cdots \times [-M, M] \subset \bigcup B((x_1/N, \ldots, x_n/N), \epsilon)$$

□

Before we begin the proof we need a Lemma.

Lemma 2.31. *Suppose $C_0 \supset C_1 \supset \cdots$ is a nested sequence of closed and bounded sets $C_k \subset \mathbb{R}^n$. Then $\cap_k C_k$ is non empty.*

Proof. Here is the proof for $n = 1$. TODO: Generalize.

Let $a_k = \inf C_k$; because C_k is closed we know that $a_k \in C_k$. By the nestedness and boundedness of C_k, we know that a_k is a non-decreasing bounded sequence and therefore has a limit a. For any fixed k, the sequence $a_n \in C_k$ for all $n \geq k$ and thus $a = \lim_{n\to\infty} a_n \in C_k$. Since k was arbitrary we have $a \in \cap_k C_k$ and we're done. □

With the Lemma in hand we can proceed to the proof of Heine-Borel.

Proof. Suppose A is closed and bounded. By boundedness there exists $N > 0$ such that $A \subset [-N, N] \times \cdots \times [-N, N]$ and by Corollary 2.28 it suffices to show that $[-N, N] \times \cdots \times [-N, N]$ is compact.

Now suppose that we are given an infinite open covering of $[-N, N] \times \cdots \times [-N, N]$ by sets A_α such that there is no finite subcover. Now bisect each side of the cube so that we can write it as a union of 2^n cubes each of side N. A_α covers each of the subcubes; if all of the subcubes had a finite subcover of A_α then by taking the union we'd have constructed a finite subcover of $[-N, N] \times \cdots \times [-N, N]$. Since we've assumed that this isn't true at least one of the subcubes has no finite subcover. Pick that cube, call it C_1 and now iterate the construction to create a nested sequence of cubes C_k where C_k has side of length $N/2^k$. Since the C_k are closed and bounded by the previous Lemma we know that the intersection $\cap_k C_k \neq \emptyset$ and therefore we can pick $x \in \cap_k C_k$. Since A_α is a cover, there exists an A such that $x \in A$. Because A is open we can in fact find a ball $B(x, r) \subset A$ for

some $r > 0$. Then for sufficiently large k, $C_k \subset B(x, r) \subset A$ which means that we have constructed a finite subcover for C_k which is a contradiction. □

Definition 2.32. Let (S, d) and (T, d') be metric spaces, a function $f : S \to T$ is said to be *uniformly continuous* if for every $\epsilon > 0$ there exists a $\delta > 0$ such that $d(x, y) < \delta$ implies $d'(f(x), f(y)) < \epsilon$.

Theorem 2.33. *Let $f : (S, d) \to (T, d')$ be a continuous function, if S is compact then f is uniformly continuous.*

Proof. The proof is by contradiction. Suppose that f is not uniformly continuous. Fix an $\epsilon > 0$, for every $n > 0$ we can find x_n and y_n such that $d(x_n, y_n) < \frac{1}{n}$ but $d'(f(x_n), f(y_n)) \geq \epsilon$. Now by compactness and Theorem 2.27 we can find a common convergence subsequence of both x_n and y_n. Let's say $\lim_{j\to\infty} x_{n_j} = x$ and $\lim_{j\to\infty} y_{n_j} = y$. Note that for every $j > 0$,

$$d(x, y) = \lim_{j\to\infty} d(x, y) \leq \lim_{j\to\infty} d(x, x_{n_j}) + d(x_{n_j}, y_{n_j}) + d(y_{n_j}, y) = 0$$

therefore $x = y$ and $f(x) = f(y)$.

Again using the triangle inequality we see

$$\lim_{j\to\infty} d'(f(x_{n_j}), f(y_{n_j})) \leq \lim_{j\to\infty} d'(f(x_{n_j}), f(x)) + d'(f(x), f(y)) + d'(f(y), f(y_{n_j})) = 0$$

which is the desired contradiction. □

Theorem 2.34. *Let $f : S \to \mathbb{R}^n$ be a continuous function, if S is compact then f is bounded.*

Proof. By the Heine-Borel Theorem and Theorem 2.29, we know that $f(S)$ is a closed bounded set. □

A related notion is that of uniform convergence of functions.

Definition 2.35. Let $f, f_n : S \to (S, d')$ be a sequence of functions. The we way that f_n converges to f *uniformly* if and only if for every $\epsilon > 0$ there exists a $N > 0$ such that for all $x \in S$, and $n > N$, $d'(f_n(x), f(x)) < \epsilon$.

One of the most important points about uniform convergence is that a uniform limit of continuous functions is continuous.

Definition 2.36. Let $f, f_n : (S, d) \to (S', d')$ be a sequence of functions where f_n are continuous. If the f_n converge to f uniformly then f is continuous.

Proof. Suppose we are given an $\epsilon > 0$ and let $x \in S$. By uniform convergence of f_n we may find an $N > 0$ such that $d'(f_n(y), f(y)) < \frac{\epsilon}{3}$ for all $n \geq N$ and $y \in S$. In particular, consider f_N. Since this function is continuous we may find $\delta > 0$ so that $d(x, y) < \delta$ implies $d'(f_N(x), f_N(y)) < \frac{\epsilon}{3}$. So by the triangle inequality, we have

$$d'(f(x), f(y)) < d'(f(x), f_N(x)) + d'(f_N(x), f_N(y)) + d'(f_N(y), f(y)) < \epsilon$$

□

2.2. **Stone Weierstrass Theorem.**

Theorem 2.37. *Let X be a compact Hausdorff space and let $A \subset C(X;\mathbb{R})$ be a subalgebra which contains a non-zero constant function. The A is dense in $C(X;\mathbb{R})$ if and only if A separates points.*

Proof. TODO: □

Corollary 2.38 (Fourier Series Approximation)**.** *For every continuous $f : \mathbb{R}^n \to \mathbb{R}$ such that $f(x+v) = f(x)$ for all $x \in \mathbb{R}$, and $v \in \mathbb{Z}^n$, for every $\epsilon > 0$ there exists constants $c_{j,k}$ and $d_{j,k}$ such that*

$$\sup_x \left| \sum_{j=0}^{n} \sum_{k=0}^{N} (c_{j,k} \sin(2k\pi x_j) + d_{j,k} \cos(2k\pi x_j)) - f(x) \right| < \epsilon$$

Proof. First we observe that there is a bijection between periodic function as in they hypothesis and functions on the topological space $T^n = S^1 \times \cdots \times S^1$ (the n-torus). Observe that if one has a uniform approximation to a function viewed as having a domain T^n then the uniform approximation applies equally well when considered as a periodic function on $\mathbb{R}^n$.

It remains to observe that T^n is compact Hausdorff, the functions $\sin(2k\pi x_j)$ an $\cos(2k\pi x_j)$ separate points and contain the constants so the Stone Weierstrass Theorem applies.

An alternative approach is a more constructive one using the Fejer kernel. □

3. Measure Theory

Measure theory is concerned with the theory of integration. Thinking intuitively for a moment, we know that we want to compute expressions of the form $\int_A f$ in which A is a set and f is a real valued function on the set A. If we take functions f that are equal to 1 on the set A, then it is clear from our intuition from elementary calculus that $\int_A 1$ should correspond the the size of A in some appropriate sense. Therefore, even if we set out to create a theory of integration we will get as a by product a theory of set measure. In fact, the development of the theory starts from the notion of set measures and develops the theory of integration using that.

Before setting out the definitions, it is worth mentioning that set theory is a weird and wild territory. Over the years, mathematicians have come up with some truly astounding constructs with sets that defy intuition. The first trivial example is to note the cardinality of Z and Z^2 is the same. A second much deeper example is the Banach-Tarski Paradox which says in effect that there is a decomposition of the unit ball in $\mathbb{R}^3$ into a finite number of pieces such that the pieces can be rearranged by only translations and rotations into two copies of the unit ball. We won't prove the Banach-Tarski paradox here, but it suffices to say that it shows you can't have all of the following in a definition of volume;

(i) Translations are volume preserving.
(ii) Rotations are volume preserving.
(iii) All sets are measurable.

By now, the time honored approach to these matters is to give up on the naive idea that all sets can be measured. Thus the definition of a measure theory comprises a definition of which sets are measureable, a means of measuring those sets and a theory of integrating suitable functions using that measure.

3.1. **Measurable Spaces.**

Definition 3.1. A non-empty collection $\mathcal{A}$ of subsets of a set Ω is called a σ-algebra if given $A, A_1, A_2, \cdots \in \mathcal{A}$ we have

(i) $A^c \in \mathcal{A}$
(ii) $\bigcup_n A_n \in \mathcal{A}$
(iii) $\bigcap_n A_n \in \mathcal{A}$

Note that this definition makes a lot of sense. Whatever our definition of the class of measurable sets is, we want to be able to perform meaningful constructions with those sets. Thus we want the set of allowable operations to be as large as possible. On the other hand, we know that we can't go beyond countable unions. For the reals once one allows points to be measurable, allowing uncountable unions would mean that every set is measurable and we already know we can't have that.

Lemma 3.2. *Let σ-algebra $\mathcal{A}$ in Ω, and $A_1, A_2, \cdots \in \mathcal{A}$,*

(i) $\Omega \in \mathcal{A}$
(ii) $\emptyset \in \mathcal{A}$

Proof. Since $\mathcal{A}$ is non empty, we can find $A \in \mathcal{A}$. Thus $\Omega = A \bigcup A^c \in \mathcal{A}$. Then taking complements shows $\emptyset \in \mathcal{A}$. □

Note that in many accounts of measure theory, the result of the above lemma is assumed as part of the definition of a σ-algebra.

Lemma 3.3. *Given a class $\mathcal{C}$ of σ-algebras on Ω, the intersection is also a σ-algebra.*

Proof. Because we have shown that every σ-algebra contains Ω, we know that the intersection is non-empty. Now let $A, A_1, A_2, \ldots$ be in every σ-algebra. Clearly every σ-algebra in the class contains $\bigcap_n A_n$, hence so does the intersection. Similarly with $\bigcup_n A_n$ and A^c. □

Note that a union of σ-algebras is not necessarily a σ-algebra. However, a union of σ-algebras generates a σ-algebra in an appropriate sense.

Definition 3.4. Given a collection $\mathcal{C}$ of subsets of Ω, we let $\sigma(\mathcal{C})$ be the smallest σ-algebra containing $\mathcal{C}$.

Note that the definition makes sense since the set of all subsets of Ω is a σ-algebra. Therefore, the class of σ-algbras containing $\mathcal{C}$ is non-empty and $\sigma(\mathcal{C})$ is the intersection of of the class by the previous lemma.

For metric spaces (and general topological spaces) there is an important σ-algebra that is associated with the topology.

Definition 3.5. Given a metric space S, the Borel σ-algebra $\mathcal{B}(S)$ is the σ-algebra generated by the open sets on S.

Lemma 3.6. *The Borel σ-algebra of $\mathbb{R}$ is generated by intervals of the form $(-\infty, x]$ for $x \in \mathbb{Q}$.*

Proof. Let $\mathcal{C}$ be the collection of all open intervals. We know that the open sets of $\mathbb{R}$ are countable unions of open intervals. Therefore, the Borel σ-algebra is generated by the set of open intervals. Now let $\mathcal{D}$ be the set of closed intervals of the form

$(-\infty, x]$ for $x \in \mathbb{Q}$. Pick an open interval (a, b) and pick a descreasing sequence or rationals $a_n \downarrow a$ and an increasing sequence of rationals $b_n \uparrow b$. Then we have

$$\begin{aligned}(a, b) &= \bigcup_{n=1}^{\infty} (a_n, b_n] \\ &= \bigcup_{n=1}^{\infty} ((-\infty, b_n] \cap (-\infty, a_n])\end{aligned}$$

which shows that $\mathcal{C} \subset \sigma(\mathcal{D})$ hence $\sigma(\mathcal{C}) \subset \sigma(\mathcal{D})$. However, since the elements of $\mathcal{D}$ are closed sets and σ-algebras are closed under set complement, we have $\mathcal{D} \subset \sigma\mathcal{C}$ and therefore

$$\mathcal{B} = \sigma(\mathcal{C}) \subset \sigma(\mathcal{D}) \subset \sigma(\mathcal{C}) = \mathcal{B}$$

and we have $\sigma(\mathcal{D}) = \mathcal{B}$. □

Next we consider how σ-algebras behave in the presence of functions. Given a function $f : S \to T$ we have the induced map on sets $f^{-1} : 2^T \to 2^S$ defined by

$$f^{-1}(B) = \{x \in S; f(x) \in B\}$$

Lemma 3.7. *For $A, B, B_1, B_2, \cdots \subset T$, then*

(i) $f^{-1}(B^c) = \left[f^{-1}(B)\right]^c$
(ii) $f^{-1} \bigcap_n B_n = \bigcap_n f^{-1} B_n$
(iii) $f^{-1} \bigcup_n B_n = \bigcup_n f^{-1} B_n$
(iv) $f^{-1}(B \setminus A) = f^{-1}(B) \setminus f^{-1}(A)$

Proof. (i)

$$\begin{aligned}f^{-1}(B^c) &= \{x \in S; f(x) \notin B\} \\ &= \{x \in S; f(x) \in B\}^c = \left[f^{-1}(B)\right]^c\end{aligned}$$

(ii)

$$\begin{aligned}f^{-1} \bigcap_n B_n &= f^{-1} \{x \in T; \forall n x \in B_n\} \\ &= \{x \in S; \forall n f(x) \in B_n\} = \bigcap_n f^{-1} B_n\end{aligned}$$

(iii)

$$\begin{aligned}f^{-1} \bigcup_n B_n &= f^{-1} \{x \in T; \exists n x \in B_n\} \\ &= \{x \in S; \exists n f(x) \in B_n\} = \bigcup_n f^{-1} B_n\end{aligned}$$

(iv) follows from (i) and (ii) by writing $B \setminus A = B \cap A^c$. □

Lemma 3.8. *Given an arbitrary function f between measurable spaces $(S, \mathcal{S})$ and $(T, \mathcal{T})$, then*

(i) $\mathcal{S}' = f^{-1}\mathcal{T}$ *is a σ-algebra on S.*
(ii) $\mathcal{T}' = \left\{A \subset T; f^{-1}(A) \in \mathcal{S}\right\}$ *is a σ-algebra on T.*

The σ-algebra denoted $\mathcal{T}'$ is often denoted $f_*\mathcal{S}$.

Proof. To show (i), let $A, A_1, A_2, \cdots \in \mathcal{S}'$. Since $\mathcal{S}' = f^{-1}\mathcal{T}$, there exist $B, B_1, B_2, \cdots \in \mathcal{T}$ such that $A = f^{-1}(B)$ and $A_i = f^{-1}(B_i)$ for $i = 1, 2, \ldots$. Now since $\mathcal{T}$ is a σ-algebra, we know that B^c, $\bigcup_n B_n$ and $\bigcap_n B_n$ are all in $\mathcal{T}$. Now using the previous lemma,

$$\begin{aligned} A^c = \left[f^{-1}(B)\right]^c &= f^{-1}(B^c) \in \mathcal{S}' \\ \bigcap_n A_n = \bigcap_n f^{-1}B_n &= f^{-1}\bigcap_n B_n \in \mathcal{S}' \\ \bigcup_n A_n = \bigcup_n f^{-1}B_n &= f^{-1}\bigcup_n B_n \in \mathcal{S}' \end{aligned}$$

Now to see (ii), first note that $\mathcal{T}'$ is non-empty since $f^{-1}(\emptyset) = \emptyset \in \mathcal{S}$. Next, pick $B, B_1, B_2, \cdots \in \mathcal{T}'$ so that $f^{-1}B, f^{-1}B_1, f^{-1}B_2 \in \mathcal{S}$. Again use the previous lemma to see

$$\begin{aligned} f^{-1}B^c = \left[f^{-1}(B)\right]^c &\in \mathcal{S} \\ f^{-1}\bigcap_n B_n = \bigcap_n f^{-1}B_n &\in \mathcal{S} \\ f^{-1}\bigcup_n B_n = \bigcup_n f^{-1}B_n &\in \mathcal{S} \end{aligned}$$

and this shows that $B^c, f^{-1}\bigcap_n B_n, f^{-1}\bigcap_n B_n \in \mathcal{T}'$. □

Lemma 3.9. *Let $f : S \to T$ be a set function and $f^{-1} : 2^T \to 2^S$ be the induced function on sets.*

(i) *f^{-1} is surjective if and only if f is injective*
(ii) *f^{-1} is injective if and only if f is surjective*
(iii) *f^{-1} is a bijection if and only if f is a bijection*

Proof. Suppose f is surjective and pick $A, B \subset T$ with $A \neq B$. Then, possibly switching the names of A and B, we have $t \in A \setminus B$. By surjectivity we know there exists an $s \in S$ such that $f(s) = t$ and therefore $s \in f^{-1}(A) \setminus f^{-1}(B)$ showing $f^{-1}(A) \neq f^{-1}(B)$. Now if f is not surjective then there exists $t \in T$ such that there is no $s \in S$ with $f(s) = t$. In this case we see that $f^{-1}(T) = S = f^{-1}(T \setminus \{t\})$ showing f^{-1} is not injective.

Suppose f is injective and let $B \subset S$ and we claim $B = f^{-1}(f(B))$. Clearly $A \subset f^{-1}(f(B))$ and if they are not equal then there exists $s \in S \setminus B$ such that $f(s) = f(b)$ for some $b \in B$ contradicting injectivity. If f is not injective then there exists $s, t \in S$ with $s \neq t$ and $f(s) = f(t)$ and clearly there can be no $A \subset T$ such that $f^{-1}(A) = \{s\}$.

The statement of (iii) is an immediate consequence of (i) and (ii). □

The definition given for $\sigma(\mathcal{C})$ for a set $\mathcal{C} \subset 2^\Omega$ as the smallest σ-algebra containing $\mathcal{C}$ may lack appeal because of the fact that it is non-constructive. It is possible to give a constructive definition of $\sigma(\mathcal{C})$ by making a transfinite recursive definition. The following makes use of the theory of ordinal numbers.

Lemma 3.10. *Let $\mathcal{C} \subset 2^\Omega$, and let ω_1 be the first uncountable ordinal and define for each countable ordinal*

(i) $\mathcal{C}_{\omega_0} = \mathcal{C}$
(ii) *For a successor ordinal α, $\mathcal{C}_\alpha$ is the set of countable unions of elements of $\mathcal{C}_{\alpha-1}$ and complements of such unions.*
(iii) *For a limit ordinal α, define $\mathcal{C}_\alpha = \bigcup_{\beta<\alpha} \mathcal{C}_\beta$.*

Then $\bigcup_{\alpha<\omega_1} \mathcal{C}_\alpha = \sigma(\mathcal{C})$.

Proof. First we show $\bigcup_{\alpha<\omega_1} \mathcal{C}_\alpha \supset \sigma(\mathcal{C})$. Since we know that $\mathcal{C} \subset \bigcup_{\alpha<\omega_1} \mathcal{C}_\alpha$, it suffices to show that $\bigcup_{\alpha<\omega_1} \mathcal{C}_\alpha$ is a σ-algebra.

It is explicit in the definition for successor ordinals, that given any $A \in \mathcal{C}_\alpha$, we have $A^c \in \mathcal{C}_{\alpha+1}$.

To show closure under set union, we suppose that we are given $A_1, A_2, \ldots$ where $A_i \in \mathcal{C}_{\alpha_i}$. We now use the fact that given a countable set of countable ordinals, there is a countable ordinal that bounds them (TODO: Prove this somewhere or find a good reference). Thus we may pick a countable ordinal $\hat{\alpha}$ such that $\alpha_i < \hat{\alpha}$ for every $i = 1, 2, \ldots$. Since $\mathcal{C}_\alpha \subset \mathcal{C}_{\alpha+1}$, we know that $A_i \in \mathcal{C}_{\hat{\alpha}}$ for all i. Now simply apply the definition of $\mathcal{C}_{\hat{\alpha}+1}$ to see $\bigcup_{i=1}^\infty A_i \in \mathcal{C}_{\hat{\alpha}+1}$. Having proven closure under complement and countable union, use De Morgan's Law to derive the countable intersection property and we are done.

Now we need to show that $\bigcup_{\alpha<\omega_1} \mathcal{C}_\alpha \subset \sigma(\mathcal{C})$. This is an easy transfinite induction on α using the properties of the σ-algebra $\sigma(\mathcal{C})$. TODO: Write this out. □

3.2. **Measurable Functions.** We've seen that arbitrary set functions can be used to create σ-algebras but when we consider functions between measurable spaces the σ-algebras are given and it makes sense to restrict our attention to a class of functions that are compatible with those σ-algebras.

Definition 3.11. A function $f : (S, \mathcal{S}) \to (T, \mathcal{T})$ is called measurable if for every $B \in \mathcal{T}$, we have $f^{-1}(B) \in \mathcal{S}$. When we want to emphasize that the measurability is with repsect to particular σ-algebras we may say that f is $\mathcal{S}/\mathcal{T}$-measurable.

Lemma 3.12. *Suppose we are given a function $f : (S, \mathcal{S}) \to (T, \mathcal{T})$ and a class of subsets $\mathcal{C} \subset 2^T$ such that $\sigma(\mathcal{C}) = \mathcal{T}$. The f is measurable if and only if $f^{-1}\mathcal{C} \subset \mathcal{S}$.*

Proof. The only if direction is trivial. So suppose $f^{-1}\mathcal{C} \subset \mathcal{S}$. Now consider $\mathcal{T}' = \{B \subset T; f^{-1}B \in S\}$. By our assumption, we have $\mathcal{C} \subset \mathcal{T}'$. Furthermore we know from Lemma 3.8 that $\mathcal{T}'$ is a σ-algebra, thus $\sigma(\mathcal{C}) \subset \mathcal{T}'$ and this shows that f is $\mathcal{S}/\mathcal{T}$ measurable. □

Lemma 3.13. *Let $f : (S, \mathcal{S}) \to (T, \mathcal{T})$ and $g : (T, \mathcal{T}) \to (U, \mathcal{U})$ be measurable. Then $g \circ f : (S, \mathcal{S}) \to (U, \mathcal{U})$ is measurable.*

Proof. This follows simply from the fact that $(g \circ f)^{-1}(B) = g^{-1}(f^{-1}(B))$ and the measurability of f and g. □

Note, from this point forward, when we refer to $\mathbb{R}$ as a measurable space, it should be assumed that we are referring to $\mathbb{R}$ with the Borel σ-algebra. Note that a function $f : (\Omega, \mathcal{A}) \to \mathbb{R}$ is measurable if and only if $\{\omega \in \Omega; f(\omega) \leq x\} \in \mathcal{A}$ for all $x \in \mathbb{R}$ (in fact it suffices to consider $x \in \mathbb{Q}$). It is also very common to consider extensions of $\mathbb{R}$ such as $\overline{\mathbb{R}} = [-\infty, \infty]$ and $\overline{\mathbb{R}}_+ = [0, \infty]$ obtained by appending points at infinity. For these spaces we take the σ-algebra generated by $\{\omega \in \Omega; f(\omega) \leq x\}$ for $x \in \overline{\mathbb{R}}$ respectively. It can be shown that there are natural topologies on each

of these compactifications and the σ-algebras defined are the Borel σ-algebras of these topologies.

We will often talk about the convergence of sequences of measurable functions. Unless we say otherwise, it should be understood that this convergence is taken pointwise.

Lemma 3.14. *Let $f_1, f_2, \ldots$ be measurable functions from $(\Omega, \mathcal{A})$ to $\overline{\mathbb{R}}$. Then $\sup_n f_n$, $\inf_n f_n$, $\limsup_n f_n$, $\liminf_n f_n$ are all measurable.*

Proof. To see measurability of $\sup_n f_n$ we suppose that $\omega \in \Omega$ is such that $\sup_n f_n(\omega) \leq x$, then x is an upper bound we have $f_n(\omega) \leq x$ for all n. On the other hand, if we assume that $\omega \in \Omega$ is such that $f_n(\omega) \leq x$ for all n then $\sup_n f_n(\omega) \leq x$ so we have

$$\left\{\omega; \sup_n f_n(\omega) \leq x\right\} = \bigcap_n \{\omega; f_n(\omega) \leq x\} \in \mathcal{A}$$

To see that $\inf_n f_n$ is measurable we use the identity $\inf_n f_n = -\sup_n(-f_n)$.

We also have the definitions

$$\limsup_{n\to\infty} f_n = \inf_n \sup_{k\geq n} f_k, \liminf_{n\to\infty} f_n = \sup_n \inf_{k\geq n} f_k$$

and the measurability of sup and inf already shown implies the measurability of lim inf and lim sup. □

We now introduce an extremely important class of measurable functions. Simple measurable functions will be used to approximate arbitrary measurable functions and in particular, will serve as the analogue of Riemann sums when we start to consider integration.

Definition 3.15. Given a set Ω and a set $A \subset \Omega$, the *indicator function* $\mathbf{1}_A$ is equal to 1 on A and 0 on A^c. A linear combination $c_1\mathbf{1}_{A_1} + \cdots + c_n\mathbf{1}_{A_n}$ is called a *simple function.*

Lemma 3.16. *A function $f : \Omega \to \mathbb{R}$ is simple if and only it takes a finite number of values. A simple function is measurable if an only if $f^{-1}(c_j)$ is measurable for each of its distinct values $c_j \in \mathbb{R}$.*

Proof. If $f = c_1\mathbf{1}_{A_1} + \cdots + c_n\mathbf{1}_{A_n}$ is simple, then since indicator functions take only the value $0, 1$ it is clear that f can have at most 2^n values.

On the other hand, if $f : \Omega \to \mathbb{R}$ only takes the finite number of distinct values $c_1, \ldots, c_n$ then clearly we may write $f = c_1\mathbf{1}_{A_1} + \cdots + c_n\mathbf{1}_{A_n}$ where $A_j = f^{-1}(c_j)$.

As regards measurability, first notice that $\mathbf{1}_A$ is measurable if and only if $A \in \mathcal{A}$. This follows from that fact that there are only four possible preimages under $\mathbf{1}_A$: $A, A^c, \Omega, \emptyset$ and each of these preimages is the preimage of a measurable subset of $\mathbb{R}$.

Similarly, if a simple function f has the distinct values $c_1, \ldots, c_n$ (including 0 if necessary) then clearly for f to be measurable it is necessary $f^{-1}(c_j)$ is measurable since points are measurable in $\mathbb{R}$. On the hand, there are 2^n possible preimages under f and they are all constructed from unions of the preimages $f^{-1}(c_j)$ so if know that $f^{-1}(c_j)$ are measurable then so is every $f^{-1}(A)$ for $A \subset \mathbb{R}$ (a stronger condition than measurability). □

Note that the representation of a simple function as a linear combination of indicator functions is not unique. However, we have just shown that a simple

function is equally well characterized as a function that takes a finite number of values. The canonical representation of a simple function is a representation such that the c_i are distinct and non-zero and the A_i are pairwise disjoint; the canonical representation is unique.

Lemma 3.17. *For any positive measurable function $f : (\Omega, \mathcal{A}) \to \overline{\mathbb{R}}_+$ there exist a sequence of simple measurable functions $f_1, f_2, \ldots$ such that $0 \leq f_n \uparrow f$.*

Proof. Define

$$f_n(\omega) = \begin{cases} k2^{-n} & \text{if } k2^{-n} \leq f(\omega) < (k+1)2^{-n} \text{ and } 0 \leq k \leq n2^n - 1. \\ n & \text{if } f(\omega) \geq n. \end{cases}$$

Note that f_n is simple since it has at most 2^n+1 values $0, \frac{1}{2^n}, \ldots, n$. f_n is measurable since $f_n^{-1}(k2^{-n}) = f^{-1}[k2^{-n}, (k+1)2^{-n})$ is measurable by measurability of f. Similarly with $f_n^{-1}(n) = f^{-1}[n, \infty)$ and Lemma 3.16. □

As an application of approximation by simple functions,

Lemma 3.18. *Let $f, g : (\Omega, \mathcal{A}) \to \mathbb{R}$ be measurable functions and let $a, b \in \mathbb{R}$. Then $af + bg$ and fg are measurable and f/g is measurable when $g \neq 0$ on Ω.*

Proof. As f and g are measurable, we can apply the previous lemma to $f_\pm = \pm((\pm f) \wedge 0)$ and $g_\pm = \pm((\pm g) \wedge 0)$ to get measurable simple functions f_n and g_n such that $\lim_{n\to\infty} f_n = f$ and $\lim_{n\to\infty} g_n = g$. Basic properties of limits show that $\lim_{n\to\infty}(af_n + bg_n) = af + bg$, $\lim_{n\to\infty} f_n g_n = fg$ and $\lim_{n\to\infty} \frac{f_n}{g_n} = \frac{f}{g}$. Thus by Lemma 3.14 we are done if we can show that each of $af_n + bg_n$, $f_n g_n$ and $\frac{f_n}{g_n}$ is measurable. In fact we will show that each of these is simple measurable.

It is easy to see that $af_n + bg_n$ are also measurable simple as are $f_n g_n$. Let f_n take the values $c_1, \ldots, c_s$ and let g_n take the values $d_1, \ldots, d_t$. Clearly the functions $af_n + bg_n$, $f_n g_n$ and $\frac{f_n}{g_n}$ are simple as each takes at most the values $ac_i + bd_j$, $c_i d_j$ and $\frac{c_i}{d_j}$ for $i = 1, \ldots, s$ and $j = 1, \ldots, t$. Measurability follows from noting that each possible value of the linear combination is created from a finite set of combinations of the values of the f_n and g_n; hence $(af_n + bg_n)^{-1}(c_j)$ is a finite union of intersections of the form $f_n^{-1}(x) \cap g_n^{-1}(y)$ where $x, y \in \mathbb{R}$ are values of f_n and g_n respectively. □

Definition 3.19. Given two measurable functions f, g on the same measurable space $(\Omega, \mathcal{A})$, we say that f is g-measurable if $\sigma(f) \subset \sigma(g)$.

The following lemma is extremely useful both conceptually and practically. In addition it's proof is a paradigmatic example of a common measure theoretic argument.

Lemma 3.20. *Let $f : (\Omega, \mathcal{A}) \to \mathbb{R}$ and $g : (\Omega, \mathcal{A}) \to (T, \mathcal{T})$ be measurable. Then f is g-measurable if and only if there exists measurable $h : (T, \mathcal{T}) \to \mathbb{R}$ such that $f = h \circ g$.*

Proof. For the if direction, assume $f = h \circ g$. Then for $B \in \mathcal{B}(\mathbb{R})$, we have $f^{-1}(B) = g^{-1}(h^{-1}(B))$. Now we know that $h^{-1}(B) \in \mathcal{T}$ and therefore, $f^{-1}(B) \in \sigma(g)$.

For the only if direction, first assume that f is an indicator function $\mathbf{1}_A$. Our assumption of g-measurability means that there exists $B \in \mathcal{T}$ such that $A = g^{-1}(B)$. If we define $h = \mathbf{1}_B$, then we have $f = h \circ g$. Now let us suppose that f is a simple function and take its canonical representation $f = c_1\mathbf{1}_{A_1} + \cdots + c_n\mathbf{1}_{A_n}$ with A_i

disjoint and c_i distinct. Since f is g-measurable, we know that there exist $B_i \in \mathcal{T}$ such that $A_i = g^{-1}(B_i)$. If we define $h = c_1\mathbf{1}_{B_1} + \cdots + c_n\mathbf{1}_{B_n}$, then $f = h \circ g$.

Now if we assume $f \geq 0$, then we know that we can find a sequence of g-measurable simple functions such that $f_n \uparrow f$. We have shown that there are h_n such that $f_n = h_n \circ g$. Define $h = \limsup_n h_n$ and then note h is g-measurable and that

$$h(g(\omega)) = \limsup_n h_n(g(\omega)) = \limsup_n f_n(\omega) = \lim_{n\to\infty} f_n(\omega) = f(\omega)$$

Lastly, for arbitrary f, we write $f = f_+ - f_-$ where $f_\pm \geq 0$ and are both g-measurable (e.g. $f_\pm = (\pm f) \wedge 0$). We find $h_\pm$ such that $f_\pm = h_\pm \circ g$ and define $h = h_+ - h_-$. □

The following definitions and lemma may seem merely technical, but in fact are an important part of the most common methodology for proving measure theoretic results.

Definition 3.21. A class $\mathcal{C}$ of subsets of a set Ω is called a λ-system if

(i) $\Omega \in \mathcal{C}$.
(ii) For all $A, B \in \mathcal{C}$ such that $A \subset B$, $B \setminus A \in \mathcal{C}$.
(iii) $A_n \uparrow A$ for $A_n \in \mathcal{C}$ the $A \in \mathcal{C}$.

Definition 3.22. A class $\mathcal{C}$ of subsets of a set Ω is called a π-system if it is closed under finite intersections.

The first observation is that the concepts of π-system and λ-system factor the conditions for being a σ-algebra.

Lemma 3.23. *If a class $\mathcal{C} \subset 2^\Omega$ is both a π-system and a λ-system, then it is a σ-algebra.*

Proof. First we show closure under set complement. Let $A \in \mathcal{C}$. Then since $\Omega \in \mathcal{C}$, we know that $A^c = \Omega \setminus A \in \mathcal{C}$. Now note that having closure under set complement together with closure under finite intersection gives closure under finite union by De Morgan's law $\{\bigcup_{i=1}^n A_i\}^c = \bigcap_{i=1}^n A_i^c$.

Let $A_1, A_2, \cdots \in \mathcal{C}$. Next we show closure under countable union. Defining $B_n = \bigcup_{i=1}^n A_i$, we know that $B_n \in \mathcal{C}$ and clearly $B_n \uparrow \bigcup_{i=1}^\infty A_i$ and therefore $\bigcup_{i=1}^\infty A_i \in \mathcal{C}$. Closure under countable intersections follows from closure under countable unions and the infinite version of De Morgan's Law. □

Theorem 3.24 (π-λ Theorem)**.** *Suppose $\mathcal{C}$ is a π-system, $\mathcal{D}$ is a λ-system such that $\mathcal{C} \subset \mathcal{D}$. Then $\sigma(\mathcal{C}) \subset \mathcal{D}$.*

Proof. The first thing to note is that the intersection of a collection of λ-systems is also a λ-system and that 2^Ω is a λ-system. Therefore, in a way entirely analogous to σ-algebras we may define the λ-system generated by a collection of sets as the intersection of all λ-systems containing the collection.

The theorem is proved for general $\mathcal{D}$ if we prove it for the special case $\mathcal{D} = \lambda(\mathcal{C})$. To see this special case, by 3.23 it suffices to show that $\lambda(\mathcal{C})$ is a π-system. A trivial induction argument shows it suffices to show closure under pairwise intersection: for every $A, B \in \lambda(\mathcal{C})$ we have $A \cap B \in \lambda(\mathcal{C})$.

By definition of π-algebra, we have closure when $A, B \in \mathcal{C}$. Now fix $C \in \mathcal{C}$ and let $\mathcal{A}_C = \{A \subset \Omega; A \cap C \in \lambda(\mathcal{C})\}$. We claim that $\mathcal{A}_C$ is a λ-system.

To see that $\Omega \in \mathcal{A}_C$ is trivial: $C \cap \Omega = C \in \mathcal{C} \subset \lambda(\mathcal{C})$. Suppose $A \supset B$ where $A, B \in \mathcal{A}_C$, then $C \cap (A \backslash B) = (C \cap A) \backslash (C \cap B) \in \lambda(\mathcal{C})$. Suppose $A_1 \subset A_2 \subset \cdots$ with $A_i \in \mathcal{A}_C$. $C \cap \bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} C \cap A_i \in \lambda(\mathcal{C})$ by distributivity of set intersection over set union and closure of λ-system under increasing unions.

Now that we know $\mathcal{A}_C$ is a λ-system containing $\mathcal{C}$ we know that $\lambda(\mathcal{C}) \subset \mathcal{A}_C$ and therefore $C \cap A \in \lambda(\mathcal{C})$ for every $A \in \lambda(\mathcal{C})$ and $C \in \mathcal{C}$.

To finish up the proof, for every $C \in \lambda(\mathcal{C})$, let $\mathcal{B}_C = \{A \in \Omega; A \cap C \in \lambda(\mathcal{C})\}$. We have just shown that $\mathcal{C} \subset \mathcal{B}_C$ and an argument exactly analogous to the one above shows that $\mathcal{B}_C$ is a λ-algebra and therefore $\lambda(\mathcal{C}) \subset \mathcal{B}_C$ proving the result. □

Though we'll see many examples of this along the way, it is worth making explicit how the Theorem 3.24 is applied. Suppose that one wishes to prove a property holds for a σ-algebra $\mathcal{A}$ of sets. A common sub-case is we'll be trying to show a property holds for the indicator functions associated with those sets (those being the most basic building blocks of measurable functions). The π-λ Theorem allows us to prove the property holds on $\mathcal{A}$ by showing

(i) The collection of all sets satisfying the property is a λ-system
(ii) There is a π-system of sets $\mathcal{P}$ that satisfies the property and $\sigma(\mathcal{P}) = \mathcal{A}$.

A proof along these lines is referred to as a *monotone class argument.*

3.3. **Measures and Integration.** Armed with a way of describing and transforming measurable sets it is finally time to measure them.

Definition 3.25. A *measure* on a measurable space $(\Omega, \mathcal{A})$ is a function $\mu : \mathcal{A} \to \overline{\mathbb{R}}_+$ satisfying

(i) $\mu(\emptyset) = 0$
(ii) $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$ for $A_1, A_2, \cdots \in \mathcal{A}$ disjoint.

A triple $(\Omega, \mathcal{A}, \mu)$ is called a *measure space.*

An important special case of measure theory occurs when the underlying space has unit measure. Many of the concepts we have already discussed have different names when discussing this special case.

Definition 3.26. A *probability space* is a measure space $(\Omega, \mathcal{A}, P)$ such that $P(\Omega) = 1$. The measure P is called the *probability measure.* Measurable sets $A \in \mathcal{A}$ are referred to as *events.* Given a measurable space $(S, \mathcal{S})$, a measurable function $\xi : \Omega \to S$ is called a *random element* in S. For the special case in which $(S, \mathcal{S}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, we call a measurable $\xi : \Omega \to \mathbb{R}$ a *random variable.*

Lemma 3.27. *Given a measure space $(\Omega, \mathcal{A}, \mu)$, and sets $A_1, A_2, \cdots \in \mathcal{A}$.*

(i) *If $A_i \uparrow A$ then $\mu A_i \uparrow \mu A$.*
(ii) *If $A_i \downarrow A$ and $\mu A_1 < \infty$ then $\mu A_i \downarrow \mu A$.*

Proof. To show (i), define $B_1 = A_1$ and $B_i = A_i \setminus A_{i-1}$ for $i > 1$. Clearly, B_i are disjoint and it is equally clear that $\bigcup_{i=1}^{n} B_i = A_n$ and $\bigcup_{n=1}^{\infty} B_n = A$. Therefore

$$\mu A_n = \mu \bigcup_{i=1}^{n} B_i = \sum_{i=1}^{n} \mu B_i \uparrow \sum_{i=1}^{\infty} \mu B_i = \mu \bigcup_{i=1}^{\infty} B_i = \mu A$$

where we have used finite and countable additivity of μ over the B_i.

To see (ii), note that $A_1 \setminus A_n \uparrow A_1 \setminus A$ and then under the finiteness assumption $\mu A_1 < \infty$, we see

$$\mu(A_1 \setminus A_n) = \mu A_1 - \mu A_n \uparrow \mu(A_1 \setminus A) = \mu A_1 - \mu A$$

Subtract μA_1 from both sides multiply by -1 to get the result. □

Lemma 3.28. *Given a measure space $(\Omega, \mathcal{A}, \mu)$, $\mu\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mu(A_i)$ for $A_1, A_2, \cdots \in \mathcal{A}$.*

Proof. First we prove finite subadditivity by an induction argument. For $n = 2$, we note that we may write disjoint unions

$$\begin{aligned} A &= (A \setminus B) \cup (A \cap B) \\ B &= (B \setminus A) \cup (A \cap B) \\ A \cup B &= (A \setminus B) \cup (B \setminus A) \cup (A \cap B) \end{aligned}$$

By finite additivity of measure and positivity of measure, we see $\mu A \cup B = \mu A + \mu B - \mu A \cap B \leq \mu A + \mu B$.

For the induction step, assume $\mu\left(\bigcup_{i=1}^{n-1} A_i\right) \leq \sum_{i=1}^{n-1} \mu(A_i)$, then use the case $n = 2$ and Lemma 3.27 to see

$$\begin{aligned} \mu\left(\bigcup_{i=1}^{n} A_i\right) &= \mu\left(\bigcup_{i=1}^{n-1} A_i \cup A_n\right) \\ &\leq \mu\left(\bigcup_{i=1}^{n-1} A_i\right) + \mu A_n \\ &\leq \sum_{i=1}^{n-1} \mu(A_i) + \mu A_n = \sum_{i=1}^{n} \mu(A_i) \end{aligned}$$

To extend the result to infinite unions, define $B_n = \bigcup_{i=1}^{n} A_i$ and note that $B_n \uparrow \bigcup_{i=1}^{\infty} A_i$ and that by finite subadditivity, $\mu B_n \leq \sum_{i=1}^{n} \mu A_i$. Taking limits we see

$$\mu \bigcup_{i=1}^{\infty} A_i = \lim_{n\to\infty} \mu B_n \leq \lim_{n\to\infty} \sum_{i=1}^{n} \mu A_i = \sum_{i=1}^{\infty} \mu A_i$$

□

Next up is the definition of integral of a measurable function on a measure space. First we proceed by defining the integral for a simple functions.

Definition 3.29. Given a canonical representation of a simple function $f = c_1 \mathbf{1}_{A_1} + \cdots + c_n \mathbf{1}_{A_n}$ we define the integral of f to be

$$\int f d\mu = \mu f = c_1 \mu A_1 + \cdots + c_n \mu A_n$$

Having the definition of the integral of a simple function in terms of the canonical representation is inconvenient at times when one is given a simple function that is not known to be in a canonical representation. It turns out that the formula above extends to any representation of the simple function as a linear combination of indicator functions. To see that we proceed in steps.

Lemma 3.30. *Given any representation of a simple function* $f = c_1 \mathbf{1}_{A_1} + \cdots + c_n \mathbf{1}_{A_n}$ *with* A_i *pairwise disjoint,*

$$\int f d\mu = c_1 \mu A_1 + \cdots + c_n \mu A_n$$

Proof. We have to construct the canonical representation of f. It is conceptually simple, but there is a bit of notation to deal with. Let $d_1, d_2, \ldots, d_m$ be the distinct values of $c_1, \ldots, c_n$. Furthermore, for each $i = 1, \ldots, m$, let $B_{i,j}$ $j = 1, \ldots, k_i$ be the set of A_n for which $c_n = d_i$. Then the canonical representation of f is

$$f = d_1 \mathbf{1}_{\bigcup_{j=1}^{k_1} B_{1,j}} + \cdots + d_m \mathbf{1}_{\bigcup_{j=1}^{k_m} B_{m,j}}$$

and then

$$\begin{aligned} \int f d\mu &= d_1 \mu \bigcup_{j=1}^{k_1} B_{1,j} + \cdots + d_m \mu \bigcup_{j=1}^{k_m} B_{m,j} \\ &= d_1 \sum_{j=1}^{k_1} \mu B_{1,j} + \cdots + d_m \sum_{j=1}^{k_m} \mu B_{m,j} \\ &= c_1 \mu A_1 + \cdots + c_n \mu A_n \end{aligned}$$

□

Lemma 3.31. *Given two simple functions* f, g*, for all* $a, b \in \mathbb{R}$*,*

$$\int (af + bg)\, d\mu = a \int f d\mu + b \int g d\mu$$

If $f \geq g$ *a.e. then we have*

$$\int f d\mu \geq \int g d\mu$$

Proof. Take the canonical representation of both f and g, $f = \sum_{i=1}^{n} c_i \mathbf{1}_{A_i}$ and $f = \sum_{i=1}^{m} d_i \mathbf{1}_{B_i}$. Furthermore define $A_0 = \Omega \setminus \bigcup_{i=1}^{n} A_i$ and $B_0 = \Omega \setminus \bigcup_{i=1}^{m} B_i$. Now consider all of the pairs $A_i \cap B_j$ and write

$$\begin{aligned} f &= \sum_{i=0}^{n} \sum_{j=0}^{m} c_i \mathbf{1}_{A_i \cap B_j} \\ g &= \sum_{i=0}^{n} \sum_{j=0}^{m} d_j \mathbf{1}_{A_i \cap B_j} \end{aligned}$$

where we have defined $c_0 = d_0 = 0$. Thus, we have the representation

$$af + bg = \sum_{i=0}^{n} \sum_{j=0}^{m} (ac_i + bd_j) \mathbf{1}_{A_i \cap B_j}$$

Since the $A_i \cap B_j$ are pairwise disjoint, we can write

$$\begin{aligned}\int af + bg &= \int \sum_{i=0}^{n}\sum_{j=0}^{m}(ac_i + bd_j)\mathbf{1}_{A_i\cap B_j} \\ &= \sum_{i=0}^{n}\sum_{j=0}^{m}(ac_i + bd_j)\mu A_i \cap B_j \\ &= a\sum_{i=0}^{n}\sum_{j=0}^{m} c_i\mu A_i \cap B_j + b\sum_{i=0}^{n}\sum_{j=0}^{m} d_j\mu A_i \cap B_j \\ &= a\int f + b\int g\end{aligned}$$

Using the same representation as above, we see that if $f \geq g$, then since the $A_i \cap B_j$ are disjoint, we must have $c_i \geq d_j$ whenever $A_i \cap B_j \neq \emptyset$. This shows $\int f \geq \int g$. □

Corollary 3.32. *Given any representation of a simple function* $f = c_1\mathbf{1}_{A_1} + \cdots + c_n\mathbf{1}_{A_n}$,

$$\int f = c_1\mu A_1 + \cdots + c_n\mu A_n$$

The corollary above is used so often that we use it without mentioning it and essentially treat it as the definition of the integral of a simple function.

Having defined integrals of simple functions, we leverage the fact that we can approximate positive measurable functions by increasing sequences of simple functions to define the integral of a positive measurable function.

Definition 3.33. Given a measurable function $f : (\Omega, \mathcal{A}, \mu) \to \overline{\mathbb{R}}_+$, we define

$$\int f = \sup_{0\leq g\leq f}\int g$$

where the supremum is taken over positive simple functions g.

Working with the supremum above is a bit inconvenient and it turns out that it suffices to work with increasing sequences of positive simple functions. To see that we first need a lemma.

Lemma 3.34. *Given a measurable function* $f : (\Omega, \mathcal{A}, \mu) \to \overline{\mathbb{R}}_+$, *a sequence* $0 \leq f_1, f_2, \ldots$ *of simple measurable functions such that* $f_n \uparrow f$ *and a simple measurable function* g *such that* $0 \leq g \leq f$, *we have* $\lim_{n\to\infty}\int f_n\, d\mu \geq \int g\, d\mu$.

Proof. Consider the case where $g = \mathbf{1}_A$ for $A \in \mathcal{A}$. Pick $\epsilon > 0$, and define

$$A_n = \{\omega \in A; f_n(\omega) \geq 1 - \epsilon\}$$

Since f_n is increasing, so is A_n. Also it is simple to see that $A_n \subset A$ since $f \geq f_n$ and $A \subset \bigcup_n A_n$ since for each $\omega \in A$ convergence of $f_n(\omega) \uparrow f(\omega)$ tells us that there is $N > 0$ such that for $n > N$, we have $|f_n(\omega) - f(\omega)| < \epsilon$, hence $A_n \uparrow A$ and $\mu A_n \uparrow \mu A = \int g d\mu$.

Now the definition of A_n, the positivity of f_n and the positivity of integration tells us that $\int f_n d\mu \geq (1-\epsilon)\mu A_n$, so taking limits we see

$$\lim_{n\to\infty}\int f_n d\mu \geq (1-\epsilon)\lim_{n\to\infty}\mu A_n = (1-\epsilon)\int g d\mu$$

Now let $\epsilon \to 0$ to get the result.

To extend the result to arbitrary positive simple functions, first consider $g = c\mathbf{1}_A$ for $c > 0$. Note that we can apply the lemma to $\mathbf{1}_A$ and the functions $\frac{1}{c}f_n \uparrow \frac{1}{c}f$, to see that $\lim_{n\to\infty} \frac{1}{c}f_n \geq \mu A$ and multiply both sides by c.

Now consider a positive simple function in canonical form $g = c_1\mathbf{1}_{A_1} + \cdots + c_m\mathbf{1}_{A_m}$. Since g is in the canonical form, $c_i > 0$ for $i = 1, \ldots, m$. Also, $A_i \cap A_j = \emptyset$ for $i \neq j$ and therefore $g\mathbf{1}_{A_i} = c_i\mathbf{1}_{A_i}$. Now apply the lemma to each $g\mathbf{1}_{A_i}$ and the family $f_n\mathbf{1}_{A_i} \uparrow f\mathbf{1}_{A_i}$ and use linearity of integral and limits. □

Corollary 3.35. *Given a measurable positive function $f : (\Omega, \mathcal{A}, \mu) \to \overline{\mathbb{R}}_+$ and any sequence of positive simple functions $0 \leq f_1, f_2, \ldots$ such that $f_n \uparrow f$,*

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

Proof. As f_n are positive simple functions with $f_n \leq f$ we know each $\int f_n \leq \int f$ and therefore $\lim_{n\to\infty} \int f_n d\mu \leq \int f d\mu$.

To see the other inequality, pick $\epsilon > 0$, and a positive simple $0 \leq g \leq f$ such that $\int f d\mu - \epsilon \leq \int g d\mu$. Apply the above lemma and we see that $\int f d\mu - \epsilon \leq \int g d\mu \leq \lim_{n\to\infty} \int f_n d\mu$. Now let $\epsilon \to 0$ to see $\int f d\mu \leq \int \lim_{n\to\infty} f_n d\mu$. □

Lemma 3.36. *Given f, g positive measurable and $a, b \geq 0$,*

$$\int (af + bg)\, d\mu = a \int f d\mu + b \int g d\mu$$

and if $f \geq g$,

$$\int f d\mu \geq \int g d\mu$$

Proof. Linearity follows by taking $0 \leq f_n \uparrow f$ and $0 \leq g_n \uparrow g$ and noting that $0 \leq af_n + bg_n \uparrow af + bg$. Now apply linearity of integral of simple functions Lemma 3.31.

Monotonicity follows immediately from noting that any simple $0 \leq h \leq g$ also satisfies $0 \leq h \leq f$. □

Perhaps the most important basic theorems of measure theory are those that describe how limits and integrals behave; in particular what happens we exchange the order of limits and integrals. There are three commonly used variants and we are now ready to state and prove the first. Before we do that we illustrate three simple examples of the things that can go wrong when we exchange the order of limits and integrals. All of these examples assume the existence of a measure λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ such that $\lambda([a, b]) = b - a$. We will prove later that such a measure exists (it is the *Lebesgue measure* on $\mathbb{R}$).

Example 3.37 (Escape to horizontal infinity)**.** Consider the sequence of functions $f_n = \mathbf{1}_{[n,n+1]}$. Note that $\lim_{n\to\infty} \int f_n\, d\lambda = 1$ but $\int \lim_{n\to\infty} f_n\, d\lambda = 0$.

Example 3.38 (Escape to vertical infinity)**.** Consider the sequence of functions $f_n = n\mathbf{1}_{[0,\frac{1}{n}]}$. Note that $\lim_{n\to\infty} \int f_n\, d\lambda = 1$ but $\int \lim_{n\to\infty} f_n\, d\lambda = 0$.

Example 3.39 (Escape to width infinity)**.** Consider the sequence of functions $f_n = \frac{1}{n}\mathbf{1}_{[0,n-1]}$. Note that $\lim_{n\to\infty} \int f_n\, d\lambda = 1$ but $\int \lim_{n\to\infty} f_n\, d\lambda = 0$.

In all cases the integral of the limit is strictly less than the limit of the integrals and in all cases some amount of *mass* has *escaped to infinty*. The limit theorems amount to proving the fact that mass can only be lost when passing to the limit of a sequence of measurable functions and establishing generally useful hypotheses that prevent mass from escaping to infinity.

Theorem 3.40. *[Monotone Convergence Theorem]Given $f, f_1, f_2, \ldots$ positive measurable functions from $(\Omega, \mathcal{A}, \mu)$ to $\overline{\mathbb{R}}_+$ such that $0 \leq f_n \uparrow f$, we have $\int f_n d\mu \uparrow \int f d\mu$.*

Proof. Choose an approximation of each f_n by an increasing sequence of positive simple functions $g_{nk} \uparrow f_n$. For each $n, k > 0$, define $h_{nk} = g_{1k} \vee \cdots \vee g_{nk}$. Note that h_{nk} is increasing in both of its subscripts. Furthermore, note that $h_{nk} \leq f_n$ because $g_{ik} \leq f_i \leq f_n$ for $i \leq n$ by the monotonicity of f_n.

We claim that $h_{kk} \uparrow f$. To see this, for every $n > 0$, $h_{kk} \geq g_{nk}$ for $k \geq n$ and therefore

$$\lim_{k\to\infty} h_{kk} \geq \lim_{k\to\infty} g_{nk} = f_n$$

By taking limits we get the inequality

$$\lim_{k\to\infty} h_{kk} \geq \lim_{n\to\infty} f_n = f$$

We get the opposite inequality because f_n increases to f, we know that for every $k > 0$, $h_{kk} \leq f_k \leq f$ and therefore $\lim_{k\to\infty} h_{kk} \leq f$.

We have an approximation of $0 \leq h_{kk} \uparrow f$ by simple functions, now we can calculate the integral of f using h_{kk}

$$\int f \, d\mu = \lim_{k\to\infty} \int h_{kk} \, d\mu \leq \lim_{k\to\infty} \int f_k \, d\mu \leq \int f \, d\mu$$

where we have used the monotonicity of the integral in both inequalities. □

Corollary 3.41. *[Tonneli's Theorem for Integrals and Sums]Given $f_1, f_2, \ldots$ positive measurable functions from $(\Omega, \mathcal{A}, \mu)$ to $\overline{\mathbb{R}}_+$, we have*

$$\int \sum_{n=1}^{\infty} f_n d\mu = \sum_{n=1}^{\infty} \int f_n \, d\mu$$

Proof. Note that the sequence partial sums $\sum_{i=1}^{n} f_i$ is increasing in $n > 0$. Now use linearity of integral and apply the Montone Convergence Theorem. □

In some cases, we may have a sequence of positive functions that are not known to be increasing. In those cases, limits may not even exists but we still have a fundamental inequality

Theorem 3.42. *[Fatou's Lemma]Given $f_1, f_2, \ldots$ positive measurable functions from $(\Omega, \mathcal{A}, \mu)$ to $\overline{\mathbb{R}}_+$, then $\int \liminf_{n\to\infty} f_n d\mu \leq \liminf_{n\to\infty} \int f_n d\mu$.*

Proof. The proof uses the Monotone Convergence Theorem. To find an increasing sequence of positive measurable functions one needn't look further than the definition $\liminf_{n\to\infty} f_n = \lim_{n\to\infty} \inf_{k\geq n} f_k$. Since $\inf_{k\geq n} f_k \uparrow \liminf_{n\to\infty} f_n$, we know by Monotone Convergence that $\lim_{n\to\infty} \int \inf_{k\geq n} f_k \, d\mu = \int \liminf_{n\to\infty} f_n \, d\mu$.

However, we have the following calculation

$$\begin{aligned}
\inf_{k\geq n} f_k \leq f_k & \qquad \text{for all } k \geq n \text{ by definition of infimum} \\
\int \inf_{k\geq n} f_k \, d\mu \leq \int f_k \, d\mu & \qquad \text{for all } k \geq n \text{ by monotonicity of integral} \\
\int \inf_{k\geq n} f_k \, d\mu \leq \inf_{k\geq n} \int f_k \, d\mu & \qquad \text{by definition of infimum} \\
\lim_{n\to\infty} \int \inf_{k\geq n} f_k \, d\mu \leq \liminf_{n\to\infty} \int f_k \, d\mu & \qquad \text{taking limits and the definition of } \liminf \\
\int \liminf_{n\to\infty} f_n d\mu = & \qquad \text{by Monotone Convergence}
\end{aligned}$$

In prose, by the definition of the infimum $\inf_{k\geq n} f_k \leq f_k$ for every $k \geq n$, therefore monotonicity of the integral yields $\int \inf_{k\geq n} f_k d\mu \leq \int f_k d\mu$ for every $k \geq n$ and hence $\int \inf_{k\geq n} f_k d\mu \leq \inf_{k\geq n} \int f_k d\mu$. Now take the limit as $n \to \infty$. □

Our last task is to eliminate the assumption of positivity in the definition of the integral.

Definition 3.43. A measurable function f on the measure space $(\Omega, \mathcal{A}, \mu)$ is *integrable* if $\int |f| \, d\mu < \infty$. For any integrable f, we define $\int f \, d\mu = \int f_+ \, d\mu - \int f_- \, d\mu$.

We've defined the integral of an integrable function in terms of a canonical decomposition $f = f_+ - f_-$. It is occasionally useful to observe that any decomposition of an integrable function as a difference of positive measurable functions can be used to calculate the integral.

Lemma 3.44. *Suppose we are given a measure space $(\Omega, \mathcal{A}, \mu)$ and an integrable function $f : \Omega \to \mathbb{R}$. Suppose $f = f_1 - f_2$ where $f_i : \Omega \to \mathbb{R}$ are positive measurable with $\int f_i \, d\mu < \infty$. Then $\int f \, d\mu = \int f_1 \, d\mu - \int f_2 \, d\mu$.*

Proof. Write $f = f_+ - f_-$ and note that $f_1 \geq f_+$ and $f_2 \geq f_-$. For example either $f_+(\omega) = 0$ or $f_+(\omega) = f(\omega)$ and we know that $f_1(\omega) = f(\omega) + f_2(\omega) \geq f(\omega)$. We also know that $f_1 - f_+ = f_2 - f_-$ and we can see that $\int (f_1 - f_+) \, d\mu = \int (f_2 - f_-) \, d\mu < \infty$. Therefore by linearity of integral

$$\begin{aligned}
\int f \, d\mu &= \int f_+ \, d\mu - \int f_- \, d\mu \\
&= \int f_+ \, d\mu + \int (f_1 - f_+) \, d\mu - \int (f_2 - f_-) \, d\mu - \int f_- \, d\mu \\
&= \int f_1 \, d\mu - \int f_2 \, d\mu
\end{aligned}$$

□

Also linearity and monotonicity of integrals extend to the integrable case. Linearity of the integral subsumes the previous result.

Lemma 3.45. *Suppose we are given a measure space $(\Omega, \mathcal{A}, \mu)$ and integrable functions $f, g : \Omega \to \mathbb{R}$. The for $a, b \in \mathbb{R}$ we have $\int (af + bg) \, d\mu = a \int f \, d\mu + b \int g \, d\mu$ and if $f \geq g$ then $\int f \, d\mu \geq \int g \, d\mu$.*

Proof. Write $f = f_+ - f_-$ and $g = g_+ - g_-$. Define

$$\hat{f}_\pm = \begin{cases} af_\pm & \text{if } a \geq 0 \\ -af_\mp & \text{if } a < 0 \end{cases}$$

It is easy to see that $\hat{f}_\pm \geq 0$, $\int \hat{f}_\pm \, d\mu < \infty$, $af = \hat{f}_+ - \hat{f}_-$ and

$$\begin{aligned} \int af \, d\mu &= \int \hat{f}_+ \, d\mu - \int \hat{f}_- \, d\mu \\ &= \begin{cases} \int af_+ \, d\mu - \int af_- \, d\mu & \text{if } a \geq 0 \\ \int -af_- \, d\mu - \int -af_+ \, d\mu & \text{if } a < 0 \end{cases} \\ &= a \int f_+ \, d\mu - a \int f_- \, d\mu = a \int f \, d\mu \end{aligned}$$

The same construction and observations are true with g and $\hat{g}_\pm$. Then $af + bg = (\hat{f}_+ + \hat{g}_+) - (\hat{f}_- + \hat{g}_-)$ and we have

$$\begin{aligned} \int (af + bg) \, d\mu &= \int (\hat{f}_+ + \hat{g}_+) \, d\mu - \int (\hat{f}_- + \hat{g}_-) \, d\mu \\ &= \int \hat{f}_+ \, d\mu - \int \hat{f}_- \, d\mu + \int \hat{g}_+ \, d\mu - \int \hat{g}_- \, d\mu \\ &= a \int f \, d\mu + b \int g \, d\mu \end{aligned}$$

To see monotonicity, observe that $f \geq g$ if and only if $f_+ \geq g_+$ and $f_- \leq g_-$. □

Lastly, it is occasionally necessary to deal with integrating measurable functions that are either infinite on a set of measure zero or undefined on a set of measure zero. This is permissible by virtue of the following Lemma.

Definition 3.46. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. We say that a property hold *almost everywhere* if the set where the property does not hold has measure zero.

Lemma 3.47. *Let $f \geq 0$ be a measurable function on $(\Omega, \mathcal{A}, \mu)$. $\int f \, d\mu = 0$ if and only if $f = 0$ almost everywhere.*

Proof. Clearly this is true by definition for indicator functions. It also is true by positivity and linearity of integral for simple functions. For arbitrary $f \geq 0$, we take and increasing approximating sequence of simple functions $f_n \uparrow f$ and note that $\int f \, d\mu = 0$ and monotonicity of integral implies $\int f \, d\mu = 0$ for each n. Therefore, $f_n = 0$ almost everywhere for each n and therefore $f_n = 0$ almost everywhere for all n by taking a countable union. This implies $f = 0$ almost everywhere. If on the other hand we assume that $f = 0$ almost everywhere, then by the increasing nature of f_n, we see that $f_n = 0$ for all n almost everywhere and therefore $\int f_n \, d\mu = 0$ for every n. By Monotone Convergence we see that $\int f \, d\mu = 0$. □

Therefore, for the definition of integrability of f can be extended to allow f to be redefined arbitrarily on a set of measure zero.

We have the following limit theorem for limits of integrable functions.

Theorem 3.48. *[Dominated Convergence Theorem]Suppose we are given $f, f_1, f_2, \ldots$ and $g, g_1, g_2, \ldots$ measurable functions on $(\Omega, \mathcal{A}, \mu)$ such that $|f_n| \leq g_n$, $\lim_{n\to\infty} f_n = f$, $\lim_{n\to\infty} g_n = g$ and $\lim_{n\to\infty} \int g_n \, d\mu = \int g \, d\mu < \infty$. Then $\lim_{n\to\infty} \int f_n \, d\mu = \int f \, d\mu$.*

Proof. The trick here is to notice that by our assumption, $g_n \pm f_n \geq 0$ and we can apply Fatou's Lemma to both sequences. Doing so we get

$$\begin{aligned}
\int g\,d\mu \pm \int f\,d\mu &= \int \lim_{n\to\infty} g_n\,d\mu \pm \int \lim_{n\to\infty} f_n\,d\mu \\
&= \int \liminf_{n\to\infty} g_n\,d\mu \pm \int \liminf_{n\to\infty} f_n\,d\mu \\
&= \int \liminf_{n\to\infty} (g_n \pm f_n)\;d\mu \\
&\leq \liminf_{n\to\infty} \int (g_n \pm f_n)\;d\mu \\
&= \liminf_{n\to\infty} \int g_n\,d\mu + \liminf_{n\to\infty} \int \pm f_n\,d\mu \\
&= \int g\,d\mu + \liminf_{n\to\infty} \int \pm f_n\,d\mu
\end{aligned}$$

Now subtract $\int g\,d\mu$ from both sides of the equation and we get two inequalities $\pm\int f\,d\mu \leq \liminf_{n\to\infty} \int \pm f_n\,d\mu$. It remains to put these two inequalities together

$$\begin{aligned}
\limsup_{n\to\infty} \int f_n\,d\mu &= -\liminf_{n\to\infty} \int -f_n\,d\mu \\
&\leq \int f\,d\mu \\
&\leq \liminf_{n\to\infty} \int f_n\,d\mu
\end{aligned}$$

and the result is proved by the obvious fact that $\liminf f_n \leq \limsup f_n$. □

Most applications of Dominated Convergence only use the special case in which the sequence g_n is constant. We call out this special case as a corollary of the general theorem.

Corollary 3.49. *Suppose we are given $f, f_1, f_2, \ldots$ and g measurable functions on $(\Omega, \mathcal{A}, \mu)$ such that $|f_n| \leq g$, $\lim_{n\to\infty} f_n = f$ and $\int g\,d\mu < \infty$. Then $\lim_{n\to\infty} \int f_n\,d\mu = \int f\,d\mu$.*

Proof. Let $g_n = g$ for all $n > 0$ and use Theorem 3.48. □

Lemma 3.50. *Suppose we are given a measure space $(\Omega, \mathcal{A}, \mu)$, a measurable space $(S, \mathcal{S})$ and measurable function $f : \Omega \to S$. The function $\mu \circ f^{-1}(A) = \mu(f^{-1}(A))$ defines a measure on $(S, \mathcal{S})$. The measure $\mu \circ f^{-1}$ is called the* push forward *of μ by f.*

Proof. Clearly, $\mu \circ f^{-1}(\emptyset) = \mu(\emptyset) = 0$. If we are given disjoint $A_1, A_2, \ldots$ then by and the fact that μ is a measure, we know

$$\begin{aligned}
\mu \circ f^{-1}\left(\bigcup_{i=1}^{\infty} A_i\right) &= \mu\left(\bigcup_{i=1}^{\infty} f^{-1}(A_i)\right) && \text{by Lemma 3.7} \\
&= \sum_{i=1}^{\infty} \mu(f^{-1}(A_i)) && \text{by countable additivity of measure} \\
&= \sum_{i=1}^{\infty} \mu \circ f^{-1}(A_i) && \text{by definition of push forward}
\end{aligned}$$

□

Definition 3.51. For a probability space $(\Omega, \mathcal{A}, P)$, a measurable space $(S, \mathcal{S})$ and a random element $\xi : \Omega \to S$, the measure $P \circ \xi^{-1}$ is called the *distribution* or *law* of ξ. We often write $\mathcal{L}(\xi)$ for the law of ξ.

Lemma 3.52. *[Change of Variables]Suppose we are given a measure space $(\Omega, \mathcal{A}, \mu)$, a measurable space $(S, \mathcal{S})$, and measurable functions $f : \Omega \to S$ and $g : S \to \mathbb{R}$, then*

$$\int (g \circ f)\, d\mu = \int g\, d(\mu \circ f^{-1})$$

Whenever either side of the equality exists, the other does and they are equal.

Proof. To begin with we assume that $g = \mathbf{1}_A$ for $A \in \mathcal{S}$. The first simple claim is that $\mathbf{1}_A \circ f = \mathbf{1}_{f^{-1}(A)}$. This is seen by unfolding definitions for an $\omega \in \Omega$:

$$\begin{aligned}
(\mathbf{1}_A \circ f)(\omega) &= \mathbf{1}_A(f(\omega)) \\
&= \begin{cases} 1 & \text{if } f(\omega) \in A \\ 0 & \text{if } f(\omega) \notin A \end{cases} \\
&= \begin{cases} 1 & \text{if } \omega \in f^{-1}(A) \\ 0 & \text{if } \omega \notin f^{-1}(A) \end{cases} \\
&= \mathbf{1}_{f^{-1}(A)}(\omega)
\end{aligned}$$

Using this fact the result of the theorem follows for $\mathbf{1}_A$ by another simple calculation

$$\begin{aligned}
\int \mathbf{1}_A\, d(\mu \circ f^{-1}) &= (\mu \circ f^{-1})(A) \\
&= \mu(f^{-1}(A)) \\
&= \int \mathbf{1}_{f^{-1}(A)}\, d\mu \\
&= \int (\mathbf{1}_A \circ f)\, d\mu
\end{aligned}$$

Next we assume that $g = c_1 \mathbf{1}_{A_1} + \cdots + c_n \mathbf{1}_{A_n}$ is a simple function. As a general property of the linearity of composition of functions we can see that

$$g \circ f = c_1(\mathbf{1}_{A_1} \circ f) + \cdots + c_n(\mathbf{1}_{A_n} \circ f)$$

Coupling this with the result for indicator functions and linearity of integral we get

$$\begin{aligned}\int g\, d(\mu \circ f^{-1}) &= \sum_{i=1}^{n} c_i \int \mathbf{1}_{A_i}\, d(\mu \circ f^{-1}) \\ &= \sum_{i=1}^{n} c_i \int (\mathbf{1}_{A_i} \circ f)\, d\mu \\ &= \int \sum_{i=1}^{n} c_i (\mathbf{1}_{A_i} \circ f)\, d\mu \\ &= \int (g \circ f)\, d\mu\end{aligned}$$

Next we suppose that g is a positive measurable function. We know that we can find an increasing sequence of positive simple functions $g_n \uparrow g$. Note that $g \circ f$ is positive measurable, $g_n \circ f$ is positive simple and $g_n \circ f \uparrow g \circ f$. Now can use the result proven for simple functions and Monotone Convergence

$$\begin{aligned}\int g\, d(\mu \circ f^{-1}) &= \lim_{n\to\infty} \int g_n\, d(\mu \circ f^{-1}) && \text{by Monotone Convergence} \\ &= \lim_{n\to\infty} \int (g_n \circ f)\, d\mu && \text{by result for simple functions} \\ &= \int (g \circ f)\, d\mu && \text{by Monotone Convergence}\end{aligned}$$

The last step is to consider an integrable g. Write it as $g = g_+ - g_-$ for $g_\pm$ positive and use linearity of the integral and the result just proven for positive functions. □

Definition 3.53. Suppose we are given a measure space $(\Omega, \mathcal{A}, \mu)$ and a positive measurable function $f : \Omega \to \mathbb{R}_+$. We define the measure $f \cdot \mu$ by the formula

$$(f \cdot \mu)(A) = \int \mathbf{1}_A \cdot f\, d\mu = \int_A f\, d\mu$$

If ν is a measure of the above form, then we say that f is a *μ-density* of ν.

Lemma 3.54. *Suppose we are given a measure space $(\Omega, \mathcal{A}, \mu)$, a positive measurable function $f : \Omega \to \mathbb{R}_+$ and and measurable function $g : \Omega \to \mathbb{R}$, then*

$$\int fg\, d\mu = \int g\, d(f \cdot \mu)$$

Whenever either side of the equality exists, the other does and they are equal.

Proof. First assume that $g = \mathbf{1}_A$ is an indicator function. The result is just the definition of the measure $f \cdot \mu$:

$$\int \mathbf{1}_A\, d(f \cdot \mu) = (f \cdot \mu)(A) = \int \mathbf{1}_A \cdot f\, d\mu$$

Next assume that $g = \sum_{i=1}^{n} c_i \mathbf{1}_{A_i}$ is a simple function. Then we can simply apply linearity of the integral

$$\begin{aligned}\int g\, d(f \cdot \mu) &= \sum_{i=1}^{n} c_i \int \mathbf{1}_{A_i}\, d(f \cdot \mu) \\ &= \sum_{i=1}^{n} c_i \int \mathbf{1}_{A_i} \cdot f\, d\mu \\ &= \int g \cdot f\, d\mu\end{aligned}$$

For a positive measurable g we pick an increasing approximation by simple functions $g_n \uparrow g$. We note that for positive f we have $g_n \cdot f$ positive (not necessarily simple) with $g_n \cdot f \uparrow g \cdot f$. Thus,

$$\begin{aligned}\int g\, d(f \cdot \mu) &= \lim_{n\to\infty} \int g_n\, d(f \cdot \mu) && \text{definition of integral} \\ &= \lim_{n\to\infty} \int g_n \cdot f\, d\mu && \text{by result for simple functions} \\ &= \int g \cdot f\, d\mu && \text{by Monotone Convergence}\end{aligned}$$

The last step is to pick an integrable $g = g_+ - g_-$ and use linearity of integral. Note also that in this case the two integrals in question are defined for exactly the same g. □

3.3.1. *Standard Machinery.* We've put together a collection of definitions and tools for talking about integration and proving theorems about integration. What is probably not clear at this point is that there are some very useful patterns for how these defintitions, lemmas and theorems are used. One such pattern is so commonplace that I have heard it called the *standard machinery*. Suppose one wants to show a result about general measurable functions. A proof of the result using the standard machinery proceeds by

(i) Demonstrating the result for indicator functions.
(ii) Arguing by linearity that the result holds for simple functions.
(iii) Showing the result holds for non-negative measurable functions by approximating by an increasing limit of simple functions and using the Monotone Convergence Theorem.
(iv) Showing the result for arbitrary functions by expressing an arbitrary measurable function as a difference of non-negative measurable functions.

The proof of Lemma 3.52 and Lemma 3.54 are examples of proofs using the standard machinery. It is a good idea to get very comfortable with such arguments as it is quite common in many texts to leave any such proof as an exercise for the reader. An important refinement of the standard machinery involves using a monotone class argument with the π-λ Theorem to demonstrate the result for all indicator functions. Recall that to do that, one shows that the collection of sets whose indicator functions satisfy the theorem is a λ-system and to then prove the result a

π-system of sets such that the π-system generates the σ-algebra of the measurable space.

3.4. **Products of Measurable Spaces.** Given a collection of measurable spaces there is a standard construction that makes the cartesian product of the spaces into a measurable space.

Definition 3.55. Suppose we are given an index set T and for each $t \in T$ we have a measurable space $(\Omega_t, \mathcal{A}_t)$. The *product σ-algebra* $\bigotimes_t \mathcal{A}_t$ on the cartesian product $\prod_t \Omega_t$ is the σ-algebra generated by all one dimensional *cylinder sets* $A_t \times \prod_{s \neq t} \Omega_s$ for $A_t \in \mathcal{A}_t$.

TODO: Show that this is the smallest σ-algebra that make the projections measurable

TODO: Show that the product of Borel σ-algebras is the Borel σ-algebra with respect to the product topology in the separable case. Note that the non-separable case is more subtle and in fact turns out to be important (especially in statistics)!

The following is an important scenario that we shall often encounter. Suppose we have a measurable space $(\Omega, \mathcal{A})$ and a collection of measurable functions $f_t : \Omega \to (S_t, \mathcal{S}_t)$. From a purely set-theoretic point of view this specification of functions is in fact equivalent to the specification of a single function $f : \Omega \to \prod_t S_t$ (i.e. if we let $\pi_s : \prod_t S_t \to S_s$ be the projections then we define $\pi_s(f(\omega)) = f_s(\omega)$).

Lemma 3.56. *Given a collection of measurable functions $f_t : \Omega \to S_t$ and the equivalent function $f : \Omega \to \prod_t S_t$ we have $\sigma(\bigwedge_t \sigma(f_t)) = \sigma(f)$.*

Proof. To see that $\sigma(\bigwedge_t \sigma(f_t)) \subset \sigma(f)$ it suffices to show that $\sigma(f_t) \subset \sigma(f)$ for all $t \in T$. This follows since for any $A_t \in \mathcal{S}_t$, we have $f_t^{-1}(A) = f^{-1}(A \times \prod_{s \neq t} \Omega_s)$. This fact also shows that $\sigma(f) \subset \sigma(\bigwedge_t \sigma(f_t))$ since the cylinder sets $A \times \prod_{s \neq t} \Omega_s$ generate $\otimes_t \mathcal{S}_t$ by Lemma 3.12. □

3.5. **Outer Measures and Lebesgue Measure on the Real Line.** To construct Lebesgue measure on the real line, one proceeds by demonstrating that one may construct a measure by first constructing a more primitive object called an outer measure and then proving that outer measure become measures when restricted to an appropriate collection of sets. Having redefined the problem as the construction of outer measure, one constructs outer measure on real line in a hands on way.

Much of this process that has broader applicability than just the real line, therefore we state and prove the results in the more general case. TODO: Come up with some intuition about outer measure (more specifically Caratheodory's characterizaiton of sets measurable with respect to an outer measure; it says in some sense that a measurable set and its complement have aren't *too* entangled with one another).

Definition 3.57. Given a set Ω, an *outer measure* is a positive function $\mu : 2^\Omega \to \overline{\mathbb{R}}_+$ satisfying

(i) $\mu(\emptyset) = 0$
(ii) If $A \subset B$, then $\mu(A) \leq \mu(B)$
(iii) Given $A_1, A_2, \cdots \subset \Omega$, then $\mu\left(\bigcup_{i=1}^\infty A_i\right) \leq \sum_{i=1}^\infty \mu(A_i)$.

Definition 3.58. Given a set Ω with outer measure μ, we say a set $A \subset \Omega$ is *μ-measurable* if for every $B \subset \Omega$,

$$\mu(B) = \mu(A \cap B) + \mu(A^c \cap B)$$

Remark 3.59. For every $A, B \subset \Omega$, we have from finite subadditivity of outer measure

$$\mu(B) = \mu((A \cap B) \cup (A^c \cap B)) \leq \mu(A \cap B) + \mu(A^c \cap B)$$

and therefore to show μ-measurability we only need to show the reverse inequality.

Lemma 3.60. *Given a set Ω with an outer measure μ, let $\mathcal{A}$ be the collection of μ-measurable sets. Then $\mathcal{A}$ is a σ-algebra and the restriction of μ to $\mathcal{A}$ is a measure.*

Proof. We first note that $A \in \mathcal{A}$ if and only if $A^c \in \mathcal{A}$ since the defining condition of $\mathcal{A}$ is symmetric in A and A^c.

Next we show $\emptyset \in \mathcal{A}$. To see this, take $B \subset \Omega$,

$$\begin{aligned} \mu(B) &= \mu(\emptyset) + \mu(B) && \text{since } \mu(\emptyset) = 0 \\ &= \mu(\emptyset \cap B) + \mu(B \cap \Omega) \end{aligned}$$

Next we show that $\mathcal{A}$ is closed under finite intersection. Pick $A, B \in \mathcal{A}$ and $E \subset \Omega$ and calculate

$$\begin{aligned} \mu(E) &= \mu(E \cap A) + \mu(E \cap A^c) && \text{since } A \in \mathcal{A} \\ &= \mu(E \cap A \cap B) + \mu(E \cap A \cap B^c) + \mu(E \cap A^c) && \text{since } B \in \mathcal{A} \\ &\geq \mu(E \cap (A \cap B)) + \mu(E \cap A \cap B^c \cup E \cap A^c) && \text{by subadditivity} \\ &\geq \mu(E \cap (A \cap B)) + \mu(E \cap (A \cap B)^c) && \text{by monotonicity of } \mu \end{aligned}$$

and we have noted that it suffices to show this inequality to show $A \cap B \in \mathcal{A}$. Now by De Morgan's Law we conclude that $\mathcal{A}$ is closed under finite union.

Now we turn to consider the behavior of μ and show that μ is finitely and countably additive over disjoint unions; in fact we show a bit more. We let $A, B \in \mathcal{A}$ and let $E \subset \Omega$ be disjoint.

$$\begin{aligned} \mu(E \cap (A \cup B)) &= \mu(E \cap (A \cup B) \cap A) + \mu(E \cap (A \cup B) \cap A^c) && \text{since } A \in \mathcal{A} \\ &= \mu(E \cap A) + \mu(E \cap B) && \text{by set algebra} \end{aligned}$$

It is easy to see that one can do induction to extend the above result to all finite disjoint unions. Now let $A_1, A_2, \cdots \in \mathcal{A}$ and $E \subset \Omega$. Define $U_n = \bigcup_{i=1}^n A_i$ and $U = \bigcup_{i=1}^\infty A_i$.

$$\begin{aligned} \mu(E \cap U) &\geq \mu(E \cap U_n) && \text{by monotonicity} \\ &= \sum_{i=1}^n \mu(E \cap A_i) && \text{by finite additivity and disjointness of } A_i \end{aligned}$$

Now take the limit we have $\mu(E \cap U) \geq \sum_{i=1}^\infty \mu(E \cap A_i)$. Applying subadditivity of μ we get the opposite inequality and we have shown

$$\mu(E \cap \bigcup_{i=1}^\infty A_i) = \sum_{i=1}^\infty \mu(E \cap A_i)$$

In particular, we can take $E = \Omega$ to show that μ is countably additive over disjoint unions.

Having shown how to calculate μ over countable disjoint unions, we can show that $U \in \mathcal{A}$. For every $n > 0$,

$$\begin{aligned}\mu(E) &= \mu(E \cap U_n) + \mu(E \cap U_n^c) \\ &\geq \sum_{i=1}^{n} \mu(E \cap A_i) + \mu(E \cap U) \qquad \text{by subadditivity and monotonicity}\end{aligned}$$

Take the limit and use the previous claim to see

$$\begin{aligned}\mu(E) &\geq \sum_{i=1}^{\infty} \mu(E \cap A_i) + \mu(E \cap U) \\ &= \mu(E \cap U) + \mu(E \cap U^c)\end{aligned}$$

thereby showing $U \in \mathcal{A}$.

The last thing to show is that a countable union of elements of $\mathcal{A}$ are in $\mathcal{A}$. This follows from what we have shown about countable disjoint unions since we have already proven this for complements, finite unions and intersections and therefore for any $A_1, A_2, \ldots$ we can define $B_n = A_n \setminus \bigcup_{i=1}^{n-1} A_i$ so that $\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i$ with the B_i disjoint. □

To define *Lebesgue measure* on $\mathbb{R}$ we will leverage the construction above and first define an outer measure by approximating by intervals. Given an interval $I \subset \mathbb{R}$, let $|I|$ be length of I.

Theorem 3.61. *[Lebesgue Measure]There exists a unique measure λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R})$ such that $\lambda(I) = |I|$ for all intervals $I \subset \mathbb{R}$.*

Before we begin the proof of the theorem we need first construct an outer measure.

Lemma 3.62. *[Lebesgue Outer Measure]Define the function $\lambda : 2^{\mathbb{R}} \to \mathbb{R}$ defined by*

$$\lambda(A) = \inf_{\{I_k\}} \sum_k |I_k|$$

where the infimum ranges over countable covers of A by intervals. Then λ is an outer measure. In addition, $\lambda(I) = |I|$ for every interval $I \subset \mathbb{R}$.

Proof. It is clear that λ is positive and $\lambda(\emptyset) = 0$. It is also clear that λ is increasing since for any $A \subset B \subset \mathbb{R}$ any cover of B is also a cover of A.

To see subadditivity, take $A_1, A_2, \cdots \subset \mathbb{R}$. Pick $\epsilon > 0$ and then for each A_n we take a countable cover by intervals $I_{n1}, I_{n2}, \ldots$ such that $\lambda(A_n) \geq \sum_{k=1}^{\infty} |I_{nk}| - \frac{\epsilon}{2^n}$. Then, the collection of intervals I_{nk} for $n, k > 0$ is an countable cover of $\bigcup_{i=1}^{\infty} A_i$ and therefore

$$\begin{aligned}\lambda\left(\bigcup_{i=1}^{\infty} A_i\right) &\leq \sum_{n=1}^{\infty}\sum_{k=1}^{\infty} |I_{nk}| \\ &\leq \sum_{n=1}^{\infty} \left(\lambda(A_n) + \frac{\epsilon}{2^n}\right) \\ &= \sum_{n=1}^{\infty} \lambda(A_n) + \epsilon\end{aligned}$$

Now let $\epsilon \to 0$ and we have proven subadditivity.

To prove that $\lambda(I) = |I|$, we first consider intervals of the form $I = [a, b]$ with $a < b$. The family of intervals $(a - \epsilon, b + \epsilon)$ for $\epsilon > 0$ shows that $\lambda I \leq |I|$ so we only need to show the opposite inequality. Suppose we are given a countable cover by open intervals $I_1, I_2, \ldots$. We need to show that $|I| \leq \sum_{k=1}^{\infty} |I_k|$. By the Heine-Borel Theorem (Theorem 2.30), there is a finite subcover $I_1, \ldots, I_n$ and it suffices to show that $|I| \leq \sum_{k=1}^{n} |I_k|$ for the finite subcover.

For finite covers we can proceed by induction. To begin, consider a cover by a single interval. For any $J \supset I$ we know that $|J| \geq |I|$.

For the induction step, assume that $\inf_{\{I_k\}} \sum_{k=1}^{n} |I_k| = |I|$ where the infimum is over covers by n intervals. Take a cover of I by $n+1$ intervals $I_1, \ldots, I_{n+1}$. There exists an I_k such that $b \in I_k$. If we write $I_k = (a_k, b_k)$, then the rest of the I_j form a cover of $[a, a_k]$.

$$\begin{aligned} |I| &= (b - a_k) + (a_k - a) \\ &\leq |I_k| + \sum_{m \neq k} |I_m| && \text{by induction hypothesis applied to } [a, a_k] \\ &= \sum_{m} |I_m| \end{aligned}$$

It remains to eliminate the restriction to bounded closed intervals. Clearly every cover of $[a, b]$ by open intervals is a cover of (a, b). On the other hand, every countable cover of (a, b) can be extended to a countable cover of $[a, b]$ by adding at most two arbitrarily small intervals of the form $(a - \epsilon, a + \epsilon)$ and $(b - \epsilon, b + \epsilon)$. An *epsilon of room* argument shows that $\lambda(a, b) = \lambda[a, b]$. Monotonicity of λ shows the same is true for half open intervals.

TODO: Show that outer measure of infinite intervals is infinite. □

Definition 3.63. A subset $A \subset \mathbb{R}$ is *Lebesgue measurable* if A is λ-measurable with respect to the Lebesgue outer measure.

Lemma 3.64. *Every Borel measurable $A \subset \mathbb{R}$ is also Lebesgue measurable.*

Proof. Since we know that the collection of Lebesgue measurable sets is a σ-algebra, and we know that the Borel algebra on $\mathbb{R}$ is generated by intervals of the form $(-\infty, x]$, it suffices to show that each such interval is Lebesgue measurable.

Take an interval $I = (-\infty, x]$, a set $E \subset \mathbb{R}$ and $\epsilon > 0$. Pick a countable covering $I_1, I_2, \ldots$ of E by open intervals so that $\lambda(E) + \epsilon \geq \sum_{k=1}^{\infty} |I_k|$.

$$\begin{aligned} \lambda(E) + \epsilon &\geq \sum_{k=1}^{\infty} |I_k| \\ &= \sum_{k=1}^{\infty} |I_k \cap I| + \sum_{k=1}^{\infty} |I_k \cap I^c| \\ &= \sum_{k=1}^{\infty} \lambda(I_k \cap I) + \sum_{k=1}^{\infty} \lambda(I_k \cap I^c) \\ &\geq \lambda\left(\bigcup_{k=1}^{\infty} I_k \cap I\right) + \lambda\left(\bigcup_{k=1}^{\infty} I_k \cap I^c\right) && \text{by subadditivity} \\ &\geq \lambda(E \cap I) + \lambda(E \cap I^c) \end{aligned}$$

where the last line holds because $I_k \cap I$ is a countable cover of $E \cap I$ and similarly for $E \cap I^c$. Now let $\epsilon \to 0$ to get the result.

TODO: Actually $I_k \cap I$ are half open intervals. The proof needs to be extended to handle this fact. Presumably an $\frac{\epsilon}{2^n}$ argument works here. Note most definitions of Lebesgue outer measure do not restrict to open covers (then you have to pay the cost of the $\frac{\epsilon}{2^n}$ argument to apply Heine Borel). □

Lemma 3.65 (Uniqueness of measure)**.** *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space with μ a finite measure. Suppose ν is a finite measure on $(\Omega, \mathcal{A})$ such that there is a π-system $\mathcal{C}$ such that $\sigma(\mathcal{C}) = \mathcal{A}$, $\Omega \in \mathcal{C}$ and for all $A \in \mathcal{C}$ we have $\mu(A) = \nu(A)$, then $\mu = \nu$.*

If we assume that μ a σ-finite measure and ν is a σ-finite measure such that there exists a partition $\Omega = \Omega_1 \cup \Omega_2 \cup \cdots$ with $\mu(\Omega_n) = \nu(\Omega_n) < \infty$, the result holds as well.

Proof. First we assume that μ (and then by hypothesis ν) is finite. We apply a monotone class argument. Consider the collection $\mathcal{D}$ of $A \in \mathcal{A}$ such that $\mu(A) = \nu(A)$. We claim that this collection is a λ-system. Since we have assumed $\mu(\Omega) = \nu(\Omega)$ we have that $\Omega \in \mathcal{D}$. Now suppose $A \subset B \in \mathcal{D}$. By additivity of measure and finiteness of μ and ν,

$$\mu(B \setminus A) = \mu(B) - \mu(A) = \nu(B) - \nu(A) = \nu(B \setminus A)$$

Now we assume $A_1 \subset A_2 \subset \cdots \in \mathcal{D}$. By continuity of measure (Lemma 3.27)

$$\mu\left(\bigcup_i A_i\right) = \lim_{n\to\infty} \mu(A_i) = \lim_{n\to\infty} \nu(A_i) = \nu\left(\bigcup_i A_i\right)$$

Application of the π-λ Theorem (Theorem 3.24) together with the fact that $\sigma(\mathcal{C}) = \mathcal{A}$ shows that equality holds on all of $\mathcal{A}$.

Now we handle to the σ-finite case. We a partition $\Omega = \Omega_1 \cup \Omega_2 \cup \cdots$ such that $\mu(\Omega_n) = \nu(\Omega_n) < \infty$ for all n. Denote μ_n and ν_n the restriction of μ and ν to the set Ω_n. We note that μ_n and ν_n each satisfy the hypothesis of the lemma for the finite measure case (e.g. $\mu_n(A) = \mu(\Omega_n \cap A)$). Therefore we can conclude that $\mu_n = \nu_n$ on all of $\mathcal{A}$ for all n. For any $A \in \mathcal{A}$ define $A_n = \cup_{k=1}^n \Omega_k \cap A$ note that

$$\mu(A_n) = \sum_{k=1}^n \mu_k(A) = \sum_{k=1}^n \nu_k(A) = \nu(A_n)$$

and $A_1 \subset A_2 \subset \cdots$ with $\cup_{n=1}^\infty A_n = A$. Now apply continuity of measure (Lemma 3.27) to see that $\mu(A) = \nu(A)$. □

TODO: Do we need to assume that there is a partition with $\mu(\Omega_n) = \nu(\Omega_n)$ or can it be derived from the fact that $\sigma(\mathcal{C}) = \mathcal{A}$. Is suspect it can be derived but the applications we have in mind it is trivial to generate the partition by hand.

Now we are ready to prove the existence and uniqueness of Lebesgue measure (Theorem 3.61).

Proof. The existence of Lebesgue measure clearly follows from Lemma 3.60 applied to the outer measure constructed in Lemma 3.62. The fact that the σ-algebra of the restriction contains the Borel sets follows from Lemma 3.64.

It remains to show uniqueness. Now clearly the collection of intervals is closed under finite intersections hence is a π-system that generates $\mathcal{B}(\mathbb{R})$. Furthermore, $\mathbb{R} = \cup_{n=-\infty}^{\infty}(n, n+1]$ so we may apply Lemma 3.65 to get uniqueness. □

Definition 3.66. A measure space $(\Omega, \mathcal{A}, \mu)$ is *σ-finite* if there exists a countable partition $\Omega = \Omega_1 \cup \Omega_2 \cup \cdots$ such that $\mu(\Omega_i) < \infty$.

3.5.1. *Abstract Version of Caratheodory Extension.* The construction of Lebesgue measure we have given actually has a broad generalization which we present here.

Definition 3.67. A non-empty collection $\mathcal{A}_0$ of subsets of a set Ω is called a *Boolean algebra* if given any $A, B \in \mathcal{A}_0$ we have

(i) $A^c \in \mathcal{A}_0$
(ii) $A \cup B \in \mathcal{A}_0$
(iii) $A \cap B \in \mathcal{A}_0$

Note that it is trivial induction argument to extend the closure properties to arbitrary finite unions and intersections.

Definition 3.68. A *pre-measure* on a Boolean algebra $(\Omega, \mathcal{A}_0)$ is a function $\mu_0 : \mathcal{A}_0 \to \overline{\mathbb{R}}_+$ such that

(i) $\mu_0(\emptyset) = 0$
(ii) For any $A_1, A_2, \ldots \in \mathcal{A}_0$ such that the A_n are disjoint and $\cup_{n=1}^{\infty} A_n \in \mathcal{A}_0$, we have $\mu_0(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu_0(A_n)$.

Lemma 3.69. *A pre-measure is finitely additive and montonic. That is to say given any disjoint $A_1, \ldots, A_n \in \mathcal{A}_0$ we have $\mu_0(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mu_0(A_i)$ and given $A \subset B$ with $A, B \in \mathcal{A}_0$, we have $\mu_0(A) \leq \mu_0(B)$.*

Proof. Finite additivity follows by extending the finite sequence to an infinite sequence by appending copies of the emptyset and using the fact that $\mu_0(\emptyset) = 0$. Monotonicity follows from finite additivity by writing $B = A \cup B \setminus A$ so that $\mu_0(B) = \mu_0(A) + \mu_0(B \setminus A) \geq \mu_0(A)$. □

Our goal is to show that any pre-measure on a Boolean algebra $\mathcal{A}_0$ may be extended to a measure on a σ-algebra containing $\mathcal{A}_0$. We proceed in four steps

1) Define an outer measure μ^* from μ_0
2) Show that all sets in $\mathcal{A}_0$ are μ^*-measurable.
3) Show that for all sets $A \in \mathcal{A}_0$, $\mu^*(A) = \mu_0(A)$.
4) Use the Caratheodory restriction to create a σ-algebra and measure.

Lemma 3.70. *Given a pre-measure μ_0 on a Boolean algebra $(\Omega, \mathcal{A}_0)$ then the set function $\mu^* : 2^{\Omega} \to \overline{\mathbb{R}}_+$ defined by*

$$\mu^*(A) = \inf\{\sum_{n=1}^{\infty} \mu_0(A_n) \mid A \subset \cup_{n=1}^{\infty} A_n \text{ and } A_n \in \mathcal{A}_0 \text{ for all } n\}$$

is an outer measure.

Proof. Because $\mu_0(\emptyset)$ and $\emptyset \subset \emptyset$ we see that $\mu^*(\emptyset) = 0$.

Suppose we are given $A \subset B$. Then if we have a cover $B \subset \cup_{n=1}^{\infty} B_n$ where $B_n \in \mathcal{A}_0$, then this is also a cover of A. Therefore $\mu^*(A)$ is an infimum over a larger collection of covers than that used in calculating $\mu^*(B)$ hence $\mu^*(A) \leq \mu^*(B)$ (we could actually pick an ϵ and an approximating cover as below then let $\epsilon \to 0$).

Now to show subadditivity. Let $A_1, A_2, \ldots$ be a sequence of arbitrary subsets of Ω. If any $\mu^*(A_n) = \infty$ then we automatically know $\mu^*(\cup_{n=1}^\infty A_n) \leq \sum_{n=1}^\infty \mu^*(A_n)$, so we may assume that all $\mu^*(A_n) < \infty$. Let $\epsilon > 0$ be given and for each n we pick $B_{1n}, B_{2n}, \ldots$ such that $A_n \subset \cup_{m=1}^\infty B_{mn}$ and $\sum_{m=1}^\infty \mu_0(B_{mn}) \leq \mu^*(A_n) + \frac{\epsilon}{2^n}$. Now, we also have that $\cup_{n=1}^\infty A_n \subset \cup_{n=1}^\infty \cup_{m=1}^\infty B_{mn}$ and therefore we know that $\mu^*(\cup_{n=1}^\infty A_n) \leq \sum_{n=1}^\infty \sum_{m=1}^\infty \mu_0(B_{mn}) \leq \sum_{n=1}^\infty \mu^*(A_n) + \epsilon$. Since ϵ was arbitrary, we have $\mu^*(\cup_{n=1}^\infty A_n) \leq \sum_{n=1}^\infty \mu^*(A_n)$ so subadditivity is proven. □

Lemma 3.71. *Given a pre-measure μ_0 on a Boolean algebra $(\Omega, \mathcal{A}_0)$ and the outer measure μ^* constructed in Lemma 3.70, if $A \in \mathcal{A}_0$ then A is μ^*-measurable.*

Proof. Let $A \in \mathcal{A}_0$ and $B \subset \Omega$ and we have to show $\mu^*(B) \geq \mu^*(A \cap B) + \mu^*(A^c \cap B)$. Pick $B_1, B_2, \ldots$ such that $B_n \in \mathcal{A}_0$ for all n and $\sum_{n=1}^\infty \mu_0(B_n) \leq \mu^*(B) + \epsilon$. By finite additivity of μ_0 and the fact that $A, B_n \in \mathcal{A}_0$, we can write $\mu_0(B_n) = \mu_0(A \cap B_n) + \mu_0(A^c \cap B_n)$ and therefore $\sum_{n=1}^\infty \mu_0(A \cap B_n) + \sum_{n=1}^\infty \mu_0(A^c \cap B_n) \leq \mu^*(B) + \epsilon$. On the other hand, we know that $A \cap B \subset \cup_{n=1}^\infty A \cap B_n$ so $\mu^*(A \cap B) \leq \sum_{n=1}^\infty \mu_0(A \cap B_n)$ and similarly with A^c. Therefore $\mu^*(A \cap B) + \mu^*(A^c \cap B) \leq \mu^*(B) + \epsilon$. Take the limit as ϵ goes to zero and we are done. □

Lemma 3.72. *Given a pre-measure μ_0 on a Boolean algebra $(\Omega, \mathcal{A}_0)$ and the outer measure μ^* constructed in Lemma 3.70, if $A \in \mathcal{A}_0$ then $\mu^*(A) = \mu_0(A)$.*

Proof. Suppose we are given $A \in \mathcal{A}_0$. Since A is a singleton cover of itself, we know that $\mu^*(A) \leq \mu_0(A)$. It remains to show $\mu_0(A) \leq \mu^*(A)$. If $\mu^*(A) = \infty$ then this is trivally true so we may assume $\mu^*(A) < \infty$. Let $\epsilon > 0$ be given and pick $A_1, A_2, \ldots \in \mathcal{A}_0$ such that $A \subset \cup_{n=1}^\infty A_n$ and $\sum_{n=1}^\infty \mu_0(A_n) \leq \mu^*(A) + \epsilon$. Our goal now is to shrink each of the A_n so that we wind up with a partition of A. Then we will be able to apply the countable additivity of pre-measures.

First, we convert the cover by A_n into a disjoint cover of A. Let $B_1 = A_1$ and then define $B_n = A_n \setminus (A_1 \cup \cdots \cup A_{n-1})$ for $n > 1$. By construction, the B_n are disjoint and $\cup_{i=1}^n B_i = \cup_{i=1}^n A_i$. Furthermore $B_n \subset A_n$ so by monotonicity of μ_0 we have $\mu_0(B_n) \leq \mu_0(A_n)$. Now have $A \subset \cup_{n=1}^\infty B_n$ with B_n disjoint, $B_n \in \mathcal{A}_0$ for all n and $\sum_{n=1}^\infty \mu_0(B_n) \leq \mu^*(A) + \epsilon$.

Lastly we convert the disjoint cover B_n into a partitioning of A. Consider $C_n = B_n \cap A$. We still have $C_n \in \mathcal{A}_0$, C_n disjoint and montonicity implies $\sum_{n=1}^\infty \mu_0(C_n) \leq \mu^*(A) + \epsilon$. But now we have $\cup_{n=1}^\infty C_n = A \in \mathcal{A}_0$ so we may apply countable additivity of premeasure to conclude $\mu_0(A) = \sum_{n=1}^\infty \mu_0(C_n) \leq \mu^*(A) + \epsilon$. Once again, ϵ was arbitrary so let it go to zero and we are done. □

TODO: construction that takes us from a semiring to a Boolean algebra. It is often convenient to start a construction of a measure with a collection of sets that is so small that it doesn't even form a Boolean algebra. For example when constructing Lebesgue measure on $\mathbb{R}$ we were really motivated by a desire that the measure of an interval $(a, b]$ should be $b - a$, yet the set of such intervals on $\mathbb{R}$ is not a Boolean algebra.

Definition 3.73. A set $\mathcal{D} \subset 2^\Omega$ is called a *semiring* if

(i) $\emptyset \in \mathcal{D}$
(ii) if $A, B \in \mathcal{D}$ then $A \cap B \in \mathcal{D}$
(iii) if $A, B \in \mathcal{D}$ then there exist disjoint $C_1, \ldots, C_n \in \mathcal{D}$ such that $A \setminus B = \cup_{j=1}^n C_j$

Example 3.74. The set of intervals $(a, b]$ is a semiring.

TODO: Other constructions of semirings (e.g. products)

Definition 3.75. A set $\mathcal{R} \subset 2^\Omega$ is called a *ring* if

(i) $\emptyset \in \mathcal{R}$
(ii) if $A, B \in \mathcal{R}$ then $A \cup B \in \mathcal{R}$
(iii) if $A, B \in \mathcal{R}$ then $A \setminus B \in \mathcal{R}$

Lemma 3.76. *If $\mathcal{D}$ is a semiring then $\mathcal{R} = \{\cup_{j=1}^n C_j \mid C_j \in \mathcal{D}$ and the C_j are disjoint$\}$ is a ring. Furthermore it is the smallest ring containing $\mathcal{D}$.*

Proof. The fact that $\emptyset \in \mathcal{R}$ is immediate. Suppose we are given $\cup_{i=1}^n A_i$ and $\cup_{j=1}^m B_j$ in $\mathcal{R}$. Then we have

$$(\cup_{i=1}^n A_i) \cap \left(\cup_{j=1}^m B_j\right) = \cup_{i=1}^n \cup_{j=1}^m A_i \cap B_j \tag{1}$$

which is in $\mathcal{R}$ because each $A_i \cap B_j \in \mathcal{D}$ and they are disjoint by the disjointness since each of A_i and B_j is a disjoint set of sets.

We also have

$$(\cup_{i=1}^n A_i) \setminus \left(\cup_{j=1}^m B_j\right) = (\cup_{i=1}^n A_i) \cap \left(\cup_{j=1}^m B_j\right)^c \tag{2}$$

$$= \cup_{i=1}^n \cap_{j=1}^m A_i \cap B_j^c \tag{3}$$

$$= \cup_{i=1}^n \cap_{j=1}^m A_i \setminus B_j \tag{4}$$

and we know that each $A_i \setminus B_j \in \mathcal{D}$ and we know that $\mathcal{D}$ is closed under finite intersections thus $\cap_{j=1}^m A_i \setminus B_j \in \mathcal{D}$. Furthermore by disjointness of A_i we have that $\cap_{j=1}^m A_i \setminus B_j$ are disjoint and therefore we have shown that $(\cup_{i=1}^n A_i) \setminus \left(\cup_{j=1}^m B_j\right) \in \mathcal{R}$.

To see that $\mathcal{R}$ is the smallest ring containing $\mathcal{D}$ note simply that it is a ring and any ring containing $\mathcal{D}$ must contain all of the finite disjoint unions of elements in $\mathcal{D}$. □

To connect up the concept of rings with that of Boolean algebras we have the following result.

Lemma 3.77. *Let $\mathcal{R}$ be a ring and define $\mathcal{R}^c = \{A^c \mid A \in \mathcal{R}\}$. Then $\mathcal{A} = \mathcal{R} \cup \mathcal{R}^c$ is a Boolean algebra and is the Boolean algebra generated by $\mathcal{R}$. If $\mathcal{R}$ is a σ-ring then $\mathcal{R} \cup \mathcal{R}^c$ is the σ-algebra generated by $\mathcal{R}$.*

Proof. Since Boolean algebras are closed under set complement it suffices to show that $\mathcal{A} = \mathcal{R} \cup \mathcal{R}^c$ is a Boolean algebra (respectively σ-algebra). Closure under set complement is immediate from construction. Closure under set intersection follows from handling the three possible cases

(i) if $A, B \in \mathcal{R}$ then $A \cap B \in \mathcal{R} \subset \mathcal{A}$ since $\mathcal{R}$ is a ring.
(ii) if $A \in \mathcal{R}$ and $B \in \mathcal{R}^c$ then $A \cap B = A \cap (B^c)^c = A \setminus B^c \in \mathcal{R} \subset \mathcal{A}$ since $B^c \in \mathcal{R}$ and $\mathcal{R}$ is a ring.
(iii) if $A, B \in \mathcal{R}^c$ then $A \cap B = (A^c \cup B^c)^c \in \mathcal{R}^c \subset \mathcal{A}$ since $A^c, B^c \in \mathcal{R}$ and $\mathcal{R}$ is a ring.

Closure under finite set union follows as usual from De Morgan's Law.

Now if $\mathcal{R}$ is a σ-ring then

TODO: Finish □

We have the following result for σ-rings that is analagous to Lemma 3.8 proven for σ-algebras.

Lemma 3.78. *Given an arbitrary set function $f : S \to T$ and σ-rings $\mathcal{S}$ and $\mathcal{T}$ on S and T respectively*

(i) *$\mathcal{S}' = f^{-1}\mathcal{T}$ is a σ-ring on S.*
(ii) *$\mathcal{T}' = \left\{A \subset T; f^{-1}(A) \in \mathcal{S}\right\}$ is a σ-ring on T.*

Proof. The proof of Lemma 3.8 shows closure under countable union and intersection. From these two facts, closure under set difference follows by writing $B \setminus A = B \cap A^c$. □

TODO: We have proven abstract Caratheodory construction in the language of Boolean algebras; fill in a gap that shows that a countably additive function on a ring actually defines a premeasure as defined above.

Lemma 3.79. *Let μ be an additive function on a semiring $\mathcal{D}$. Let $\mu(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mu(A_i)$ for any disjoint $A_1, \ldots, A_n \in \mathcal{D}$. Then μ is well defined and finitely additive on the ring $\mathcal{R}$ generated by $\mathcal{D}$. If μ is countably additive on $\mathcal{D}$ then μ is countably additive on $\mathcal{R}$ and extends to a measure on σ-algebra generated by $\mathcal{D}$.*

Proof. □

3.5.2. *Product Measures and Fubini's Theorem.* Prior to showing how to construct product measures, we need a technical lemma.

Lemma 3.80 (Measurability of Sections)**.** *Let $(S, \mathcal{S}, \mu)$ be a measure space with μ a σ-finite measure, let $(T, \mathcal{T})$ be a measurable space and $f : S \times T \to \mathbb{R}_+$ be a positive $\mathcal{S} \otimes \mathcal{T}$-measurable function. Then*

(i) *$f(s,t)$ is an $\mathcal{S}$-measurable function of $s \in S$ for every fixed $t \in T$.*
(ii) *$\int f(s,t)\, d\mu(s)$ is $\mathcal{T}$-measurable for as a function of $t \in T$.*

Proof. To see (i) and (ii), let us first assume that μ is a bounded measure. The proof uses the standard machinery. First assume that $f(s,t) = \mathbf{1}_{B\times C}$ for $B \in \mathcal{S}$ and $C \in \mathcal{T}$. Then note that for fixed $t \in T$, $f(s,t) = \mathbf{1}_B$ if $t \in C$ and $f(s,t) = 0$ otherwise; in both cases we see that f is $\mathcal{S}$-measurable. Also we calculate, $\int \mathbf{1}_{B\times C}(s,t)\, d\mu(s) = \mathbf{1}_C(t) \int \mathbf{1}_B(s)\, d\mu(s) = \mu(B)\mathbf{1}_C(t)$ which clearly $\mathcal{T}$-measurable since $\mu(B) < \infty$.

Observe that the set of sets $B \times C$ is a π-system. Let

$$\mathcal{H} = \{A \in \mathcal{S} \otimes \mathcal{T} \mid \mathbf{1}_A(s,t) \text{ is } \mathcal{S}\text{-measurable for every fixed } t \in T \text{ and } \int \mathbf{1}_A(s,t)\, d\mu(s) \text{ is } \mathcal{T}\text{-measurable } \}$$

and we claim that $\mathcal{H}$ is a λ-system. Clearly $S \times T \in \mathcal{H}$ from what we have already shown. Suppose next that $A \subset B$ are both in $\mathcal{H}$. Note that $\mathbf{1}_{B\setminus A} = \mathbf{1}_B - \mathbf{1}_A$ so each section is a difference of $\mathcal{S}$-measurable functions hence $\mathcal{S}$-measurable. Similarly,

$$\int \mathbf{1}_{B\setminus A}(s,t)\, d\mu(s) = \int \mathbf{1}_B(s,t)\, d\mu(s) - \int \mathbf{1}_A(s,t)\, d\mu(s)$$

is a difference of $\mathcal{T}$-measurable function hence $\mathcal{T}$-measurable.

Lastly, suppose that $A_1 \subset A_2 \subset \cdots \in \mathcal{H}$. Then $\mathbf{1}_{A_i} \uparrow \mathbf{1}_{\bigcup A_i}$ and this statement is true when considering each function as a function on $S \times T$ but also for every section with fixed $t \in T$. Hence every section is a increasing limit of $\mathcal{S}$-measurable functions and therefore $\mathcal{S}$-measurable. Also we can apply Montone Convergence Theorem to see that

$$\int \mathbf{1}_{\bigcup A_i}(s,t)\, d\mu(s) = \lim_{n\to\infty} \int \mathbf{1}_{A_i}(s,t)\, d\mu(s)$$

which shows $\mathcal{T}$-measurability. Now the π-λ Theorem shows that $\mathcal{H} = \mathcal{S} \otimes \mathcal{T}$ and we have the result for all indicators.

Next, linearity of taking sections and integrals shows that all simple functions also satisfy the theorem. Lastly for a general positive $f(s,t)$ we take an increasing sequence of simple functions $f_n \uparrow f$. Again, the limit is taken pointwise so every section of f is the limit of the sections of f_n each of which has been shown $\mathcal{S}$-measurable. As the limit of $\mathcal{S}$-measurable functions, we see that every section f is also $\mathcal{S}$-measurable. Since for a fixed $t \in T$, $f_n(s,t)$ is increasing as a function of s alone we apply the Monotone Convergence Theorem to see that

$$\int f(s,t)\, d\mu(s) = \lim_{n\to\infty} \int f_n(s,t)\, d\mu(s)$$

which shows $\mathcal{T}$-measurability of $\int f(s,t)\, d\mu(s)$ since it is a limit of $\mathcal{T}$-measurable functions.

Now let μ be a σ-finite measure on S. Then there is a disjoint partition $S_1, S_2, \ldots$ of S such that $\mu S_n < \infty$. Thus, $\mu_n(A) = \mu(A \cap S_n)$ defines a bounded measure and we know from Lemma 3.54 that for any measurable g, $\int g\, d\mu_n = \int g \mathbf{1}_{S_n}\, d\mu$. Putting these observations together,

$$\begin{aligned}
\int f(s,t)\, d\mu(s) &= \int f(s,t) \sum_{n=1}^{\infty} \mathbf{1}_{S_n}(s)\, d\mu(s) && \text{since } S_n \text{ is a partition of } S \\
&= \sum_{n=1}^{\infty} \int f(s,t) \mathbf{1}_{S_n}(s)\, d\mu(s) && \text{by Corollary 3.41} \\
&= \sum_{n=1}^{\infty} \int f(s,t)\, d\mu_n(s)
\end{aligned}$$

Since each μ_n is bounded, we have proven that each $\int f(s,t)\, d\mu_n(s)$ is $\mathcal{T}$-measurable hence the same is true for the partial sums by linearity and then the infinite sum by taking a limit. □

TODO: Come up with an example of a non-measurable function for which all sections are measurable.

Theorem 3.81 (Fubini-Tonelli Theorem)**.** *Let $(S, \mathcal{S}, \mu)$ and $(T, \mathcal{T}, \nu)$ be two σ-finite measure spaces. There exists a unique measure $\mu \otimes \nu$ on $(S \times T, \mathcal{S} \otimes \mathcal{T})$ satisfying*

$$(\mu \otimes \nu)(B \times C) = \mu B \cdot \nu C \qquad \text{for all } B \in \mathcal{S},\ C \in \mathcal{T}.$$

In addition if $f : S \times T \to \mathbb{R}_+$ is a positive measurable function then

$$\int f(s,t)\, d(\mu \otimes \nu) = \int \left[\int f(s,t)\, d\nu(t) \right] d\mu(s) = \int \left[\int f(s,t)\, d\mu(s) \right] d\nu(t)$$

This last sequence of equalities also holds if $f : S \times T \to \mathbb{R}$ is measurable and integrable with respect to $\mu \otimes \nu$.

Proof. Note that the class of sets of the form $A \times B$ for $A \in \mathcal{S}$ and $B \in \mathcal{T}$ is clearly a π-system and generates $\mathcal{S} \otimes \mathcal{T}$ by definition of the product σ-algebra. Furthermore by σ-finiteness of both μ and ν we can construct a disjoint partition $S \times T = \cup_i \cup_j S_i \times T_j$ with $\mu(S_i)\nu(T_j) < \infty$. Therefore we can apply Lemma 3.65 to see that the property $(\mu \otimes \nu)(A \times B) = \mu(A)\nu(B)$ uniquely determines $\mu \otimes \nu$.

To show existence of such a measure, define

$$(\mu \otimes \nu)(A) = \int \left[\int \mathbf{1}_A(s,t)\, d\nu(t) \right] d\mu(s)$$

The fact that the iterated integrals are well defined follows from Lemma 3.80. To see that it is a measure, first note that it is simple to see $(\mu \otimes \nu)(\emptyset) = 0$.

To prove countable additivity, suppose we are given disjoint $A_1, A_2, \cdots \in \mathcal{S} \otimes \mathcal{T}$. By disjointness, we know $\mathbf{1}_{\bigcup_{i=1}^{\infty} A_i} = \sum_{i=1}^{\infty} \mathbf{1}_{A_i}$. Now because indicator functions and the inner integrals are positive, we can interchange integrals and sums twice (Corollary 3.41) and get

$$\begin{aligned}
(\mu \otimes \nu)(\bigcup_{i=1}^{\infty} A_i) &= \int \left[\int \mathbf{1}_{\bigcup_{i=1}^{\infty} A_i}(s,t)\, d\nu(t) \right] d\mu(s) \\
&= \int \left[\int \sum_{i=1}^{\infty} \mathbf{1}_{A_i}(s,t)\, d\nu(t) \right] d\mu(s) \\
&= \sum_{i=1}^{\infty} \int \left[\int \mathbf{1}_{A_i}(s,t)\, d\nu(t) \right] d\mu(s)
\end{aligned}$$

It is also clear that for $A = B \times C$ with $B \in \mathcal{S}$ and $C \in \mathcal{T}$,

$$\begin{aligned}
(\mu \otimes \nu)(B \times C) &= \int \left[\int \mathbf{1}_B(s) \mathbf{1}_C(t)\, d\nu(t) \right] d\mu(s) \\
&= \int \mathbf{1}_B(s)\, d\mu(s) \cdot \int \mathbf{1}_C(t)\, d\nu(t) \\
&= \mu B \cdot \nu C
\end{aligned}$$

Therefore we have proven the existence of the product measure.

The argument proving existence of the product measure applies equally well if we reverse the order of μ and ν and shows that

$$(\mu \otimes \nu)(B \times C) = \int \left[\int \mathbf{1}_{B \times C}(s,t)\, d\nu(t) \right] d\mu(s) = \int \left[\int \mathbf{1}_{B \times C}(s,t)\, d\mu(s) \right] d\nu(t)$$

which proves that the integrals are equal for indicator functions of sets of the form $B \times C$ and therefore for all indicator functions by the montone class argument we used at the beginning of the proof. At this point, the standard machinery can be deployed. Linearity of integrals easily shows that the equality extends to simple functions. Lastly suppose we have a positive measurable function $f(s,t) : S \times T \to \overline{R}_+$ with a sequence of positive simple functions $f_n(s,t) \uparrow f(s,t)$. By the Monotone Convergence Theorem and monotonicity of integral we know that

$$\begin{aligned}
0 \leq \int f_n(s,t)\, d\mu(s) &\uparrow \int f(s,t)\, d\mu(s) \\
0 \leq \int f_n(s,t)\, d\nu(t) &\uparrow \int f(s,t)\, d\nu(t)
\end{aligned}$$

and therefore we have

$$\begin{aligned}
\int f(s,t)\,d(\mu\otimes\nu) &= \lim_{n\to\infty}\int f_n(s,t)\,d(\mu\otimes\nu) && \text{by definition of integral of } f\\
&= \lim_{n\to\infty}\int\left[\int f_n(s,t)\,d\mu(s)\right]d\nu(t) && \text{by Tonelli for simple functions}\\
&= \int\left[\int f(s,t)\,d\mu(s)\right]d\nu(t) && \text{by Monotone Convergence on } \int f_n\,d\mu(s)
\end{aligned}$$

It is worth pointing out explicitly that even if $f(s,t)$ is never equal to infinity, the integrals may be equal to infinity on all of S or T and it is critical that we have phrased the theory of integration for positive functions in terms of functions with values in $\overline{\mathbb{R}}_+$.

TODO: Clean up the following argument; it has all right details but is more than a bit ragged. Particularly annoying is that this is the first time we've talked about defining integrals for signed functions that take infinite values on a set of measure zero.

Now assume that f is integrable with respect to $\mu\otimes\nu$: $\int |f(s,t)|\,d(\mu\otimes\nu) < \infty$. We write $f = f_+ - f_-$ and note that $\int f_\pm(s,t)\,d(\mu\otimes\nu) < \infty$ and use Tonelli's Theorem just proven to see that

$$\int f_\pm(s,t)\,d(\mu\otimes\nu) = \int\left[\int f_\pm(s,t)\,d\nu(t)\right]d\mu(s) = \int\left[\int f_\pm(s,t)\,d\mu(s)\right]d\nu(t) < \infty$$

The finiteness of the iterated integrals implies that the integrands are almost surely finite and therefore we see that each section $\int f_\pm\,d\mu(s)$ and $\int f_\pm\,d\nu(t)$ is almost surely finite. The trick is that being almost surely finite isn't good enough when trying to calculate the iterated integrals of f and we might run into the awkward situation in which there is a $t\in T$ such that *both* $\int f_+\,d\mu(s)$ and $\int f_+\,d\mu(s)$ are infinite. However define $N_S = \{s\in S \mid \int |f|\,d\nu(t) = \infty\}$ and $N_T = \{t\in T \mid \int |f|\,d\mu(s) = \infty\}$. We have noted that N_S is a μ-null set and that N_T is a ν-null set hence $N_S\times N_T$ is a $(\mu\otimes\nu)$-null set. We modify f so that it is zero on $N_S\times N_T$ by defining $\tilde{f}(s,t) = (1-\mathbf{1}_{N_S\times N_T})f(s,t)$. Note the following

$$\begin{aligned}
\int \tilde{f}d(\mu\otimes\nu) &= \int f d(\mu\otimes\nu)\\
\int \tilde{f}d\mu(s) &= \begin{cases}\int f d\mu(s) & \text{if } t\notin N_T\\ 0 & \text{if } t\in N_T\end{cases}\\
\int \tilde{f}d\nu(t) &= \begin{cases}\int f d\nu(t) & \text{if } s\notin N_S\\ 0 & \text{if } s\in N_S\end{cases}
\end{aligned}$$

Now we can write $\tilde{f} = \tilde{f}_+ - \tilde{f}_-$ and apply Tonelli's Theorem to see

$$\begin{aligned}
\int \tilde{f} d(\mu \otimes \nu) &= \int \tilde{f}_+ d(\mu \otimes \nu) - \int \tilde{f}_- d(\mu \otimes \nu) \\
&= \int \left[\int \tilde{f}_+ d\mu(s) \right] d\nu(t) - \int \left[\int \tilde{f}_- d\mu(s) \right] d\nu(t) \\
&= \int \left[\int \tilde{f}_+ d\mu(s) - \int \tilde{f}_- d\mu(s) \right] d\nu(t) \\
&= \int \left[\int \tilde{f} d\mu(s) \right] d\nu(t)
\end{aligned}$$

But we know $\int \left[\int \tilde{f} d\mu(s) \right] d\nu(t) = \int \left[\int f d\mu(s) \right] d\nu(t)$ so we get the result for f as well. □

TODO: Royden has some exercises that demonstrate how each of these hypotheses is necessary (e.g. Counterexample to Fubini for non-integrable f). Incorporate them.

Example 3.82. Define the measure space $(\mathbb{N}, 2^{\mathbb{N}}, \mu)$ where $\mu(A) = \text{card}(A)$. μ is called the *counting measure*. Consider the function

$$f(s,t) = \begin{cases} 2 - 2^{-s+1} & \text{if } s = t \\ -2 + 2^{-s+1} & \text{if } s = t+1 \\ 0 & \text{otherwise} \end{cases}$$

on $(\mathbb{N} \times \mathbb{N}, 2^{\mathbb{N} \times \mathbb{N}}, \mu \otimes \mu)$. Since $\mu \otimes \mu$ is the counting measure on $\mathbb{N} \times \mathbb{N}$ it is easy to see that

$$\int |f(s,t)| \, d(\mu \otimes \mu) = \sum_{s=1}^{\infty} \sum_{t=1}^{\infty} |f(s,t)| = \infty$$

so f is not integrable. However in this case both of the iterated integrals are defined. For fixed t,

$$\int f(s,t) \, d\mu(s) = \sum_{s=1}^{\infty} f(s,t) = 2^{-t} - 2^{-t+1} = -2^{-t}$$

hence

$$\int \left[\int f(s,t) \, d\mu(s) \right] d\mu(t) = \sum_{t=1}^{\infty} -2^{-t} = -1$$

For fixed s,

$$\int f(s,t) \, d\mu(s) = \sum_{t=1}^{\infty} f(s,t) = \begin{cases} 1 & \text{if } s = 1 \\ 0 & \text{otherwise} \end{cases}$$

and therefore

$$\int \left[\int f(s,t) \, d\mu(t) \right] d\mu(s) = 1$$

This example shows that the positivity of f is a necessary condition in Tonelli's Theorem and that the assumption of integrability is necessary in Fubini's Theorem.

TODO Outer measures, Caratheodory construction, Lesbegue Measure (existence and uniqueness), Product Measures and Fubini's Theorem, Radon-Nikodym Theorem and Fundamental Theorem of Calculus, Differential Change of Variables for Lebesgue Measure on $\mathbb{R}^n$ (useful for calculations involving probability densities).

Lemma 3.83 (Translation Invariance of Lebesgue Measure)**.** *Suppose μ is a measure on $\mathbb{R}^n$ which is translation invariant and for which $\mu([0,1]^n) = 1$, then $\mu = \lambda^n$.*

Proof. Suppose we are given a translation invariant measure μ such that $\mu([0,1]^n) = 1$. By writing boxes as a union of cubes and using finite and countable additivity together with translation invariance it is easy to see that for any box $\mathcal{I}_1 \times \cdots \times \mathcal{I}_n$ where each $\mathcal{I}_k$ has rational endpoints that we have

$$\begin{aligned}\mu(\mathcal{I}_1 \times \cdots \times \mathcal{I}_n) &= |\mathcal{I}_1| \cdots |\mathcal{I}_n| \\ &= \lambda^n(\mathcal{I}_1 \times \cdots \times \mathcal{I}_n)\end{aligned}$$

Now fix $\mathcal{I}_2, \ldots, \mathcal{I}_n$ and consider $\nu(A) = \frac{1}{|\mathcal{I}_2| \cdots |\mathcal{I}_n|} \mu(A \times \mathcal{I}_2 \times \cdots \times \mathcal{I}_n)$ as a function of $A \in \mathcal{B}(\mathbb{R})$. It is easy to see that this is a Borel measure and we have already seen that $\nu(\mathcal{I}) = |\mathcal{I}|$ for all rational intervals (hence all intervals by countable additivity). Therefore $\nu = \lambda$ is Lebesgue measure on $\mathcal{B}(\mathbb{R})$ and we have for every $B_1 \in \mathcal{B}(\mathbb{R})$,

$$\mu(B_1 \times \mathcal{I}_2 \times \cdots \times \mathcal{I}_n) = \lambda^n(B_1 \times \mathcal{I}_2 \times \cdots \times \mathcal{I}_n)$$

Now iterate the argument $2, \cdots, n$ fixing all but the i^{th} argument to extend to all cylinder sets $B_1 \times \cdots \times B_n$ and we apply the uniqueness of product measures.

Now it remains to show that λ^d is indeed translation invariant. TODO □

Corollary 3.84. *Lebesgue measure λ^n on $\mathbb{R}^n$ is invariant under orthogonal transformations.*

Proof. Suppose we are given an orthogonal transformation P. We claim that the measure $\lambda_P^n(A) = \lambda^a(PA)$ is translation invariant. To see this, assume we are given $h \in \mathbb{R}^n$ and note that

$$\begin{aligned}\lambda_P^n(A+h) &= \lambda^n(PA + Ph) && \text{linearity of } P \\ &= \lambda^n(PA) && \text{translation invariance of } \lambda^n \\ &= \lambda_P^n(A) && \text{definition of } \lambda_P^n\end{aligned}$$

Therefore we know that $\lambda_P^n = c\lambda^n$ for some constant $c > 0$. Take the unit ball $B^n \subset \mathbb{R}^n$ and notice that $PB^n = B^n$ to see that in fact $c = 1$. □

Corollary 3.85. *[Linear Change of Variables]For an arbitrary linear transformation $T : \mathbb{R}^n \to \mathbb{R}^n$, $\lambda^n(TA) = |\det T| \, \lambda^n(A)$ for all measurable A.*

Proof. Note that by the Singular Value Decompostion, we can write $T = UDV$ with U, V orthogonal. By the rotation invariance of λ^n, we are reduced to the case of a diagonal matrix. In that case, the result is easy. TODO write down the easy stuff too! □

3.6. **Radon-Nikodym Theorem and Differentiation.** We have seen the construction of measures by integration of a density. A productive line of inquiry is to ask if one can characterize measures that arise through this construction and those that cannot arise through this construction. As it turns out an precise answer may be given for σ-finite measures; this is the content of the Radon-Nikodym Theorem.

If one restricts attention to $\mathbb{R}$ and considers the Fundamental Theorem of Calculus for Riemann integrals

$$\frac{d}{dx}\int_0^x f(y)\,dy = f(x)$$

one can surmise that there is a connection between the considerations of the Radon-Nikodym Theorem and the theory of differentiation of integrals. This is indeed the case and we will prove the extension of the Fundamental Theorem of Calculus to Lebesgue integrals using the Radon-Nikodym Theorem. Note that it is probably more traditional to explore the theory of differention of functions of a real variable without using the more abstract Radon-Nikodym Theorem but if one intends to cover both one can save some time by proceeding in the way we have chosen (stolen unabashedly from Kallenberg).

The first step is to develop a couple of tools that may be used to compare two measures. The trick is that if one takes the difference of two measure, one does not get a measure. However there is a clever observation that helps to repair the defect.

Definition 3.86. A *bounded signed measure* on a measurable space $(\Omega, \mathcal{A})$ is a bounded function $\nu : \mathcal{A} \to \mathbb{R}$ such that for every disjoint $A_1, A_2, \cdots \in \mathcal{A}$ such that $\sum_{n=1}^{\infty} |\nu(A_n)| < \infty$, we have $\nu(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \nu(A_n)$

Definition 3.87. Two measures μ and ν on a measurable space $(\Omega, \mathcal{A})$ are said to be *mutually singular* if there exists $A \in \mathcal{A}$ such that $\mu A = 0$ and $\nu A^c = 0$. We often write $\mu \perp \nu$.

Example 3.88. Lebesgue measure and any Dirac measure on $\mathbb{R}$ are mutually singular.

Example 3.89. Let f, g be positive measurable functions on $\mathbb{R}$ such that $\int f \wedge g\, d\lambda = 0$. Then $f \cdot \lambda$ and $g \cdot \lambda$ are mutually singular.

Definition 3.90. Given two measures μ and ν on a measurable space $(\Omega, \mathcal{A})$ we say that ν is *absolutely continuous* with respect to μ if for every $A \in \mathcal{A}$ such that $\mu A = 0$ we also have $\nu A = 0$. We often write $\nu \ll \mu$.

Example 3.91. Let f be a positive measurable function on the measure space $(\Omega, \mathcal{A}, \mu)$, then $f \cdot \mu$ is absolutely continuous with respect to μ. We shall soon see that this is the only way to construct absolutely continuous measures.

Theorem 3.92. *[Hahn Decomposition]Given a bounded signed measure ν on a measurable space $(\Omega, \mathcal{A})$ there are unique bounded mutually singular positive measures ν_+ and ν_- such that $\nu = \nu_+ - \nu_-$.*

Proof. Let $c = \sup_{A \in \mathcal{A}} \nu(A)$. The first claim is that there is a $A_+ \in \mathcal{A}$ such that $\nu A_+ = c$. To see this, first we note the following crude bound. Suppose we are given $A, A' \in \mathcal{A}$ such that $\nu A \geq c - \epsilon$ and $\nu A' \geq c - \epsilon'$. Then

$$\begin{aligned}
\nu(A \cup A') &= \nu A + \nu A' - \nu A \cap A' \\
&\geq \nu A + \nu A' - c && \text{by bound on } \nu \\
&\geq c - \epsilon - \epsilon' && \text{by bounds on } A, A'
\end{aligned}$$

Now approximate the supremum by taking $A_1, A_2, \cdots \in \mathcal{A}$ such that $\nu A_n \geq c - 2^{-n}$ and apply the bound above with countably additivity to see

$$\nu \bigcup_{i=n+1}^{\infty} A_i \geq c - \sum_{i=n+1}^{\infty} 2^{-i} = c - 2^{-n}$$

There is something a bit confusing about this bound; namely as n is increasing the sets are getting smaller but the bound on the measure is increasing. TODO: It is probably worth sorting out exactly what this is telling us (I think it is just that all of the tails are equal up to null sets and of measure c). Let $A_+ = \bigcap_{n=1}^{\infty} \bigcup_{i=n+1}^{\infty} A_i$ and note by countable additivity and the boundedness of ν (see proof of Lemma 3.27) we have

$$\nu A_+ = \lim_{n\to\infty} \nu \bigcup_{i=n+1}^{\infty} A_i \geq c$$

By the definition of c we see that $\nu A_+ = c$. Now define $A_- = A_+^c$ and define the restrictions

$$\nu_+ B = \nu(A_+ \cap B)$$
$$\nu_- B = \nu(A_- \cap B)$$

TODO: prove decomposition property and uniqueness. □

Theorem 3.93 (Radon-Nikodym Theorem)**.** *Let μ, ν be σ-finite measures on the measurable space $(\Omega, \mathcal{A})$. There exist unique measures $\nu_a \ll \mu$ and $\nu_s \perp \mu$ such that $\nu = \nu_a + \nu_s$. Furthermore, there is a unique positive measurable $f : \Omega \to \mathbb{R}$ such that $\nu_a = f \cdot \mu$.*

Proof. TODO □

In addition to the product measure construction we have just seen there is another important construction for $\mathbb{R}$.

Definition 3.94. A measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is called *locally finite* if $\mu(I) < \infty$ for every finite interval $I \subset \mathbb{R}$.

Lemma 3.95 (Lebesgue-Stieltjes Measure)**.** *There is a 1-1 correspondence between locally finite measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and nondecreasing right continuous functions $F : \mathbb{R} \to \mathbb{R}$ such that $F(0) = 0$ given by*

$$\mu((a, b]) = F(b) - F(a)$$

Proof. Suppose we are given a locally finite measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Define

$$F(x) = \begin{cases} \mu(0, x] & \text{if } x > 0 \\ -\mu(x, 0] & \text{if } x < 0 \\ 0 & \text{if } x = 0 \end{cases}$$

Local finiteness of μ implies that F is well defined. Monotonicity of μ implies that F is nondecreasing. Continuity of measure implies that F is right continuous. Clearly,

$$\mu(a, b] = F(b) - F(a)$$

and furthermore F is the unique function that satisfies this property.

On the other hand, given an F that is nondecreasing, right continuous and satisfies $F(0) = 0$ we define a generalized inverse by

$$G(y) = \inf\{x \in \mathbb{R} \mid F(x) \geq y\}$$

Note that if $y < w$ then $\{x \in \mathbb{R} \mid F(x) \geq w\} \subset \{x \in \mathbb{R} \mid F(x) \geq y\}$ which shows that G is a nondecreasing function. The fact that G is nondecreasing implies that $G^{-1}(-\infty, y] = (-\infty, x]$ for some $x \in \mathbb{R}$ and therefore G is a measurable function. Furthermore,

$$G(F(x)) = \inf\{s \in \mathbb{R} \mid F(s) \geq F(x)\} \leq x$$

and on the other hand since

$$G(y) = \inf\{x \in \mathbb{R} \mid F(x) \geq y\}$$

we can find a sequence $x_n \downarrow G(y)$ such that $F(x_n) \geq y$ and therefore by right continuity of F we now that $F(G(y)) = \lim_{n\to\infty} F(x_n) \geq y$.

Together these two facts show that $G(y) \leq c$ if and only if $y \leq F(c)$. In one direction suppose $y \leq F(c)$, then applying G to both sides and using the nondecreasing nature of G, we get $G(y) \leq G(F(c)) \leq c$. In the other direction, we assume $G(y) \leq c$ and apply F to both sides and to see

$$F(c) \geq F(G(y)) \geq y$$

In a similar way, we see that $c < G(y)$ if and only if $F(c) < y$.

Now we can finish the proof by defining $\mu = (\lambda \circ G^{-1})$ where λ is Lebesgue measure on $\mathbb{R}$. We observe that this is an inverse to the construction of F given above.

$$\begin{aligned}\mu(a, b] &= \lambda\left(\{y \in \mathbb{R} \mid a < G(y) \leq b\}\right) \\ &= \lambda(F(a), F(b)] = F(b) - F(a)\end{aligned}$$

Uniqueness of measure μ with this property follow by Lemma 3.65. □

Note the choice of the normalizing condition $F(0) = 0$ is somewhat arbitrary albeit a natural choice when considering arbitrary locally finite measures on $\mathbb{R}$. We will see later that for finite measures, and probability measures in particular, it is more useful to pick a different normalization $\lim_{x\to-\infty} F(x) = 0$.

By the description of all measures on $\mathbb{R}$ as Lebesgue-Stieltjes measures, we have set the stage for the translation of results about measures into results about nondecreasing, right continuous functions. In particular, if we apply the Radon-Nikodym Theorem to we see that any such F may be written as $F = F_a + F_s$ which represent the absolutely continuous and singular parts of the decomposition respectively. If one unwinds the defining property of F_a from the Lebesgue-Stieltjes integral, one sees that in the absolutely continuous case, $F_a(x) = \int_0^x f \, d\lambda$ for an appropriate density f.

Theorem 3.96 (Fundamental Theorem Of Calculus)**.** *Let any nondecreasing, right continuous function $F(x) = \int_0^x f \, d\lambda + F_s(x)$ is differentiable a.e. with derivative $F' = f$.*

Proof. TODO □

Corollary 3.97 (Integration By Parts)**.** *Suppose f and g are absolutely continuous functions. Then*

$$\int_a^b f'gd\lambda = f(b)g(b) - f(a)g(a) - \int_a^b fg'd\lambda$$

Lemma 3.98. *Let $\mathcal{I}$ be an arbitrary collection of open intervals of $\mathbb{R}$. Let $G = \bigcup_{I\in\mathcal{I}} I$ and suppose that $\lambda G < \infty$. Then there exists disjoint $I_1, \ldots, I_n$ such that $\sum_{i=1}^n |I_i| \geq \frac{\lambda G}{4}$.*

Proof. TODO □

Lemma 3.99. *Let μ be a locally finite measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and let $F(x) = \mu(0, x]$. Let $A \in \mathcal{B}$ be a set with $\mu A = 0$, then $F' = 0$ almost everywhere λ on A.*

Proof. The intuition behind the proof is that the derivative $F'(x)$ represents the ratio of μ-measure and λ-measure for arbitrarily small intervals around $x \in \mathbb{R}$. For $x \in A$, we expect the μ-measure and therefore the derivative to be 0. Since A may not contain any honest intervals, there is some finesse required to make the intuition rigorous.

First pick $\delta > 0$ and and open set $G_\delta \supset A$ such that $\mu G_\delta < \delta$.

TODO: Prove that such G_δ exists; this is a fact for arbitrary Borel σ-algebras.

For each $\epsilon > 0$, let

$$\begin{aligned} A_\epsilon &= \{x \in A \mid \sup_{h\to 0} \frac{F(x+h) - F(x-h)}{h} > \epsilon\} \\ &= \{x \in A \mid \sup_{h\to 0} \frac{\mu(x-h, x+h]}{2\epsilon} > |(x-h, x+h]|\} \end{aligned}$$

TODO: Prove that A_ϵ is measurable.

For any $x \in A_\epsilon$ we can pick $h > 0$ small enough so that $I_x = (x-h, x+h] \subset G_\delta$. Note that $A_\epsilon \subset \bigcup_{x\in A_\epsilon} I_x$. By the previous Lemma 3.98 we pick a finite disjoint set $I_1, \ldots, I_n$ and note that

$$\begin{aligned} \lambda A_\epsilon &\leq \lambda \bigcup_{x\in A_\epsilon} I_x \\ &\leq 4\sum_{k=1}^n |I_k| \\ &\leq 4\sum_{k=1}^n \frac{\mu I_k}{2\epsilon} \\ &\leq \frac{\delta}{2\epsilon} \end{aligned}$$

Now $\delta > 0$ was arbitrary so we see that $\lambda A_\epsilon = 0$. Since $\epsilon > 0$ was arbitrary so we see that $F'(x) = 0$ almost everywhere on A since the set of points where $F' \neq 0$ is a countable union of A_ϵ (e.g. take $\bigcup_n A_{\frac{1}{n}}$). □

3.7. **Approximation By Smooth Functions.** In this section we discuss a technique for approximating arbitrary measurable and integrable functions by smooth functions.

To start, we establish the existence of an infinitely differentiable function which is supported on the interval $[-1, 1]$.

Lemma 3.100. *The function*

$$f(x) = \begin{cases} e^{\frac{-1}{1-x^2}} & |x| < 1 \\ 0 & |x| \geq 1 \end{cases}$$

is compactly supported on $[-1,1]$ *and has continuous derivatives of all orders.*

Proof. It is clear from the definition that $f(x)$ is compactly supported on $[-1,1]$. To see that it has continuous derivatives of all orders we use an induction to prove that for every $n \geq 0$, there exists a polynomial $P_n(x)$ and a nonnegative integer N_n such that

$$f^{(n)}(x) = \frac{P_n(x)}{(1-x^2)^{N_n}} e^{\frac{-1}{1-x^2}}$$

Clearly this is true for $n = 0$. Supposing that it is true for $n > 0$, we calculate using the induction hypothesis, the product rule and chain rule

$$\begin{aligned} f^{(n+1)}(x) &= \frac{d}{dx} \frac{P_n(x)}{(1-x^2)^{N_n}} e^{\frac{-1}{1-x^2}} \\ &= \frac{(1-x^2)^{N_n} P_n'(x) - P_n(x) N_n (1-x^2)^{N_n - 1}}{(1-x^2)^{2N_n}} e^{\frac{-1}{1-x^2}} + \frac{P_n(x)}{(1-x^2)^{N_n}} \frac{-1}{1-x^2} \frac{-2x}{(1-x^2)^2} e^{\frac{-1}{1-x^2}} \end{aligned}$$

which shows the result after creating a common denominator.

It is clear that the derivatives are continuous away from $-1, 1$ so it remains to show $\lim_{x \to -1^+} f^{(n)}(x) = 0$ and $\lim_{x \to 1^-} f^{(n)}(x) = 0$.

Take the former limit. We write $f^{(n)}(x) = \frac{P_n(x)}{(1-x)^{N_n}(1+x)^{N_n}} e^{\frac{-1}{1-x^2}}$ and note that

TODO: Show $\lim_{x \to -1} \frac{1}{(1+x)^M} e^{\frac{-1}{1-x^2}} = 0$ for all $M \geq 0$. □

TODO: What is $\int f(x)$?

Lemma 3.101. *Let* $\rho(x)$ *be a positive function in* $C_c^\infty(\mathbb{R})$ *such that* $\rho(x)$ *is supported on* $[-1,1]$ *and* $\int_{-\infty}^{\infty} \rho(x)\,dx = \int_{-1}^{1} \rho(x)\,dx = 1$. *Let* $f : \mathbb{R} \to \mathbb{R}$ *be a continuous function. Define*

$$f_n(x) = n \int_{-n}^{n} \rho(n(x-y)) f(y) dy$$

Then $f_n \in C_c^\infty(\mathbb{R})$, $f_n^{(m)}(x) = n \int_{-n}^{n} \rho^{(m)}(n(x-y)) f(y) dy$ *and* f_n *converges to* f *uniformly on compact sets. Furthermore, if* f *is bounded then* $\|f_n\|_\infty \leq \|f\|_\infty$.

Proof. First note that because $\rho(x)$ and all of its derivatives are compactly supported, they are also bounded. In particular, there is an $M > 0$ such that $|\rho'(x)| \leq M$. To clean up the notation a little bit, define $\rho_n(y) = n\rho(ny)$ so we have

$$f_n(x) = \int_{-n}^{n} \rho_n(x-y) f(y) dy$$

Since the support of $\rho_n(x)$ is contained in $[-\frac{1}{n}, \frac{1}{n}]$, if we fix $x \in \mathbb{R}$ and view $\rho_n(x-y)$ as a function of y, its support is contained in $[x - \frac{1}{n}, x + \frac{1}{n}]$. Thus the support of $f_n(x)$ is contained in $[-n - \frac{1}{n}, n + \frac{1}{n}]$.

To examine the derivative of $f_n(x)$, pick $h > 0$ and consider the difference quotient

$$\frac{f_n(x+h) - f_n(x)}{h} = \frac{1}{h}\int_{-n}^{n} (\rho_n(x+h-y) - \rho_n(x-y))f(y)dy$$

Taylor's Theorem tells us that $\frac{1}{h}(\rho_n(x+h-y) - \rho_n(x-y)) = \rho_n'(c)$ for some $c \in [x+h-y, x-y]$. Therefore, $\left|\frac{1}{h}(\rho_n(x+h-y) - \rho_n(x-y))f(y)\right| \leq M\,|f(y)|$ and by integrability of $f(y)$ on the interval $[-n,n]$ (i.e. the integrability of $f(y)\cdot \mathbf{1}_{[-n,n]}(y)$ which follows rom the boundedness of $f(y)$ on the compact set $[-n,n]$) we may use Dominated Convergence to conclude that

$$\begin{aligned}
f_n'(x) &= \lim_{h\to 0} \frac{f_n(x+h) - f_n(x)}{h} \\
&= \lim_{h\to 0} \frac{1}{h}\int_{-n}^{n} (\rho_n(x+h-y) - \rho_n(x-y))f(y)dy \\
&= \int_{-n}^{n} \lim_{h\to 0} \frac{1}{h}(\rho_n(x+h-y) - \rho_n(x-y))f(y)dy \\
&= \int_{-n}^{n} \rho_n'(x-y)f(y)dy
\end{aligned}$$

Continuity of $f_n'(x)$ follows from the continuity of $f(y)$ and $\rho_n'(x-y)$ and Dominated Convergence as above. A simple induction extends the result to derivatives of arbitrary order.

Next we show the convergence. Pick a compact set $K \subset \mathbb{R}$ and $\epsilon > 0$, Since f is uniformly continuous on K, there is a $\delta > 0$ such that for any $x \in K$ we have $|x-y| \leq \delta$ implies $|f(x) - f(y)| \leq \epsilon$. Pick $N_1 > 0$ such that $\frac{1}{n} < \delta$ for all $n \geq N_1$. The hypothesis $\int_{-\infty}^{\infty} \rho(y)\,dy = \int_{-1}^{1} \rho(y)\,dy = 1$ and simple change of variables shows $\int_{-\infty}^{\infty} \rho_n(x-y)\,dy = \int_{x-\frac{1}{n}}^{x+\frac{1}{n}} \rho_n(x-y)\,dy = 1$ for all $x \in \mathbb{R}$ and $n > 0$. Pick $N_2 > 0$ so that for all $n > N_2$, we have $K \subset [-n+\frac{1}{n}, n-\frac{1}{n}]$. Therefore we can write $f(x) = \int_{-n}^{n} \rho_n(x-y)f(x)\,dy = 1$ for any $x \in K$ and $n > N_2$. We have for any $n \geq \max(N_1, N_2)$

$$\begin{aligned}
|f_n(x) - f(x)| &= \left|\int_{-n}^{n} (\rho_n(x-y)f(y) - \rho_n(x-y)f(x))\,dy\right| \\
&= \left|\int_{x-\frac{1}{n}}^{x+\frac{1}{n}} (\rho_n(x-y)f(y) - \rho_n(x-y)f(x))\,dy\right| && \text{since } n > N_2 \\
&\leq \int_{x-\frac{1}{n}}^{x+\frac{1}{n}} \rho_n(x-y)\,|f(y) - f(x)|\,dy \\
&\leq \epsilon \int_{x-\frac{1}{n}}^{x+\frac{1}{n}} \rho_n(x-y)\,dy && \text{since } \frac{1}{n} < \delta \\
&\leq \epsilon && \text{since } \rho_n \text{ is positive and } \int_{-\infty}^{\infty} \rho_n(x)\,dx = 1
\end{aligned}$$

The last thing to prove is the norm inequality in case f is bounded.

$$\begin{aligned} |f_n(x)| &\leq n \int_{-n}^{n} \rho(n(x-y))\,|f(y)|\,dy \qquad \text{because } \rho \text{ is positive} \\ &\leq n\|f\|_\infty \int_{-\infty}^{\infty} \rho(n(x-y))dy = \|f\|_\infty \end{aligned}$$

□

Approximation by convolution with a compactly supported bump function is usually sufficient for our purposes, however it is also useful to replace the bump function with Gaussians.

We will need the following fact that is a standard exercise from multivariate calculus

Lemma 3.102. $\int_{-\infty}^{\infty} e^{-x^2/2}\,dx = \sqrt{2\pi}$.

Proof. By Tonelli's Theorem,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\,dxdy = \int_{-\infty}^{\infty} e^{-x^2/2}\,dx \int_{-\infty}^{\infty} e^{-y^2/2}\,dy = \left(int_{-\infty}^{\infty} e^{-x^2/2}\,dx\right)^2$$

However, if we switch to polar coordinates and Tonelli's Theorem,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\,dxdy = \int_0^{2\pi}\int_0^{\infty} e^{-r^2/2} r\,drd\theta = \int_0^{2\pi} d\theta = 2\pi$$

and we are done. □

Now we can see that we may uniformly approximate compactly supported continuous functions by convolution with Gaussians.

Lemma 3.103. *Define* $\rho(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ *and let* $\rho_n(x) = n\rho(nx)$. *Let* $f \in C_c(\mathbb{R})$ *then define* $f_n(x) = (f * \rho_n)(x)$. *Then* $f_n(x) \in C_c^\infty(\mathbb{R})$ *and* f_n *converges to* f *uniformly.*

Proof. The proof is rather similar to that in the preceeding Lemma 3.101. By simple change of variables and Lemma 3.102 we see that $\int_{-\infty}^{\infty} \rho_n(y)\,dy = \int_{-\infty}^{\infty} \rho_n(x-y)\,dy = 1$ and therefore we have the trivial identity $f(x) = \int_{-\infty}^{\infty} f(x)\rho_n(x-y)\,dy$. Because f has compact support, we know that f is uniformly continuous, so given $\epsilon > 0$ we can find $\delta > 0$ such that $|x-y| < \delta$ implies $|f(x) - f(y)| < \epsilon$. Similarly, by compact support f is bounded and we may assume $f(x) < M$ for some $M > 0$. Assume we are given $\epsilon > 0$ then take $\delta > 0$ as above and for any $n > 0$ we have Therefore

$$\begin{aligned} |f * \rho_n(x) - f(x)| &= \left| \int_{-\infty}^{\infty} \rho_n(x-y)(f(y) - f(x))\,dy \right| \\ &\leq \int_{|x-y|<\delta} \rho_n(x-y)\,|f(y) - f(x)|\,dy + \int_{|x-y|\geq\delta} \rho_n(x-y)\,|f(y) - f(x)|\,dy \\ &\leq \epsilon + 2M \int_{|x-y|\geq\delta} \rho_n(x-y)\,dy \end{aligned}$$

Now we consider the last term and change integration variables

$$\begin{aligned}\int_{|x-y|\geq\delta}\rho_n(x-y)\,dy &= \int_{|y|\geq\delta}\rho_n(y)\,dy\\ &= \frac{1}{\sqrt{2\pi}}\int_{|y|\geq n\delta}e^{-y^2/2}\,dy\\ &\leq \frac{2}{\sqrt{2\pi}}\int_{n\delta}^{\infty}\frac{y}{n\delta}e^{-y^2/2}\,dy\\ &= \frac{2}{n\delta\sqrt{2\pi}}e^{-n^2\delta^2/2}\end{aligned}$$

One point here is the elementary fact that $\lim_{n\to\infty}\frac{2}{n\delta\sqrt{2\pi}}e^{-n^2\delta^2/2}=0$ but the second point is that this limit does not depend on x. Thus we may pick $N>0$ independent of x, such that $\int_{|x-y|\geq\delta}\rho_n(x-y)\,dy<\frac{\epsilon}{2M}$ for $n>N$ and therefore

$$|f*\rho_n(x)-f(x)|<2\epsilon$$

which proves the uniform convergence of $f*\rho_n$. □

3.8. **Daniell-Stone Integrals.** We record some required facts about σ-rings that are completely analogous to corresponding facts about σ-algebras.

Lemma 3.104. *Let X be a topological space and let $\mathcal{B}(X)$ be the Borel σ-algebra on X. If A is a Borel set then $\{B\in\mathcal{B}(X)\mid B\subset A\}$ is a σ-ring of sets in X.*

Proof. Clearly it contains the empty set and is closed under countable union. To see that it is closed under set difference simply note $B\setminus C=B\cap C^c\subset B\subset A$ and is clearly a Borel set of X. □

Note that in fact the set of sets in the previous Lemma is the Borel σ-algebra of A with the induced topology.

Lemma 3.105. *The σ-ring of Borel sets of $\mathbb{R}$ that do not contain 0 is generated by intervals $(-\infty,-c)$ and (c,∞) with $c>0$.*

Proof. As noted above the σ-ring in the statement of the Lemma is the σ-algebra of $\mathbb{R}\setminus\{0\}$ in the induced topology. We know that open sets of $\mathbb{R}$ are precisely countable disjoint unions of open intervals (Lemma 2.16). For any open interval (a,b) we either have $(a,b)\subset\mathbb{R}\setminus\{0\}$ or $a<0<b$ hence $(a,b)\cap\mathbb{R}\setminus\{0\}=(a,0)\cup(0,b)$. We conclude that the open sets of $\mathbb{R}\setminus\{0\}$ are countable disjoint unions of open intervals none of which contains 0. Now one can adapt the proof of Lemma 3.6 to get the result. □

TODO: Introduce notation for the σ-ring generated by a set of sets.

Lemma 3.106. *Let $f:S\to T$ be a set mapping and let $\mathcal{C}\subset 2^T$, then the σ-ring generated by $f^{-1}(\mathcal{C})$ is the same as the pullback of the σ-ring generated by $\mathcal{C}$.*

Proof. It is clear that the σ-ring generated by $f^{-1}(\mathcal{C})$ is contained in the pullback of the σ-ring generated by $\mathcal{C}$. To see the reverse conclusion, pushforward the σ-ring generated by $f^{-1}(\mathcal{C})$; this is equal to $\{A\subset T\mid f^{-1}(A)$ is in the σ-ring generated by $f^{-1}(\mathcal{C})\}$ and is itself a σ-ring (Lemma 3.78). It clearly contains $\mathcal{C}$ and therefore the σ-ring generated by $\mathcal{C}$ as well. Therefore the pullback of the σ-ring generated by $\mathcal{C}$ is contained in σ-generated by $f^{-1}(\mathcal{C})$ and we are done. □

It turns out that having a countably additive set function on a σ-ring is almost the same thing as having a measure on the generated σ-algebra. This fact is made precise by the following result.

Lemma 3.107. *Let $\mathcal{R}$ be a σ-ring on a set S and let $\mu : \mathcal{R} \to \overline{\mathbb{R}}_+$ be a function that is countably additive on disjoint sets. Let $\mu_*(E) = \sup\{\mu(A) \mid A \subset E \text{ and } A \in \mathcal{R}\}$ be the inner measure defined by μ on all of 2^S. Let $\mathcal{A} = \mathcal{R} \cup \mathcal{R}^c$ be the σ-algebra generated by $\mathcal{R}$.*

(i) *If we define $\tilde{\mu}(A) = \mu(A)$ for $A \in \mathcal{R}$ and $\tilde{\mu}(A) = \infty$ for $A \in \mathcal{R}^c$ then $\tilde{\mu}$ is a measure on $\mathcal{A}$.*
(ii) *For any $b \in \overline{\mathbb{R}}_+$ if we define $\tilde{\mu}(A) = \mu(A)$ for $A \in \mathcal{R}$ and $\tilde{\mu}(A) = \mu_*(A) + b$ for $A \in \mathcal{R}^c$ then $\tilde{\mu}$ is a measure on $\mathcal{A}$.*
(iii) *Every measure on $\mathcal{A}$ that extends μ on $\mathcal{R}$ is of the above form.*
(iv) *μ has a unique extension to $\mathcal{A}$ if and only if $\mathcal{R} = \mathcal{A}$ or $\mu_*(A) = \infty$ for every $A \in \mathcal{R}^c$.*

Proof. There is nothing to prove if $\mathcal{R} = \mathcal{A}$ so we assume otherwise. Note that in this case there are no disjoint sets in $\mathcal{R}^c$ (if $A, B \in \mathcal{R}^c$ satisfy $A \cap B = \emptyset$ then taking complements $A^c \cup B^c = S$ which shows $S \in \mathcal{R}$ which implies $\mathcal{R} = \mathcal{A}$).

To prove the that the proposed set functions are measures we only need to show countable additivity over all of $\mathcal{A}$. By the above comment we can assume that we have $A_1, A_2, \cdots \in \mathcal{R}$ and $A_0 \in \mathcal{R}^c$ which are all disjoint. Recall that $\cup_{i=0}^{\infty} A_i \in \mathcal{R}^c$. For (i) we have

$$\begin{aligned} \infty = \tilde{\mu}(\cup_{i=0}^{\infty} A_i) & \qquad \text{by definition of } \tilde{\mu} \text{ on } \mathcal{R}^c \\ = \sum_{i=0}^{\infty} \mu(A_i) & \qquad \text{since } \tilde{\mu}(A_0) = \infty \end{aligned}$$

For (ii) things are a little more complicated. First we handle the case of $b = 0$. Since for any $A \in \mathcal{R}$ we have $\mu_*(A) = \mu(A)$ we simplify notation and let the extension be denoted by μ_*. Note that for any $\epsilon > 0$ we can find $B_0 \in \mathcal{R}$ such that $B_0 \subset A_0$ and $\mu(B_0) \geq \mu_*(A_0) - \epsilon$. Then if we define $B_i = A_i$ for $i = 1, 2, \ldots$ we have the B_i are all disjoint sets in $\mathcal{R}$ and $\cup_{i=0}^{\infty} B_i \subset \cup_{i=0}^{\infty} A_i$. Therefore

$$\begin{aligned} \mu_*(\cup_{i=0}^{\infty} A_i) &= \sup\{\mu(C) \mid C \subset \cup_{i=0}^{\infty} A_i \text{ and } C \in \mathcal{R}\} \\ &\geq \mu(\cup_{i=0}^{\infty} B_i) \\ &= \sum_{i=0}^{\infty} \mu(B_i) \\ &\geq \sum_{i=0}^{\infty} \mu_*(A_i) - \epsilon \end{aligned}$$

Since ϵ was arbitrary we conclude $\mu_*(\cup_{i=0}^{\infty} A_i) \geq \sum_{i=0}^{\infty} \mu_*(A_i)$.

To see the other inequality, for any $\epsilon > 0$ we can pick $C \in \mathcal{R}$ such that $C \subset \cup_{i=0}^{\infty} A_i$ and $\mu(C) \geq \mu_*(\cup_{i=0}^{\infty} A_i) - \epsilon$. Since $A_0 \in \mathcal{R}^c$ there is a $B_0 \in \mathcal{R}$ such that $A_0 = B_0^c$ and therefore $C \cap A_0 = C \cap B_0^c = C \setminus B_0 \in \mathcal{R}$. Because $A_i \in \mathcal{R}$ for $i = 1, 2, \ldots$ we know that $A_i \cap C \in \mathcal{R}$ for $i = 1, 2, \ldots$. Putting these two observations together we know can write $C = \cup_{i=0}^{\infty} C_i$ where each $C_i = C \cap A_i \in \mathcal{R}$,

$C_i \subset A_i$ and C_i are disjoint. Now applying the definition of μ_* and countable additivity and monotonicity of μ we see

$$\mu_*(\cup_{i=0}^{\infty} A_i) - \epsilon \leq \mu(C) = \sum_{i=0}^{\infty} \mu(C_i) \leq \sum_{i=0}^{\infty} \mu_*(A_i)$$

Since $\epsilon > 0$ was arbitrary we conclude $\mu_*(\cup_{i=0}^{\infty} A_i) \leq \sum_{i=0}^{\infty} \mu_*(A_i)$ and therefore we have proven $\mu_*(\cup_{i=0}^{\infty} A_i) = \sum_{i=0}^{\infty} \mu_*(A_i)$.

Now we extend the argument to see that defining $\tilde{\mu}(A) = \mu_*(A) + b$ on $\mathcal{R}^c$ also defines a measure. Once again only countable additivity needs to be shown. As noted $\cup_{i=0}^{\infty} A_i \in \mathcal{R}^c$ so using what we have just proven for μ_*,

$$\tilde{\mu}(\cup_{i=0}^{\infty} A_i) = \mu_*(\cup_{i=0}^{\infty} A_i) + b = \mu_*(A_0) + \sum_{i=1}^{\infty} \mu(A_i) + b = \sum_{i=0}^{\infty} \tilde{\mu}(A_i)$$

To see (iii) we must show that every extension of μ to $\mathcal{A}$ has the form $\mu_* + b$ on $\mathcal{R}^c$ for a particular $b \in \overline{\mathbb{R}}_+$. Let $\tilde{\mu}$ be an extension of μ to $\mathcal{A}$. Suppose we have $A_1, A_2 \in \mathcal{R}^c$. From monotonicity we know that $\mu_*(A) \leq \tilde{\mu}(A)$ for every $A \in \mathcal{R}^c$. So there exists constants $b_1, b_2 \in \overline{\mathbb{R}}_+$ such that $\tilde{\mu}(A_i) = \mu_*(A_i) + b_i$ for $i = 1, 2$ and we need to show that $b_1 = b_2$. In addition since $A_1 \cup A_2 \in \mathcal{R}^c$, there is a b such that $\tilde{\mu}(A_1 \cup A_2) = \mu_*(A_1 \cup A_2) + b$. Note that $A_2 \setminus A_1 = A_2 \cap A_1^c = A_1^c \setminus A_2^c \in \mathcal{R}$ therefore

$$\mu_*(A_1 \cup A_2) + b = \tilde{\mu}(A_1 \cup A_2) = \tilde{\mu}(A_1) + \tilde{\mu}(A_2 \setminus A_1) = \mu_*(A_1) + b_1 + \mu_*(A_2 \setminus A_1)$$

which implies $b = b_1$ since μ_* is a measure. The same argument shows that $b = b_2$ hence we see that $b_1 = b_2$ and we are done.

The claim in (iv) is direct consequence of what we have shown. If $\mu_*(A) \neq \infty$ for some $A \in \mathcal{R}^c$ then we have constructed a uncountably infinite number of distinct extension of μ given by $\mu_* + b$ on $\mathcal{R}^c$. On the other hand if $\mu_*(A) = \infty$ for all $A \in \mathcal{R}^c$ then we know any extension must be of the form $\mu_* + b$ on $\mathcal{R}^c$ but these are all equal to ∞ so the uniqueness of the extension is established. □

Example 3.108. It is instructive to consider the scenario of the previous Lemma in the context of the specific example of the σ-ring generated by taking the set of Borel sets on $\mathbb{R}$ that do not contain 0 and Lebesgue measure. We are clearly in the non-unique case with this example and the different extensions correspond to putting point masses with different weights at 0.

We have developed tools that enable us to define measures based on more primitive set functions and this has allowed us to create very important measures such as Lebesgue measure on $\mathbb{R}$. There is another broad class of results that exist that allow one to construct measures. The basic observation is that a measure begets an integral that is a linear function from measurable functions to the extended reals hence it makes sense to pose the question of when a linear functional on some set of measurable functions arises from a measure. Being in possession of such results we are in a position to construct measures by constructing linear functionals instead. In all cases the results in the space make some assumptions about the space of measurable functions on which the functional is defined. In this section we consider the first result in this class; one that is distinguished by the fact that it works on general spaces that do not possess any topological structure.

Definition 3.109. Let $\mathcal{L}$ be a real vector space of real valued functions on a set Ω. We say $\mathcal{L}$ is a *vector lattice* if given any $f, g \in \mathcal{L}$ we have $f \vee g \in \mathcal{L}$ and $f \wedge g \in \mathcal{L}$.

Proposition 3.110. *If $\mathcal{L}$ is a real vector space of real valued functions on a set Ω such that for any $f, g \in \mathcal{L}$ we have $f \vee g \in \mathcal{L}$ then $\mathcal{L}$ is a vector lattice.*

Proof. Simply note that $f \wedge g = -(-f \vee -g)$. □

Definition 3.111. Given a set Ω and a vector lattice $\mathcal{L}$ of real functions on Ω a *pre-integral* is a linear function $I : \mathcal{L} \to \mathbb{R}$ such that

(i) if $f \in \mathcal{L}$ and $f \geq 0$ then $I(f) \geq 0$
(ii) if $f_1, f_2, \cdots \in \mathcal{L}$ such that $f_n \downarrow 0$ then $I(f_n) \downarrow 0$.

To construct a measure that corresponds to a pre-integral we make an intermediate step using the interpretation of an integral as the area under a curve. This will provide us with a measure on the product space $\Omega \times \mathbb{R}$ and then we will show how we restrict this measure in an appropriate way to construct the measure that generates an integral equivalent to I.

Theorem 3.112. *Let $\mathcal{L}$ be a vector lattice of functions on a set Ω with a pre-integral I. For any $f, g \in \mathcal{L}$ such that $f \leq g$ we define*

$$[f, g) = \{(\omega, t) \in \Omega \times \mathbb{R} \mid f(\omega) \leq t < g(\omega)\}$$

, $\mathcal{D} = \{[f, g) \mid f, g \in \mathcal{L}$ such that $f \leq g\}$ and $\nu([f, g)) = I(g - f)$. Then ν is countably additive and extends to a measure on the σ-algebra generated by $\mathcal{D}$.

Proof. The proof proceeds by showing that $\mathcal{D}$ is a semiring, that ν is countably additive on $\mathcal{D}$ and by applying Lemma (TODO:). □

Theorem 3.113. *Let I be a pre-integral on a Stone vector lattice $\mathcal{L}$. Then on the σ-algebra generated by the lattice $\mathcal{L}$ there is a measure μ such that $I(f) = \int f \, d\mu$ for all $f \in \mathcal{L}$. Futhermore the measure μ is uniquely determined on the σ-ring generated by $\mathcal{L}$.*

Proof. We proceed by first defining our measure on the σ-ring $\mathcal{R}$ generated by the functions $\mathcal{L}$. This can be extended (not necessarily uniquely) to a measure on the σ-algebra using Lemma 3.107. Because we have arranged for all of the functions in $\mathcal{L}$ to be $\mathcal{R}$ measurable their integrals will not depend on the extension of μ to a full σ-algebra and their integrals will be determined by the values of μ on $\mathcal{R}$ alone.

Claim 1: $\mathcal{R}$ is generated by sets of the form $f^{-1}(1, \infty)$ for $f \in \mathcal{L}$.

Note that for $c > 0$,

$$f^{-1}(c, \infty) = \{\omega \in \Omega \mid f(\omega) \geq c\} = \{\omega \in \Omega \mid (f/c)(\omega) \geq 1\} = (f/c)^{-1}(1, \infty)$$

and since $\mathcal{L}$ is a Stone lattice (a fortiori a real vector space) we know that $f/c \in \mathcal{L}$. A similar argument shows that for $c > 0$, $f^{-1}(-\infty, -c) = (-f/c)^{-1}(1, \infty)$. We know that intervals $(-\infty, -c)$ and (c, ∞) generate the σ-ring on $\mathbb{R} \setminus \{0\}$, therefore for any $f \in \mathcal{L}$, we have $f^{-1}(\mathcal{B}(\mathbb{R} \setminus \{0\})$ is the σ-ring generated by sets $f^{-1}(c, \infty)$ and $f^{-1}(-\infty, -c)$ for $c > 0$ (Lemma 3.106) which are the same as the sets $(f/c)^{-1}(1, \infty)$ for $c \neq 0$. Thus the σ-ring generated by $\cup_{f \in \mathcal{L}} f^{-1}(\mathcal{B}(\mathbb{R} \setminus \{0\})$ is contained in the σ-ring generated by $\cup_{f \in \mathcal{L}} f^{-1}(1, \infty)$.

Claim 2: We can define a measure μ on the σ-algebra generated by $\mathcal{L}$.

If suffices to define a countably additive set function on the σ-ring $\mathcal{R}$ (Lemma 3.107). We define the measure by embedding $\mathcal{R}$ as sub-σ-ring in σ-algbra $\mathcal{A}$ constructed in Theorem 3.112. To see this, suppose that we have a set $A = f^{-1}(1,\infty)$ with $f \in \mathcal{L}$ and $f \geq 0$. For arbitrary $c > 0$, we define

$$f_n(\omega) = n(f(\omega) - f(\omega) \wedge 1) \wedge c = \begin{cases} 0 & \text{if } \omega \notin A \\ n(f(\omega) - 1) \wedge c & \text{if } \omega \in A \end{cases}$$

and observe that $f_n \in \mathcal{L}$ and $f_n \uparrow c\mathbf{1}_A$. Applying this observation to graphs of f_n in $\Omega \times \mathbb{R}$ we see that $A \times [0,c) = [0, c\mathbf{1}_A) = \cup_{n=1}^{\infty}[0, f_n)$ which shows that $A \times [0,c) \in \mathcal{A}$ for all $c > 0$. From this it follows that $A \times [0,c) \in \mathcal{A}$ for all $A \in \mathcal{R}$. To see this note that for a fixed $c > 0$, the set $\mathcal{R}_c = \{A \times [0,c) \mid A \in \mathcal{R}\}$ is a σ-ring and the set $\{A \subset \Omega \mid A \times [0,c) \in \mathcal{R}_c\}$ is a σ-ring (it can be constructed as a pushforward under an appropriately constructed map or one can see it directly) that contains sets of the form $f^{-1}(1,\infty)$. Thus, $\mathcal{R} \subset \{A \subset \Omega \mid A \times [0,c) \in \mathcal{R}_c\}$.

Having shown that $\mathcal{R}_c$ is a σ-ring in $\mathcal{A}$, we take $c = 1$ and define $\mu(A) = \nu(A \times [0,1)$. That this is countably additive follows from the fact that ν is a measure, so we can extend μ to the σ-algebra $\mathcal{R} \cup \mathcal{R}^c$ in any way we chose.

Now we show how to compute integrals of functions $f \in \mathcal{L}$ with respect to μ and show that they agree with the pre-integral I. Claim 3: □

It should be remarked that one can develop a good deal of measure an integration theory starting from some of the concepts introduced in this section; indeed for a short period of time it was fashionable to do this instead of taking the approach of developing the theory of σ-algebras, measure and integral in the way we have done. Alas, that fashion has passed so we content ourselves with the most streamlined presentation of these ideas we know that gives us Theorem 3.113.

4. Integrals

$$\int_0^\infty e^{-x^2} dx = \frac{\sqrt{\pi}}{2}$$

$$\int_0^\infty x^{2n} e^{-x^2} dx = \frac{\sqrt{\pi}\,(2n-1)!!}{2^{n+1}}$$

$$\int_0^\infty x^{2n+1} e^{-x^2} dx = \frac{n!}{2}$$

$$\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} dx$$

5. Inequalities

From time to time in these notes we'll have a need for some simple inequalities for elementary functions. The following Lemma collects them in one place since they are all proven by use of basic calculus.

Lemma 5.1. *The following inequalities hold:*

(i) $1 + x \leq e^x$ *for all* $x \in \mathbf{R}$.
(ii) $e^x \leq 1 + 2x$ *for all* $x \in [0,1]$.

(iii) $e^x \leq 1 + x + x^2$ *for all* $x \leq 1$.
(iv) $\frac{1}{2}(e^x + e^{-x}) \leq e^{x^2/2}$ *for all* $x \in \mathbf{R}$.
(v) $1 - \frac{x^2}{2} \leq \cos(x)$ *for all* $x \in \mathbf{R}$.
(vi) $x + \log(1-x) \leq 0$ *for all* $x \in [0, 1)$.
(vii) $e^{-x} \leq 1 - (1 - e^{-1})x$ *for all* $x \in [0, 1]$.

Proof. Note that for $x \geq 0$ we can consider $f(x) = e^x - x - 1$ and note that $f(0) = 0$ and moreover we can see that $f(x)$ has a global minimum at $x = 0$ since $f'(x) = e^x - 1$ vanishes precisely at $x = 0$ and $f''(x) = e^x$ is strictly positive.

In a similar vein to show (ii), define $f(x) = 1 + 2x - e^x$ and notice that $f(x)$ has a global maximum at $x = \ln(2)$ and no other local maximum. Thus, it suffices to validate the inequality at the endpoints $x = 0$ and $x = 1$ which is obvious.

To show (iv) we just manipulate series expansions.

$$\begin{aligned}\frac{1}{2}\left(e^x + e^{-x}\right) &= \frac{1}{2}\left(\sum_{n=0}^{\infty} \frac{x^n}{n!} + \sum_{n=0}^{\infty} \frac{(-x)^n}{n!}\right) \\ &= \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!} \\ &\leq \sum_{n=0}^{\infty} \frac{x^{2n}}{2^n n!} = e^{\frac{x^2}{2}}\end{aligned}$$

To show (v), define $f(x) = \frac{x^2}{2} - 1 + \cos(x)$. Calculate the first derivative $f'(x) = x - \sin(x)$. The function $f'(x) = 0$ if and only if $x = 0$. Clearly this is true if $x = 0$, and clearly $f'(x) \neq 0$ for $|x| > 1$. For the interval, $|x| \leq 1$ calculate the second derivative $f''(x) = 1 - \cos(x)$ and note that it is strictly positive for $|x| \leq 1$ and $x \neq 0$. Thus, $f'(x)$ is strictly increasing on the intervals $[-1, 0)$ and $(0, 1]$ and therefore $f'(x) \neq 0$ on these intervals as well. Note also that this argument shows that $f'(x)$ changes sign at $x = 0$ which shows that $f(0) = 0$ is a global minimum.

To show (vi), define $f(x) = x + \log(1 - x)$ and differentiate to see that $f'(x) = 1 - \frac{1}{1-x} = \frac{-x}{1-x} < 0$ for $x \in (0, 1)$. Therefore $f(x) \leq f(0) = 0$ for $x \in [0, 1)$.

To show (vii), let $a = 1 - e^{-1}$ and $f(x) = 1 - ax - e^{-x}$. Take first derivative $f'(x) = -a + e^{-x}$ which has a zero at $x = -\ln a \approx 0.5$. Furthermore $f''(x) = -e^{-x} < 0$ so we have a global maximum at $x = -\ln a$, therefore to show $f(x) \geq 0$ for $x \in [0, 1]$ it suffices to show it at the endpoints: $f(0) = f(1) = 0$. □

6. Probability

Here we begin to focus on the special case of probability spaces. The development of measure theoretic probabilty begins with the assumptions that we are given a

Definition 6.1. A *probability space* is a measure space $(\Omega, \mathcal{A}, P)$ such that $\mathbf{P}\{\Omega\} = 1$.

Given a measurable function $\xi : \Omega \to (S, \mathcal{S})$ we will refer to ξ as a *random element* of S. The special case of a measurable function $\xi : \Omega \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is called a *random variable*. For a random element ξ, by Lemma 3.50 we can push forward the probability measure to get a measure $(P \circ \xi^{-1})$ called the *distribution* or *law* of

ξ. One sometimes writes $\mathcal{L}(\xi)$ to denote the distribution of ξ and one writes $\xi \stackrel{d}{=} \eta$ to denote that ξ and η have the same distribution.

In probability theory the existence of a probability space is critical to the formal development of the theory however it is almost always the case that one is only concerned with results that don't depend on the exact choice of probability space. To make this statement more precise we introduce

Definition 6.2. A probability space $(\Omega', \mathcal{A}', P')$ is an *extension* of $(\Omega, \mathcal{A}, P)$ if there is a surjective measurable map $\pi : \Omega' \to \Omega$ such that $P = P' \circ \pi^{-1}$.

A result is considered properly *probabilistic* if it is preserved under extension of sample space. Note that this is a cultural statement and not a mathematical theorem. As an example of a probabilistic concept, we have the ability to talk about an *event* A and its probability $\mathbf{P}\{A\}$ since given any π we can unambiguously refer to $\pi^{-1}(A)$ as the same event in Ω' and we know that probability is preserved. As an example of a non-probabilistic concept we have the cardinality of an event.

In keeping with the philosophy that probabilistic results are invariant under extension of the underlying probability space, we will follow common practice and try to avoid explicit mention of the underlying probability space in many definitions and results.

Definition 6.3. Given a random vector $\xi = (\xi_1, \ldots, \xi_n)$ in $\mathbb{R}^n$ we define the *distribution function* to be

$$F(x_1, \ldots, x_n) = \mathbf{P}\{\cap_{i=1}^n (\xi_i \leq x_i)\}$$

Lemma 6.4. *Let ξ and η be random vectors in $\mathbb{R}^n$ with distribution functions F and G, then $\xi \stackrel{d}{=} \eta$ if and only if $F = G$.*

Proof. This follows from Lemma 3.65 by noting that sets of the form $(-\infty, x_1] \times \cdots \times (-\infty, x_n]$ form a π-system that contains $\mathbb{R}^n$. □

The construction of Lebesgue-Stieltjes measure shows that every Borel measure on $\mathbb{R}$ is determined uniquely by its distribution function.

Lemma 6.5. *Probability measures of $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ are in one to one correspondence with $F : \mathbb{R} \to \mathbb{R}$ that are right continuous, nondecreasing such that $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$ via the mapping $F(x) = \boldsymbol{P}\{(-\infty, x]\}$.*

Proof. Clearly any probability measure is locally finite so we apply Lemma 3.95 to create a 1-1 correspondence with $\hat{F}$, right continuous and nondecreasing such that $\mathbf{P}\{(a, b]\} = \hat{F}(b) - \hat{F}(a)$. Now define $F(x) = \hat{F}(x) + \mathbf{P}\{(-\infty, 0]\}$. □

Definition 6.6. The *expectation* of a random variable ξ on a probability space $(\Omega, \mathcal{A}, P)$ is defined to be

$$\mathbf{E}\,[\xi] = \int \xi \, dP$$

A very useful corollary to the abstract change of variables Lemma 3.52 is the following

Lemma 6.7 (Expectation Rule)**.** *Let ξ be a random variable and $f : \mathbb{R} \to \mathbb{R}$ be a Borel measurable function. Then*

$$\boldsymbol{E}[f(\xi)] = \int f \, d(P \circ \xi^{-1})$$

In particular,

$$\boldsymbol{E}[\xi] = \int x \, d(P \circ \xi^{-1})$$

Proof. This is just a restatement of Lemma 3.52 for the special case of random variables and measurable functions on $\mathbb{R}$. □

The following lemma is useful for relating tail bounds and expectations.

Lemma 6.8. *Let ξ be a positive random variable with finite expectation. Then $\boldsymbol{E}[\xi] = \int_0^\infty \boldsymbol{P}\{\xi \geq \lambda\} d\lambda$.*

Proof. This is just an application of Tonelli's Theorem,

$$\begin{aligned}
\int_0^\infty \mathbf{P}\{\xi \geq \lambda\} d\lambda &= \int_0^\infty \left[\int \mathbf{1}_{\xi \geq \lambda} dP \right] d\lambda \\
&= \int \left[\int_0^\infty \mathbf{1}_{\xi \geq \lambda} d\lambda \right] dP \\
&= \int \left[\int_0^\xi d\lambda \right] dP \\
&= \int \xi \, dP \\
&= \mathbf{E}[\xi]
\end{aligned}$$

□

Lemma 6.9 (Cauchy Schwartz Inequality)**.** *Let ξ and η satisfy $\boldsymbol{E}\left[\xi^2\right], \boldsymbol{E}\left[\eta^2\right] < \infty$ then $\xi\eta$ is integrable and $\boldsymbol{E}[\xi\eta]^2 \leq \boldsymbol{E}\left[\xi^2\right] \boldsymbol{E}\left[\eta^2\right]$.*

Proof. Since we have both $0 \leq (\xi + \eta)^2$ and $0 \leq (\xi - \eta)^2$ we have $|\xi\eta| \leq \frac{1}{2}(\xi^2 + \eta^2)$ which shows that $\xi\eta$ is integrable.

There are a host of different proofs of Cauchy Schwartz inequality. Here is perhaps the simplest one. Note that for all $t \in \mathbb{R}$, $0 \leq \mathbf{E}\left[(t\xi + \eta)^2\right] = \mathbf{E}\left[\xi^2\right] t^2 + 2\mathbf{E}[\xi\eta]\, t + \mathbf{E}\left[\eta^2\right]$. The quadratic formula implies that $\sqrt{4\mathbf{E}\left[\xi^2\right]\mathbf{E}\left[\eta^2\right] - (2\mathbf{E}[\xi\eta])^2} \geq 0$ which in turn implies the result.

The proof we just provided is probably the slickest one available but has the disadvantage of being very specific to the quadratic case. There is a different proof of Cauchy Schwartz that we provide that involves two steps that have a broader application. The idea is to derive Cauchy Schwartz from the trival fact that for all real numbers x, y we have $xy \leq \frac{x^2}{2} + \frac{y^2}{2}$ (which we used when showing integrability of $\xi\eta$). Applying this fact to ξ and η we see that

$$\mathbf{E}[\xi\eta] \leq \frac{\mathbf{E}\left[\xi^2\right]}{2} + \frac{\mathbf{E}\left[\eta^2\right]}{2}$$

To finish the proof, we apply a *normalization trick* by defining $\hat{\xi} = \frac{\xi}{\sqrt{\mathbf{E}[\xi^2]}}$ and $\hat{\eta} = \frac{\eta}{\sqrt{\mathbf{E}[\eta^2]}}$ so that $\mathbf{E}\left[\hat{\xi}^2\right] = \mathbf{E}\left[\hat{\eta}^2\right] = 1$. Now we apply the above bound and

linearity of expectation to see that

$$\frac{1}{\sqrt{\mathbf{E}\left[\xi^2\right]}\sqrt{\mathbf{E}\left[\eta^2\right]}}\mathbf{E}\left[\xi\eta\right] = \mathbf{E}\left[\hat{\xi}\hat{\eta}\right] \leq 1$$

which yields the result. □

Applications of Cauchy Schwatz are ubiquitous in analysis. Only slightly less common are applications of the following generalization. First a definition

Definition 6.10. Given any $p > 0$ and random variable ξ, the L^p *norm* of ξ is

$$\|\xi\|_p = \left(\mathbf{E}\left[|\xi|^p\right]\right)^{\frac{1}{p}}$$

Lemma 6.11 (Hölder Inequality)**.** *Given $p, q, r > 0$ such that $\frac{1}{r} = \frac{1}{p} + \frac{1}{q}$ and random variables ξ and η, we have*

$$\|\xi\eta\|_r \leq \|\xi\|_p \|\eta\|_q$$

Proof. We start by assuming that $r = 1$. The proof here is a direct generalization of the second proof we provided for Cauchy Schwartz. To get started we need to find a generalization of the simple fact that $xy \leq \frac{x^2}{2} + \frac{y^2}{2}$.

The inequality we need is called Young's Inequality and is derived from the following fact. Let f be an continuous increasing function $f : [0, c] \to \mathbb{R}$ such that $f(0) = 0$. Then the area interpretation of integral tells us that for $0 \leq a \leq c$ and $0 \leq b \leq f(c)$ we have

$$ab \leq \int_0^a f(x)\,dx + \int_0^b f^{-1}(x)\,dx$$

with equality if and only if $b = f(a)$.

For our case, we first assume that $r = 1$. Define $f(x) = x^{p-1}$ then observe that $f^{-1}(x) = x^{q-1}$ since $1 = \frac{1}{p} + \frac{1}{q}$ is equivalent to $(p-1)(q-1) = 1$. Therefore we have Young's Inequality, $ab \leq \frac{a^p}{p} + \frac{b^q}{q}$.

Now applying the normalization trick by defining $\hat{\xi} = \frac{|\xi|}{\|\xi\|_p}$ and $\hat{\eta} = \frac{|\eta|}{\|\eta\|_q}$ so that $\|\hat{\xi}\|_p = \|\hat{\eta}\|_q = 1$. We now apply Young's Inequality to $\hat{\xi}$ and $\hat{\eta}$ to see

$$\frac{1}{\|\hat{\xi}\|_p \|\hat{\eta}\|_q}\mathbf{E}\left[|\xi\eta|\right] = \mathbf{E}\left[\hat{\xi}\hat{\eta}\right] \leq \frac{1}{p} + \frac{1}{q} = 1$$

Lastly we generalize to general $r > 0$. Given $\frac{1}{r} = \frac{1}{p} + \frac{1}{q}$ we define $\hat{p} = \frac{p}{r}$ and $\hat{q} = \frac{q}{r}$ so that $1 = \frac{1}{\hat{p}} + \frac{1}{\hat{q}}$ and

$$\mathbf{E}\left[|\xi\eta|^r\right] \leq \|\xi^r\|_{\hat{p}} \|\eta^r\|_{\hat{q}} = \|\xi\|_p^r \|\eta\|_q^r$$

Taking r^{th} roots we are done. □

Corollary 6.12. *For $p > r > 0$ and any random variable ξ, we have $\|\xi\|_r \leq \|\xi\|_p$.*

Proof. Define $q = \frac{p-r}{pr} > 0$ and apply Hölder's Inequality to see that $\|\xi\|_r \leq \|\xi\|_p \|1\|_q = \|\xi\|_p$. □

It worth noting that the corollary above is generally true on finite measure spaces but fails for non-finite measure spaces (e.g. consider $f(x) = \frac{1}{x}$ which has finite L^p norm on $[1, \infty)$ for $p > 1$ but infinite L^1 norm on $[1, \infty)$).

6.1. **Convexity and Jensen's Inequality.**

Definition 6.13. A function $\varphi : \mathbb{R}^n \to \mathbb{R}$ is said to be *convex* if for all $x, y \in \mathbb{R}^n$ and $t \in [0,1]$, we have

$$\varphi(tx + (1-t)y) \leq t\varphi(x) + (1-t)\varphi(y)$$

φ is said to be *strictly convex* if it is convex and for all $t \in (0,1)$,

$$\varphi(tx + (1-t)y) < t\varphi(x) + (1-t)\varphi(y)$$

TODO: Convex functions are continuous

Convex functions are almost surely differentiable.

Lemma 6.14. *Let $\varphi : [a,b] \to \mathbb{R}$ be convex. Then for every $a < x < b$, we have*

$$\frac{\varphi(x) - \varphi(a)}{x - a} \leq \frac{\varphi(b) - \varphi(a)}{b - a} \leq \frac{\varphi(b) - \varphi(x)}{b - x}$$

If φ is strictly convex then the inequalities may be replaced by strict inequalities.

Proof. Note that we can write $x = ta + (1-t)b$ with $t = \frac{b-x}{b-a} \in [0,1]$. So applying the definition of convexity we know that $\varphi(x) \leq t\varphi(a) + (1-t)\varphi(b)$ and using the fact that $1 - t = \frac{x-a}{b-a}$ we get

$$\frac{\varphi(x) - \varphi(a)}{x - a} \leq \frac{t\varphi(a) + (1-t)\varphi(b) - \varphi(a)}{x - a} = \frac{1-t}{x-a}(\varphi(b) - \varphi(a)) = \frac{\varphi(b) - \varphi(a)}{b - a}$$

and in a similar way,

$$\frac{\varphi(b) - \varphi(x)}{b - x} \geq \frac{\varphi(b) - t\varphi(a) - (1-t)\varphi(b)}{b - x} = \frac{t}{b-x}(\varphi(b) - \varphi(a)) = \frac{\varphi(b) - \varphi(a)}{b - a}$$

It is clear from the definition of strict convexity that the inequalities above may be replaced by strict inequalities if φ is strictly convex. □

Lemma 6.15. *Let $\varphi : [a,b] \to \mathbb{R}$ be a convex function, then for every $x \in (a,b)$, $D^-\varphi(x)$ and $D^+\varphi(x)$ exist and furthermore for $a < x < y < b$ we have*

$$D^-\varphi(x) \leq D^+\varphi(x) \leq \frac{\varphi(y) - \varphi(x)}{y - x} \leq D^-\varphi(y) \leq D^+\varphi(y)$$

If φ is strictly convex then we have

$$D^+\varphi(x) < \frac{\varphi(y) - \varphi(x)}{y - x} < D^-\varphi(y)$$

Proof. Lemma 6.14 shows that for $a < x < b$ and $h > 0$, $\frac{\varphi(x+h)-\varphi(x)}{h}$ is an increasing function of h bounded below by $\frac{\varphi(x)-\varphi(a)}{x-a}$. Thus $D^+\varphi(x) = \lim_{h \downarrow 0} \frac{\varphi(x+h)-\varphi(x)}{h}$ is a decreasing limit hence exists. Similarly $\frac{\varphi(x-h)-\varphi(x)}{-h} = \frac{\varphi(x)-\varphi(x-h)}{h}$ is an decreasing function of h bounded above by $\frac{\varphi(b)-\varphi(x)}{b-x}$. Thus $D^-\varphi(x) = \lim_{h \downarrow 0} \frac{\varphi(x-h)-\varphi(x)}{-h}$ is a bounded increasing limit hence exists.

The inequalities follow directly from Lemma 6.14. For example, since $D^+\varphi(x) = \lim_{h \downarrow 0} \frac{\varphi(x+h)-\varphi(x)}{h}$ and for all $x < x + h < y$, we have $\frac{\varphi(x+h)-\varphi(x)}{h} \leq \frac{\varphi(y)-\varphi(x)}{y-x}$ we get $D^+\varphi(x) \leq \frac{\varphi(y)-\varphi(x)}{y-x}$. In the strictly convex case, we know that for any w with $x < w < y$ we have by what we have just shown and another application of Lemma 6.14

$$D^+\varphi(x) \leq \frac{\varphi(w) - \varphi(x)}{w - x} < \frac{\varphi(y) - \varphi(x)}{y - x}$$

The case of $D^-\varphi(y)$ follows analogously. □

Corollary 6.16. *Let $\varphi : [a,b] \to \mathbb{R}$ be convex then for $x \in (a,b)$ there exists constants $A, B \in \mathbb{R}$ such that $Ay + B \leq \varphi(y)$ for all $y \in [a,b]$ and $Ax + B = \varphi(x)$. If φ is strictly convex then we may assume that $Ay + B < \varphi(y)$ for $y \neq x$.*

Proof. By Lemma 6.15 we can pick $D^-(x) \leq A \leq D^+(x)$. Also by that result we know that for all $h > 0$, in fact we have

$$\frac{\varphi(x) - \varphi(x-h)}{h} \leq A \leq \frac{\varphi(x+h) - \varphi(x)}{h}$$

which gives the result upon clearing denominators and defining $B = \varphi(x)$. Once again, the strictly convex case follows easily. □

TODO: Extend this to $\mathbb{R}^n$ (presumably this can be done by taking partial Dini Derivatives.

Theorem 6.17 (Jensen's Inequality)**.** *Let ξ be a random vector in $\mathbb{R}^n$ and $\varphi : \mathbb{R}^n \to \mathbb{R}$ be a convex function such that ξ and $\varphi(\xi)$ are integrable. Then*

$$\varphi(\boldsymbol{E}[\xi]) \leq \boldsymbol{E}[\varphi(\xi)]$$

If φ is strictly convex then we have $\varphi(\boldsymbol{E}[\xi]) = \boldsymbol{E}[\varphi(\xi)]$ if and only if $\xi = \boldsymbol{E}[\xi]$ a.s.

Proof. We use the fact that for every $x \in \mathbb{R}^n$ we have a subdifferential $\langle a, y\rangle + b$ that satisfies

$$\langle a, y\rangle + b \leq \varphi(y)$$
$$\langle a, x\rangle + b = \varphi(x)$$

In particular, choose such an $a, b \in \mathbb{R}^n$ for the choice $x = \mathbf{E}[\xi]$. Then by monotonicity and linearity of integral

$$\begin{aligned}\mathbf{E}[\varphi(\xi)] &\geq \mathbf{E}[\langle a, \xi\rangle + b] \\ &= \langle a, \mathbf{E}[\xi]\rangle + b = \varphi(\xi)\end{aligned}$$

which gives the result.

If φ is strictly convex then when $\xi \neq \mathbf{E}[\xi]$, we have

$$0 < \varphi(\xi) - \varphi(\mathbf{E}[\xi]) - \langle a, \xi - \mathbf{E}[\xi]\rangle$$

Thus if $\varphi(\mathbf{E}[\xi]) = \mathbf{E}[\varphi(\xi)]$ using linearity of expectation

$$\begin{aligned}\mathbf{E}[(\varphi(\xi) - \varphi(\mathbf{E}[\xi]) - \langle a, \xi - \mathbf{E}[\xi]\rangle); \xi \neq \mathbf{E}[\xi]] &= \mathbf{E}[\varphi(\xi) - \varphi(\mathbf{E}[\xi]) - \langle a, \xi - \mathbf{E}[\xi]\rangle] \\ &= 0\end{aligned}$$

from which we conclude $\mathbf{1}_{\xi \neq \mathbf{E}[\xi]} = 0$ a.s. □

7. Independence

Definition 7.1. Given a measure space $(\Omega, \mathcal{A}, P)$, a set T and a collection of σ-algebras $\mathcal{F}_t$ for $t \in T$, we say that the $\mathcal{F}_t$ are *k-ary independent* if for every finite subset $t_1, \ldots, t_n \in T$ with $n \leq k$ and every $A_{t_i} \in \mathcal{F}_{t_i}$ we have $\mathbf{P}\{A_{t_1} \cap \cdots \cap A_{t_n}\} = \mathbf{P}\{A_{t_1}\} \cdots \mathbf{P}\{A_{t_n}\}$. We say that $\mathcal{F}_t$ are *independent* if the $\mathcal{F}_t$ are k-ary independent for every $k > 0$. It is common to refer to independent events as *jointly independent* or *mutually independent* events when it is desirable to provide emphasis that we are not considering k-ary independence for some particular value of k. Furthermore, 2-ary independent events are often referred to as *pairwise independent* events.

Definition 7.2. Given a probability space $(\Omega, \mathcal{A}, P)$, a set T and a collection of random elements $\xi_t : (\Omega, \mathcal{A}) \to (S_t, \mathcal{S}_t)$ for $t \in T$, we say that the ξ_t are *independent* if the σ-algebras $\sigma(\xi_t)$ are independent.

Example 7.3. Given two sets $A, B \in \mathcal{A}$ it is easy to see that $\sigma(A)$ and $\sigma(B)$ are independent if and only if $\mathbf{P}\{A \cap B\} = \mathbf{P}\{A\} \cdot \mathbf{P}\{B\}$ thus the notion of independence of σ-algebras generalizes the simple notion of independence from elementary probability.

Example 7.4. Consider the space of triples $\{(0,0,0),(0,1,1),(1,0,1),(1,1,0)\}$ with a uniform distribution. Let ξ_1, ξ_2, ξ_3 be the coordinate functions. Note that each of ξ_i is uniformly distributed and that each joint distribution (ξ_i, ξ_j) for $i \neq j$ is uniformly distributed as well. This shows that the ξ_i are pairwise independent. On the other hand, note that joint distribution (ξ_1, ξ_2, ξ_3) is also uniformly distributed hence does not equal the product of the marginal distributions hence the ξ_i are not jointly independent. Intuitively the source of the dependence is clear; we have arranged the sample space so that specifying two coordinate values determines the value of the third coordinate. Note this example can also be framed in a more elementary way in terms of events. Consider the events $A_1 = \{(0,0,0),(0,1,1)\}$, $A_2 = \{(0,0,0),(1,0,1)\}$ and $A_3 = \{(0,1,1),(1,0,1)\}$. Note that the events are pairwise independent but not independent.

Lemma 7.5. *Suppose we are given a finite collection of random elements $\xi_1, \ldots, \xi_n$ in measurable spaces $S_1, \ldots, S_n$ with distributions $\mu_1, \ldots, \mu_n$. The ξ_i are independent if and only if the distribution of $(\mu_1, \ldots, \mu_n)$ on $S_1 \times \cdots \times S_n$ is $\mu_1 \otimes \cdots \otimes \mu_n$.*

Proof. If we assume that joint distribution of ξ_i is $\mu_1 \otimes \cdots \otimes \mu_n$ the clearly ξ_i are independent since

$$\begin{aligned}\mathbf{P}\{\xi_1^{-1}(B_1) \cap \cdots \cap \xi_n^{-1}(B_n)\} &= \mathbf{P}\{(\xi_1, \ldots, \xi_n)^{-1}(B_1 \times \cdots \times B_n)\} \\ &= \mathbf{P}\{\xi_1^{-1}(B_1)\} \cdots \mathbf{P}\{\xi_n^{-1}(B_n)\}\end{aligned}$$

On the other hand, if we assume that the ξ_i are independent the above calculation shows that $(P \circ (\xi_1, \ldots, \xi_n)^{-1}) = \mu_1 \otimes \cdots \otimes \mu_n$ on cylinder sets which together with the finiteness of probability measures shows that they are equal everywhere by the uniqueness of product measure proved in Theorem 3.81. □

The fact that the joint distribution of indpendent random variables only depends on the distribution of the underlying random variables has the important consequence that the distribution of *sums* of independent random variables also only depends on the distribution of the underlying random variables. However we can actually be a bit more precise than that.

Definition 7.6. A *measurable group* is a group G with a σ-algebra $\mathcal{G}$ such that the group inverse is $\mathcal{G}$-measurable and the group operation is $\mathcal{G} \otimes \mathcal{G}/\mathcal{G}$-measurable.

Definition 7.7. Given two σ-finite measures μ and ν on a measurable group $(G, \mathcal{G})$, the *convolution* $\mu * \nu$ is the measure on G defined by taking the pushforward of $\mu \otimes \nu$ under the group operation.

Lemma 7.8. *Convolution of measures on a measurable group $(G, \mathcal{G})$ is associative. Furthermore, if G is Abelian, then convolution of measures is commutative and we*

have the formula

$$\mu * \nu(B) = \int \mu(B - g)\, d\nu(g) = \int \nu(B - g)\, d\mu(g)$$

Proof. First we derive the formula for the convolution of two measures as integrals. Suppose we are given σ-finite measures μ, ν and a measurable $A \in \mathcal{G}$. Define $A^2 = \{(g,h) \mid gh \in A\}$ and then the definition of the pushforward of a measure,the construction of product measure and Tonelli's Theorem we get

$$\begin{aligned}
(\mu * \nu)(A) &= (\mu \otimes \nu)(A^2) \\
&= \int \int \mathbf{1}_{A^2}(g,h)\, d(\mu \otimes \nu)(g,h) \\
&= \int \left[\int \mathbf{1}_{A^2}(g,h)\, d\mu(g) \right] d\nu(h) \\
&= \int \left[\int \mathbf{1}_{A^2}(g,h)\, d\nu(h) \right] d\mu(g)
\end{aligned}$$

Now consider the inner integral for a fixed $h \in G$ and define for each such fixed h the right translation Ah^{-1} and note that as a function of g alone, $\mathbf{1}_{A^2}(g,h) = \mathbf{1}_{Ah^{-1}}(g)$. Similarly, for fixed g we introduce the left translation $g^{-1}A$ and have $\mathbf{1}_{A^2}(g,h) = \mathbf{1}_{g^{-1}A}(h)$. Substituting into the integrals above,

$$(\mu * \nu)(A) = \int \mu(A \cdot g^{-1})\, d\nu(g) = \int \nu(g^{-1} \cdot A)\, d\mu(g)$$

In particular, if G is Abelian then $g^{-1} \cdot A = A \cdot g^{-1}$ and we have the formula above.

To see the associativity is an application of Tonelli's Theorem with a bit of messy notation. Suppose we are given σ-finite measures μ_1, μ_2, μ_3 and a measurable $A \in \mathcal{G}$. Define $A^3 = \{(g,h,k) \mid ghk \in A\}$ and note that for fixed h, k we have $\mathbf{1}_{A^3}(g,h,k) = \mathbf{1}_{Ak^{-1}h^{-1}}(g)$ and for fixed g, h we have $\mathbf{1}_{A^3}(g,h,k) = \mathbf{1}_{k^{-1}g^{-1}A}(k)$ Now applying this observation and the integral formula above

$$\begin{aligned}
((\mu_1 * \mu_2) * \mu_3)(A) &= \int (\mu_1 * \mu_2)(Ak^{-1})\, d\mu_3(k) \\
&= \int \int \mu_1(Ak^{-1}h^{-1})\, d\mu_2(h) d\mu_3(k) \\
&= \int \int \int \mathbf{1}_{A^3}(g,h,k)\, d\mu_1(g) d\mu_2(h) d\mu_3(k) \\
&= \int \int \int \mathbf{1}_{A^3}(g,h,k)\, d\mu_3(k) d\mu_2(h) d\mu_1(g) \\
&= \int \int \mu_3(h^{-1}g^{-1}A)\, d\mu_2(h) d\mu_1(g) \\
&= \int (\mu_2 * \mu_3)(g^{-1}A)\, d\mu_1(g) \\
&= (\mu_1 * (\mu_2 * \mu_3))(A)
\end{aligned}$$

□

Definition 7.9. A measure μ on a measurable group $(G, \mathcal{G})$ is said to be *left invariant* if for every $g \in G$ and $A \in \mathcal{G}$, $\mu(g \cdot A) = \mu(A)$. A measure is said to be

right invariant if for every $g \in G$ and $A \in \mathcal{G}$, $\mu(A \cdot g) = \mu(A)$. A measure that is both right invariant and left invariant is said to be *invariant*.

Lemma 7.10. *Let λ be an invariant measure on a measurable Abelian group $(G, \mathcal{G})$ and let $\mu = f \cdot \lambda$ and $\nu = g \cdot \lambda$ be measures which have densities with respect to λ. Then $\mu * \nu$ has the λ-density*

$$(f * g)(x) = \int f(x-y)g(y)\, d\lambda(y)$$

Proof. By the integral formula for convolution, given $A \in \mathcal{G}$,

$$\begin{aligned}
(\mu * \nu)(A) &= \int \mu(A-y)\, d\nu(y) \\
&= \int\int \mathbf{1}_{A-y}(x) f(x) g(y)\, d\lambda(x) d\lambda(y) \\
&= \int\int \mathbf{1}_{A}(x+y) f(x) g(y)\, d\lambda(x) d\lambda(y) \\
&= \int\int \mathbf{1}_{A}(x) f(x-y) g(y)\, d\lambda(x) d\lambda(y) \\
&= \int \mathbf{1}_A(x) \left[\int f(x-y) g(y)\, d\lambda(y) \right] d\lambda(x) \\
&= ((f * g) \cdot \lambda)(A)
\end{aligned}$$

□

Example 7.11. Let ξ and η be independent $N(0,1)$ random variables. Then $\xi + \eta$ is an $N(0,2)$ random variable. From Corollary 7.10, we know $\xi + \eta$ has density given by the convolution of Gaussian densities.

$$\frac{1}{2\pi} \int e^{\frac{-(x-y)^2}{2}} e^{\frac{-y^2}{2}} dy = \frac{1}{2\pi} \int e^{-(y^2 - xy + \frac{1}{2}x^2)} dy = \frac{1}{2\pi} e^{\frac{-x^2}{4}} \int e^{-(y-\frac{x}{2})^2} dy = \frac{1}{\sqrt{4\pi}} e^{\frac{-x^2}{4}}$$

Lemma 7.12. *Suppose we are given two π-systems $\mathcal{S}$ and $\mathcal{T}$ in a probability space $(\Omega, \mathcal{A}, P)$ such that $\boldsymbol{P}\{A \cap B\} = \boldsymbol{P}\{A\}\boldsymbol{P}\{B\}$ for all $A \in \mathcal{S}$ and $B \in \mathcal{T}$. Then $\sigma(\mathcal{S})$ and $\sigma(\mathcal{T})$ are independent.*

Proof. This is simply a pair of monotone class arguments. First pick arbitrary element $A \in \mathcal{A}$. We define $\mathcal{C} = \{B \in \mathcal{A} \mid \mathbf{P}\{A \cap B\} = \mathbf{P}\{A\}\mathbf{P}\{B\}\}$. We claim that $\mathcal{C}$ is a λ-system. First it is clear that $\Omega \in \mathcal{C}$. Next assume that $B, C \in \mathcal{C}$ with $C \supset B$. Then $C \setminus B \in \mathcal{C}$ because

$$\begin{aligned}
\mathbf{P}\{A \cap (C \setminus B)\} &= \mathbf{P}\{(A \cap C) \setminus (A \cap B)\} \\
&= \mathbf{P}\{A \cap C\} - \mathbf{P}\{A \cap B\} \\
&= \mathbf{P}\{A\}\mathbf{P}\{C\} - \mathbf{P}\{A\}\mathbf{P}\{B\} \\
&= \mathbf{P}\{A\}\left(\mathbf{P}\{C\} - \mathbf{P}\{B\}\right) = \mathbf{P}\{A\}\mathbf{P}\{C \setminus B\}
\end{aligned}$$

Next assume that $B_1 \subset B_2 \subset \cdots$ with $B_i \in \mathcal{C}$. We have $\cup_{n=1}^{\infty} B_n \in \mathcal{C}$ by the calculation

$$\begin{aligned}
\mathbf{P}\{A \cap \cup_{n=1}^{\infty} B_n\} &= \mathbf{P}\{\cup_{n=1}^{\infty} A \cap B_n\} && \text{by DeMorgan's Law} \\
&= \lim_{n\to\infty} \mathbf{P}\{A \cap B_n\} && \text{by Continuity of Measure} \\
&= \lim_{n\to\infty} \mathbf{P}\{A\}\mathbf{P}\{B_n\} && \text{since } B_n \in \mathcal{C} \\
&= \mathbf{P}\{A\} \lim_{n\to\infty} \mathbf{P}\{B_n\} \\
&= \mathbf{P}\{A\}\mathbf{P}\{\cup_{n=1}^{\infty} B_n\} && \text{by Continuity of Measure}
\end{aligned}$$

Our assumption is that if we pick $A \in \mathcal{S}$, then $\mathcal{T} \subset \mathcal{C}$ so the π-λ Theorem (Theorem 3.24) shows that $\sigma(\mathcal{T}) \subset \mathcal{C}$. Since our choice of $A \in \mathcal{S}$ can be arbitrary, we know for every for every $A \in \mathcal{S}$ and every $B \in \sigma(\mathcal{T})$ we have $\mathbf{P}\{A \cap B\} = \mathbf{P}\{A\}\mathbf{P}\{B\}$.

It remains to extend $\mathcal{S}$ to $\sigma(\mathcal{S})$. This is done in exactly the same way. Pick a $B \in \sigma(\mathcal{T})$ and define $\mathcal{D}\{A \in \mathcal{A} \mid \mathbf{P}\{A \cap B\} = \mathbf{P}\{A\}\mathbf{P}\{B\}\}$. We have shown that $\mathcal{D}$ is a λ-system and that $\mathcal{S} \subset \mathcal{D}$ hence the π-λ Theorem gives us $\mathcal{D} \supset \sigma(\mathcal{S})$. Since $B \in \sigma(\mathcal{T})$ was arbitrary we have shown independence of $\sigma(\mathcal{S})$ and $\sigma(\mathcal{T})$. □

Lemma 7.13. *Let $\mathcal{A}_t$ for $t \in T$ be an independent family of σ-algebras on Ω. The for any disjoint partition $\mathcal{T}$ of T we have $\sigma(\bigcup_{s\in S} \mathcal{A}_s)$ are independent where $S \in \mathcal{T}$.*

Proof. For S and element of the partition of T, let $\mathcal{C}_S$ be the set of all finite intersections of elements from $\cup_{s\in S}\mathcal{A}_s$. Clearly each $\mathcal{C}_S$ is a π-system that generates $\sigma(\bigcup_{s\in S} \mathcal{A}_s)$. Moreover, the independence of the $\mathcal{A}_t$ for all $t \in T$ shows that the $\mathcal{C}_S$ are independent π-systems by associativity of finite intersection of sets and multiplication in $\mathbb{R}$. Thus Lemma 7.12 shows the result. □

Note that the previous lemma can be taken as demonstrating that independence of sets cannot be destroyed by applying the operations of complementation, countable union and countable intersection. The property of independence is also very robust in the sense that it cannot be destroyed by composition with any measurable mapping.

Lemma 7.14. *A finite collection of random elements $\xi_1, \ldots, \xi_n$ in measurable spaces $(S_1, \mathcal{S}_1), \cdots, (S_n, \mathcal{S}_n)$ is independent if and only $f_1 \circ \xi_1, \cdots, f_n \circ \xi_n$ is independent for every measurable $f_1, \cdots, f_n$.*

Proof. The reverse implication is clear because the identity on every $(S_i, \mathcal{S}_i)$ is measurable.

Now if ξ_i are independent then by definition $\sigma(\xi_i)$ are independent σ-algebras. But for any measurable f_i, $\sigma(f_i \circ \xi_i) \subset \sigma(\xi_i)$ and therefore the $f_1 \circ \xi_1, \cdots, f_n \circ \xi_n$ are independent. □

Implicit in a few of the above proofs is the fact that independence among groups of independent objects can be reduced to checking independence of finite subsets within the groups. Here is a codification of this fact stated in the simple case of checking pairwise independence.

Lemma 7.15. *Let $\mathcal{F}_t$ and $\mathcal{G}_s$ be sets of σ-algebras. Then $\sigma(\bigcup_{t\in T} \mathcal{F}_t)$ is independent of $\sigma(\bigcup_{s\in S} \mathcal{G}_s)$ if an only if for every finite subset $T' \subset T$ and $S' \subset S$, we have $\sigma(\bigcup_{t\in T'} \mathcal{F}_t)$ is independent of $\sigma(\bigcup_{s\in S'} \mathcal{G}_s)$*

Proof. One direction of this is trivial. For the other direction suppose we have independence over each of the finite subsets. To prove the result note that set of finite intersections of elements of $\bigcup_{t\in T}\mathcal{F}_t$ is a π-system that generates $\sigma(\bigcup_{t\in T}\mathcal{F}_t)$ (and similarly with S). Our assumption tells us that these π-systems are independent hence we appeal to Lemma 7.12. □

Lemma 7.16. *A finite collection of random elements $\xi_1,\dots,\xi_n$ in measurable spaces $(S_1,\mathcal{S}_1),\cdots,(S_n,\mathcal{S}_n)$ is independent if and only if*

$$\boldsymbol{E}[f_1(\xi_1)\cdots f_n(\xi_n)] = \boldsymbol{E}[f_1(\xi_1)]\cdots\boldsymbol{E}[f_n(\xi_n)]$$

for all $f_i : S_n \to \mathbb{R}$ that are either bounded measurable or positive measurable.

Proof. Note that for the special case $f_i = \mathbf{1}_{A_i}$ for Borel sets $A_i \in \mathcal{B}(\mathbb{R})$, $f_i(\xi_i) = \mathbf{1}_{f_i^{-1}(A_i)}$ and therefore the claim is equivalent to the definition of independence as we can see by the following calculation

$$\begin{aligned}\mathbf{E}[f_1(\xi_1)\cdots f_n(\xi_n)] &= \mathbf{E}\left[\mathbf{1}_{f_1^{-1}(A_1)}\cdots\mathbf{1}_{f_n^{-1}(A_n)}\right]\\ &= \mathbf{P}\{f_1^{-1}(A_1)\cap\cdots\cap f_n^{-1}(A_n)\}\\ &= \mathbf{P}\{f_1^{-1}(A_1)\}\cdots\mathbf{P}\{f_n^{-1}(A_n)\}\\ &= \mathbf{E}[f_1(\xi_1)]\cdots\mathbf{E}[f_n(\xi_n)]\end{aligned}$$

Therefore if we assume the result for all positive or bound measurable f then we certainly have independence.

On the other hand if we assume independence of the ξ_i then we know that the desired result holds for f_i that are indicator functions. It remains to apply the standard machinery to derive the result for more general f_i.

For f_i simple functions we simply use linearity of expectation. If we write $f_i = c_{1,i}\mathbf{1}_{A_{1_i,i}} + \cdots + c_{m_i,i}\mathbf{1}_{A_{m_i,i}}$ then

$$\begin{aligned}\mathbf{E}[f_1(\xi_1)\cdots f_n(\xi_n)] &= \sum_{k_1=1}^{m_1}\cdots\sum_{k_n=1}^{m_n} c_{k_1,1}\cdots c_{k_n,n}\mathbf{E}\left[\mathbf{1}_{A_{k_1,1}}(\xi_1)\cdots\mathbf{1}_{A_{k_n,n}}(\xi_n)\right]\\ &= \sum_{k_1=1}^{m_1}\cdots\sum_{k_n=1}^{m_n} c_{k_1,1}\cdots c_{k_n,n}\mathbf{E}\left[\mathbf{1}_{A_{k_1,1}}(\xi_1)\right]\cdots\mathbf{E}\left[\mathbf{1}_{A_{k_n,n}}(\xi_n)\right]\\ &= \sum_{k_1=1}^{m_1} c_{k_1,1}\mathbf{E}\left[\mathbf{1}_{A_{k_1,1}}(\xi_1)\right]\cdots\sum_{k_n=1}^{m_n} c_{k_n,n}\mathbf{E}\left[\mathbf{1}_{A_{k_n,n}}(\xi_n)\right]\\ &= \mathbf{E}[f_1(\xi_1)]\cdots\mathbf{E}[f_n(\xi_n)]\end{aligned}$$

To show the result for positive f, first start by assuming that f_1 is positive and $f_2,\cdots,f_n$ are simple. Pick $f_{i,1}$ increasing simple functions such that $f_{i,1}\uparrow f_1$. Then we have $f_{i,1}f_2\cdots f_n \uparrow f_1f_2\cdots f_n$ we have

$$\begin{aligned}\mathbf{E}[f_1(\xi_1)\cdots f_n(\xi_n)] &= \lim_{i\to\infty}\mathbf{E}[f_{i,1}(\xi_1)\cdots f_n(\xi_n)] && \text{by Monotone Convergence}\\ &= \lim_{i\to\infty}\mathbf{E}[f_{i,1}(\xi_1)]\cdots\mathbf{E}[f_n(\xi_n)] && \text{result for simple functions}\\ &= \mathbf{E}[f_1(\xi_1)]\cdots\mathbf{E}[f_n(\xi_n)] && \text{by Monotone Convergence}\end{aligned}$$

Having shown the result for f_1 positive and $f_2,\cdots,f_n$ simple just iterate with Monotone Convergence as above to see the result for all $f_1,\cdots,f_n$ positive.

For f_i bounded, first write $f_1 = f_1^+ - f_1^-$ with $f_1^\pm \geq 0$ and bounded and assume that $f_2, \ldots, f_n$ are positive and bounded. Note that $f_1^\pm \circ \xi$ is integrable by the boundedness of $f_1^\pm$. Therefore by linearity of expectation and the fact that we have proven the result for positive f_i

$$\begin{aligned}\mathbf{E}\left[f_1(\xi_1)f_2(\xi_2)\cdots f_n(\xi_n)\right] &= \mathbf{E}\left[f_1^+(\xi_1)f_2(\xi_2)\cdots f_n(\xi_n)\right] - \mathbf{E}\left[f_1^-(\xi_1)f_2(\xi_2)\cdots f_n(\xi_n)\right] \\ &= \mathbf{E}\left[f_1^+(\xi_1)\right]\mathbf{E}\left[f_2(\xi_2)\right]\cdots\mathbf{E}\left[f_n(\xi_n)\right] \\ &\quad - \mathbf{E}\left[f_1^-(\xi_1)\right]\mathbf{E}\left[f_2(\xi_2)\right]\cdots\mathbf{E}\left[f_n(\xi_n)\right] \\ &= \mathbf{E}\left[f_1(\xi_1)\right]\mathbf{E}\left[f_2(\xi_2)\right]\cdots\mathbf{E}\left[f_n(\xi_n)\right]\end{aligned}$$

Now perform induction on i to get the final result. □

Example 7.17. TODO: Find an example where this fails for integrable f. I'm pretty sure the crux is to find f that is integrable for which $f \circ \xi$ is not. In any case if one finds such a pair, then the result doesn't really even make sense since not all of the expectations are defined.

Corollary 7.18. *Suppose f, g are independent integrable random variables then fg is integrable and $\boldsymbol{E}[fg] = \boldsymbol{E}[f]\,\boldsymbol{E}[g]$.*

Proof. By Lemma 7.16, independence of f, g and positivity and measurability of $|x|$, we see that

$$\mathbf{E}\left[|fg|\right] = \mathbf{E}\left[|f|\cdot|g|\right] = \mathbf{E}\left[|f|\right]\mathbf{E}\left[|g|\right] < \infty$$

showing integrability of fg.

This argument also shows that $\mathbf{E}[fg] = \mathbf{E}[f]\,\mathbf{E}[g]$ for positive f, g. To extend to to integrable f, g write $f = f_+ - f_-$ and $g = g + -g_-$ and use linearity of expectation

$$\begin{aligned}\mathbf{E}[fg] &= \mathbf{E}[f_+g_+] - \mathbf{E}[f_+g_-] - \mathbf{E}[f_-g_-] + \mathbf{E}[f_-g_-] \\ &= \mathbf{E}[f_+]\,\mathbf{E}[g_+] - \mathbf{E}[f_+]\,\mathbf{E}[g_-] - \mathbf{E}[f_-]\,\mathbf{E}[g_-] + \mathbf{E}[f_-]\,\mathbf{E}[g_-] \\ &= \left(\mathbf{E}[f_+] - \mathbf{E}[f_-]\right)\left(\mathbf{E}[g_+] - \mathbf{E}[g_-]\right) \\ &= \mathbf{E}[f]\,\mathbf{E}[g]\end{aligned}$$

□

Example 7.19. This is an example of random variables ξ and η such that $\mathbf{E}[\xi\eta] = \mathbf{E}[\xi]\cdot\mathbf{E}[\eta]$ (are *uncorrelated*) but ξ and η are not independent.

Consider the sample space $\Omega = \{1, 2, 3\}$ with uniform distribution. A random variable $\xi : \Omega \to \mathbb{R}$ is just a vector in $\mathbb{R}^3$. Let $\xi = (1, -1, 0)$ and let $\eta = (-1, -1, 2)$. Note that $\mathbf{E}[\xi] = \mathbf{E}[\eta] = \mathbf{E}[\xi\eta] = 0$ and therefore ξ and η are uncorrelated. On the other hand ξ and η are not independent; for example

$$0 = \mathbf{P}\{\xi = 1 \wedge \eta = 2\} \neq \mathbf{P}\{\xi = 1\}\mathbf{P}\{\eta = 2\} = \frac{1}{9}$$

Definition 7.20. Given a sequence of events A_n the event that A_n occurs *infinitely often* is the set $\bigcap_{n=1}^\infty \bigcup_{k=n}^\infty A_k = \limsup_{n\to\infty} A_n$. The probability that A_n occurs infinitely often is often written $\mathbf{P}\{A_n \text{ i.o.}\}$.

Theorem 7.21. *[Borel Cantelli Theorem]Let $(\Omega, \mathcal{A}, P)$ be a probability space and let $A_1, A_2, \cdots \in \mathcal{A}$.*

(i) *If $\sum_{i=1}^\infty \boldsymbol{P}\{A_i\} < \infty$ then $\boldsymbol{P}\{A_i \text{ i.o.}\} = 0$.*

(ii) *If the A_i are independent and $\boldsymbol{P}\{A_i \text{ i.o.}\} = 0$, then we have $\sum_{i=1}^{\infty} \boldsymbol{P}\{A_i\} < \infty$. More precisely, if $\sum_{i=1}^{\infty} \boldsymbol{P}\{A_i\} = \infty$ then $\boldsymbol{P}\{A_i \text{ i.o.}\} = 1$.*

Proof. To prove (i) we observe that the convergence of $\sum_{i=1}^{\infty} \mathbf{P}\{A_i\}$ implies that the partial sums converge to zero, $\lim_{n\to\infty} \sum_{i=n}^{\infty} \mathbf{P}\{A_i\} = 0$. Now we apply a union bound (subadditivity of measure) and use continuity of measure to see that

$$\mathbf{P}\{A_n \text{ i.o.}\} = \lim_{n\to\infty} \mathbf{P}\{\bigcup_{k=n}^{\infty} A_k\} \leq \lim_{n\to\infty} \sum_{k=n}^{\infty} \mathbf{P}\{A_k\} = 0$$

To see (ii), first observe the simple calculation

$$\begin{aligned}
\mathbf{P}\{\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k\} &= \lim_{n\to\infty} \mathbf{P}\{\bigcup_{k=n}^{\infty} A_k\} && \text{by continuity of measure} \\
&= \lim_{n\to\infty} \lim_{m\to\infty} \mathbf{P}\{\bigcup_{k=n}^{m} A_k\} && \text{by continuity of measure} \\
&= \lim_{n\to\infty} \lim_{m\to\infty} \left(1 - \mathbf{P}\{\bigcap_{k=n}^{m} A_k^c\}\right) && \text{by DeMorgan's law} \\
&= \lim_{n\to\infty} \lim_{m\to\infty} \left(1 - \prod_{k=n}^{m} \mathbf{P}\{A_k^c\}\right) && \text{by independence} \\
&= 1 - \lim_{n\to\infty} \lim_{m\to\infty} \left(\prod_{k=n}^{m} (1 - \mathbf{P}\{A_k\})\right)
\end{aligned}$$

Now we recall the elementary bound $1 + x \leq e^x$ for $x \in \mathbb{R}$ from Lemma 5.1 and assume that $\sum_{n=1}^{\infty} \mathbf{P}\{A_n\} = \infty$. By the calculation above we have

$$\begin{aligned}
\mathbf{P}\{\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k\} &= 1 - \lim_{n\to\infty} \lim_{m\to\infty} \left(\prod_{k=n}^{m} (1 - \mathbf{P}\{A_k\})\right) \\
&\geq 1 - \lim_{n\to\infty} \lim_{m\to\infty} \left(\prod_{k=n}^{m} e^{-\mathbf{P}\{A_k\}}\right) \\
&= 1 - \lim_{n\to\infty} \lim_{m\to\infty} e^{-\sum_{k=n}^{m} \mathbf{P}\{A_k\}} \\
&= 1
\end{aligned}$$

But of course we know that $\mathbf{P}\{\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k\} \leq 1$ so in fact we have shown that $\mathbf{P}\{A_n \text{ i.o.}\} = 1$. □

Example 7.22. Here is a somewhat synthetic example that shows when A_n are dependent it is possible to have $\mathbf{P}\{A_n \text{ i.o.}\} = 0$ while $\sum_{n=1}^{\infty} \mathbf{P}\{A_n\} = \infty$. Take $([0,1], \mathcal{B}([0,1]), \lambda$ as the measure space. Take the intervals $[0, \frac{1}{n}]$ in a sequence such that $[0, \frac{1}{n}]$ occurs n times (e.g. $[0,1], [0,\frac{1}{2}], [0,\frac{1}{2}], [0,\frac{1}{3}], [0,\frac{1}{3}], [0,\frac{1}{3}], \dots$). Clearly $\{A_n \text{ i.o.}\} = \{0\}$. On the other hand it is clear that $\sum_{n=1}^{\infty} \mathbf{P}\{A_n\} = \infty$.

Example 7.23. This is a more probabilistic example. Consider a game in which there is a n-sided die for each $n = 2, 3, \dots$. In the n^{th} round of the game, one rolls the n-sided die. If one gets a 1 then one stops the game else one continues to play. Let A_n be the event that the player is still alive at round n. It is clear

that player has a probability of $\frac{1}{2}\cdots\frac{n-1}{n}=\frac{1}{n}$ of being alive at round n. It is also clear that the probability the player never loses is bounded by $\frac{1}{n}$ for all n hence is 0. The probability the player never loses is the same as $\mathbf{P}\{A_n \text{ i.o.}\}$ on the other hand, $\sum_{n=1}^{\infty}\mathbf{P}\{A_n\}=\sum_{n=1}^{\infty}\frac{1}{n}=\infty$.

The Borel Cantelli Theorem tells us that $\mathbf{P}\{A_n \text{ i.o.}\}$ can only take the values 0 and 1 when the A_n are independent events (and in fact gives us a test for determining which alternative holds). The 0/1 dichotomy is a general feature of sequences of independent events and describing the nature this dichotomy motivates the following definitions.

Definition 7.24. Let $\mathcal{A}_n$ be a sequence of σ-algebras on a space Ω. The *tail σ-algebra* $\mathcal{T}_\infty$ is defined to be

$$\mathcal{T}_\infty=\bigcap_{n=1}^{\infty}\sigma\left(\bigcup_{k=n}^{\infty}\mathcal{A}_k\right)$$

Theorem 7.25 (Kolmogorov's $0-1$ Law)**.** *Let $\mathcal{A}_n$ be a sequence of independent σ-algebras on a probability space $(\Omega,\mathcal{A},P)$ such that $\mathcal{A}_n\subset\mathcal{A}$ for all $n>0$. Then for every $T\in\mathcal{T}_\infty$ we have $\boldsymbol{P}\{T\}=0$ or $\boldsymbol{P}\{T\}=1$.*

Proof. Let $\mathcal{T}_n=\sigma\left(\bigcup_{k=n}^{\infty}\mathcal{A}_k\right)$ and $\mathcal{S}_n=\sigma\left(\bigcup_{k=1}^{n-1}\mathcal{A}_k\right)$. Then by Lemma 7.13 we see that $\mathcal{T}_n$ and $\mathcal{S}_n$ are independent. Therefore for $A\in\mathcal{T}_n$ and $B\in\mathcal{S}_n$ we have $\mathbf{P}\{A\cap B\}=\mathbf{P}\{A\}\mathbf{P}\{B\}$.

Now pick $A\in\mathcal{T}_\infty$, then by the above observation we have $\mathbf{P}\{A\cap B\}=\mathbf{P}\{A\}\mathbf{P}\{B\}$ for $B\in\bigcup_{n=1}^{\infty}\mathcal{S}_n$. Since $\mathcal{S}_1\subset\mathcal{S}_1\subset\cdots$, we can easily see that $\bigcup_{n=1}^{\infty}\mathcal{S}_n$ is a π-system. Given $B_1,B_2\in\bigcup_{n=1}^{\infty}\mathcal{S}_n$ there exist n_1,n_2 such that $B_i\in\mathcal{S}_{n_i}$ for $i=1,2$. Then define $n=\max(n_1,n_2)$ and $B_i\in\mathcal{S}_n$ for $i=1,2$ and therefore $B_1\cap B_2\in\mathcal{S}_n\subset\bigcup_{n=1}^{\infty}\mathcal{S}_n$. Applying Lemma 7.12 we conclude that $\mathcal{T}_\infty$ and $\sigma(\bigcup_{n=1}^{\infty}\mathcal{S}_n)$ are independent. Note that for every $n>0$, $\mathcal{T}_n\subset\sigma(\bigcup_{n=1}^{\infty}\mathcal{S}_n)$ hence the same is true of their intersection $\mathcal{T}_\infty$. We may conclude that for any $A\in\mathcal{T}_\infty$ we have

$$\mathbf{P}\{A\}=\mathbf{P}\{A\cap A\}=\mathbf{P}\{A\}\mathbf{P}\{A\}$$

which shows that $\mathbf{P}\{A\}=0$ or $\mathbf{P}\{A\}=1$. □

Tail algebras arise naturally in various limiting processes involving random variables. In the case in which the random variables are independent, the limits have various kinds of almost sure properties that can be derived from Kolmogorov's $0-1$ Law. Here are a few examples.

Corollary 7.26. *Let (S,d) be a complete metric space and let ξ_n be a sequence of independent random elements in S. Then either ξ_n converges almost surely or diverges almost surely.*

Proof. Let $\mathcal{T}_n=\sigma\left(\cup_{k\geq n}\sigma(\xi_k)\right)$ and let $\mathcal{T}=\cap_{n=1}^{\infty}\mathcal{T}_n$ be the tail σ-algebra. By Kolmogorov's $0-1$ Law it suffices to show that the event that ξ_n converges is $\mathcal{T}$-measurable. Since S is complete, we know that ξ_n converges if and only if for every $\epsilon>0$ there exists $N>0$ such that $d(\xi_m,\xi_n)<\epsilon$. With that in mind, for every $m>0$, $n>0$ and $\epsilon>0$ define

$$A_{n,m,\epsilon}=\{d(\xi_m,\xi_n)<\epsilon\}$$

which is $\sigma\left(\sigma(\xi_m) \cup \sigma(\xi_n)\right)$-measurable.

To prove convergence it suffices to demonstrate it for any sequence of $\epsilon_k \to 0$. So in particular if we choose $\epsilon_k = \frac{1}{k}$ we see that the event that ξ_n converges is

$$\bigcap_{k=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{m,n \geq N} A_{m,n,\frac{1}{k}}$$

Note that each $\bigcap_{m,n\geq N} A_{m,n,\frac{1}{k}}$ is $\mathcal{T}_N$-measurable and $A_{N+1} \subset A_N$ hence $\bigcup_{N=1}^{\infty} \bigcap_{m,n\geq N} A_{m,n,\frac{1}{k}}$ is $\mathcal{T}$-measurable. Taking the countable union of $\mathcal{T}$-measurable sets we see the event that ξ_n converges is $\mathcal{T}$-measurable. □

Corollary 7.27. *Let ξ_n be a sequence of independent random variables. Then $\limsup_{n\to\infty} \xi_n$ and $\liminf_{n\to\infty} \xi_n$ are almost surely constant.*

Proof. Because $\liminf_n \xi_n = -\limsup_n -\xi_n$ if suffices to show the result for $\limsup_n \xi_n$. Let $\mathcal{T}$ be the tail σ-algebra of $\sigma(\xi_n)$ and let $\mathcal{T}_n = \sigma\left(\cup_{k\geq n}\sigma(\xi_k)\right)$. By Kolmogorov's 0-1 Law, it suffices to show that $\limsup_{n\to\infty} \xi_n$ is $\mathcal{T}$-measurable.

By definition, $\limsup_{n\to\infty} \xi_n = \lim_{n\to\infty} \sup_{k\geq n} \xi_k$. The term $\sup_{k\geq n} \xi_k$ is $\mathcal{T}_n$-measurable by 3.14 and when taking the limit of the sequence we can ignore any finite prefix of the sequence. Therefore we can express the limit as a limit of $\mathcal{T}_n$-measurable functions for $n > 0$ arbitrary. This shows that $\limsup_{n\to\infty} \xi_n$ is $\mathcal{T}_n$-measurable for all $n > 0$ hence $\mathcal{T}$-measurable. □

Corollary 7.28. *Let ξ_n be a sequence of independent random variables. Then $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{\infty} \xi_k$ almost surely diverges or almost sure converges. If it converges then the limit is almost surely constant.*

Proof. Note that $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \xi_k = \lim_{n\to\infty} \frac{1}{n} \sum_{k=m}^{n} \xi_k$ for any $m > 0$. Pick such an $m > 0$ and note that every finite partial sum $\frac{1}{n} \sum_{k=m}^{n} \xi_k$ is $\mathcal{T}_m$-measurable hence so is the limit $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \xi_k$. Since $m > 0$ was arbitrary we know that $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \xi_k$ is $\mathcal{T}$-measurable. □

The Borel Cantelli Theorem is a very useful technique in demonstrating the almost sure convergence of sequences of random variables. The following simple version of the Strong Law of Large Numbers illustrates the technique with a minimum of distractions.

Lemma 7.29. *Let $\xi, \xi_1, \xi_2, \ldots$ be independent identically distributed random variables with $\boldsymbol{E}\left[\xi^4\right] < \infty$, then $\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \xi_k = \boldsymbol{E}[\xi]$ a.s.*

Proof. First note that it suffices to show the result when $\mathbf{E}[\xi] = 0$ since we can just compute

$$0 = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} (\xi_k - \mathbf{E}[\xi]) = \lim_{n\to\infty} (\frac{1}{n} \sum_{k=1}^{n} \xi_k) - \mathbf{E}[\xi]$$

Furthermore by Corollary 6.12 the finite 4^{th} moment of ξ implies finiteness of the first four moments, hence $\mathbf{E}\left[(\xi - \mathbf{E}[\xi])^4\right] < \infty$.

Now assuming that ξ_k have mean zero we fix $\epsilon > 0$ and apply Markov bounding to see

$$\begin{aligned}
\mathbf{P}\{\left|\sum_{k=1}^{n}\xi_k\right| > n\epsilon\} &= \mathbf{P}\{\left(\sum_{k=1}^{n}\xi_k\right)^4 > n^4\epsilon^4\} \\
&\leq \frac{\mathbf{E}\left[\left(\sum_{k=1}^{n}\xi_k\right)^4\right]}{n^4\epsilon^4} && \text{by Markov's inequality} \\
&= \frac{\sum_{k=1}^{n}\mathbf{E}\left[\xi_k^4\right] + 6\sum_{k=1}^{n}\sum_{l=k+1}^{n}\mathbf{E}\left[\xi_k^2\xi_l^2\right]}{n^4\epsilon^4} && \text{by independence and zero mean} \\
&= \frac{\sum_{k=1}^{n}\mathbf{E}\left[\xi_k^4\right] + 6\sum_{k=1}^{n}\sum_{l=k+1}^{n}\sqrt{\mathbf{E}\left[\xi_k^4\right]\mathbf{E}\left[\xi_l^4\right]}}{n^4\epsilon^4} && \text{by Cauchy Schwartz} \\
&= \frac{\mathbf{E}\left[\xi^4\right](n+3(n^2-n))}{n^4\epsilon^4} \leq \frac{3\mathbf{E}\left[\xi^4\right]}{n^2\epsilon^4}
\end{aligned}$$

And therefore $\sum_{n=1}^{\infty}\mathbf{P}\{|\sum_{k=1}^{n}\xi_k| > n\epsilon\} < \infty$. Now we can apply Borel Cantelli to see that $\mathbf{P}\{\frac{1}{n}|\sum_{k=1}^{n}\xi_k| > \epsilon \text{ i.o.}\} = 0$.

By the above argument, for every $m \in \mathbb{N}$ we get an event A_m with $\mathbf{P}\{A_m\} = 0$ such that for every $\omega \notin A_m$ there is $N_{\omega,m}$such that $\frac{1}{n}|\sum_{k=1}^{n}\xi_k(\omega)| \leq \frac{1}{m}$ for $n > N_{\omega,m}$. Let $A = \cup m = 1^{\infty}A_m$ and note that by countable subadditivity $\mathbf{P}\{A\} = 0$. Furthermore, for every $\epsilon > 0$, $\omega \in A$ we pick $m > \frac{1}{\epsilon}$ and then for $n > N_{\omega,m}$ we have $\frac{1}{n}|\sum_{k=1}^{n}\xi_k(\omega)| \leq \frac{1}{m} < \epsilon$ for $n > N_{\omega,m}$ giving the result. □

The proof above demonstrates a general pattern in applications of Borel Cantelli in which one applies it a countably infinite number of times and still derive an almost sure result. We'll prove more refined versions of the Strong Law of Large Numbers later and those will also use Borel Cantelli but with more complications.

It will prove to be important to be able to construct random variables with prescribed distributions. In particular, we will soon need to be able to construct independent random variables with prescribed distributions. The standard way of constructing them is to use product spaces, however we have only developed product spaces of finitely many factors. Rather than developing the full fledged theory of infinitary products, we provide a mechanism which suffices for the construction of countably many random variables with prescribed distributions; in fact we show that it is possible to do so on the probability space $([0,1], \mathcal{B}([0,1]), \lambda)$. First proceed by noticing that there is ready source of independence waiting for us to harvest. Given $x \in [0,1]$ we can take the unique binary expansion $x = 0.\xi_1\xi_2\cdots$ which has the property that $\sum_{n=1}^{\infty}\xi_n = \infty$ (here we are resolving the ambiguity between expansions that have a tail of 1's and those with a tail of 0's).

Lemma 7.30. *Let $\xi_n : [0,1] \to [0,1]$ be defined by taking the n^{th} digit of the binary expansion of $x \in [0,1]$. Then ξ_n is a measurable function. Let $\vartheta : [0,1] \to [0,1]$, then ϑ has a uniform distribution if and only if $\xi_n \circ \vartheta$ comprise an independent sequence of Bernoulli random variables with probability $\frac{1}{2}$.*

Proof. To see the measurability of ξ_n we first define the *floor function* to be $\lfloor x \rfloor = \sup\{n \in \mathbb{Z} \mid n \leq x\}$. Then define

$$\xi(x) = \begin{cases} 0 & \text{if } x - \lfloor x \rfloor \in [0, \frac{1}{2}) \\ 1 & \text{if } x - \lfloor x \rfloor \in [\frac{1}{2}, 1) \end{cases}$$

It is clear that ξ is a measurable function since $\xi^{-1}(0) = \cup_n [n, n + \frac{1}{2})$ and $\xi^{-1}(1) = \cup_n [n + \frac{1}{2}, n + 1)$. Now define

$$\xi_n(x) = \xi(2^{n-1}x) \qquad \text{for } n \in \mathbb{N} \text{ and } x \in \mathbb{R}$$

and notice that ξ_n give the binary expansion of $x \in \mathbb{R}$. By measurability of ξ we see that ξ_n are also measurable.

Now suppose that ϑ is a $U(0,1)$ random variable on $[0,1]$ and consider $\xi_n \circ \vartheta$. For every $(k_1, \ldots, k_n) \in \{0,1\}^n$, let $q = \sum_{j=1}^n \frac{k_j}{2^j}$ we clearly have

$$\mathbf{P}\{\cap_{j \leq n} \{\xi_j(\vartheta(x)) = k_j\}\} = \mathbf{P}\{\vartheta(x) \in [q, q + \frac{1}{2^n})\}\} = \frac{1}{2^n}$$

and summing over $(k_1, \ldots, k_{n-1})$ we see

$$\mathbf{P}\{\xi_n(\vartheta(x)) = k_n\} = \sum_{(k_1,\ldots,k_{n-1}) \in \{0,1\}^{n-1}} \mathbf{P}\{\cap_{j \leq n} \{\xi_j(\vartheta(x)) = k_j\}\} = \frac{1}{2}$$

which shows that each $\xi_n \circ \vartheta$ is a Bernoulli random variable with probability $\frac{1}{2}$.

In a similar vein, given $n_1, \ldots, n_m$ and $k_{n_j} \in \{0,1\}$, let $n = \sup(n_1, \ldots, n_m)$ for $j = 1, \ldots, m$ and $A_n = \{(l_1, \ldots, l_n) \mid l_{n_j} = k_{n_j} \text{ for } j = 1, \ldots, m\}$ and we have

$$\begin{aligned} \mathbf{P}\{\cap_{j=1}^m \{\xi_{n_j}(\vartheta(x)) = k_{n_j}\}\} &= \sum_{(k_1,\ldots,k_n) \in A_n} \mathbf{P}\{\cap_{j \leq n} \{\xi_j(\vartheta(x)) = k_j\}\} \\ &= 2^{n-m} \frac{1}{2^n} = \frac{1}{2^m} \end{aligned}$$

which shows that $\xi_{n_j} \circ \vartheta$ are independent.

Next, suppose that we know $\xi_n \circ \vartheta$ is an independent Bernoulli sequence with probability $\frac{1}{2}$. Let $\tilde{\vartheta}$ be a $U(0,1)$ random variable (e.g. $\tilde{\vartheta}(x) = x$) and then we know from the first part of the Lemma that $\xi_n \circ \tilde{\vartheta}$ is also a Bernoulli sequence with probability $\frac{1}{2}$.

Because of the independence of each the sequences and the fact that the elementwise the two sequences have the same distribution we know that the distribution of the sums is just the convolution of the distributions of the terms in the sequence, hence $\sum \xi_n \circ \vartheta \stackrel{d}{=} \sum \xi_n \circ \tilde{\vartheta}$. Thus we have shown that $\sum \xi_n \circ \vartheta$ is also $U(0,1)$. $\square$

Lemma 7.31. *There exist measurable functions $f_1, f_2, \ldots$ on $[0,1]$ such that whenever ϑ is a $U(0,1)$ random variable, the sequence $f_n \circ \vartheta$ is a family of independent $U(0,1)$ random variables.*

Proof. Let $\xi_n \circ \vartheta$ denote the binary expansion of ϑ from Lemma 7.30. By the result of that Lemma, we know that the $\xi_n \circ \vartheta$ are an i.i.d. sequence of Bernoulli random variables with probability $\frac{1}{2}$. Now choose any bijection between $\mathbb{N}$ and $\mathbb{N}^2$ (e.g. the diagonal mapping). With this relabeling of of the constructed family we now have a sequence $\xi_{n,m} \circ \vartheta$ of i.i.d. Bernoulli random variables. Define $f_n(x) = \sum_{m=1}^\infty \frac{\xi_{n,m}(x)}{2^m}$ and apply Lemma 7.30 a second time to see that each $f_n \circ \vartheta$ is a $U(0,1)$ random

variable. Furthermore, $f_n \circ \vartheta$ is $\sigma(\cup_m \sigma(\xi_{n,m} \circ \vartheta))$-measurable so by Lemma 7.13 we see that the $f_n \circ \vartheta$ are independent. □

Theorem 7.32. *For any probability measures $\mu_1, \mu_2, \ldots$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ there exist independent random variables $f_1, f_2, \ldots$ on $([0,1], \mathcal{B}([0,1]), \lambda)$ such that $\mathcal{L}(f_n) = \mu_n$.*

Proof. Define $\vartheta(x) = x$ which is clearly a $U(0,1)$-random variable on $[0,1]$ and use Lemma 7.31 to construct ϑ_n, a sequence of independent $U(0,1)$ random variables. Let F_n be the distribution function of the probability measure μ_n and let $G_n(y) = \sup\{x \in \mathbb{R} \mid F(x) \geq y\}$ be the generalized inverse of F_n. By the proof of Theorem 3.95, we know that $\mathcal{L}(G_n \circ \vartheta_n) = \mu_n$ and by Lemma 7.14 we know that $G_n \circ \vartheta_n$ are still independent. □

8. Convergence of Random Variables

TODO: a.s. convergence, convergence in probability and weak convergence (convergence in distribution), tightness of distribution.

Definition 8.1. Let (S, d) be a σ-compact metric space with the Borel σ-algebra and let ξ_n be a sequence of random elements in S. Let ξ be a random element in S.

(i) ξ_n *converges almost surely* to ξ if for almost every $\omega \in \Omega$, $\xi_n(\omega)$ converges to $\xi(\omega)$ in S. We write $\xi_n \xrightarrow{a.s.} \xi$ to denote almost sure convergence.

(ii) ξ_n *converges in probability* to ξ if for any $\epsilon > 0$ we have

$$\lim_{n\to\infty} \mathbf{P}\{\{\omega : d(\xi_n(\omega), \xi(\omega)) > \epsilon\}\} = 0$$

We write $\xi_n \xrightarrow{P} \xi$ to denote convergence in probability.

(iii) ξ_n *converges in distribution* to ξ if, for every bounded continuous function $f : S \to \mathbb{R}$, one has

$$\lim_{n\to\infty} \mathbf{E}\left[f(\xi_n)\right] = \mathbf{E}\left[f(\xi)\right].$$

We write $\xi_n \xrightarrow{d} \xi$ to denote convergence in distribution.

(iv) ξ_n has a *tight sequence of distributions* if, for every $\epsilon > 0$, there exists a compact subset K of S such that $\mathbf{P}\{\xi_n \in K\} \geq 1 - \epsilon$ for sufficiently large n.

TODO: Note that convergence in distribution is really a property of the distribution of the random variables and not the random variables themselves.

For the case of random variables there is another strong form of convergence that is quite useful.

Definition 8.2. If $\xi, \xi_1, \xi_2, \ldots$ are random variables then ξ_n *converges in L^p* to ξ if $\lim_{n\to\infty} \mathbf{E}\left[|\xi_n - \xi|^p\right] = 0$. We write $\xi_n \xrightarrow{L^p} \xi$ to denote convergence in L^p. We may also call convergence in L^p *convergence in p^{th} mean.*

TODO: Motivation for concept of almost sure convergence via Law of Large Numbers. Think of modeling coin tossing using random variables. The n^{th} coin flip is represented as a Bernoulli random variable ξ_n where $\xi_n(\omega) = 1$ means that the coin lands with heads. The *empirical probability* of heads is in n trials is $S_n = \frac{1}{n}\sum_{k=1}^{n} \xi_k$. Now our intuition is that S_n converges to $1/2$ in some appropriate

sense. Now the simple minded notion of pointwise convergence that we used in the development of measure theory (e.g. in all of the limit theorems) is too strong for this scenario. Clearly, it is theoretically possible for a person to toss a coin an infinite number of times and get only heads. It is possible by extremely improbable; so improbable in fact that its probability is zero.

TODO: Motivation for concept of convergence in probability. Motivation for convergence in mean is pretty clear.

There is also some useful technical intuition around how one might prove that sequences converge almost surely. The idea is implicit in the definitions but is useful to take the time to call it out and make it perfectly explicit; we will see it time and again. If one looks at the contrapositive of almost sure convergence, it means that there is probability zero that a sequence of random elements does not converge. The property of not converging is that there exists an $\epsilon > 0$ such that for all $N > 0$, $d(\xi, \xi_n) \geq \epsilon$ for all $n > N$. Converting the logic in set operations, let $A_{N,\epsilon}$ be the event that $d(\xi, \xi_n) \geq \epsilon$ for all $n > N$. Convergence fails precisely on the event $\cup_{\epsilon>0} \cap_{N=1}^{\infty} A_{N,\epsilon}$, so almost sure convergence means that $\mathbf{P}\{\cup_{\epsilon>0} \cap_{N=1}^{\infty} A_{N,\epsilon}\} = 0$. TODO: Note that one can restrict ϵ to a countable subset of $\mathbb{R}$ (e.g. $\mathbb{Q}$ or $\frac{1}{n}$). Note that the same reasoning applies when handling almost sure Cauchy sequences as well.

Almost sure convergence is such a simple notion that it seems there may be nothing worth explaining about it. However the following result ties in the definition of almost sure convergence with the idea of events happening infinitely often that we encountered when discussing indpedendence. The connection proves to be quite powerful and we'll soon see that it make the Borel-Cantelli Lemma a useful tool for proving almost sure convergence.

Lemma 8.3. *Let $\xi, \xi_1, \xi_2, \ldots$ be random elements in the metric space (S, d), then $\xi_n \overset{a.s.}{\to} \xi$ if and only if for every $\epsilon > 0$, $\boldsymbol{P}\{d(\xi_n, \xi) \geq \epsilon \text{ i.o.}\} = 0$ if and only if for every $\epsilon > 0$, $\lim_{n\to\infty} \boldsymbol{P}\{\sup_{m\geq n} d(\xi_m, \xi) > \epsilon\} = 0$.*

Proof. By definition if $\xi_n \overset{a.s.}{\to} \xi$ there is a set $A \subset \Omega$ such that $\mathbf{P}\{A\} = 1$ and for all $\epsilon > 0$ and $\omega \in A$ there exists $N_{\epsilon,\omega} \geq 0$ such that $d(\xi_n(\omega), \xi(\omega)) < \epsilon$ when $n \geq N_{\epsilon,\omega}$. In particular for $\omega \in A$, $d(\xi_n, \xi) \geq \epsilon$ finitely often. Therefore $\{d(\xi_n, \xi) \geq \epsilon \text{ i.o.}\} \subset A^c$ and $\mathbf{P}\{d(\xi_n, \xi) \geq \epsilon \text{ i.o.}\} \leq \mathbf{P}\{A\} = 0$.

In the opposite direction, let $A_\epsilon = \{d(\xi_n, \xi) \geq \epsilon \text{ i.o.}\}$ and by assumption $\mathbf{P}\{A_\epsilon\} = 0$. The event that ξ_n does not converge to ξ is precisely $A = \cup_{\epsilon>0} A_\epsilon$ and we might think we are done. Unfortunately $\cup_{\epsilon>0} A_\epsilon$ is an uncountable union and we can't conclude that $\mathbf{P}\{A\} = 0$. We resolve this by noting that in fact $A = \cup_n A_{\frac{1}{n}}$ which is a countable union of sets of measure zero; hence has measure zero.

TODO: Fix inconsistency in use of $\geq$ and $>$.

To see the second equivalence, just unfold the definition of events happening infinitely often and use continuity of measure

$$\begin{aligned}
\mathbf{P}\{d(\xi_n, \xi) > \epsilon \text{ i.o.}\} &= \mathbf{P}\{\cap_{n=1}^{\infty} \cup_{m=n}^{\infty} \{d(\xi_m, \xi) > \epsilon\}\} \\
&= \lim_{n\to\infty} \mathbf{P}\{\cup_{m=n}^{\infty} \{d(\xi_m, \xi) > \epsilon\}\} \\
&= \lim_{n\to\infty} \mathbf{P}\{\sup_{m\geq n} d(\xi_m, \xi) > \epsilon\}
\end{aligned}$$

□

Lemma 8.4. *Let $\xi, \xi_1, \xi_2, \ldots$ be random elements in the metric space (S, d). If $\xi_n \overset{a.s.}{\to} \xi$ then $\xi_n \overset{P}{\to} \xi$.*

Proof. By Lemma 8.3 and continuity of measure, if $\xi_n \overset{a.s.}{\to} \xi$ then we know that for each $\epsilon > 0$,

$$0 = \mathbf{P}\{d(\xi_n, \xi) \geq \epsilon \text{ i.o.}\} = \lim_{n\to\infty} \mathbf{P}\{\cup_{k\geq n} d(\xi_k, \xi) \geq \epsilon\}$$

Now clearly we have $\mathbf{P}\{d(\xi_n, \xi) \geq \epsilon\} \leq \mathbf{P}\{\cup_{k\geq n} d(\xi_k, \xi) \geq \epsilon\}$ so convergence in probability follows.

Here is an alternative approach that currently has a hole in the argument. Is it worth patching the hole? Suppose there exists $\epsilon, \delta > 0$ for which there is a subsequence $n_j \to \infty$ and $\mathbf{P}\{d(\xi_{n_j}, \xi) > \epsilon\} \geq \delta > 0$. We claim that $\mathbf{P}\{\cap_j \{d(\xi_{n_j}, \xi) > \epsilon\}\} > 0$ (is this really true?). Note $\cap_j \{d(\xi_{n_j}, \xi) > \epsilon\} \subset \{\omega \mid \xi_{n_j}(\omega) \text{ does not converge to } \xi(\omega)\}$ hence ξ_n does not converge on a set of positive measure. □

Example 8.5. [Sequence converging in probability but not almost surely]Consider the $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ with Lebesgue measure. For a sequence of intervals $I_n \subset \mathbb{R}$ observe that $\mathbf{1}_{I_n} \overset{P}{\to} 0$ if and only if $|I_n| \to 0$. For every $n > 0$ consider the events $A_{n,j} = [\frac{j-1}{n}, \frac{j}{n}]$ for $j = 1, \ldots, n$. Now consider the sequence of random variables obtained by taking the lexicographic order of pairs (n, j) for $n > 0$ and $j = 1, \ldots, n$ and the indicator functions $\mathbf{1}_{A_{n,j}}$; call the resulting sequence f_m. Note that $f_m \overset{P}{\to} 0$ by the above discussion. On the the other hand, the sequence does not converge pointwise anywhere on $[0, 1]$ because for every $x \in [0, 1]$, we can see $\limsup_{m\to\infty} f_m(x) = 1$ but $\liminf_{m\to\infty} f_m(x) = 0$.

Lemma 8.6. *Let $\xi, \xi_1, \xi_2, \ldots$ be random variables, if $\xi_n \overset{L^p}{\to} \xi$, then $\xi_n \overset{P}{\to} \xi$.*

Proof. This is a simple application of Markov's Inequality (Lemma 14.1)

$$\mathbf{P}\{|\xi_n - \xi| > \epsilon\} = \mathbf{P}\{|\xi_n - \xi|^p > \epsilon^p\} \leq \frac{\mathbf{E}\left[|\xi_n - \xi|^p\right]}{\epsilon^p}$$

but the right hand side converges to 0 by assumption. □

Example 8.7 (Sequence converging in probability but in mean)**.** To see that a sequence of random elements can converge in probability but not in mean we can modify Example 8.5. Using the notation from that example, define the random variables $n\mathbf{1}_{A_{n,j}}$ and order them lexicographically into the sequence f_m. Note that point behind rescaling is that we have arrange for $\mathbf{E}\left[n\mathbf{1}_{A_{n,j}}\right] = 1$. The argument that the $f_m \overset{P}{\to} 0$ follows essentially unchanged; convergence in probability is insensitive the rescaling of the random variables. On the other hand, it is clear that $\mathbf{E}[f_m] = 1$ for all $m > 0$ and therefore f_m do not converge in mean to 0.

There are few useful characterization of convergence in probability that are important tools to have. The first provides a characterization of convergence in probability as a convergence of expectations. Because of the previous example, we know that convergence in probability does not control the behavior of random elements on arbitrarily small sets hence it alone is not capable of controlling the values of expectations. Adding in such control as an explicit extra condition we can tie the concepts together.

Lemma 8.8. *Let $\xi, \xi_1, \xi_2, \ldots$ be random elements in the metric space (S, d). $\xi_n \xrightarrow{P} \xi$ if and only if $\lim_{n\to\infty} \boldsymbol{E}[d(\xi_n, \xi) \wedge 1] = 0$.*

Proof. Suppose that $\xi_n \xrightarrow{P} \xi$. We pick $\epsilon > 0$ and $N > 0$ such that $\mathbf{P}\{d(\xi_n, \xi) > \epsilon\} < \epsilon$ for $n > N$. Now write

$$\begin{aligned} d(\xi_n, \xi) \wedge 1 &= d(\xi_n, \xi) \wedge 1 \cdot \mathbf{1}_{d(\xi_n,\xi)>\epsilon} + d(\xi_n, \xi) \wedge 1 \cdot \mathbf{1}_{d(\xi_n,\xi)\leq\epsilon} \\ &\leq \mathbf{1}_{d(\xi_n,\xi)>\epsilon} + \epsilon \end{aligned}$$

Taking expectations we see

$$\mathbf{E}\left[d(\xi_n, \xi) \wedge 1\right] \leq \mathbf{P}\{d(\xi_n, \xi) > \epsilon\} + \epsilon \leq 2\epsilon \qquad \text{for } n > N.$$

Suppose that $\lim_{n\to\infty} \mathbf{E}\left[d(\xi_n, \xi) \wedge 1\right] = 0$. First note that in proving convergence in probability, it suffices to consider $\epsilon < 1$ since for any $\epsilon < \epsilon'$ we have $\mathbf{P}\{d(\xi_n, \xi) > \epsilon')\} \leq \mathbf{P}\{d(\xi_n, \xi) > \epsilon)\}$. So pick $0 < \epsilon < 1$ and use Markov's Inequality (Lemma 14.1) to see

$$\lim_{n\to\infty} \mathbf{P}\{d(\xi_n, \xi) > \epsilon)\} = \lim_{n\to\infty} \mathbf{P}\{d(\xi_n, \xi) \wedge 1 > \epsilon)\} \leq \lim_{n\to\infty} \frac{\mathbf{E}\left[d(\xi_n, \xi) \wedge 1\right]}{\epsilon} = 0$$

□

As an example of how this Lemma is can be used, note that it provides a quick alternative proof to Lemma 8.4: If $\xi_n \xrightarrow{a.s.} \xi$ then $d(\xi_n, \xi) \wedge 1 \xrightarrow{a.s.} 0$ and Dominated Convergence implies $\mathbf{E}\left[d(\xi_n, \xi) \wedge 1\right] \to 0$.

The relationship between almost sure convergence and convergence in probability can be made even tighter than Lemma 8.4.

Lemma 8.9. *Suppose (S, d) is a metric space and let $\xi, \xi_1, \xi_2, \ldots$ be random elements in S. Then $\xi_n \xrightarrow{P} \xi$ if and only for every subsequence $N' \subset \mathbb{N}$ there is a further subsequence $N'' \subset N'$ such that $\lim_{n\in N''} \xi_n = \xi$ a.s.*

Proof. Let $\xi_n \xrightarrow{P} \xi$. By Lemma 8.8, we know that $\lim_{n\to\infty} \mathbf{E}\left[d(\xi_n, \xi) \wedge 1\right] = 0$. Thus we can pick $n_k > 0$ such that $\mathbf{E}\left[d(\xi_{n_k}, \xi) \wedge 1\right] < \frac{1}{2^k}$. Therefore

$$\sum_{k=1}^{\infty} \mathbf{E}\left[d(\xi_{n_k}, \xi) \wedge 1\right] = \mathbf{E}\left[\sum_{k=1}^{\infty} d(\xi_{n_k}, \xi) \wedge 1\right] < \infty$$

where we have used Tonelli's Theorem 3.41. Finiteness of the second integral implies $\sum_{k=1}^{\infty} d(\xi_{n_k}, \xi) \wedge 1 < \infty$ almost surely and convergence of the sum implies that the terms $d(\xi_{n_k}, \xi) \wedge 1 \xrightarrow{a.s.} 0$ which in turn implies $d(\xi_{n_k}, \xi) \xrightarrow{a.s.} 0$

Here is an alternative proof of the first implication using Borel-Cantelli. Pick a sequence $n_1, n_2, \ldots$ such that $\mathbf{P}\{d(\xi_{n_k}, \xi) > \frac{1}{k}\} < \frac{1}{2^k}$. Then the sets $A_k = \{\omega \mid d(\xi_{n_k}(\omega), \xi(\omega)) > \frac{1}{k}\}$ satisfy $\sum_{k=1}^{\infty} \mu A_k < \infty$ and we can apply Borel-Cantelli to conclude that $\mu(A_k i.o.) = 0$. Thus $\omega \notin A_k i.o.$ we pick $N_1 > 0$ such that $\omega \notin A_k$ for $k > N_1$ and given $\epsilon > 0$, we pick $N_2 > \frac{1}{\epsilon}$. Then for $k > \max(N_1, N_2)$ we see that $d(\xi_{n_k}(\omega), \xi(\omega)) \leq \frac{1}{k} < \epsilon$ and we have shown that $\xi_{n_k} \xrightarrow{a.s.} \xi$.

To prove the converse, suppose that ξ_n does not converge in probability to ξ. The definitions tell us that we can find $\epsilon > 0$, $\delta > 0$ and a subsequence N' such that $\mathbf{P}\{d(\xi_{n_k}, \xi) > \epsilon\} > \delta$ for all $n \in N'$. We claim that there is no subsequence of N'' for which $\xi_n \xrightarrow{a.s.} \xi$ along N''. The claim is verified by using the fact (shown in the proof of Lemma 8.4) that convergence almost surely means that $\mathbf{P}\{\cup_{k\geq n}\{d(\xi_k, \xi) > \epsilon\}\} \to 0$ for all $\epsilon > 0$. For our chosen ϵ, along any subsequence $N'' \subset N'$ every tail

event $\cup_{k\in N'',k\geq n}\{d(\xi_k,\xi)>\epsilon\}$ contains only events with probability greater than δ hence cannot converge to 0. □

The previous lemma has a nice side effect which is a proof that the property of convergence in probability does not actually depend on the choice of metric.

Corollary 8.10. *Let $\xi,\xi_1,\xi_2,\dots$ be a random elements in a metrizable space S. The property $\xi_n \xrightarrow{P} \xi$ does not depend on the choice of metric d.*

The previous lemma also gives us a very simply proof the extremely useful Continuous Mapping Theorem for convergence in probability.

Lemma 8.11. *Let $\xi,\xi_1,\xi_2,\dots$ be a random elements in a metric space (S,d) such that $\xi_n \xrightarrow{P} \xi$. Let (T,d') be a metric space and let $f: S\to T$ be a continuous function, then $f(\xi_n)\xrightarrow{P} f(\xi)$.*

Proof. Pick a subsequence $N'\subset\mathbb{N}$ and note that by Lemma 8.9 we know there exists a subsequence $N''\subset N'$ such that $\xi_n \xrightarrow{a.s.} \xi$ along N''. By the continuity of f, we know that $f(\xi_n)\xrightarrow{a.s.} f(\xi)$ along N'' hence another application of Lemma 8.9 shows that $f(\xi_n)\xrightarrow{P} f(\xi)$. □

The full power of the Continuous Mapping Theorem for convergence in probability is only fully appreciated in conjuction with the following useful characterization of convergence in probability in product spaces. It is important to reinforce that the following Lemma fails in the case of convergence in distribution and one of the best uses of convergence in probability is a way of getting around that latter limitation.

Lemma 8.12. *Let $\xi,\xi_1,\xi_2,\dots$ and $\eta,\eta_1,\eta_2,\dots$ be random sequences in (S,d) and (T,d') respectively. Then $(\xi_n,\eta_n)\xrightarrow{P}(\xi,\eta)$ if an only if $\xi_n\xrightarrow{P}\xi$ and $\eta_n\xrightarrow{P}\eta$.*

Proof. Note that by Corollary 8.10 we may work with any metric on $S\times T$. We choose the metric $d''((x,w),(y,z)) = d(x,y)+d'(w,z)$. First we assume that $(\xi_n,\eta_n)\xrightarrow{P}(\xi,\eta)$. Then we know that for every $\epsilon>0$, we have

$$\lim_{n\to\infty}\mathbf{P}\{d''((\xi_n,\eta_n),(\xi,\eta))>\epsilon\}=0$$

By our choice of metric d'' we can see that $d(\xi_n,\xi)\leq d''((\xi_n,\eta_n),(\xi,\eta))$ and $d'(\eta_n,\eta)\leq d''((\xi_n,\eta_n),(\xi,\eta))$ and therefore we can conclude that $\xi_n\xrightarrow{P}\xi$ and $\eta_n\xrightarrow{P}\eta$.

On the other hand if we assume that $\xi_n\xrightarrow{P}\xi$ and $\eta_n\xrightarrow{P}\eta$ the for every $\epsilon>0$ we have the union bound

$$\mathbf{P}\{d''((\xi_n,\eta_n),(\xi,\eta))>\epsilon\}\leq\mathbf{P}\{d(\xi_n,\xi)>\frac{\epsilon}{2}\}+\mathbf{P}\{d'(\eta_n,\eta)>\frac{\epsilon}{2}\}$$

which shows the converse. □

Corollary 8.13. *Let $\xi,\xi_1,\xi_2,\dots$ and $\eta,\eta_1,\eta_2,\dots$ be sequences of random variables such that $\xi_n\xrightarrow{P}\xi$ and $\eta_n\xrightarrow{P}\eta$, then*

(i) $\xi_n+\eta_n\xrightarrow{P}\xi+\eta$
(ii) $\xi_n\eta_n\xrightarrow{P}\xi\eta$
(iii) $\xi_n/\eta_n\xrightarrow{P}\xi/\eta$ *if* $\eta\neq 0$ *a.e.*

Proof. By Lemma 8.12 we know that $(\xi_n, \eta_n) \xrightarrow{P} (\xi, \eta)$ in $\mathbb{R}^2$. By continuity of algebraic operations and the Continuous Mapping Theorem the result holds. □

8.1. **The Weak Law Of Large Numbers.**

Theorem 8.14 (Weak Law of Large Numbers)**.** *Let $\xi_1, \xi_2, \ldots$ be independent and identically distributed random variables with*

$$\mu = \boldsymbol{E}[\xi_i] < \infty$$

Then

$$\frac{1}{n}\sum_{k=1}^{n} \xi_k \xrightarrow{P} \mu$$

Proof. If is worth first proving the result with the additional assumption of finite variance, so assume $\sigma^2 = \mathbf{Var}(\xi_j) < \infty$. The first thing to note is that it suffices to assume that $\mu = 0$. For we can replace ξ_j by $\xi_j - \mu$. Now define $\hat{S}_n = \frac{1}{n}\sum_{k=1}^{n} \xi_k$ and note that by linearity of expectation, $\mathbf{E}\left[\hat{S}_n\right] = 0$ and by independence,

$$\mathbf{Var}\left(\hat{S}_n\right) = \frac{1}{n^2}\sum_{k=1}^{n} \mathbf{E}\left[\xi_k^2\right] = \frac{\sigma^2}{n}$$

Pick $\epsilon > 0$ and using Markov Inequality (Lemma 14.1)

$$\mathbf{P}\{\left|\hat{S}_n\right| > \epsilon\} = \mathbf{P}\{\hat{S}_n^2 > \epsilon^2\} \leq \frac{\mathbf{Var}\left(\hat{S}_n\right)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

so $\lim_{n\to\infty} \mathbf{P}\{\left|\hat{S}_n\right| > \epsilon\} = 0$ and thus $\hat{S}_n \xrightarrow{P} 0$.

Now to extend the result to eliminate the finite variance assumption we use a version of a *truncation argument*. One leverages the fact that by Lemma 7.16, independence of random variables is preserved under arbitrary measurable transformations. In particular, for every $N > 0$, define $f_N(x) = x \cdot \mathbf{1}_{|x|\leq N}$ which is easily seen to be measurable and define

$$\xi_{i,\leq N} = f_{\leq N} \circ \xi_i$$
$$\xi_{i,>N} = \xi_i - \xi_{i,\leq N}$$

We first establish some simple facts about the behavior of the truncation sequences $\xi_{i,\leq N}$ and $\xi_{i,>N}$. Since ξ_i are integrable we have the bound

$$\mathbf{Var}(\xi_{i,\leq N}) = \mathbf{E}\left[\xi_{i,\leq N}^2\right] - \mathbf{E}[\xi_{i,\leq N}]^2 \leq \mathbf{E}\left[\xi_{i,\leq N}^2\right] \leq N\mathbf{E}[|\xi_i|] < \infty$$

which shows that $\xi_{i,\leq N}$ has finite variance. Let $\mu_N = \mathbf{E}[\xi_{i,\leq N}]$.

Next note that integrability of ξ_i implies that $|\xi_i| < \infty$ a.s. hence $\lim_{N\to\infty} \xi_{i,>N} = \lim_{N\to\infty} |\xi_{i,>N}| = 0$ a.s. Since $|\xi_{i,>N}| < |\xi_i|$, we can apply Dominated Convergence Theorem and linearity of expectation to see that

$$\lim_{N\to\infty} \mathbf{E}[\xi_{i,>N}] = \lim_{N\to\infty} \mathbf{E}[|\xi_{i,>N}|] = 0$$
$$\lim_{N\to\infty} \mathbf{E}[\xi_{i,\leq N}] = \mathbf{E}[\xi_i] - \lim_{N\to\infty} \mathbf{E}[\xi_{i,>N}] = \mathbf{E}[\xi_i]$$

Now we stitch these observations together to provide the proof of the Weak Law of Large Numbers. Suppose we are given $\epsilon > 0$ and $\delta > 0$. Pick N large enough so that

$$|\mathbf{E}[\xi_{i,\leq N}] - \mathbf{E}[\xi_i]| < \frac{\epsilon}{3}$$
$$\mathbf{E}[|\xi_{i,>N}|] < \frac{\epsilon\delta}{3}$$

It is important to note these two bounds depend only on the underlying distribution of ξ_i and therefore by the identically distributed assumption on the ξ_i if we pick N so the above properties are satisified for a single i, in fact the properties are satisified uniformly for all $i > 0$.

Using the triangle inequality and a union bound (i.e. the general fact that $\{|a+b| \geq \epsilon\} \subset \{|a| \geq \frac{\epsilon}{2}\} \cup \{|b| \geq \frac{\epsilon}{2}\}$) we have

$$\begin{aligned}\mathbf{P}\{\left|\frac{\sum_{i=1}^n \xi_i}{n} - \mu\right| \geq \epsilon\} &= \mathbf{P}\{\left|\frac{\sum_{i=1}^n \xi_{i,\leq N}}{n} - \mu_N + \mu_N - \mu + \frac{\sum_{i=1}^n \xi_{i,>N}}{n}\right| \geq \epsilon\} \\ &\leq \mathbf{P}\{\left|\frac{\sum_{i=1}^n \xi_{i,\leq N}}{n} - \mu_N\right| \geq \frac{\epsilon}{3}\} \\ &\quad + \mathbf{P}\{|\mu_N - \mu| \geq \frac{\epsilon}{3}\} + \mathbf{P}\{\left|\frac{\sum_{i=1}^n \xi_{i,>N}}{n}\right| \geq \frac{\epsilon}{3}\}\end{aligned}$$

Consider each of the three terms in turn. The first term we apply Chebyshev bounding

$$\mathbf{P}\{\left|\frac{\sum_{i=1}^n \xi_{i,\leq N}}{n} - \mu_N\right| \geq \frac{\epsilon}{3}\} \leq \frac{9\mathbf{Var}\left(\frac{\sum_{i=1}^n \xi_{i,\leq N}}{n}\right)}{\epsilon^2} \leq \frac{9N\mathbf{E}[|\xi_1|]}{n\epsilon^2} < \delta$$

provided we choose $n > \frac{9N\mathbf{E}[|\xi_1|]}{\delta\epsilon^2}$. The second term is 0 since we have assumed N large enough so that $|\mu_N - \mu| < \frac{\epsilon}{3}$. The third term we use a Markov bound

$$\mathbf{P}\{\left|\frac{\sum_{i=1}^n \xi_{i,>N}}{n}\right| \geq \frac{\epsilon}{3}\} \leq \frac{3\mathbf{E}\left[\left|\frac{\sum_{i=1}^n \xi_{i,>N}}{n}\right|\right]}{\epsilon} \leq \frac{3\mathbf{E}[|\xi_{i,>N}|]}{\epsilon} < \delta$$

□

It is worth examining the proof above to see that we didn't use the full strength of the identical distribution property. Really all we used was the fact that were able to provide bounds on the expectation of the tails of the sequences *uniformly*. As an exercise, it is worth noting that the above proof goes through almost unchanged provided we merely assume that ξ_n are independent and uniformly integrable.

Example 8.15. The following is an example of a how the Weak Law of Large Numbers can fail despite having a sequence of independent random variables with bounded first moment.

Let η_n be a sequence of independent Bernoulli random variables with the rate of η_n equal to $\frac{1}{2^n}$. Now define $\xi_n = 2^n\eta_n$ and $S_n = \frac{1}{n}\sum_{k=1}^n \xi_k$. It is helpful to think in Computer Science terms and consider $\sum_{k=1}^n \xi_k$ to be a random n-bit positive integer in which bit k has probability $\frac{1}{2^k}$ of being set. Note that $\mathbf{E}[\xi_n] = \mathbf{E}[|\xi_n|] = 1$ and therefore $\mathbf{E}[S_n] = 1$. On the other hand we proceed to show that S_n does not converge in probability to 1. We do this by constructing a subsequence S_{n_k} such

that $\lim_{k\to\infty} \mathbf{P}\{S_{n_k} < \frac{1}{2}\} = 1$ (note the choice of the constant $\frac{1}{2}$ is somewhat arbitrary; any positive constant would do).

Consider the subsequence S_{2^k} and the complementary event

$$\{S_{2^k} \geq \frac{1}{2}\} = \{\sum_{n=1}^{2^k} \xi_n \geq 2^{k-1}\} = \bigcup_{m=k-1}^{2^k} \{\xi_m \neq 0\}$$

Taking expectations, we get

$$\begin{aligned}\mathbf{P}\{S_{2^k} \geq \frac{1}{2}\} &\leq \sum_{m=k-1}^{2^k} \mathbf{P}\{\xi_m \neq 0\} \\ &= \sum_{m=k-1}^{2^k} \frac{1}{2^m} = \frac{1}{2^{k-1}} \cdot 2 \cdot (1 - 2^{2^k - k + 1}) < \frac{1}{2^{k-2}}\end{aligned}$$

which is enough to show by taking complements that $\lim_{k\to\infty} \mathbf{P}\{S_{2^k} < \frac{1}{2}\} = 1$.

TODO: Discussion about what is going on here. Essentially, the averages here have a distribution which is peaking around 0 but has enough of a possibility of rare events happening (with exponentially large impact) to move the mean of the averages up to 1. Thus the distribution is concentrating around 0 which is NOT the mean!

TODO: Question: does this sequence converge in distribution? I'd guess is converges to the Dirac measure at 0.

TODO: Other weak law "counterexamples" such as Cauchy distributions. Varadhan mentions that one can tweak a Cauchy distribution so that it has no mean but the sequence of averages converges in probability.

8.2. **The Strong Law Of Large Numbers.** This is the most common approach to proving of the Strong Law of Large Numbers. The proof requires the development of some tools for proving the almost sure converges of infinite sums of independent random variables.

TODO: Observe how this next result is related to second moment bounds (Chebyshev applied to sums).

Lemma 8.16 (Kolmogorov's Maximal Inequality)**.** *Let $\xi_1, \xi_2, \ldots$ be independent random variables with $\boldsymbol{E}\left[\xi_n^2\right] < \infty$ for all $n > 0$. The for every $\epsilon > 0$, we have*

$$\boldsymbol{P}\{\sup_n \left|\sum_{k=1}^{n} \xi_k - \boldsymbol{E}\left[\xi_k\right]\right| \geq \epsilon\} < \frac{1}{\epsilon^2} \sum_{k=1}^{\infty} \boldsymbol{Var}(\xi_k)$$

Proof. It is clear we may assume that $\mathbf{E}\left[\xi_n\right] = 0$ for all $n > 0$.

Before we start in on the result to be proven, we need an small observation. To clean up notation a bit we define $S_n = \sum_{k=1}^{n} \xi_k$. Pick $N > n > 0$ and observe $0 \leq (S_N - S_n)^2 = S_N^2 - 2S_N S_n + S_n^2 = S_N^2 - S_n^2 - 2(S_N - S_n)S_n$ and therefore $S_N^2 - S_n^2 \geq 2(S_N - S_n)S_n$. Now using the fact that by Lemma 7.13 we know $S_N - S_n$ is independent of S_n. Therefore for any $A_n \in \sigma(S_n)$ we have

$$\mathbf{E}\left[S_N^2 - S_n^2; A_n\right] \geq 2\mathbf{E}\left[(S_N - S_n)S_n; A_n\right] = 2\mathbf{E}\left[S_N - S_n\right]\mathbf{E}\left[S_n; A_n\right] = 0$$

which gives us

$$\mathbf{E}\left[S_N^2; A_n\right] \geq \mathbf{E}\left[S_n^2; A_n\right]$$

by linearity of expectation.

Now we start in on the inequality to be proven. Note that by continuity of measure, we know that

$$\mathbf{P}\{\sup_n |S_n| \geq \epsilon\} = \lim_{N\to\infty} \mathbf{P}\{\sup_{n\leq N} |S_n| \geq \epsilon\}$$

so it suffices to show for every $N > 0$

$$\mathbf{P}\{\sup_{n\leq N} |S_n| \geq \epsilon\} \leq \frac{1}{\epsilon^2}\sum_{k=1}^{N} \mathbf{E}\left[\xi_k^2\right] = \frac{1}{\epsilon^2}\mathbf{E}\left[S_N^2\right]$$

Consider $\mathbf{P}\{\sup_{n\leq N} |S_n| \geq \epsilon\}$. Define the event $A_n = \{|S_k| < \epsilon \text{ for } 1 \leq k < n \text{ and } |S_n| \geq \epsilon$ and note that A_n is $\sigma(\xi_n)$-measurable and we have the disjoint union

$$\{\sup_{n\leq N} |S_n| \geq \epsilon\} = A_1 \cup \cdots \cup A_N$$

and therefore

$$\begin{aligned}
\mathbf{P}\{\sup_{n\leq N} |S_n| \geq \epsilon\} &= \sum_{k=1}^{N} \mathbf{P}\{A_k\} && \text{by additivity of measure} \\
&\leq \frac{1}{\epsilon^2}\sum_{k=1}^{N} \mathbf{E}\left[S_k^2; A_k\right] && |S_k| \geq \epsilon \text{ on the event } A_k \\
&\leq \frac{1}{\epsilon^2}\sum_{k=1}^{N} \mathbf{E}\left[S_N^2; A_k\right] && \\
&= \frac{1}{\epsilon^2}\mathbf{E}\left[S_N^2; \sup_{n\leq N} |S_n| \geq \epsilon\right] && \text{by additivity of measure} \\
&\leq \frac{1}{\epsilon^2}\mathbf{E}\left[S_N^2\right] && \text{positivity of } S_N^2
\end{aligned}$$

and the result is proved. □

The previous lemma gives us a criterion for almost sure convergence of sums of square integrable random variables with finite variance.

Lemma 8.17 (Kolmogorov One-Series Criterion)**.** *Let $\xi_1, \xi_2, \ldots$ be independent square integrable random variables. If $\sum_{n=1}^{\infty} \boldsymbol{Var}(\xi_n) < \infty$ then $\sum_{n=1}^{\infty} (\xi_n - \boldsymbol{E}[\xi_n])$ converges a.s.*

Proof. We may clearly assume that $\mathbf{E}[\xi_n] = 0$ for all $n > 0$. Define $S_n = \sum_{k=1}^{n} \xi_k$.

Before giving a proper proof, it might be worth looking a simple heuristic argument to give some intuition why this result should be true. For every $N > 0$,

$$\begin{aligned}
\mathbf{P}\{\left|\sum_{n=1}^{\infty} \xi_n\right| > N\} &\leq \frac{\mathbf{Var}\left(\sum_{n=1}^{\infty} \xi_n\right)}{N^2} && \text{by Chebeshev's Inequality} \\
&= \frac{\sum_{n=1}^{\infty} \mathbf{E}\left[\xi_n^2\right]}{N^2} && \text{by independence and zero mean}
\end{aligned}$$

and therefore we know that

$$\sum_{N=1}^{\infty} \mathbf{P}\{\left|\sum_{n=1}^{\infty} \xi_n\right| > N\} \leq \sum_{n=1}^{\infty} \mathbf{E}\left[\xi_n^2\right] \sum_{N=1}^{\infty} \frac{1}{N^2} < \infty$$

so Borel Cantelli implies $\mathbf{P}\{|\sum_{n=1}^{\infty}\xi_n| > N \text{ i.o.}\} = 0$ which implies almost sure convergence. The problem with this argument is that we have manipulated the series as if we knew it converged which is what we are trying to prove (is this really the problem, or is the problem that we are dealing with conditional convergence so showing the almost sure boundedness of the sum doesn't imply convergence; in that case this argument is completely irrelevant). Kolmogorov's Maximal Inequality gives us a way to make a more rigorous argument.

Pick $\epsilon > 0$ and for every $N > 0$ define $A_{N,\epsilon} = \{\sup_{n>N}|S_n - S_N| \geq \epsilon\}$. Applying Lemma 8.16 to the sequence ξ_n for $n = N+1, N+2, \ldots$, we know that

$$\mathbf{P}\{A_{N,\epsilon}\} = \mathbf{P}\{\sup_{n>N}|S_n - S_N| \geq \epsilon\} \leq \frac{1}{\epsilon^2}\sum_{n=N+1}^{\infty}\mathbf{E}\left[\xi_n^2\right]$$

and by the convergence of $\sum_{n=1}^{\infty}\mathbf{E}\left[\xi_n^2\right]$ we know that

$$\lim_{N\to\infty}\mathbf{P}\{A_{N,\epsilon}\} \leq \lim_{N\to\infty}\frac{1}{\epsilon^2}\sum_{n=N+1}^{\infty}\mathbf{E}\left[\xi_n^2\right] = 0$$

which by subadditivity of measure tells us that $\mathbf{P}\{\cap_{N=1}^{\infty}A_{N,\epsilon}\} = 0$. Now, for every $n > 0$ define $B_n = \cap_{N=1}^{\infty}A_{N,\frac{1}{n}}$, define $B = \cup_n B_n$ and note that by countable additivity of measure, $\mathbf{P}\{B\} = 0$.

We show that S_n converges for all $\omega \notin B$. Pick $\omega \notin B$. Assume we are given $\epsilon > 0$ and pick $n > 0$ such that $\frac{1}{n} < \epsilon$. We know $\omega \notin B_n$ and therefore for some $N > 0$, $\omega \notin A_{N,\frac{1}{n}}$ which implies that $|S_k - S_N| < \frac{1}{n} < \epsilon$ for all $k > N$. This shows that $S_n(\omega)$ is a Cauchy sequence for every $\omega \notin B$ and by completeness of $\mathbb{R}$ this shows that S_n is almost surely convergent.

Here is a more concise variant of the same basic argument. Pick $\epsilon > 0$ and applying Lemma 8.16 to the sequence ξ_n for $n = N+1, N+2, \ldots$, we know that

$$\mathbf{P}\{\sup_{n>N}|S_n - S_N| \geq \epsilon\} \leq \frac{1}{\epsilon^2}\sum_{n=N+1}^{\infty}\mathbf{E}\left[\xi_n^2\right]$$

and by the convergence of $\sum_{n=1}^{\infty}\mathbf{E}\left[\xi_n^2\right]$ we know that

$$\lim_{N\to\infty}\mathbf{P}\{\sup_{n>N}|S_n - S_N| \geq \epsilon\} \leq \lim_{N\to\infty}\frac{1}{\epsilon^2}\sum_{n=N+1}^{\infty}\mathbf{E}\left[\xi_n^2\right] = 0$$

which shows that $\sup_{n>N}|S_n - S_N| \xrightarrow{P} 0$. Now by Lemma 8.9 we know that a subsequence of $\sup_{n>N}|S_n - S_N|$ converges to 0 a.s. However, as $\sup_{n>N}|S_n - S_N|$ is nonincreasing in N (TODO: I don't see this; in fact I don't think it is true without a positivity assumption), the almost sure converge of the subsequence implies the almost sure converge of the entire sequence. The convergence $\sup_{n>N}|S_n - S_N| \xrightarrow{a.s.} 0$ is just the statement that S_n is almost sure Cauchy which by completeness of $\mathbb{R}$ says that S_n converges almost surely. □

Having just proven a convergence criterion for a sequence of partial sums of independent random variables, we should ask ourselves how this can help us establish criteria for the sequence of averages that the Strong Law of Large Numbers refers to. The key result here has nothing to do with probability.

Lemma 8.18. *Let $a_1, a_2, \ldots$ and $b_1, b_2, \ldots$ be sequences of real numbers. Define $\Delta a_n = a_{n+1} - a_n$ and $\Delta b_n = b_{n+1} - b_n$, then for every $n > m > 0$,*

$$\sum_{k=m}^{n} a_k \Delta b_k = a_{n+1}b_{n+1} - a_m b_m - \sum_{k=m}^{n} b_{k+1}\Delta a_k$$

Proof. Note that we have the *product rule*

$$\begin{aligned}\Delta(a \cdot b)_k &= a_{k+1}b_{k+1} - a_k b_k \\ &= a_{k+1}b_{k+1} - a_k b_{k+1} + a_k b_{k+1} - a_k b_k \\ &= a_k \Delta b_k + b_{k+1}\Delta a_k\end{aligned}$$

and therefore

$$\begin{aligned}a_{n+1}b_{n+1} - a_m b_m &= \sum_{k=m}^{n} \Delta(a \cdot b)_k \\ &= \sum_{k=m}^{n} a_k \Delta b_k + \sum_{k=m}^{n} b_{k+1}\Delta a_k\end{aligned}$$

□

Lemma 8.19. *Let $0 = b_0 \leq b_1 \leq b_2 \leq \ldots$ be a non-decreasing sequence of positive real numers such that $\lim_{n\to\infty} b_n = \infty$ and define $\beta_n = b_n - b_{n-1}$ for $n > 0$. If $s_1, s_2, \ldots$ is a sequence of real numbers with $\lim_{n\to\infty} s_n = s$ then*

$$\lim_{n\to\infty} \frac{1}{b_n} \sum_{k=1}^{n} \beta_k s_k = s$$

In particular, if $x_1, x_2, \ldots$ are real numbers, then if $\sum_{n=1}^{\infty} \frac{x_n}{b_n} < \infty$ then $\lim_{n\to\infty} \frac{1}{b_n} \sum_{k=1}^{n} x_k < \infty$.

Proof. To see the first part of the Lemma, note that for any constant $s \in \mathbb{R}$, $\frac{1}{b_n} \sum_{k=1}^{n} \beta_k s = s$ and therefore we may assume that $s = 0$.

Pick an $\epsilon > 0$ and then select $N_1 > 0$ such that $|s_k| < \frac{\epsilon}{2}$ for all $k \geq N_1$. Define $M = \sup_{n\geq 1} |s_n|$ and then because $\lim_{n\to\infty} b_n = \infty$ we can pick $N_2 > 0$ such that $\frac{b_{N_1} M}{b_n} < \frac{\epsilon}{2}$ for all $n > N_2$. Now for every $n > \max(N_1, N_2)$,

$$\begin{aligned}\left|\frac{1}{b_n}\sum_{k=1}^{n} \beta_k s_k\right| &\leq \left|\frac{1}{b_n}\sum_{k=1}^{N_1} \beta_k s_k\right| + \left|\frac{1}{b_n}\sum_{k=N_1+1}^{n} \beta_k s_k\right| \\ &\leq \frac{b_{N_1} M}{b_n} + \frac{(b_n - b_{N_1})\epsilon}{2b_n} \leq \epsilon\end{aligned}$$

and we are done.

To see the second part of the Lemma, define $s_0 = 0$ and $s_n = \sum_{k=1}^{n} \frac{x_k}{b_k}$, now apply summation by parts to see

$$\begin{aligned}\frac{1}{b_n}\sum_{k=1}^{n} \Delta b_{k-1} s_{k-1} &= \frac{1}{b_n}\left(b_n s_n - b_0 s_0 - \sum_{k=1}^{n} b_k \Delta s_{k-1}\right) \\ &= s_n - \frac{1}{b_n}\sum_{k=1}^{n} x_k\end{aligned}$$

so we can take limits and apply the first part of this Lemma to find

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{b_n}\sum_{k=1}^{n}x_k &= \lim_{n\to\infty}s_n - \lim_{n\to\infty}\frac{1}{b_n}\sum_{k=1}^{n}\Delta b_{k-1}s_{k-1}\\ &= s - s = 0\end{aligned}$$

□

Corollary 8.20. *Assume that* $0 \le b_1 \le b_2 \le \cdots$ *and* $\lim_{n\to\infty} b_n = \infty$ *and let* $\xi_1, \xi_2, \ldots$ *be independent square integrable random variables. If* $\sum_{n=1}^{\infty} \frac{\boldsymbol{Var}(\xi_n)}{b_n^2} < \infty$ *then*

$$\frac{1}{b_n}\sum_{k=1}^{n}(\xi_k - \boldsymbol{E}[\xi_k]) \overset{a.s.}{\to} 0$$

Theorem 8.21 (Strong Law of Large Numbers)**.** *Let* $\xi, \xi_1, \xi_2, \ldots$ *be independent and identically distributed random variables. Then if* ξ_1 *is integrable*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\xi_k = \boldsymbol{E}[\xi] \quad \text{a.s.}$$

Conversely if $\frac{1}{n}\sum_{k=1}^{n}\xi_k$ *converges on a set of positive measure, then* ξ_1 *is integrable.*

Proof. First, one makes the standard reduction to the case in which $\mathbf{E}[\xi_n] = 0$ for all $n > 0$.

Next we apply a truncation argument by defining

$$\eta_n = \xi_{n,\le n} = \xi_n \cdot \mathbf{1}_{[0,n]}(|\xi_n|)$$

Note

$$\begin{aligned}\sum_{n=1}^{\infty}\mathbf{P}\{\eta_n \neq \xi_n\} &= \sum_{n=1}^{\infty}\mathbf{P}\{|\xi_n| > n\}\\ &\le \sum_{n=1}^{\infty}\int_{n-1}^{n}\mathbf{P}\{|\xi_n| \ge \lambda\}d\lambda \qquad \text{since } \mathbf{P}\{|\xi_n| \ge \lambda\} \text{ is decreasing}\\ &= \int_0^{\infty}\mathbf{P}\{|\xi| \ge \lambda\}d\lambda \qquad \text{by i.i.d.}\\ &= \mathbf{E}[|\xi|] < \infty \qquad \text{by Lemma 6.8}\end{aligned}$$

Now we apply Borel Cantelli to conclude that $\mathbf{P}\{\eta_n \neq \xi_n \text{ i.o.}\} = 0$. Stated conversely, $\mathbf{P}\{\text{there exists } N > 0 \text{ such that } \xi_n \le n \text{ for all } n > N\} = 1$.

Next define $\overline{\eta}_n = \frac{1}{n}\sum_{k=1}^{n}\eta_k$ and $\overline{\xi}_n = \frac{1}{n}\sum_{k=1}^{n}\xi_k$. We claim that $\lim_{n\to\infty}\overline{\eta}_n = 0$ a.s. if and only if $\lim_{n\to\infty}\overline{\xi}_n = 0$ a.s.

For almost all $\omega \in \Omega$ we can pick $N_\omega > 0$ such that $\xi_n(\omega) = \eta_n(\omega)$ for all $n > N_\omega$. Let $C_\omega = \sum_{k=1}^{N_\omega}(\eta_k(\omega) - \xi_k(\omega))$ so that for $n > N_\omega$, we have $\lim_{n\to\infty}\overline{\eta}_n(\omega) = \lim_{n\to\infty}\overline{\xi}_n(\omega) + \frac{C_\omega}{n}$ and therefore $\lim_{n\to\infty}\overline{\eta}_n(\omega) = \lim_{n\to\infty}\overline{\xi}_n(\omega)$.

Therefore it suffices to show $\lim_{n\to\infty}\overline{\eta}_n = 0$ a.s. Although we no longer have $\mathbf{E}[\eta_n] = 0$ because we have truncated ξ_n, the *average* of the means of η_n is 0. This

follows from noting that $\lim_{n\to\infty} \xi_{\leq n} = \xi$ and $|\xi_{\leq n}| \leq |\xi|$ so

$$\begin{aligned} 0 &= \mathbf{E}[\xi] \\ &= \lim_{n\to\infty} \mathbf{E}[\xi_{\leq n}] && \text{by Dominated Convergence} \\ &= \lim_{n\to\infty} \mathbf{E}[\xi_{n,\leq n}] && \text{by i.i.d.} \\ &= \lim_{n\to\infty} \mathbf{E}[\eta_n] \end{aligned}$$

and therefore by application of Lemma 8.19

$$\frac{1}{n}\sum_{k=1}^{n} \mathbf{E}[\eta_n] = \lim_{n\to\infty} \mathbf{E}[\eta_n] = 0$$

Therefore if we can show that $\sum_{n=1}^{\infty} \frac{\mathbf{Var}(\eta_n)}{n^2} < \infty$, then by Corollary 8.20 we can conclude

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \eta_k = \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \mathbf{E}[\eta_k] = 0 \text{ a.s.}$$

and we'll be done.

To show the desired bound we'll need the elementary fact that $C = \sup_{n>0} n\sum_{k=n}^{\infty} \frac{1}{k^2} < \infty$. This can be seen by viewing the sum as lower Riemann sum for an integral bounding

$$n\sum_{k=n}^{\infty} \frac{1}{k^2} \leq n\int_{n-1}^{\infty} \frac{dx}{x^2} = \frac{n}{n-1} \leq 2$$

Now we can finish the proof

$$\begin{aligned} \sum_{n=1}^{\infty} \frac{\mathbf{Var}(\eta_n)}{n^2} &\leq \sum_{n=1}^{\infty} \frac{\mathbf{E}\left[\eta_n^2\right]}{n^2} \\ &= \sum_{n=1}^{\infty} \frac{\mathbf{E}\left[\xi_n^2; |\xi_n| \leq n\right]}{n^2} \\ &= \sum_{n=1}^{\infty}\sum_{k=1}^{n} \frac{\mathbf{E}\left[\xi^2; k-1 \leq |\xi| \leq k\right]}{n^2} \\ &= \sum_{k=1}^{\infty} \mathbf{E}\left[\xi^2; k-1 \leq |\xi| \leq k\right] \sum_{n=k}^{\infty} \frac{1}{n^2} \\ &\leq \sum_{k=1}^{\infty} \frac{C}{k} \mathbf{E}\left[\xi^2; k-1 \leq |\xi| \leq k\right] \\ &\leq C\sum_{k=1}^{\infty} \frac{k}{k} \mathbf{E}\left[|\xi|; k-1 \leq |\xi| \leq k\right] = C\mathbf{E}\left[|\xi|\right] < \infty \end{aligned}$$

It remains to show the converse result; namely that if $\overline{\xi}_n$ converges on a set of positive measure then ξ is integrable. First, note by Corollary 7.28, we know that $\overline{\xi}_n$ converges almost surely.

$$\begin{aligned}\lim_{n\to\infty}\frac{\xi_n}{n} &= \lim_{n\to\infty}\left(\overline{\xi}_n - \frac{n-1}{n}\overline{\xi}_{n-1}\right)\\ &= \lim_{n\to\infty}\overline{\xi}_n - 1\cdot\lim_{n\to\infty}\overline{\xi}_n = 0 \text{ a.s.}\end{aligned}$$

and therefore if we define $A_n = \{|\xi_n| \geq n\}$ then we know that $\mathbf{P}\{A_n \text{ i.o.}\} = 0$ (in particular for each ω for which $\lim_{n\to\infty}\frac{\xi_n(\omega)}{n} = 0$ and any $\epsilon > 0$, we can find $N > 0$ such that $|\xi_n(\omega)| < \epsilon n$ for all $n > N$; just choose $\epsilon < 1$). But we also know that ξ_n are independent and therefore by Lemma 7.14 the A_n are independent so Borel Cantelli implies $\sum_{n=1}^{\infty}\mathbf{P}\{A_n\} < \infty$. But now we can apply a tail bound

$$\begin{aligned}\mathbf{E}\left[|\xi|\right] &= \int_0^\infty \mathbf{P}\{|\xi| \geq \lambda\}\, d\lambda && \text{by Lemma 6.8}\\ &\leq \sum_{n=0}^{\infty}\mathbf{P}\{|\xi| \geq n\} && \text{bounding by an upper Riemann sum}\\ &= 1 + \sum_{n=1}^{\infty}\mathbf{P}\{A_n\} < \infty && \text{by i.i.d.}\end{aligned}$$

□

Proof. The following proof uses a different truncation argument (one closer to the WLLN argument we presented) and is taken from Tao.

TODO: Understand that proof better and write it down completely.

So to apply Borel Cantelli we need so find a sequence N_j such that

$$\sum_{j=1}^{\infty} n_j \mathbf{P}\{\xi > N_j\} < \infty$$

$$\sum_{j=1}^{\infty}\frac{1}{n_j}\mathbf{E}\left[\xi_{\leq N_j}\right] < \infty$$

We show that both sums are finite if we choose $N_j = n_j$. In both cases this follows by establishing pointwise bounds in terms of ξ. For the first sum we use Tonelli's Theorem to exchange sums and expectations

$$\begin{aligned}\sum_{j=1}^{\infty} n_j \mathbf{P}\{\xi > n_j\} = \sum_{j=1}^{\infty} n_j \mathbf{E}\left[\mathbf{1}_{\xi > n_j}\right] &= \mathbf{E}\left[\sum_{j=1}^{\infty} n_j \mathbf{1}_{\xi > n_j}\right]\\ &= \mathbf{E}\left[\sum_{n_j < \xi} n_j\right]\end{aligned}$$

TODO: Fill this in. Essentially the idea is that we have an approximately geometric series so the above is $O(\xi)$.

For the second sum,

$$\sum_{j=1}^{\infty}\frac{1}{n_j}\mathbf{E}\left[\xi_{\leq n_j}\right] \leq \frac{1}{n_1}\mathbf{E}\left[\xi\right]\sum_{j=1}^{\infty} c^{-j} = \frac{c\mathbf{E}\left[\xi\right]}{n_1(c-1)} < \infty$$

□

Theorem 8.22 (Strong Law of Large Numbers (Finite Variance Case))**.** *Let $\xi_1, \xi_2, \ldots$ be independent and identically distributed random variables. Let*

$$\mu = \boldsymbol{E}[\xi_i] \text{ and } \sigma^2 = \boldsymbol{Var}(\xi_j)^2 < \infty$$

Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \xi_k = \mu \quad \text{a.s. and in } L^2$$

Proof. First note that by replacing ξ_n with $\xi_n - \mu$ it suffices to prove the Theorem with $\mu = 0$.

Next it is convenient to define the terms $S_n = \sum_{k=1}^{n} \xi_k$ and $\eta_n = \frac{S_n}{n}$. and observe that by linearity $\mathbf{E}[S_n] = \mathbf{E}[\eta_n] = 0$ and by independence

$$\begin{aligned} \mathbf{Var}(\eta_n) &= \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} \mathbf{E}[\xi_j \xi_k] \\ &= \frac{1}{n^2} \sum_{k=1}^{n} \mathbf{E}\left[\xi_k^2\right] = \frac{\sigma^2}{n} \end{aligned}$$

By taking the limit we see that $\lim_{n\to\infty} \mathbf{Var}(\eta_n) = 0$ which implies that $\eta_n \to 0$ in L^2.

To see almost sure convergence we first pass to a subsequence. Consider the subsequence η_{n^2} and note by the above variance calculation and Corollary 3.41 that

$$\mathbf{E}\left[\sum_{n=1}^{\infty} \eta_{n^2}^2\right] = \sum_{n=1}^{\infty} \mathbf{E}\left[\eta_{n^2}^2\right] = \sum_{n=1}^{\infty} \frac{\sigma^2}{n^2} < \infty$$

Finiteness of the first expectation implies that $\sum_{n=1}^{\infty} \eta_{n^2}^2 < \infty$ almost surely which in turn implies that $\lim_{n\to\infty} \eta_{n^2}^2 = 0$ and $\lim_{n\to\infty} \eta_{n^2} = 0$ almost surely . It remains to prove almost sure convergence for the entire sequence.

Pick an arbitrary $n > 0$ and define $p(n) = \lfloor \sqrt{n} \rfloor$ so that $p(n)$ is the integer satisfying $(p(n))^2 \le n < (p(n)+1)^2$. Then we have

$$\eta_n - \frac{p(n)^2}{n} \eta_{p(n)^2} = \frac{1}{n} \sum_{k=p(n)^2+1}^{n} \xi_k$$

and calculating variances as before,

$$\begin{aligned} \mathbf{Var}\left(\eta_n - \frac{p(n)^2}{n}\eta_{p(n)^2}\right) &= \mathbf{E}\left[\left(\eta_n - \frac{p(n)^2}{n}\eta_{p(n)^2}\right)^2\right] \\ &= \frac{1}{n^2} \sum_{k=p(n)^2+1}^{n} \mathbf{E}\left[\xi_k^2\right] \\ &= \frac{\sigma^2(n - p(n)^2)}{n^2} \\ &< \frac{\sigma^2(2p(n)+1)}{n^2} \le \frac{3\sigma^2}{n^{\frac{3}{2}}} \end{aligned}$$

This bound tells us that

$$\mathbf{E}\left[\sum_{n=1}^{\infty}\left(\eta_n - \frac{p(n)^2}{n}\eta_{p(n)^2}\right)^2\right] = \sum_{n=1}^{\infty}\mathbf{E}\left[\left(\eta_n - \frac{p(n)^2}{n}\eta_{p(n)^2}\right)^2\right] < \infty$$

which as before tells us that

$$\sum_{n=1}^{\infty}\left(\eta_n - \frac{p(n)^2}{n}\eta_{p(n)^2}\right)^2 < \infty$$

almost surely and

$$\lim_{n\to\infty}\left(\eta_n - \frac{p(n)^2}{n}\eta_{p(n)^2}\right) = 0$$

almost surely.

Since we have already proven $\eta_{p(n)^2} \overset{a.s.}{\to} 0$ and we can see by definition that $0 < \frac{p(n)}{n} \leq 1$ we conclude that $\eta_n \overset{a.s.}{\to} 0$. □

8.2.1. *Empirical Distributions and the Glivenko-Cantelli Theorem.* Here is a simple application of the Strong Law of Large Numbers that has important applications in statistics. Consider the process of making a sequence of independent observations for purpose of inferring a statement about an underlying distribution of a random variable. A basic statistical methodology is to use the distribution of ones sample as an approximation to the unknown distribution. We aim to give a demonstration of why this methodology is sound. First we make precise what we mean by the distribution of the sample.

Definition 8.23. Given independent random variables $\xi_1, \xi_2, \ldots$, for each $n > 0$ and $x \in \mathbb{R}$, we define the *empirical distribution function* to be

$$\hat{F}_n(x, \omega) = \frac{1}{n}\sum_{k=1}^{n}\mathbf{1}_{\xi_k \leq x}(\omega)$$

Note that the empirical distribution function depends on both x and $\omega \in \Omega$ but it is customary to omit mention of the argument ω and simply write $\hat{F}_n(x)$. In general we will follow this custom but on occasion where we feel it is important enough for clarity we'll include it as we did in the definition. In the statistical context we've alluded to each ξ_k represents the value of the k^{th} observation. The empirical distribution of n samples is the distribution function of the *empirical measure* obtained by placing an equally weighted point mass at the value of each observation.

Lemma 8.24. *Let $\xi_1, \xi_2, \ldots$ be i.i.d. random variables with distribution function $F(x)$ and empirical distribution functions $\hat{F}_1(x), \hat{F}_2(x), \ldots$. Then for each $x \in \mathbb{R}$,*

$$\lim_{n\to\infty}\hat{F}_n(x) = F(x) \text{ a.s.}$$

and in addition

$$\lim_{n\to\infty}\lim_{y\to x^-}\hat{F}_n(y) = \lim_{y\to x^-}F(y) \text{ a.s.}$$

Proof. This statement is a simple application of the Strong Law of Large Numbers. First note that for every $x \in \mathbb{R}$, by Lemma 7.14, the functions $\mathbf{1}_{\xi_n \leq x}$ are independent. Because the ξ_n are identically distributed the same follows for $\mathbf{1}_{\xi_n \leq x}$.

Lastly, the functions $\mathbf{1}_{\xi_n \leq x}$ are bounded and therefore integrable so we can apply the Strong Law of Large Numbers to conclude that

$$\lim_{n\to\infty} \hat{F}_n(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_{\xi_k \leq x} = \mathbf{E}\left[\mathbf{1}_{\xi_1 \leq x}\right] = F(x) \text{ a.s.}$$

To see the almost sure pointwise convergence of the left limits, first note that for every $x \in \mathbb{R}$, we have

$$\lim_{n\to\infty} \mathbf{1}_{(-\infty, x-\frac{1}{n}]}(y) = \begin{cases} 1 & \text{if } y < x \\ 0 & \text{if } y \geq x \end{cases} = \mathbf{1}_{(-\infty,x)}(y)$$

Therefore,

$$\begin{aligned} F(x-) &= \lim_{n\to\infty} F(x - \frac{1}{n}) && \text{by the existence of left limits in } F(x) \\ &= \mathbf{E}\left[\mathbf{1}_{\xi \leq x-\frac{1}{n}}\right] \\ &= \mathbf{E}\left[\lim_{n\to\infty} \mathbf{1}_{\xi \leq x-\frac{1}{n}}\right] && \text{by Dominated Convergence Theorem} \\ &= \mathbf{E}\left[\mathbf{1}_{\xi<x}\right] \end{aligned}$$

By the same argument,

$$\begin{aligned} \hat{F}_m(x-) &= \lim_{n\to\infty} \hat{F}_m(x - \frac{1}{n}) \\ &= \lim_{n\to\infty} \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_{\xi_i \leq x-\frac{1}{n}} \\ &= \sum_{i=1}^{m} \mathbf{1}_{\xi_i<x} \end{aligned}$$

As in the pointwise argument above, the family $\mathbf{1}_{\xi_i<x}$ is an i.i.d. family of integrable random variables so using the above computations and the Strong Law of Large Numbers we see that

$$\lim_{n\to\infty} \hat{F}_n(x-) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbf{1}_{\xi_i<x} = \mathbf{E}\left[\mathbf{1}_{\xi<x}\right] = F(x-) \text{ a.s.}$$

□

In fact, with a little more work leveraging properties of distribution functions, we can prove that the empirical distribution function converges uniformly.

Theorem 8.25 (Glivenko-Cantelli Theorem)**.** *Let $\xi_1, \xi_2, \ldots$ be i.i.d. random variables with distribution function $F(x)$ and empirical distribution functions $\hat{F}_1(x), \hat{F}_2(x), \ldots$. Then,*

$$\lim_{n\to\infty} \sup_x \left|\hat{F}_n(x) - F(x)\right| = 0 \text{ a.s.}$$

Proof. TODO: Give an intuitive idea of the proof (the notation is messy and a bit opaque). Essentially we use the properties of distribution functions (cadlag property and the compactness of the range) to establish that if two distrubtion functions are close at a carefully selected finite number of points then they are uniformly close.

By leveraging the boundedness (compactness) of the range of the distribution function, we can get some nice uniform bounds on the growth of that distribution function. Compare the following construction with Lemma 3.95. Let

$$G(y) = \inf\{x \in \mathbb{R} \mid F(x) \geq y\}$$

be the generalized left continuous inverse of $F(x)$. For each positive integer $m > 0$, consider the partition $x_{k,m} = G(\frac{k}{m})$ for $k = 1, \ldots, m-1$. We observe the following facts: by the definition of $G(y)$, for $x < x_{k,m}$, we have $F(x) < \frac{k}{m}$ and by right continuity of $F(x)$ and the definition of $G(y)$, $F(G(y)) \geq y$, so in particular $F(x_{k,m}) \geq \frac{k}{m}$. These two facts provide the following statements

$$\begin{aligned} F(x_{k+1,m}-) - F(x_{k,m}) &\leq \frac{1}{m} && \text{for } 1 \leq k < m-1 \\ F(x_{1,m}-) &\leq \frac{1}{m} \\ F(x_{m-1,m}) &\geq 1 - \frac{1}{m} \end{aligned}$$

Now, for each $m > 0$, $n > 0$ and $\omega \in \Omega$, define

$$D_{n,m}(\omega) = \max(\max_k \left|\hat{F}_n(x_{m,k}, \omega) - F(x_{k,m})\right|, \max_k \left|\hat{F}_n(x_{m,k}-, \omega) - F(x_{k,m}-)\right|)$$

and we proceed to use this quantity to bound the distance between $\hat{F}_n(x, \omega)$ and $F(x)$.

First, observe the bound for $x < x_{k,m}$ for $1 \leq k \leq m-1$,

$$\begin{aligned} \hat{F}_n(x, \omega) &\leq \hat{F}_n(x_{k,m}-, \omega) \\ &\leq F(x_{k,m}-) + D_{n,m}(\omega) && \text{by definition of } D_{n,m}(\omega) \\ &\leq F(x) + \frac{1}{m} + D_{n,m}(\omega) \end{aligned}$$

and for $x \geq x_{k,m}$ for $1 \leq k \leq m-1$

$$\begin{aligned} \hat{F}_n(x, \omega) &\geq \hat{F}_n(x_{k,m}, \omega) \\ &\geq F(x_{k,m}) - D_{n,m}(\omega) \\ &\geq F(x) - \frac{1}{m} - D_{n,m}(\omega) \end{aligned}$$

When we put these together for $x \in [x_{k,m}, x_{k+1,m})$ for $1 \leq k < m-1$ and we have

$$\sup_{x_{1,m} \leq x < x_{m-1,m}} \left|\hat{F}_n(x, \omega) - F(x)\right| < \frac{1}{m} + D_{n,m}(\omega)$$

It remains to complete the picture of what happens when $x < x_{1,m}$ and $x \geq x_{m-1,m}$.

For $-\infty < x < x_{1,m}$, we have

$$\begin{aligned} \hat{F}_n(x, \omega) &\geq 0 \\ &\geq F(x) - \frac{1}{m} \\ &\geq F(x) - \frac{1}{m} - D_{n,m}(\omega) \end{aligned}$$

and lastly we have for $x \geq x_{m-1,m}$,

$$\begin{aligned}\hat{F}_n(x,\omega) &\leq 1 \\ &\leq F(x) + \frac{1}{m} \\ &\leq F(x) + \frac{1}{m} + D_{n,m}(\omega)\end{aligned}$$

which allows us to extend for all $x \in \mathbb{R}$,

$$\sup_x \left|\hat{F}_n(x,\omega) - F(x)\right| < \frac{1}{m} + D_{n,m}(\omega)$$

Now for each m, $\lim_{n\to\infty} D_{n,m} = 0$ a.s. by Lemma 8.24 and by taking a countable union of sets of probability zero, we have for all $m > 0$, $\lim_{n\to\infty} D_{n,m} = 0$ a.s. Therefore by taking the limit as $m \to \infty$ and $n \to \infty$, we have result. □

We now take a short digression into statistics to show how the Glivenko-Cantelli Theorem can be used. The approach taken in demonstrating the result below has far reaching generalizations; don't let the epsilons and deltas distract you from appreciating the conceptual framework.

Definition 8.26. Let P be a Borel probability measure on $\mathbb{R}$ with distribution function $F(x) = \mathbf{E}\left[\mathbf{1}_{(-\infty,x]}\right]$. We define the *median* of P to be $\text{Med}(P) = \inf_x\{F(x) \geq \frac{1}{2}\}$. If ξ is a random variable then we will often write $\text{Med}(\xi)$ for the median of the distribution of ξ.

Lemma 8.27. *Let $\xi_1, \xi_2, \ldots$ be i.i.d. random variables and distribution function $F(x)$. Suppose that $F(x) > \frac{1}{2}$ for all $x > \text{Med}\,\xi$. The sample median $\lim_{n\to\infty} \text{Med}(P_n) = \text{Med}(\xi)$ a.s.; one says that the sample median is a* strongly consistent *estimator of* $\text{Med}(\xi)$.

Proof. The key to the proof is viewing the median as a functional on the space of distribution functions. The Glivenko-Cantelli Theorem tells us that empirical distributions functions converge uniformly so what we need to prove convergence of the sample medians is a continuity property of the median functional. We develop the required continuity property in bare handed way without talking about metric spaces or topologies.

Suppose we have two Borel probability measures P and Q with distribution functions $F_P(x)$ and $F_Q(x)$ with $F_P(x) > \frac{1}{2}$ for $x > \text{Med}(P)$. Given $\epsilon > 0$, pick $\delta > 0$ such that

$$\begin{aligned}F_P(\text{Med}(P) - \epsilon) &< \text{Med}(P) - \delta \\ F_P(\text{Med}(P) + \epsilon) &> \text{Med}(P) + \delta\end{aligned}$$

We claim that if Q satisfies $\sup_x |F_P(x) - F_Q(x)| \leq \delta$ then $|\text{Med}(P) - \text{Med}(Q)| \leq \epsilon$.

To see this first note that

$$F_P(\text{Med}(Q)) \geq F_Q(\text{Med}(Q)) - \delta \geq \frac{1}{2} - \epsilon$$

which implies that $\text{Med}(Q) \geq \text{Med}(P) - \epsilon$ by choice of δ and the increasing nature of $F_P(x)$. Secondly note that for any $x < \text{Med}(Q)$ we have

$$F_P(x) \leq F_Q(x) + \delta < \frac{1}{2} + \epsilon$$

which implies $x < \text{Med}(P) + \epsilon$ and therefore by arbitraryness of x, we have $\text{Med}(Q) \leq \text{Med}(P) + \epsilon$ and we are done with the claim.

Now as per our plan we couple the continuity just proven with Glivenko-Cantelli to derive the result. □

Note that the value $\sup_x \left|\hat{F}_n(x) - F(x)\right|$ is called the *Kolmogorov-Smirnov statistic* and is used in the nonparametric *Kolmogorov-Smirnov Test* for goodness of fit. The Glivenko-Cantelli Theorem tells us that this is a consistent estimator of goodness of fit, however the test itself requires information on the rate of convergence. The most common result in this area is *Donsker's Theorem*. Mention the DKW Inequality too; weak forms of this can be established using the Pollard proof of Glivenko Cantelli which the one that generalizes to Vapnik-Chervonenkis families. We can develop that proof after we do some exponential inequalities.

TODO: Mention that there are generalizations of these results in the closely related fields of Empirical Process Theory and Statistical Learning Theory. One of the goals of such generalizations is to prove consistency of more general statistics derived from the empirical measure.

8.3. **Convergence In Distribution.** Our first goal is to establish that convergence in distribution is implied by convergence in probability.

Lemma 8.28. *Let $\xi, \xi_1, \xi_2, \ldots$ be a random elements in a metric space (S, d) such that $\xi_n \xrightarrow{P} \xi$, then $\xi_n \xrightarrow{d} \xi$.*

Proof. Pick a bounded continous function $f : S \to \mathbb{R}$, then $\mathbf{E}[f(\xi_n)]$. By Lemma 8.11 we know that $f(\xi_n) \xrightarrow{P} f(\xi)$. Because f is bounded, we know that $f(\xi_n)$ and $f(\xi)$ are integrable and therefore $f(\xi_n) \xrightarrow{L^1} f(\xi)$ which implies the result. □

Example 8.29 (Sequence converging in distribution but not in probability)**.** Consider the binary expansion of real numbers in $[0, 1]$, $x = 0.\xi_1\xi_2\cdots$ and consider each ξ_i as a random variable on the probability space $([0, 1], \mathcal{B}([0, 1]), \lambda)$. We claim that ξ_i converge in distribution to the uniform distribution on $\{0, 1\}$ but that the ξ_i diverge in probability. We know from Lemma 7.30 that the ξ_i are i.i.d. Bernoulli random variables with rate $\frac{1}{2}$ so the convergence in distribution follows. If the ξ_i converge in probability, there is a subsequence that converges almost surely.

By independence of the ξ_i, we know that for any $i \neq j$

$$\begin{aligned}\mathbf{P}\{\xi_i \neq \xi_j\} &= \mathbf{P}\{\xi_i = 0 \text{ and } \xi_j = 1\} + \mathbf{P}\{\xi_i = 1 \text{ and } \xi_j = 0\} \\ &= \mathbf{P}\{\xi_i = 0\}\mathbf{P}\{\xi_j = 1\} + \mathbf{P}\{\xi_i = 1\}\mathbf{P}\{\xi_j = 0\} = \frac{1}{2}\end{aligned}$$

and therefore for $i \neq j$,

$$\mathbf{E}[d(\xi_i, \xi_j) \wedge 1] = \mathbf{E}[d(\xi_i, \xi_j)] = \mathbf{P}\{\xi_i \neq \xi_j\} = \frac{1}{2}$$

and we conclude that ξ_i has no subsequence that is Cauchy in probability and hence ξ_i does not converge in probability.

Example 8.30 (Sequence converging in distribution but diverging in mean)**.** Let ξ_n be random variable which takes the value n^2 with probability $\frac{1}{n}$ and takes the

value 0 with probability $\frac{n-1}{n}$. Note that $\lim_{n\to\infty} \xi_n = \lim_{n\to\infty} n = \infty$. On the other hand, if we let f be a bounded continuous function then

$$\begin{aligned}\lim_{n\to\infty} \mathbf{E}\left[f(\xi_n)\right] &= \lim_{n\to\infty} \frac{n-1}{n} f(0) + \lim_{n\to\infty} \frac{1}{n} f(n^2) \\ &= f(0)\end{aligned}$$

where we have used the boundedness of f. Therefore, $\xi_n \xrightarrow{d} \delta_0$ even though it diverges in mean.

Lemma 8.31. *Let ξ_n be a sequence of real valued random variables that converge in distribution to a random variable ξ that is almost surely a constant, then ξ_n converges to ξ in probability as well.*

Proof. Suppose that ξ_n converges in distribution to $c \in \mathbb{R}$. Note that the function $f(x) = |x - c| \wedge 1$ is bounded and continuous and therefore we know

$$\lim_{n\to\infty} \mathbf{E}\left[|\xi_n - c| \wedge 1\right] = \mathbf{E}\left[|c - c| \wedge 1\right] = 0$$

which, by Lemma 8.8, shows that ξ_n converges to c in probability as well. □

The definition we have given for convergence in distribution has the advantage of applying to general random elements in metric spaces but that comes at the cost of being a bit abstract. It is worth connecting the abstract definition with more direct criteria that apply for random variables.

In fact the first equivalence is for discrete random variables. Given that our definition of convergence in distribution is in terms of metric spaces, we must be specific about the metric that we put on the range a discrete random variable. For discussing convergence in distribution the primary feature that we are concerned with is the defintion of continuous functions. If we put a metric

$$d(x, y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{if } x = y \end{cases}$$

then all functions are continuous. Note that the same is true if we consider the induced metric $\mathbb{Z} \subset \mathbb{R}$.

Lemma 8.32. *Let $\xi, \xi_1, \xi_2, \ldots$ be a sequence of discrete random variables with countable range S. Then $\xi_n \xrightarrow{d} \xi$ if and only if for every $x \in S$, we have $\lim_{n\to\infty} \boldsymbol{P}\{\xi_n = x\} = \boldsymbol{P}\{\xi = x\}$.*

Proof. First let's assume that $\xi_n \xrightarrow{d} \xi$. From the discussion preceeding the Lemma, we know that for any bounded function $f : S \to \mathbb{R}$, we have $\lim_{n\to\infty} \mathbf{E}\left[f(\xi_n)\right] = \mathbf{E}\left[f(\xi)\right]$. In particular, for each $x \in S$, we may take $f(y) = \mathbf{1}_x(y)$ in which case we have $\lim_{n\to\infty} \mathbf{P}\{\xi_n = x\} = \mathbf{P}\{\xi = x\}$ as required.

So now assume the converse. In the following, it is helpful to label the elements of S using the natural numbers. Note that we can cast our assumption as saying that for every $x_j \in S$,

$$\lim_{n\to\infty} \mathbf{E}\left[\mathbf{1}_{x_j}(\xi_n)\right] = \mathbf{E}\left[\mathbf{1}_{x_j}(\xi)\right]$$

Furthermore, any bounded function can be written as a linear combination $f(y) = \sum_{j=1}^{\infty} f_j \cdot \mathbf{1}_{x_j}(y)$. By linearity of expectation and our assumption it is trivial to see

that for any finite linear combination $f_N(y) = \sum_{j=1}^N f_j \cdot \mathbf{1}_{x_j}(y)$, we in fact have

$$\lim_{n\to\infty} \mathbf{E}\left[f_N(\xi_n)\right] = \mathbf{E}\left[f_N(\xi)\right]$$

and our task is to extend this to general infinite sums. Let $M > 0$ be a bound for f defined as above.

Pick an $\epsilon > 0$. Since $\sum_{j=1}^\infty \mathbf{P}\{\xi = x_j\} = 1$ we can find $J > 0$ such that $\sum_{j=1}^J \mathbf{P}\{\xi = x_j\} > 1 - \epsilon$. For each $j = 1, \ldots, J$ we can find $N_j > 0$ such that $|\mathbf{P}\{\xi = x_j\} - \mathbf{P}\{\xi = x_j\}| < \frac{\epsilon}{J}$ for $n > N_j$. Now take $N = \max(N_1, \ldots, N_J)$ and then we have for all $n > N$, $\sum_{j=1}^J \mathbf{P}\{\xi_n = x_j\} > 1 - 2\epsilon$. If we let $f_j = f(x_j)$ for each $x_j \in S$, then we have the following calculation

$$\begin{aligned}
|\mathbf{E}\left[f(\xi_n) - f(\xi)\right]| &\leq \sum_{j=1}^J f_j \left|\mathbf{P}\{\xi_n = x_j\} - \mathbf{P}\{\xi = x_j\}\right| + \left|\sum_{j=J+1}^\infty f_j \mathbf{P}\{\xi_n = x_j\}\right| + \left|\sum_{j=J+1}^\infty f_j \mathbf{P}\{\xi = x_j\}\right| \\
&\leq \sum_{j=1}^J |f_j| \frac{\epsilon}{J} + 2M\epsilon + M\epsilon < 4M\epsilon
\end{aligned}$$

Since $\epsilon > 0$ was arbitrary we have $\lim_{n\to\infty} \mathbf{E}\left[f(\xi_n)\right] = \mathbf{E}\left[f(\xi)\right]$ and we are done. □

In the case of general random variables, we can also characterize convergence in distribution by looking at pointwise convergence of distribution functions and using a proof similar in spirit to that used above for discrete random variables, but it comes with a subtle twist.

Lemma 8.33. *Let $\xi, \xi_1, \xi_2, \ldots$ be sequence of random variables with distribution functions $F(x), F_1(x), F_2(x), \ldots$. If $\xi_n \xrightarrow{d} \xi$ then $\lim_{n\to\infty} F_n(x) = F(x)$ for all $x \in \mathbb{R}$ such that F is continuous at x. Conversely, if $\lim_{n\to\infty} F_n(x) = F(x)$ on a dense subset of $\mathbb{R}$ then $\xi_n \to \xi$.*

Proof. Let us first assume that $\xi_n \to \xi$. Consider a function $\mathbf{1}_{(-\infty,x]}$ for $x \in \mathbb{R}$ so that $F(x) = \mathbf{E}\left[\mathbf{1}_{(-\infty,\xi]}\right]$ and $F_n(x) = \mathbf{E}\left[\mathbf{1}_{(-\infty,\xi_n]}\right]$. Note that we cannot just apply the definition of convergence in distribution to derive the result because $\mathbf{1}_{(-\infty,x]}$ is not continuous; so our goal is to extend to defining property of convergence in distribution to a particular class of discontinuous functions. The way to do this is to approximate by continuous functions. To this end, define for each integer $x \in \mathbb{R}$, $m > 0$ the following bounded continuous approximations of the indicator function $\mathbf{1}_{(-\infty,x]}$:

$$f^+_{x,m}(y) = \begin{cases} 1 & \text{if } y \leq x \\ m(x-y)+1 & \text{if } x < y < x + \frac{1}{m} \\ 0 & \text{if } x + \frac{1}{m} \leq y \end{cases}$$

and

$$f^-_{x,m}(y) = \begin{cases} 1 & \text{if } y \leq x - \frac{1}{m} \\ m(x-y) & \text{if } x - \frac{1}{m} < y < x \\ 0 & \text{if } x \leq y \end{cases}$$

and note that $f^-_{x,m}(y) < \mathbf{1}_{(-\infty,x]}(y) < f^+_{x,m}(y)$ and

$$\begin{aligned}
\mathbf{E}\left[f^-_{x,m}(\xi)\right] &= \lim_{n\to\infty} \mathbf{E}\left[f^-_{x,m}(\xi_n)\right] \\
&\leq \liminf_{n\to\infty} \mathbf{E}\left[\mathbf{1}_{(-\infty,x]}(\xi_n)\right] \\
&= \liminf_{n\to\infty} F_n(x) \\
&\leq \limsup_{n\to\infty} F_n(x) \\
&= \limsup_{n\to\infty} \mathbf{E}\left[\mathbf{1}_{(-\infty,x]}(\xi_n)\right] \\
&\leq \limsup_{n\to\infty} \mathbf{E}\left[f^+_{x,m}(\xi_n)\right] \\
&= \lim_{n\to\infty} \mathbf{E}\left[f^+_{x,m}(\xi_n)\right] = \mathbf{E}\left[f^+_{x,m}(\xi)\right]
\end{aligned}$$

But we also can see that for every $x, y \in \mathbb{R}$, $\lim_{m\to\infty} f^-_{x,m}(y) = \mathbf{1}_{(-\infty,x)}(y)$ and $\lim_{m\to\infty} f^+_{x,m}(y) = \mathbf{1}_{(-\infty,x]}(y)$. By application of Dominated Convergence, we see that $\lim_{m\to\infty} \mathbf{E}\left[f^-_{x,m}(\xi)\right] = F(x-)$ and $\lim_{m\to\infty} \mathbf{E}\left[f^+_{x,m}(\xi)\right] = F(x)$ so if x is a point of continuity of F then $F(x-) = F(x)$ which shows $\liminf_{n\to\infty} F_n(x) = \limsup_{n\to\infty} F_n(x) = F(x)$.

Now let's assume that we have a dense set $D \subset \mathbb{R}$ with $\lim_{n\to\infty} F_n(x) = F(x)$ for all $x \in D$. Pick a bounded continuous function $f : \mathbb{R} \to \mathbb{R}$ and we must show $\lim_{n\to\infty} \mathbf{E}\left[f(\xi_n)\right] \to \mathbf{E}\left[f(\xi)\right]$. We will again make an approximation argument. To see how to proceed, recast our hypothesis as the statement that $\lim_{n\to\infty} \mathbf{E}\left[\mathbf{1}_{(-\infty,x]}(\xi_n)\right] \to \mathbf{E}\left[\mathbf{1}_{(-\infty,x]}(\xi)\right]$ for every $x \in D$ and note that by taking sums of functions of the form $\mathbf{1}_{(-\infty,x]}(y)$ allows us to create step functions. So, the idea of the proof is to carefully approximate f by step functions so that we may leverage our hypothesis.

We pick $\epsilon > 0$. First it is helpful to allow ourselves to concentrate on a finite subinterval of the reals. As F is a distribution function, we know $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{n\to\infty} F(x) = 1$ and therefore by density of D we may find $r, s \in D$ such that $F(r) \leq \frac{\epsilon}{2}$ and $F(s) \geq 1 - \frac{\epsilon}{2}$. Because $\lim_{n\to\infty} F_n(x) = F(x)$ for $x \in D$, we can find and $N_1 > 0$ such that $F_n(r) \leq \epsilon$ and $F_n(s) \geq 1 - \epsilon$ for $n > N_1$.

Now we turn our attention to the approximation of f and note that by compactness of $[r, s]$ we know that we can find a finite partition $r_0 = r < r_1 < \cdots < r_{m-1} < r_m = s$ such that $r_j \in D$ and $|f(r_j) - f(r_{j-1})| \leq \epsilon$ for $1 \leq j \leq m$. To see this we know that f is uniformly continuous on $[r, s]$ and therefore there exists $\delta > 0$ such that for any $x, y \in [r, s]$ with $|x - y| < \delta$ we have $|f(x) - f(y)| < \epsilon$. We construct r_j inductively starting with $r_0 = r$. Using uniform continuity as above and the density of D, given r_{j-1} we can find r_j with $r_{j-1} + \frac{\delta}{2} \leq r_j < r_{j-1} + \delta$ and we know that $|f(r_j) - f(r_{j-1})| < \epsilon$. In less than $\lceil \frac{2(s-r)}{\delta} \rceil$ steps we have $|r_j - s| < \delta$ and we terminate the construction. Having constructed the partition, define the step function

$$g(y) = \sum_{j=1}^{m} f(r_j)\left(\mathbf{1}_{(-\infty,r_j]}(y) - \mathbf{1}_{(-\infty,r_{j-1}]}(y)\right) = \sum_{j=1}^{m} f(r_j)\mathbf{1}_{(r_{j-1},r_j]}(y)$$

and note that by construction we have $|f(y) - g(y)| \leq \epsilon$ for all $r \leq y \leq s$.

So now we estimate

$$|\mathbf{E}[f(\xi_n)] - \mathbf{E}[f(\xi)]| \leq |\mathbf{E}[f(\xi_n)] - \mathbf{E}[g(\xi_n)]| + |\mathbf{E}[g(\xi_n)] - \mathbf{E}[g(\xi)]| + |\mathbf{E}[g(\xi)] - \mathbf{E}[f(\xi)]|$$

and consider each term on the left hand side. By boundedness of f we pick $M > 0$ such that $f(x) \leq M$ for all $x \in \mathbb{R}$ and note that since $g(y) = 0$ for $y \leq r$ and $y > s$,

$$\begin{aligned}|\mathbf{E}[f(\xi_n)] - \mathbf{E}[g(\xi_n)]| &\leq |\mathbf{E}[f(\xi_n); \xi_n \leq r]| + |\mathbf{E}[f(\xi_n) - g(\xi_n); r < \xi_n \leq s]| + |\mathbf{E}[f(\xi_n); \xi_n > s]| \\ &\leq \epsilon M + \epsilon + \epsilon M = \epsilon(2M+1)\end{aligned}$$

and similarly,

$$|\mathbf{E}[f(\xi)] - \mathbf{E}[g(\xi)]| \leq \frac{\epsilon}{2}M + \epsilon + \frac{\epsilon}{2}M = \epsilon(M+1)$$

Now leveraging the fact that $\lim_{n\to\infty} F_n(r_j) = F(r_j)$ for every $0 \leq j \leq m$ and the finiteness of this set, we can pick $N_2 > 0$ such that $|F_n(r_j) - F(r_j)| \leq \frac{\epsilon}{2mM}$ for all $n > N_2$ and all $0 \leq j \leq m$. Using this fact and the definition of g,

$$\begin{aligned}|\mathbf{E}[g(\xi_n)] - \mathbf{E}[g(\xi)]| &= \left|\sum_{j=1}^{m} f(r_j)\left(\mathbf{E}\left[\mathbf{1}_{(-\infty,r_j]}(\xi_n)\right] - \mathbf{E}\left[\mathbf{1}_{(-\infty,r_{j-1}]}(\xi_n)\right] - \mathbf{E}\left[\mathbf{1}_{(-\infty,r_j]}(\xi)\right] + \mathbf{E}\left[\mathbf{1}_{(-\infty,r_{j-1}]}(\xi)\right]\right)\right| \\ &= \left|\sum_{j=1}^{m} f(r_j)\left(F_n(r_j) - F_n(r_{j-1}) - F(r_j) + F(r_{j-1})\right)\right| \\ &\leq \sum_{j=1}^{m} |f(r_j)|\left(|F_n(r_j) - F(r_j)| + |F_n(r_{j-1}) - F(r_{j-1})|\right) \\ &\leq \epsilon\end{aligned}$$

for every $n > N_2$.

Putting these three bounds together we have for $n > N_1 \wedge N_2$, $|\mathbf{E}[f(\xi_n)] - \mathbf{E}[f(\xi)]| \leq (3M+3)\epsilon$ and we are done. □

Example 8.34. Let ξ_n be a $U(-\frac{1}{n}, \frac{1}{n})$ random variable and let $\xi = 0$ a.s., then $\xi_n \xrightarrow{d} \xi$. Note that the distribution function of ξ_n is

$$F_n(x) = \begin{cases} 1 & \text{if } x \geq \frac{1}{n} \\ \frac{1}{2}(nx+1) & \text{if } -\frac{1}{n} < x < \frac{1}{n} \\ 0 & \text{if } x \leq \frac{1}{n} \end{cases}$$

Then it is clear that $\lim_{n\to\infty} F_n(x) = 0$ for $x < 0$ and $\lim_{n\to\infty} F_n(x) = 1$ for $x > 0$. Since the distribution function of δ_0 is $\mathbf{1}_{[0,\infty)}$ we apply Lemma 8.33 to conclude convergence in distribution. Note that $\lim_{n\to\infty} F_n(0) = \frac{1}{2} \neq F(0) = 1$. It is also worth noting that the pointwise limit of F_n isn't actually a distribution function (e.g. is not right continuous at 0). TODO: Is convergence in distribution easy to prove directly using the defintion?

The theory of convergence in distribution is rather vast and can be studied at many different levels of generality and sophistication. For example, we have stated the basic definitions on general metric spaces and for some of most basic foundations it is no more difficult to prove things in metric spaces than in a more concrete case such as random variables or vectors. However it soon becomes wise to temporarily drop the generality and concentrate on the special case of random vectors (e.g. to prove probably the most famous result of probability: the Central Limit Theorem).

At some point it becomes necessary to return to the general case but at that point one needs to be prepared to bring more powerful tools to the table as the theory becomes much more subtle.

In this section we start the program and deal with those first results in the theory of weak convergence that can be simply dealt with in the context of general metric spaces.

One of the key features of dealing with probability measures (and to a lesser extent measures in general) is that they are very *well behaved* when viewed as functionals (i.e. linear mappings from functions to $\mathbb{R}$). We've left that statement deliberately vague for the moment since is properly understood within the context of the general theory of distributions. What we want to begin exploring is a side effect of this good behavior: namely that weak convergence of probability measures can be characterized by using many different classes of functions other than the bounded continuous ones. In one direction one can prove results that tell us that to prove weak convergence it is not necessary to test with all bounded continuous functions but one only need use some subset of these. In fact, in the case of random variables and random vectors, it is only necessary to test with compactly supported infinitely differentiable functions (which we won't prove quite yet since we're still dealing with general metric spaces). In another direction, knowing that one has a weakly convergent sequence of probability measures one can extend the convergence with test functions to use statements about some classes of discontinuous functions (e.g. indicator functions). Combining both directions, one can characterize weak convergence by testing against certain classes of discontinuous functions.

Our first foray into the plasticity of weak convergence of probability measures is the following set of conditions that characterize weak convergence of Borel probability measures on metric spaces. Before we state the Theorem we need a couple of quick definitions.

Definition 8.35. Let μ be a Borel probability measure on a metric space S. We say that a subset $A \subset S$ is a μ-continuity set if $\mu(\partial A) = 0$.

Definition 8.36. Let (S, d) and (S', d') be metric spaces. We say $f : S \to S'$ is *Lipschitz continuous* if there exists a $C \geq 0$ such that $d(f(x), f(y)) \leq Cd(x, y)$ for all $x, y \in S$. We often such a C a *Lipschitz constant*.

It is often convenient to refer to a Lipschitz continuous function as being Lipschitz.

Example 8.37. As examples of continuous functions that fail to be Lipschitz continous consider $f(x) = x^2$ on $\mathbb{R}$ and $\sin(1/x)$ on $(0, \infty)$. Note that x^2 is Lipschitz on any compact set. This latter fact can be generalized to show that any continuously differentiable function can be shown to be Lipschitz on any compact set.

Lemma 8.38. *A Lipschitz function f is uniformly continuous.*

Proof. Let C be a Lipschitz constant for f. The for $\epsilon > 0$, let $\delta = \frac{\epsilon}{C}$. □

As an example of Lipschitz function that we'll make use of in the next Theorem, consider the following.

Lemma 8.39. *Let $F \subset S$ be a closed subset and define $f(x) = d(x, F) = \inf_{y \in F} d(x, y)$. Then $f(x)$ is Lipschitz with Lipschitz constant 1.*

Proof. Let $\epsilon > 0$, $x, y \in S$ and pick a $z \in F$ such that $f(x) \leq d(x,z) \leq f(x) + \epsilon$. By the triangle inequality, we have

$$f(y) \leq d(y,z) \leq d(x,z) + d(x,y) \leq f(x) + d(x,y) + \epsilon$$

The argument is symmetric in x and y so we also have that

$$f(x) \leq f(y) + d(x,y) + \epsilon$$

and therefore $|f(x) - f(y)| \leq d(x,y) + \epsilon$. Since ϵ was arbitrary let it go to 0 and we are done. □

Lemma 8.40. *Let $f, g : S \to \mathbb{R}$ be Lipschitz with Lipschitz constants C_f and C_g respectively. Then both $f \wedge g$ and $f \vee g$ are Lipschitz with Lipschitz constants $C_f \vee C_g$.*

Proof. The proof is elementary but long winded; we only do the case of $f \wedge g$. Pick $x, y \in S$ and consider $|(f \wedge g)(x) - (f \wedge g)(y)|$. We break the analysis down into four cases.

Case (i): Suppose $(f \wedge g)(x) \geq (f \wedge g)(y)$ and $f(y) \leq g(y)$.

$$|(f \wedge g)(x) - (f \wedge g)(y)| = (f \wedge g)(x) - f(y) \leq f(x) - f(y) \leq C_f d(x,y)$$

Case (ii): Suppose $(f \wedge g)(x) \geq (f \wedge g)(y)$ and $g(y) \leq f(y)$.

$$|(f \wedge g)(x) - (f \wedge g)(y)| = (f \wedge g)(x) - g(y) \leq g(x) - g(y) \leq C_g d(x,y)$$

Case (iii): Suppose $(f \wedge g)(y) \geq (f \wedge g)(x)$ and $f(x) \leq g(x)$.

$$|(f \wedge g)(x) - (f \wedge g)(y)| = (f \wedge g)(y) - f(x) \leq f(y) - f(x) \leq C_f d(x,y)$$

Case (iv): Suppose $(f \wedge g)(y) \geq (f \wedge g)(x)$ and $g(x) \leq f(x)$.

$$|(f \wedge g)(x) - (f \wedge g)(y)| = (f \wedge g)(y) - g(x) \leq g(y) - g(x) \leq C_g d(x,y)$$

Thus we see $|(f \wedge g)(x) - (f \wedge g)(y)| \leq (C_f \vee C_g) d(x,y)$.

The case of $f \vee g$ follows in a similar way. □

Theorem 8.41 (Portmanteau Theorem)**.** *Let μ and μ_n be a sequence of Borel probability measures on a metric space S. The following are equivalent*

(i) *μ_n converge in distribution to μ.*
(ii) *$\boldsymbol{E}_n[f] \to \boldsymbol{E}[f]$ for all bounded Lipschitz functions f.*
(iii) *$\limsup_{n\to\infty} \mu_n(C) \leq \mu(C)$ for all closed sets C*
(iv) *$\liminf_{n\to\infty} \mu_n(U) \geq \mu(U)$ for all open sets U*
(v) *$\lim_{n\to\infty} \mu_n(A) = \mu(A)$ for all μ-continuity sets A.*

Before we begin the proof, we pay particular attention to the fact that one does not have equality in the case of indicator functions. What this is saying is that mass can move out to the boundary during limiting processes of distributions. In the case of open sets that mass can be lost (to the boundary) whereas in the case of closed sets, it can magically appear in the limit. An example here is the limit of point masses $\delta_{\frac{1}{n}}$. It is elementary that $\delta_{\frac{1}{n}} \xrightarrow{d} \delta_0$ but if one considers the open set $(0,1)$, then $\delta_{\frac{1}{n}}(0,1) = 1$ but $\delta_0(0,1) = 0$. In a similar way, take the closed set $\{0\}$ and we see $\delta_{\frac{1}{n}}\{0\} = 0$ but $\delta_0\{0\} = 1$. The statement in (v) neatly captures the idea that the only way we fail to converge with indicator functions is when mass appears on the boundary of the set; if we rule out that possiblity assuming the set is a continuity set then we have convergence when the corresponding indicator function is used as the test function.

Proof. Note that (i) implies (ii) is trivial since a bounded Lipschitz function is also bounded and continuous.

(ii) implies (iv): Suppose we have $U \subset S$ an open set. Let $f_n(x) = (nd(x, U^c)) \wedge 1$. By Lemma 8.39 and Lemma 8.40 we know that $f_n(x)$ is Lipschitz with constant n. It is trivial to see that $f_n(x)$ is increasing. Furthermore $\lim_{n\to\infty} f_n(x) = \mathbf{1}_U(x)$. This can be seen by noting that if $x \in U$, then by taking a ball $B(x, r) \subset U$, we know that $d(x, U^c) \geq r$ and therefore $f_n(x) = 1$ for $n \geq \frac{1}{r}$. On the other hand, it is trivial that $f_n(x) = 0$ for all $x \in U^c$ and all n. Armed with these facts we prove (iv)

$$\begin{aligned}
\mu(U) &= \lim_{n\to\infty} \mathbf{E}[f_n] && \text{by Monotone Convergence Theorem} \\
&= \lim_{n\to\infty} \lim_{m\to\infty} \mathbf{E}_m[f_n] && \text{by (ii)} \\
&\leq \lim_{n\to\infty} \liminf_{m\to\infty} \mathbf{E}_m[\mathbf{1}_U] && \text{since } f_n \leq \mathbf{1}_U \\
&= \liminf_{m\to\infty} \mu_m(U)
\end{aligned}$$

(iii) is equivalent to (iv): Assume (iii) and use the fact $\liminf_{n\to\infty} f_n = -\limsup_{n\to\infty} -f_n$ and (iv) to calculate for an open set U,

$$\liminf_{n\to\infty} \mu_n(U) = -\limsup_{n\to\infty} -\mu_n(U) = -\limsup_{n\to\infty} \mu_n(U^c) + 1 \geq -\mu(U^c) + 1 = \mu(U)$$

The proof that (iv) implies (iii) follows in an analogous way.

(iv) implies (i). Suppose $f \geq 0$ continuous, then for every $\lambda \in \mathbb{R}$, we know that $\{f > \lambda\} = f^{-1}((\lambda, \infty))$ is an open subset of S. Because of that we we may use Lemma 6.8, Fatou's Lemma (Theorem 3.42) and (iii) to see

$$\begin{aligned}
\int f \, d\mu &= \int_0^\infty \mathbf{P}_\mu\{f > \lambda\} \, d\lambda \\
&\leq \int_0^\infty \liminf_{n\to\infty} \mathbf{P}_{\mu_n}\{f > \lambda\} \, d\lambda \\
&\leq \liminf_{n\to\infty} \int_0^\infty \mathbf{P}_{\mu_n}\{f > \lambda\} \, d\lambda \\
&= \liminf_{n\to\infty} \int f \, d\mu_n
\end{aligned}$$

Now we play the same trick as in the proof of Dominated Convergence. Suppose f is bounded and continuous and suppose $|f| \leq c$. By what we have just shown,

$$\begin{aligned}
\int f \, d\mu = -c + \int (c + f) \, d\mu &\leq -c + \liminf_{n\to\infty} \int (c + f) \, d\mu_n = \liminf_{n\to\infty} \int f \, d\mu_n \\
-\int f \, d\mu = -c + \int (c - f) \, d\mu &\leq -c + \liminf_{n\to\infty} \int (c - f) \, d\mu_n = -\limsup_{n\to\infty} \int f \, d\mu_n
\end{aligned}$$

Therefore

$$\limsup_{n\to\infty} \int f \, d\mu_n \leq \int f \, d\mu \leq \liminf_{n\to\infty} \int f \, d\mu_n$$

which implies $\lim_{n\to\infty} \int f \, d\mu_n = \int f \, d\mu$ and (i) is proven.

(iii) and (iv) imply (v). Pick a μ-continuity set A. The first thing to note is that $\mu(A) = \mu(\overline{A}) = \mu(\text{int}(A))$ because they all differ by a subset of ∂A. Now on the one hand,

$$\liminf_{n\to\infty} \mu_n(A) \geq \liminf_{n\to\infty} \mu_n(\text{int}(A)) \geq \mu(\text{int}(A)) = \mu(A)$$

On the other hand,

$$\limsup_{n\to\infty} \mu_n(A) \leq \limsup_{n\to\infty} \mu_n(\overline{A}) \leq \mu(\overline{A}) = \mu(A)$$

which shows that $\lim_{n\to\infty} \mu_n(A) = \mu(A)$.

(v) implies (iii). Pick a closed set and for every $\epsilon > 0$ consider the closed ϵ-neighborhood $F_\epsilon = \{x \mid d(x, F) \leq \epsilon\}$. Note that $\partial F_\epsilon \subset \{x \mid d(x, F) = \epsilon\}$ since if $d(x, F) < \epsilon$ then by continuity of the function $f(y) = d(y, F)$ we can find a ball $B(x, r)$ such that $d(y, F) < \epsilon$ for every $y \in B(x, r)$; thus proving x is in the interior of F_ϵ. The fact that $\partial F_\epsilon \subset \{x \mid d(x, F) = \epsilon\}$ shows that the ∂F_ϵ are disjoint.

Next note that $\mu(\partial F_\epsilon) \neq 0$ for at most a countable number of ϵ. For every $n \geq 1$, there can only be a finite number F_ϵ with $\mu(\partial F_\epsilon) \geq \frac{1}{n}$ because of the disjointness of F_ϵ and the countable additivity of μ. So the set of all ϵ with $\mu(\partial F_\epsilon) > 0$ is a countable union of finite set and therefore countable. Now the complement of a countable set in $\mathbb{R}$ is dense (Lemma 2.17) hence F_ϵ is a μ-continuity set for a dense set of ϵ.

Now deriving (iii) is easy. Pick a decreasing sequence of ϵ_m such that $\lim_{m\to\infty} \epsilon_m = 0$ and each F_{ϵ_m} is a μ-continuity set. Therefore by subadditivity of measure and and our hypothesis, for each m

$$\limsup_{n\to\infty} \mu_n(F) \leq \lim_{n\to\infty} \mu_n(F_{\epsilon_m}) = \mu(F_{\epsilon_m})$$

However, by continuity of measure, we know that

$$\limsup_{n\to\infty} \mu_n(F) \leq \lim_{m\to\infty} \mu(F_{\epsilon_m}) = \mu_n(F)$$

and we're done. □

Definition 8.42. Given metric spaces (S, d) and (S', d') and a map $g : S \to S'$, the set of discontinuity points D_g is the set of $x \in S$ such that for every $\epsilon > 0$ and $\delta > 0$ there exists $y \in S$ such that $d(x, y) < \delta$ and $d'(g(x), g(y)) > \epsilon$.

Theorem 8.43 (Continuous Mapping Theorem)**.** *Let ξ_n and ξ be random elements in a metric space S. Let S' be a metric space such that there exists a map $g : S \to S'$ with the property that the $\boldsymbol{P}\{\xi \in D_g\} = 0$. Then*

(i) *If ξ_n converges in distribution to ξ then $g(\xi_n)$ converges in distribution to $g(\xi)$.*
(ii) *If ξ_n converges in probability to ξ then $g(\xi_n)$ converges in probability to $g(\xi)$.*
(iii) *If ξ_n converges a.s. to ξ then $g(\xi_n)$ converges a.s. to $g(\xi)$.*

Proof. TODO: This proof makes the assumption that g is continuous. This is a big simplification for the distribution case in particular. Provide the proof with the weaker assumption.

To prove (i), suppose we are given a bounded continuous $f : S' \to \mathbb{R}$. Then $f \circ g : S \to \mathbb{R}$ is also bounded and continuous hence

$$\lim_{n\to\infty} \int f(g(\xi_n))\, d\mu = \int f(g(\xi))\, d\mu$$

which shows that $g(\xi_n) \xrightarrow{d} g(\xi)$.

To prove (ii), for every $\epsilon, \delta > 0$, define

$$B_\delta^\epsilon = \{x \in S \mid \exists\, y \in S \text{ with } d(x,y) < \delta \text{ and } d'(g(x), g(y)) \geq \epsilon\}$$

Note that for $\delta' < \delta$ and fixed ϵ we have $B_{\delta'}^\epsilon \subset B_\delta^\epsilon$. Continuity of g implies that $\bigcap_{m=1}^\infty B_{\frac{1}{m}}^\epsilon = \emptyset$; and therefore by continuity of measure (Lemma 3.27) we know that $\lim_{m\to\infty} \mathbf{P}\{\xi \in B_{\frac{1}{m}}^\epsilon\} = 0$.

Now fix $\epsilon, \gamma > 0$ and note that for all $n, m > 0$, we have the bound

$$\mathbf{P}\{d'(g(\xi_n), g(\xi)) \geq \epsilon\} \leq \mathbf{P}\{d(\xi_n, \xi) \geq \frac{1}{m}\} + \mathbf{P}\{\xi \in B_{\frac{1}{m}}^\epsilon\}$$

By the previous observation, we can find an $m > 0$ such that $\mathbf{P}\{\xi \in B_{\frac{1}{m}}^\epsilon\} < \frac{\gamma}{2}$. Having picked such an $m > 0$, since ξ_i converges to ξ in probability, we can find $N > 0$ such that $\mathbf{P}\{d(\xi_n, \xi) \geq \frac{1}{m}\} < \frac{\gamma}{2}$ for all $n > N$.

To prove (iii), simply note that by continuity of g, $\xi_n(\omega) \to \xi(\omega)$ implies $g(\xi_n(\omega)) \to g(\xi(\omega))$. □

The following result is a basic tool in the theory of asymptotic statistics. We state and prove it here because it is a straightforward application of the Portmanteau Theorem, but we'll wait until we've proven the Central Limit Theorem to give examples of how it is applied.

Theorem 8.44 (Slutsky's Theorem)**.** *Let ξ_n and η_n be two sequences of random elements in (S, d) such that $d(\xi_n, \eta_n) \xrightarrow{P} 0$. If ξ is a random element in (S, d) such that $\xi_n \xrightarrow{d} \xi$ in distribution, then $\eta_n \xrightarrow{d} \xi$.*

Proof. By the Portmanteau Theorem (Theorem 8.41) it suffices to show $\mathbf{E}[f(\eta_n)] \to \mathbf{E}[f(\xi)]$ for all bounded Lipschitz functions $f : S \to \mathbb{R}$. Pick such an f and $M, K > 0$ such that $|f(x)| \leq M$ and $|f(x) - f(y)| \leq Kd(x, y)$. Then if we pick $\epsilon > 0$,

$$\begin{aligned}
\lim_{n\to\infty} |\mathbf{E}[f(\eta_n)] - \mathbf{E}[f(\xi_n)]| &\leq \lim_{n\to\infty} \mathbf{E}[|f(\eta_n) - f(\xi_n)|] \\
&\leq \lim_{n\to\infty} \mathbf{E}\left[|f(\eta_n) - f(\xi_n)|\, \mathbf{1}_{d(\eta_n,\xi_n)\leq\epsilon}\right] + \mathbf{E}\left[|f(\eta_n) - f(\xi_n)|\, \mathbf{1}_{d(\eta_n,\xi_n)>\epsilon}\right] \\
&\leq \epsilon K + 2M \lim_{n\to\infty} \mathbf{P}\{d(\eta_n, \xi_n) > \epsilon\} \\
&= \epsilon K
\end{aligned}$$

Since ϵ was arbitrary, we have $\lim_{n\to\infty} \mathbf{E}[f(\eta_n)] = \lim_{n\to\infty} \mathbf{E}[f(\xi_n)] = \mathbf{E}[f(\xi)]$ and we are done. □

Corollary 8.45 (Slutsky's Theorem)**.** *Let ξ_n and η_n be two sequences of random elements in (S, d). If ξ is a random element in (S, d) such that ξ_n converges to ξ in distribution and $c \in S$ is a constant such that η_n converges to c in probability, then for every continuous function f, $f(\xi_n, \eta_n)$ also converges to $f(\xi, c)$ in distribution.*

Proof. The critical observation here is that with the assumptions above the random element (ξ_n, η_n) converges to (ξ, c) in distribution. Then we can apply the Continuous Mapping Theorem (Theorem 8.43) to derive the result. To see $(\xi_n, \eta_n) \xrightarrow{d} (\xi, c)$, first note that $d((\xi_n, \eta_n), (\xi_n, c)) = d(\eta_n, c) \xrightarrow{P} 0$ by assumption. Therefore by the previous lemma, it suffices to show that $(\xi_n, c) \xrightarrow{d} (\xi, c)$. Pick a continuous bounded function $f : S \times S \to \mathbb{R}$ and note that $f(-, c) : S \to \mathbb{R}$ is also continuous and bounded. Therefore $\lim_{n\to\infty} \mathbf{E}[f(\xi_n, c)] = \mathbf{E}[f(\xi, c)]$. □

8.4. **Uniform Integrability.** In this section we introduce the technical notion of uniform integrability of a family of random variables. Informally uniform integrability is the property that the tails of the family of integrable random variables can be simultaneously bounded in expectation. Practically one implication of this property is that one can use a single truncation parameter to approximate all of the random variables in a uniformly integrable family. As an application of this fact we'll observe that the truncation argument proof of the Weak Law of Large Numbers extends from i.i.d. sequences of random variables to uniformly integrable sequences of random variables. It also worth noting that the property of uniform integrability figures prominently in martingale theory.

Definition 8.46. A collection of random variables ξ_t for $t \in T$ is *uniformly integrable* if and only if $\lim_{M\to\infty} \sup_{t\in T} \mathbf{E}[|\xi_t| ; |\xi_t| > M] = 0$.

Example 8.47. A sequence of identically distributed variables ξ_n is uniformly integrable. This can be seen easily by defining $g(x) = |x| \mathbf{1}_{|x|>M}$ and noting that

$$\mathbf{E}[|\xi_n| ; |\xi_n| > M] = \mathbf{E}[g(\xi_n)] = \int g(x)\, d\xi_n$$

by Lemma 3.52 which shows that the expectation is independent of n since $d\xi_n$ is independent of n.

Lemma 8.48. *The random variables ξ_t for $t \in T$ are uniformly integrable if and only if*

(i) $\sup_{t\in T} \boldsymbol{E}[|\xi_t|] < \infty$
(ii) *For every $\epsilon > 0$ there exists $\delta > 0$ such that if $\boldsymbol{P}\{A\} < \delta$ then $\boldsymbol{E}[|\xi_t| ; A] < \epsilon$ for all $t \in T$.*

Proof. First we assume uniform integrability of ξ_t. To prove (i), pick $M > 0$ such that $\mathbf{E}[|\xi_t| ; |\xi_t| > M] < 1$ for all $t \in T$. Then for $t \in T$,

$$\begin{aligned} \mathbf{E}[|\xi_t|] &= \mathbf{E}[|\xi_t| ; |\xi_t| \le M] + \mathbf{E}[|\xi_t| ; |\xi_t| > M] \\ &\le M + 1 \end{aligned}$$

To show (ii), pick $\epsilon > 0$, $M > 0$ such that $\mathbf{E}[|\xi_t| ; |\xi_t| > M] < \frac{\epsilon}{2}$ and $\delta < \frac{\epsilon}{2M}$. Then

$$\begin{aligned} \mathbf{E}[|\xi_t| ; A] &= \mathbf{E}[|\xi_t| ; A \wedge |\xi_t| \le M] + \mathbf{E}[|\xi_t| ; A \wedge |\xi_t| > M] \\ &\le M\delta + \mathbf{E}[|\xi_t| ; |\xi_t| > M] \le \epsilon \end{aligned}$$

Now assume (i) and (ii). Pick $\epsilon > 0$ and $\delta > 0$ as in (ii) and let $M > 0$ be such that $\mathbf{E}[|\xi_t|] \le M$ for all $t \in T$. Pick $N > \frac{M}{\delta}$ and note that

$$\mathbf{P}\{|\xi_t| > N\} \le \frac{\mathbf{E}[|\xi_t|]}{N} \le \frac{M}{N} < \delta$$

so by (ii), $\mathbf{E}[|\xi_t| ; |\xi_t| > N] < \epsilon$ and uniform integrability is proven. □

Here is a simple result that illustrates how the conditions for uniform integrability in the previous Lemma can often be more convenient than the definition.

Lemma 8.49. *Suppose $|\xi_t|^p$ and $|\eta_t|^p$ are both uniformly integrable families of random variables. Then for every $a, b \in \mathbb{R}$, $|a\xi_t + b\eta_t|^p$ is uniformly integrable.*

Proof. By Lemma 8.48 we know that $\sup_t \mathbf{E}\left[|\xi_t|^p\right] < \infty$ and $\sup_t \mathbf{E}\left[|\eta_t|^p\right] < \infty$; equivalently $\sup_t \|\xi_t\|_p < \infty$ and $\sup_t \|\eta_t\|_p < \infty$. Now by the triangle inequality/Minkowski's inequality $\sup_t \|a\xi_t + b\eta_t\|_p \leq a \sup_t \|\xi_t\|_p + b \sup_t \|\eta_t\|_p < \infty$. Thus condition (i) of Lemma 8.48 is shown.

To see condition (ii) of Lemma 8.48, suppose $\epsilon > 0$ is given. By this same Lemma applied to $|\xi_t|^p$ and $|\eta_t|^p$ pick a $\delta > 0$ such that for all A with $\mathbf{P}\{A\} < \delta$ we have $\mathbf{E}\left[|\xi_t|^p ; A\right] \leq \frac{\epsilon}{2^p a^p}$ and $\mathbf{E}\left[|\eta_t|^p ; A\right] \leq \frac{\epsilon}{2^p b^p}$ for all t. Then we have

$$\begin{aligned}
\mathbf{E}\left[|a\xi_t + b\eta_t|^p ; A\right] &= \|a\xi_t \mathbf{1}_A + b\eta_t \mathbf{1}_A\|_p^p \\
&\leq (a\|\xi_t \mathbf{1}_A\|_p + b\|\eta_t \mathbf{1}_A\|_p)^p \\
&\leq \left(a\frac{\epsilon^{1/p}}{2a} + b\frac{\epsilon^{1/p}}{2b}\right)^p = \epsilon
\end{aligned}$$

□

Here is an example that shows that the condition (i) of the previous Lemma is not sufficient to guarantee uniform integrability.

Example 8.50. Here we demonstrate a sequence ξ_n with $\sup_n \mathbf{E}\left[|\xi_n|\right] < \infty$ but ξ_n is not uniformly integrable. Consider the sequence ξ_n constructed in Example 8.15. Recall for that sequence, $\mathbf{E}\left[|\xi_n|\right] = 1$ for all $n > 0$. On the other hand, for any $M > 0$ and $n > 0$ we have

$$\mathbf{E}\left[|\xi_n| ; |\xi_n| > M\right] = \begin{cases} 0 & \text{if } 2^n \leq M \\ 1 & \text{if } 2^n > M \end{cases}$$

and therefore for all $M > 0$ we have $\sup_n \mathbf{E}\left[|\xi_n| ; |\xi_n| > M\right] = 1$.

TODO: Show convergence result for uniformly integrable sequences.

While we have shown that convergence in probability is strictly weaker than convergence in mean, it turns out that adding the condition of uniform integrability is precisely what is needed to make them equivalent. Before proving that result we have a Lemma that illustrates the connection between uniform integrability and convergence of means.

Lemma 8.51. *Let $\xi, \xi_1, \xi_2, \ldots$ be positive random variables such that $\xi_n \xrightarrow{d} \xi$, then $\boldsymbol{E}[\xi] \leq \liminf_{n\to\infty} \boldsymbol{E}[\xi_n]$. Moreover, $\boldsymbol{E}[\xi] = \lim_{n\to\infty} \boldsymbol{E}[\xi_n] < \infty$ if and only if ξ_n are uniformly integrable.*

Proof. To see the first inequality, note that for any $R \geq 0$, the function

$$f_R(x) = \begin{cases} R & x > R \\ x & 0 \leq x \leq R \\ 0 & x < 0 \end{cases}$$

is bounded and continuous and for fixed x, $f_R(x)$ is increasing in R. The first inequality follows:

$$\begin{aligned}
\mathbf{E}[\xi] &= \lim_{R\to\infty} \mathbf{E}[f_R(\xi)] && \text{by Monotone Convergence Theorem}\\
&= \lim_{R\to\infty} \lim_{n\to\infty} \mathbf{E}[f_R(\xi_n)] && \text{because } \xi_n \xrightarrow{d} \xi\\
&\le \liminf_n \mathbf{E}[\xi_n] && \text{because } f_R(x) \le x \text{ for all } x \ge 0
\end{aligned}$$

An alternative derivation is:

$$\begin{aligned}
\mathbf{E}[\xi] &= \int \mathbf{P}\{\xi > \lambda\}\, d\lambda && \text{by Lemma 6.8}\\
&\le \int \liminf_n \mathbf{P}\{\xi_n > \lambda\}\, d\lambda && \text{by Portmanteau Lemma 8.41}\\
&\le \liminf_n \int \mathbf{P}\{\xi_n > \lambda\}\, d\lambda && \text{by Fatou's Lemma (Theorem 3.42)}\\
&= \liminf_n \mathbf{E}[\xi_n] && \text{by Lemma 6.8}
\end{aligned}$$

Now assume that ξ_n is uniformly integrable. Then by what we have just proven and Lemma 8.48 we have

$$\mathbf{E}[\xi] \le \liminf_n \mathbf{E}[\xi_n] \le \sup_n \mathbf{E}[\xi_n] < \infty$$

So now we use the triangle inequality to write

$$|\mathbf{E}[\xi_n] - \mathbf{E}[\xi]| \le |\mathbf{E}[\xi_n] - \mathbf{E}[f_R(\xi_n)]| + |\mathbf{E}[f_R(\xi_n)] - \mathbf{E}[f_R(\xi)]| + |\mathbf{E}[f_R(\xi)] - \mathbf{E}[\xi]|$$

We take the limit as n goes to infinity and then as R goes to infinity and consider each term on the right side in turn.

For the first term:

$$\begin{aligned}
\lim_{R\to\infty} \limsup_n |\mathbf{E}[\xi_n] - \mathbf{E}[f_R(\xi_n)]| &= \lim_{R\to\infty} \limsup_n (\mathbf{E}[\xi_n; \xi_n > R] - R\mathbf{P}\{\xi_n > R\})\\
&\le \lim_{R\to\infty} \limsup_n \mathbf{E}[\xi_n; \xi_n > R] - \lim_{R\to\infty} \liminf_n R\mathbf{P}\{\xi_n > R\}\\
&= 0
\end{aligned}$$

where in the last line we have used uniform integrability of ξ_n as well as the following

$$\lim_{R\to\infty} R \liminf_n \mathbf{P}\{\xi_n > R\} \le \lim_{R\to\infty} R \sup_n \mathbf{P}\{\xi_n > R\} \le \lim_{R\to\infty} \sup_n \mathbf{E}[\xi_n; \xi_n > R] = 0$$

The second term we have $\limsup_n |\mathbf{E}[f_R(\xi_n)] - \mathbf{E}[f_R(\xi)]| = 0$ because f_R is bounded continuous and $\xi_n \xrightarrow{d} \xi$. The third term we have $\lim_{R\to\infty} |\mathbf{E}[f_R(\xi)] - \mathbf{E}[\xi]| = 0$ by Monotone Convergence.

Putting the bounds on the three terms of the right hand side together we have $\limsup_{n\to\infty} |\mathbf{E}[\xi_n] - \mathbf{E}[\xi]| = 0$ which by positivity shows $\lim_{n\to\infty} |\mathbf{E}[\xi_n] - \mathbf{E}[\xi]| = 0$.

TODO: Here is an alternative proof of the same fact by approximating $x\mathbf{1}_{x\le R}$ from above by continuous functions. I might like this proof better (since I came up with it?)

Now assume that $\lim_{n\to\infty} \mathbf{E}[\xi_n] = \mathbf{E}[\xi] < \infty$ and we need to show uniform integrability of ξ_n. The idea is to approximate $x\mathbf{1}_{x\ge R}$ by a continuous function so that we can use the weak convergence of ξ_n. The trick is that this function

isn't bounded but is the difference between a bounded function and the function $f(x) = x$; the behavior of this latter function is covered by the hypothesis that the means converge. To make all of this precise, define the following bounded continuous function

$$g_R(x) = x \wedge (R - x)_+ = \begin{cases} 0 & \text{if } x < 0 \text{ or } x > R \\ x & \text{if } 0 \le x \le \frac{R}{2} \\ R - x & \text{if } \frac{R}{2} < x \le R \end{cases}$$

and note that

$$x - g_R(x) = \begin{cases} 0 & \text{if } x < \frac{R}{2} \\ 2x - R & \text{if } \frac{R}{2} \le x \le R \\ x & \text{if } R < x \end{cases}$$

so $x\mathbf{1}_{x \ge R} \le x - g_R(x) \le x$, and $\lim_{R\to\infty} x - g_R(x) = 0$. Putting these facts together we see

$$\begin{aligned} &\lim_{R\to\infty} \limsup_{n\to\infty} \mathbf{E}\left[\xi_n; \xi_n \ge R\right] \\ &\le \lim_{R\to\infty} \limsup_{n\to\infty} \mathbf{E}\left[\xi_n - g_R(\xi_n)\right] \\ &= \lim_{R\to\infty} \left(\lim_{n\to\infty} \mathbf{E}\left[\xi_n\right] - \lim_{n\to\infty} \mathbf{E}\left[g_R(\xi_n)\right] \right) \\ &= \lim_{R\to\infty} \mathbf{E}\left[\xi\right] - \mathbf{E}\left[g_R(\xi)\right] && \text{by assumption and } \xi_n \xrightarrow{d} \xi \\ &= \lim_{R\to\infty} \mathbf{E}\left[\xi - g_R(\xi)\right] = 0 && \text{by Dominated Convergence} \end{aligned}$$

□

Converge in mean and convergence of means become equivalent in the presence of almost sure convergence.

Lemma 8.52. *Suppose $\xi, \xi_1, \xi_2, \ldots$ are random variables*

(i) $\xi_n \xrightarrow{L^p} \xi$ *implies* $\|\xi_n\|_p \to \|\xi\|_p$

(ii) *If* $\xi_n \xrightarrow{a.s.} \xi$ *and* $\|\xi_n\|_p \to \|\xi\|_p$ *then* $\xi_n \xrightarrow{L^p} \xi$

Proof. To see (i), suppose $\xi_n \xrightarrow{L^p} \xi$ and note that by the triangle inequality,

$$\lim_{n\to\infty} \|\xi_n\|_p \le \lim_{n\to\infty} \|\xi_n - \xi\|_p + \|\xi\|_p = \|\xi\|_p$$

and

$$\|\xi\|_p \le \lim_{n\to\infty} \|\xi_n - \xi\|_p + \lim_{n\to\infty} \|\xi_n\|_p = \lim_{n\to\infty} \|\xi_n\|_p$$

therefore $\lim_{n\to\infty} \|\xi_n\|_p = \|\xi\|_p$.

To see (ii), if $\xi_n \xrightarrow{a.s.} \xi$ and $\|\xi_n\|_p \to \|\xi\|_p$ then we know that $|\xi_n - \xi|^p \xrightarrow{a.s.} 0$ and we have the bound

$$|\xi_n - \xi|^p \le (|\xi_n| + |\xi|)^p \le 2^p \max(|\xi_n|^p, |\xi|^p) \le 2^p \left(|\xi_n|^p + |\xi|^p\right)$$

and our assumption tells us that $\lim_{n\to\infty} 2^p \mathbf{E}\left[|\xi_n|^p + |\xi|^p\right] = 2^{p+1} \|\xi\|_p^p < \infty$. Therefore we can apply Dominated Convergence (Theorem 3.48) to conclude that $\lim_{n\to\infty} \|\xi_n - \xi\|_p = 0$. □

To summarize and complete the discussion, we have the following

TODO: Fix the statement here; this is taken from Kallenberg but it feels imprecise to me (e.g. the equivalence of (ii) and (iii) doesn't really require convergence in probability but only convergence in distribution; (i) implies convergence in probability (by Markov)). The only new content here is the extension of (ii) implies (i) to the context of almost sure convergence to convergence in probability by the argument along subsequences).

Lemma 8.53. *Let $\xi, \xi_1, \xi_2, \ldots$ be random variables in L^p for $p > 0$ and suppose $\xi_n \xrightarrow{P} \xi$. Then the following are equivalent:*

(i) $\xi_n \xrightarrow{L^p} \xi$
(ii) $\|\xi_n\|_p \to \|\xi\|_p$
(iii) *The sequence of random variables $|\xi_1|^p, |\xi_2|^p, \ldots$ is uniformly integrable.*

Proof. To see (i) implies (ii), this is the first part of Lemma 8.52.

Note that since $\xi_n \xrightarrow{P} \xi$ implies $\xi_n \xrightarrow{d} \xi$ we know that (ii) and (iii) are equivalent by Lemma 8.51.

To see that (ii) implies (i), suppose that $\|\xi_n - \xi\|_p$ does not converge to zero. They there exists an $\epsilon > 0$ and a subsequence $N' \subset \mathbb{N}$ such that $\|\xi_n - \xi\|_p \geq \epsilon$ along N'. Since $\xi_n \xrightarrow{P} \xi$ by Lemma 8.9 there is a further subsequence $N'' \subset N'$ such that $\xi_n \xrightarrow{a.s.} \xi$ along N''. However, Lemma 8.52 tells us that $\|\xi_n - \xi\|_p$ converges to 0 along N'' which is a contradiction.

An alternative argument is to show that (iii) implies (i) directly. Since we have $|\xi_n|^p$ is uniformly integrable and trivially the singleton collection $|\xi|^p$ is uniformly integrable, it follows from Lemma 8.49 that $|\xi_n - \xi|^p$ is uniformly integrable. Now suppose $\epsilon > 0$ is given and take $R > \epsilon$ so that by use of convergence in probability and uniform integrability we get

$$\begin{aligned}
\lim_{n\to\infty} \mathbf{E}\left[|\xi_n - \xi|^p\right] &= \lim_{R\to\infty} \limsup_{n\to\infty} \left(\mathbf{E}\left[|\xi_n - \xi|^p ; |\xi_n - \xi|^p \leq \epsilon\right] + \mathbf{E}\left[|\xi_n - \xi|^p ; \epsilon < |\xi_n - \xi|^p < R\right] + \mathbf{E}\left[|\xi_n - \xi|^p ; |\xi_n \right.\right. \\
&\leq \epsilon + \lim_{R\to\infty} \lim_{n\to\infty} R\mathbf{P}\{\epsilon < |\xi_n - \xi|^p\} + \lim_{R\to\infty} \sup_n \mathbf{E}\left[|\xi_n - \xi|^p ; |\xi_n - \xi|^p \geq R\right] \\
&= \epsilon
\end{aligned}$$

and since $\epsilon > 0$ was arbitrary, we have $\lim_{n\to\infty} \mathbf{E}\left[|\xi_n - \xi|^p\right] = 0$. □

TODO: Show how the proof of the Weak Law of Large Numbers applies to uniformly integrable sequences not just i.i.d.

Lemma 8.54. *ξ_t is uniformly integrable if and only if there exists a convex and increasing $f : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\lim_{x\to\infty} \frac{f(x)}{x} = \infty$ and $\sup_t \boldsymbol{E}[f(|\xi_t|)] < \infty$.*

Proof. Suppose we have $f : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\lim_{x\to\infty} \frac{f(x)}{x} = \infty$ and $\sup_t \mathbf{E}\left[f(|\xi_t|)\right] < \infty$ (it doesn't have to be increasing or convex). Let $\epsilon > 0$ be given and pick $R > 0$ such that $\frac{f(x)}{x} \geq \frac{\sup_t \mathbf{E}[f(|\xi_t|)]}{\epsilon}$ for $x \geq R$. Then for all $t \in T$,

$$\mathbf{E}\left[|\xi_t| ; |\xi_t| \geq R\right] \leq \frac{\epsilon}{\sup_t \mathbf{E}\left[f(|\xi_t|)\right]} \mathbf{E}\left[f(|\xi_t|); |\xi_t| \geq R\right] \leq \epsilon$$

thus $\lim_{R\to\infty} \sup_t \mathbf{E}\left[|\xi_t| ; |\xi_t| \geq R\right] = 0$ and uniform integrability is shown.

The key step to finding f is the following observation. Suppose we are given an increasing $f : \mathbb{R}_+ \to \mathbb{R}_+$, then if we use Lemma 6.8 then for any positive ξ we

$$\mathbf{E}\left[f(\xi)\right] = \int_0^\infty \mathbf{P}\{f(\xi) \geq \lambda\}\, d\lambda = \int_{f^{-1}(0)}^{f^{-1}(\infty)} \mathbf{P}\{\xi \geq \eta\} f'(\eta)\, d\eta \quad \text{letting } f(\eta) = \lambda$$

so the problem of finding f can be recast as finding a function g such that $\int \mathbf{P}\{\xi \geq \eta\} g(\eta)\, d\eta < \infty$ and $\lim_{x\to\infty} g(x) = \infty$. Though the computation above isn't rigorous since we haven't justified the change of variables in the integral, this idea tells us that we should assume f of the form $f(x) = \int_0^x g(y)\, dy$ and for such an f we can rigorously calculate using Tonneli's Theorem

$$\mathbf{E}\left[f(\xi)\right] = \mathbf{E}\left[\int_0^{|\xi|} g(y)\, dy\right] = \mathbf{E}\left[\int_0^\infty g(y)\mathbf{1}_{|\xi|\geq y}\, dy\right] = \int_0^\infty g(y)\mathbf{P}\{|\xi| \geq y\}\, dy < \infty$$

Furthermore, $\lim_{x\to\infty} \frac{f(x)}{x} = \infty$ by L'Hopital's Rule (TODO: can do this without differentiation) Moreover if $g(x)$ is increasing then we know that $f(x)$ is convex.

So our goal is to find $g(x)$ such that $\lim_{x\to\infty} g(x) = \infty$ and $\sup_t \int_0^\infty \mathbf{P}\{|\xi_t| \geq \eta\} g(\eta)\, d\eta < \infty$.

The existence of $g(x)$ for any positive integrable $\phi(x)$ can be established by the following explicit construction. Let

$$g(x) = \frac{1}{\sqrt{\int_x^\infty \phi(x)\, dx}}$$

and note that Dominated Convergence shows $\lim_{x\to\infty} g(x) = \infty$ and the Fundamental Theorem of Calculus (Theorem 3.96) shows that (TODO: this also requires the Chain Rule which isn't trivial in this context)

$$g(x)\phi(x) = -2\frac{d}{dx}\sqrt{\int_x^\infty \phi(x)\, dx}$$

and therefore

$$\int_0^\infty g(x)\phi(x)\, dx = 2\sqrt{\int_0^\infty \phi(x)\, dx} < \infty$$

Now suppose that ξ_t is uniformly integrable.

$$\begin{aligned}
\mathbf{E}\left[|\xi_t|\,;|\xi_t| \geq R\right] &= \int_0^\infty \mathbf{P}\{|\xi_t|\,\mathbf{1}_{|\xi_t|\geq R} \geq \lambda\}\, d\lambda \\
&= \int_R^\infty \mathbf{P}\{|\xi_t| \geq \lambda\}\, d\lambda + \int_0^R \mathbf{P}\{|\xi_t| \geq R\}\, d\lambda \\
&= \int_R^\infty \mathbf{P}\{|\xi_t| \geq \lambda\}\, d\lambda + R\mathbf{P}\{|\xi_t| \geq R\}
\end{aligned}$$

and since $\lim_{R\to\infty} \sup_t \mathbf{E}\left[|\xi_t|\,;|\xi_t| \geq R\right] = 0$ we also get $\lim_{R\to\infty} \sup_t \int_R^\infty \mathbf{P}\{|\xi_t| \geq \lambda\}\, d\lambda = 0$ which shows that if we define

$$g(x) = \frac{1}{\sqrt{\sup_t \int_x^\infty \mathbf{P}\{|\xi_t| \geq \lambda\}\, d\lambda}}$$

then we have

$$\lim_{x\to\infty} g(x) = \infty$$

and moreover for any $t \in T$,

$$g(x) \leq \frac{1}{\sqrt{\int_x^\infty \mathbf{P}\{|\xi_t| \geq \lambda\}\, d\lambda}}$$

so by the previous construction we know that

$$\int_0^\infty \mathbf{P}\{|\xi_t| \geq x\} g(x)\, dx \leq \int_0^\infty \mathbf{P}\{|\xi_t| \geq x\} \frac{1}{\sqrt{\int_x^\infty \mathbf{P}\{|\xi_t| \geq \lambda\}\, d\lambda}}\, dx < \infty$$

TODO: Finish and address any issues related to the fact that we only have almost everywhere differentiability of an integral in Lebesgue theory (e.g. chain rule, u-substitution) (also is L'Hopital valid). □

9. Lindeberg's Central Limit Theorem

The Law of Large Numbers tells us that when we are given i.i.d. random variables ξ_i with finite expectation, we have almost sure convergence of $\frac{1}{n}\sum_{k=1}^n \xi_k = \mathbf{E}[\xi_i]$. Using different notation we can say,

$$\sum_{k=1}^n \xi_k - n\mathbf{E}[\xi_i] = o(n)$$

From one point of view, the Central Limit Theorem arises from asking the question about whether $o(n)$ can be replaced by $o(n^p)$ or $\mathcal{O}(n^p)$ for $p < 1$. In this sense the Central Limit Theorem gives some information about the rate of convergence of the sums $\frac{1}{n}\sum_{k=1}^n \xi_k$ to their limit.

First some intuition about the Central Limit Theorem. Let's assume that we have a sequence of i.i.d. random variables ξ_i such that ξ_i has moments of all orders (a much stronger assumption than one needs for the CLT). We also assume

$$\mathbf{E}[\xi_i] = 0, \mathbf{E}\left[\xi_i^2\right] = 1$$

Consider the following computation of the moments of the partial sums of ξ_i. Let $S_n = \xi_1 + \cdots + \xi_n$.

$$\begin{aligned}
\mathbf{E}\left[S_n^{m+1}\right] &= \mathbf{E}\left[(\xi_1 + \cdots + \xi_n)(\xi_1 + \cdots + \xi_n)^m\right] \\
&= \sum_{i=1}^n \mathbf{E}\left[\xi_i(\xi_n + S_{n-1})^m\right] \\
&= n\mathbf{E}\left[\xi_n(\xi_n + S_{n-1})^m\right] \text{ TODO: don't know how to prove this step} \\
&= n\sum_{j=0}^m \binom{m}{j} \mathbf{E}\left[\xi_n^{j+1}\right] \mathbf{E}\left[S_{n-1}^{m-j}\right] \\
&= nm\mathbf{E}\left[S_{n-1}^{m-1}\right] + n\sum_{j=2}^m \binom{m}{j} \mathbf{E}\left[\xi_n^{j+1}\right] \mathbf{E}\left[S_{n-1}^{m-j}\right]
\end{aligned}$$

Now define $\hat{S}_n = S_n/\sqrt{n}$, and divide both sides of the above by $n^{\frac{m+1}{2}}$ and we see

$$\mathbf{E}\left[\hat{S}_n^{m+1}\right] = m\mathbf{E}\left[\hat{S}_n^{m-1}\right] + \sum_{j=2}^{m}\binom{m}{j}\frac{1}{n^{\frac{j-1}{2}}}\mathbf{E}\left[\xi_n^{j+1}\right]\mathbf{E}\left[\hat{S}_{n-1}^{m-j}\right]$$

An induction on m together with the observation that $\mathbf{E}\left[\hat{S}_n^0\right] = 1$ and $\mathbf{E}\left[\hat{S_n}\right] = 0$ shows that

$$\lim_{n\to\infty}\mathbf{E}\left[\hat{S}_n^{2m+1}\right] = 0$$
$$\lim_{n\to\infty}\mathbf{E}\left[\hat{S}_n^{2m}\right] = \prod_{j=1}^{m}(2j-1) = \frac{(2m)!}{2^m m!}$$

We can recognize that these are the moments of the standard normal distribution.

The above argument is one path to use to see how Gaussian distributions might arise when looking at sums of i.i.d random variables but relies on an unecessarily strong set of assumptions (not to mention it ignores the fact that moments alone to not characterize a distribution).

In fact convergence to normal distributions is more general than i.i.d. variables and we look for a version that has a rather precise set of assumptions called the Lindeberg conditions. The statement of the result and the corresponding notation is unwieldy but the proof itself doesn't seem to suffer much from the added complexity. Furthermore the added generality provides a useful space to explore when examining the limits of asymptotic normality.

Theorem 9.1 (Lindeberg)**.** *Let $\xi_1, \xi_2, \ldots$ be independent square integrable random variables $\boldsymbol{E}[\xi_m] = 0$ and $\boldsymbol{E}\left[\xi_m^2\right] = \sigma_m^2 > 0$. Define*

$$S_n = \sum_{i=1}^{n}\xi_i$$
$$\Sigma_n = \sqrt{\sum_{i=1}^{n}\sigma_i^2}$$
$$\hat{S}_n = \frac{S_n}{\Sigma_n}$$
$$r_n = \max_{1\le i\le n}\frac{\sigma_i}{\Sigma_n}$$
$$g_n(\epsilon) = \frac{1}{\Sigma_n^2}\sum_{i=1}^{n}\boldsymbol{E}\left[\xi_i^2\boldsymbol{1}_{|\xi_i|\ge\epsilon\Sigma_n}\right]$$

and let $d\gamma = \frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}dx$ be the distribution of an $N(0,1)$ random variable. Now for all $\epsilon > 0$, $\varphi \in C^3(\mathbb{R};\mathbb{R})$ with bounded 2nd and 3rd derivative,

$$\left|\boldsymbol{E}\left[\varphi(\hat{S}_n)\right] - \int_{\mathbb{R}}\varphi d\gamma\right| \le \left(\frac{\epsilon}{6} + \frac{r_n}{2}\right)\|\varphi'''\|_\infty + g_n(\epsilon)\|\varphi''\|_\infty$$

and

$$r_n^2 \le \epsilon^2 + g_n(\epsilon)$$

In particular, if $\lim_{n\to\infty} g_n(\epsilon) = 0$ *for every* $\epsilon > 0$*, then*

$$\lim_{n\to\infty} \left| \boldsymbol{E}\left[\varphi(\hat{S}_n)\right] - \int_{\mathbb{R}} \varphi d\gamma \right| = 0$$

Before attacking the proof we note how everything specializes in the case of i.i.d. random variables. In this case $\Sigma_n = \sqrt{n}\sigma$, $\hat{S}_n = \frac{\sum_{i=1}^n \xi_i}{\sqrt{n}\sigma}$ and $g_n(\epsilon) = \frac{1}{\sigma^2}\mathbf{E}\left[\xi^2; |\xi| \geq \epsilon\sqrt{n}\sigma\right]$. Because $\mathbf{E}\left[\xi^2\right] < \infty$ we know that $\xi^2 < \infty$ a.s. and we have $\xi^2 \mathbf{1}_{|\xi|\geq\epsilon\sqrt{n}\sigma} \overset{a.s.}{\to} 0$. Noting $\xi^2 \mathbf{1}_{|\xi|\geq\epsilon\sqrt{n}\sigma} \leq \xi^2$, Dominated Convergence tells us that $\lim_{n\to\infty} g_n(\epsilon) = 0$.

This special case also sheds some light on aspects of the hypotheses. For example, the $\sqrt{n}$ in the denominator is the only possible choice to acheive convergence to a random variable with finite non-zero variance; it is precisely the term requires to make $\sigma(\hat{S}_n)$ converge to a finite non-zero number (in fact in the i.i.d. case it makes the sequence constant).

It is also worth spending some time understanding the nature of $g_n(\epsilon)$. First, it is clear from independence and definitions that

$$\mathbf{E}\left[\hat{S}_n^2\right] = \sum_{i=1}^n \mathbf{E}\left[\left(\frac{\xi_i}{\Sigma_n}\right)^2\right] = \frac{1}{\Sigma_n^2}\sum_{i=1}^n \sigma_i^2 = 1$$

but we can also write

$$g_n(\epsilon) = \frac{1}{\Sigma_n^2}\sum_{i=1}^n \mathbf{E}\left[\xi_i^2 \mathbf{1}_{|\xi_i|\geq\epsilon\Sigma_n}\right] = \sum_{i=1}^n \mathbf{E}\left[\left(\frac{\xi_i}{\Sigma_n}\right)^2; \left|\frac{\xi_i}{\Sigma_n}\right| \geq \epsilon\right]$$

So the $\hat{S}_n$ is the sum of ξ_i normalized to maintain a constant unit variance. Our assumption that $\lim_{n\to\infty} g_n(\epsilon) = 0$ is an assertion that in the limit, all of that unit variance is contained in a bounded region around 0. In the i.i.d. case that is clearly true because all of the unscaled ξ_n have their "energy" in a constant fashion, so rescaling is able to concentrate that energy arbitrarily close to 0. It is permissible to have the energy of the ξ_n moving off to infinity but only if it travels at a rate less than $\sqrt{n}$.

TODO: Question is it possible to satisfy the Lindeberg condition when $\lim_{n\to\infty} \Sigma_n < \infty$?

Proof. Fix an $n > 0$ and define $\hat{\xi}_m = \frac{\xi_m}{\Sigma_n}$ and $\hat{S}_n = \hat{\xi}_1 + \cdots + \hat{\xi}_n$. Note that $\mathbf{E}\left[\hat{S}_n^2\right] = 1$. Let $\eta_1, \eta_2, \ldots$ be independent $N(0,1)$ random variables that are also independent of the ξ_i. Note that we may have to extend Ω in order to arrange this (e.g. extend by $[0,1]$ and use Theorem 7.32). We rescale each η_i so that it has the same variance as $\hat{\xi}_i$; define $\hat{\eta}_i = \frac{\sigma_i \eta_i}{\Sigma_n}$ and $\hat{T}_n = \hat{\eta}_1 + \cdots + \hat{\eta}_n$. Notice that $\mathbf{E}\left[\hat{\eta}_m^2\right] = \mathbf{E}\left[\hat{\xi}_m^2\right] = \frac{\sigma_m^2}{\Sigma_n^2}$ and $\hat{T}_n$ is also a $N(0,1)$ random variable. Therefore, by the Expectation Rule (Lemma 6.7) $\int \varphi \, d\gamma = \mathbf{E}\left[\varphi(\hat{T}_n)\right]$ and we can write

$$\left|\mathbf{E}\left[\varphi(\hat{S}_n)\right] - \int_{\mathbb{R}} \varphi d\gamma\right| = \left|\mathbf{E}\left[\varphi(\hat{S}_n)\right] - \mathbf{E}\left[\varphi(\hat{T}_n)\right]\right|$$

By having arranged for $\hat{\xi}_i$ and $\hat{\eta}_i$ to have same first and second moments so one should be thinking that we have constructed a "second order approximation". TODO: What is critical is that the approximation of the individual $\hat{\xi}_i$ may not be

a good one, the approximation $\hat{S_n}$ by $\hat{T_n}$ is a good one. Find the critical point(s) in the proof where this comes to light.

The real trick of the proof is to interpolate between $\varphi(\hat{S_n})$ and $\varphi(\hat{T_n})$ by exchanging $\hat{\xi}_i$ and $\hat{\eta}_i$ one summand at a time. By varying only one summand we will then be able use Taylor's Theorem to estimate the differences between the terms. Concretely we write,

$$\begin{aligned}\varphi(\hat{S_n}) - \varphi(\hat{T_n}) &= \varphi(\hat{\xi}_1 + \cdots + \hat{\xi}_n) - \varphi(\hat{\eta}_1 + \cdots + \hat{\eta}_n) \\ &= \varphi(\hat{\xi}_1 + \cdots + \hat{\xi}_n) - \varphi(\hat{\eta}_1 + \hat{\xi}_2 + \cdots + \hat{\xi}_n) \\ &+ \varphi(\hat{\eta}_1 + \hat{\xi}_2 + \cdots + \hat{\xi}_n) - \varphi(\hat{\eta}_1 + \hat{\eta}_2 + \hat{\xi}_3 + \cdots + \hat{\xi}_n) \\ &+ \cdots \\ &+ \varphi(\hat{\eta}_1 + \cdots + \hat{\eta}_{n-1} + \hat{\xi}_n) - \varphi(\hat{\eta}_1 + \cdots + \hat{\eta}_n)\end{aligned}$$

Since we have to manipulate these terms a bit, it helps to clean up the notation by defining:

$$U_m = \begin{cases} \hat{\xi}_2 + \cdots + \hat{\xi}_n & \text{if } m = 1 \\ \hat{\eta}_1 + \cdots + \hat{\eta}_{m-1} + \hat{\xi}_{m+1} + \cdots + \hat{\xi}_n & \text{if } 1 < m < n \\ \hat{\eta}_1 + \cdots + \hat{\eta}_{n-1} & \text{if } m = n \end{cases}$$

and then we can write the above interpolation as

$$\varphi(\hat{S_n}) - \varphi(\hat{T_n}) = \sum_{m=1}^{n} \varphi(U_m + \hat{\xi}_m) - \varphi(U_m + \hat{\eta}_m)$$

Now we can take absolute values, use the triangle inequality and use linearity of expectation to see

$$\begin{aligned}\left|\mathbf{E}\left[\varphi(\hat{S_n}) - \varphi(\hat{T_n})\right]\right| &\leq \sum_{m=1}^{n} \left|\mathbf{E}\left[\varphi(U_m + \hat{\xi}_m)\right] - \mathbf{E}\left[\varphi(U_m + \hat{\eta}_m)\right]\right| \\ &= \sum_{m=1}^{n} \left|\mathbf{E}\left[\varphi(U_m + \hat{\xi}_m) - \varphi(U_m + \hat{\eta}_m)\right]\right|\end{aligned}$$

Now we focus on each term $\varphi(U_m + \hat{\xi}_m) - \varphi(U_m + \hat{\eta}_m)$ by applying Taylor's Formula (Theorem 2.19) to see

$$\varphi(U_m + x) = \varphi(U_m) + x\varphi'(U_m) + \frac{x^2}{2}\varphi''(U_m) + R_m(x)$$

where

$$R_m(x) = \int_{U_m}^{U_m + x} \frac{(U_m + x - t)^2}{2} \varphi'''(t)\, dt$$

For example,applying this expansion with $x = \hat{\xi}_m$, using linearity of expectation, independence of $\hat{\xi}_m$ and U_m and Lemma 7.16 we get

$$\begin{aligned}\mathbf{E}\left[\varphi(U_m + \hat{\xi}_m)\right] &= \mathbf{E}\left[\varphi(U_m) + \hat{\xi}_m\varphi'(U_m) + \frac{\hat{\xi}_m^2}{2}\varphi''(U_m) + R_m(\hat{\xi}_m)\right] \\ &= \mathbf{E}\left[\varphi(U_m)\right] + \frac{\sigma_m^2}{2\Sigma_n^2}\mathbf{E}\left[\varphi''(U_m)\right] + \mathbf{E}\left[R_m(\hat{\xi}_m)\right]\end{aligned}$$

and in exactly the same way because we have arrange for $\hat{\xi}_m$ and $\hat{\eta}_m$ to share the first two moments, we get

$$\mathbf{E}\left[\varphi(U_m + \hat{\eta}_m)\right] = \mathbf{E}\left[\varphi(U_m)\right] + \frac{\sigma_m^2}{2\Sigma_n^2}\mathbf{E}\left[\varphi''(U_m)\right] + \mathbf{E}\left[R_m(\hat{\eta}_m)\right]$$

Thus, $\mathbf{E}\left[\varphi(U_m + \hat{\xi}_m) - \varphi(U_m + \hat{\eta}_m)\right] = \mathbf{E}\left[R_m(\hat{\xi}_m)\right] - \mathbf{E}\left[R_m(\hat{\eta}_m)\right]$ and

$$\left|\mathbf{E}\left[\varphi(\hat{S}_n) - \varphi(\hat{T}_n)\right]\right| \leq \sum_{m=1}^{n}\left|\mathbf{E}\left[R_m(\hat{\xi}_m)\right]\right| + \sum_{m=1}^{n}\left|\mathbf{E}\left[R_m(\hat{\eta}_m)\right]\right|$$

We complete the proof by bounding each expectation above. On the one hand, there is the Lagrange Form for the remainder term (Lemma 2.20) that shows that $R_m(x) = \varphi'''(c)\frac{x^3}{6}$ for some $c \in [U_m, U_m + x]$ hence $|R_m(x)| \leq \|\varphi'''\|_\infty \frac{|x|^3}{6}$. On the other hand, sticking with the integral form of the remainder term, since $t \in [U_m, U_m + x]$ we can bound the term $(U_m + x - t)^2 \leq |x|^2$ in the integral and integrate to conclude

$$\begin{aligned}|R_m(x)| &= \int_{U_m}^{U_m+x} \frac{(U_m + x - t)^2}{2}\varphi'''(t)\,dt \leq \frac{|x|^2}{2}\int_{U_m}^{U_m+x}\varphi'''(t)\,dt \\ &= \frac{|x|^2}{2}\left(\varphi''(U_m + x) - \varphi''(x)\right) \leq \|\varphi''\|_\infty |x|^2\end{aligned}$$

With this setup, pick $\epsilon > 0$ and first consider the remainder term $R_m(\hat{\xi}_m)$ and a note that we have to be a little careful. We would like to use the stronger 3^{rd} moment bound however we have not assumed that $\hat{\xi}_m$ has a finite 3^{rd} moment. So what we do is truncate $\hat{\xi}_m$ and take a 2^{nd} moment bound over the tail (valid because of the finite variance assumption) and use a 3^{rd} moment bound on the truncated $\hat{\xi}_m$. The details follow:

$$\left|\mathbf{E}\left[R_m(\hat{\xi}_m)\right]\right| \leq \left|\mathbf{E}\left[R_m(\hat{\xi}_m); \left|\hat{\xi}_m\right| \leq \epsilon\right]\right| + \left|\mathbf{E}\left[R_m(\hat{\xi}_m); \left|\hat{\xi}_m\right| > \epsilon\right]\right|$$

We take the sum of first terms and apply the Taylor's formula bound to see

$$\begin{aligned}\sum_{m=1}^{n}\left|\mathbf{E}\left[R_m(\hat{\xi}_m); \left|\hat{\xi}_m\right| \leq \epsilon\right]\right| &\leq \frac{\|\varphi'''\|_\infty}{6}\sum_{m=1}^{n}\left|\mathbf{E}\left[\left|\hat{\xi}_m\right|^3; \left|\hat{\xi}_m\right| \leq \epsilon\right]\right| \\ &\leq \epsilon\frac{\|\varphi'''\|_\infty}{6}\sum_{m=1}^{n}\left|\mathbf{E}\left[\left|\hat{\xi}_m\right|^2\right]\right| \\ &= \epsilon\frac{\|\varphi'''\|_\infty}{6}\sum_{m=1}^{n}\frac{\sigma_m^2}{\Sigma_n^2} = \epsilon\frac{\|\varphi'''\|_\infty}{6}\end{aligned}$$

Next take the sum of the second terms to see

$$\begin{aligned}\sum_{m=1}^{n}\left|\mathbf{E}\left[R_m(\hat{\xi}_m);\left|\hat{\xi}_m\right|>\epsilon\right]\right| &\le \|\varphi''\|_\infty \sum_{m=1}^{n}\left|\mathbf{E}\left[\left|\hat{\xi}_m\right|^2;\left|\hat{\xi}_m\right|>\epsilon\right]\right| \\ &= \|\varphi''\|_\infty \frac{1}{\Sigma_n^2}\sum_{m=1}^{n}\left|\mathbf{E}\left[|\xi_m|^2;|\xi_m|>\epsilon\Sigma_n\right]\right| \\ &= \|\varphi''\|_\infty g_\epsilon(n)\end{aligned}$$

Lastly, to bound the remainder term on $\hat{\eta}_m$ we can directly appeal to the 3^{rd} moment bound because as a normal random variable $\hat{\eta}_m$ has finite moments of all orders:

$$\begin{aligned}\sum_{m=1}^{n}\left|\mathbf{E}\left[R_m(\hat{\eta}_m)\right]\right| &\le \frac{\|\varphi'''\|_\infty}{6}\sum_{m=1}^{n}\left|\mathbf{E}\left[|\hat{\eta}_m|^3\right]\right| \\ &= \frac{\|\varphi'''\|_\infty}{6}\sum_{m=1}^{n}\frac{\sigma_m^3}{\Sigma_n^3}\left|\mathbf{E}\left[|\eta_m|^3\right]\right| \\ &= \frac{r_n\|\varphi'''\|_\infty}{6}\sum_{m=1}^{n}\frac{\sigma_m^2}{\Sigma_n^2}\left|\mathbf{E}\left[|\eta_m|^3\right]\right| \\ &= \frac{r_n\|\varphi'''\|_\infty}{6}\frac{2\sqrt{2}}{\sqrt{\pi}} < \frac{r_n\|\varphi'''\|_\infty}{2}\end{aligned}$$

TODO: We used a calculation of the 3^{rd} absolute moment of the standard normal distribution ($\frac{2\sqrt{2}}{\sqrt{\pi}}$). We need to record that calculation somewhere.

The last thing to show is the bound on r_n^2. For each $n>0$ and $1\le m\le n$,

$$\begin{aligned}\frac{\sigma_m^2}{\Sigma_n^2} &= \frac{1}{\Sigma_n^2}\left(\mathbf{E}\left[\xi_m^2;|\xi_m|<\epsilon\Sigma_n\right]+\mathbf{E}\left[\xi_m^2;|\xi_m|\ge\epsilon\Sigma_n\right]\right) \\ &\le \frac{1}{\Sigma_n^2}\left(\epsilon^2\Sigma_n^2+\Sigma_n^2 g_n(\epsilon)\right) = \epsilon^2+g_n(\epsilon)\end{aligned}$$

hence $r_n^2 = \max_{1\le m\le n}\frac{\sigma_m^2}{\Sigma_n^2} \le \epsilon^2+g_n(\epsilon)$. □

Note that the Lindeberg condition is a sufficient condition but not a necessary condition for convergence to a normal distribution; but is not too far off. Thus it is useful to examine a case in which we don't satisfy the condition.

Example 9.2 (Failure of Lindeberg Condition)**.** Let ξ_n be a sequence of independent random variables such that $\xi_n = n$ with probability $\frac{1}{2n^2}$, $\xi_n = -n$ with probability $\frac{1}{2n^2}$ and $\xi_n = 0$ with probability $1-\frac{1}{n^2}$. Note that $\mathbf{Var}(\xi_n) = (-n)^2\cdot\frac{1}{2n^2}+0\cdot(1-\frac{1}{2n^2})+n^2\cdot\frac{1}{2n^2} = 1$. $\sum_{n=1}^{\infty}\mathbf{P}\{\xi_n\neq 0\} = \sum_{n=1}^{\infty}\frac{1}{n^2}<\infty$ so by Borel Cantelli, we have ξ_n are eventually 0 a.s.; hence $S_n = \sum_{i=1}^{n}\xi_i$ is bounded a.s. and $\lim_{n\to\infty}\frac{S_n}{\sqrt{n}} = 0$ a.s. Therefore, $\frac{S_n}{\sqrt{n}}$ does not converge to a Gaussian in distribution.

We know that ξ_n must not satisfy the Lindeberg condition and it is instructive to perform that calculation explicitly. Using the notation of Theorem 9.1, $\Sigma_n = \sqrt{n}$, thus for any $\epsilon>0$, and $n>\epsilon^2$, we have

$$\xi_n\cdot\mathbf{1}_{|\xi_n|>\epsilon\Sigma_n} = \xi_n\cdot\mathbf{1}_{|\xi_n|>\epsilon\sqrt{n}} = \xi_n$$

so only a finite number of summands of $\mathbf{E}\left[\xi_n^2; |\xi_n| > \epsilon\sqrt{n}\right]$ are different from 1, hence

$$\lim_{n\to\infty} g_n(\epsilon) = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\left[\xi_n^2; |\xi_n| > \epsilon\sqrt{n}\right] = 1$$

TODO: Mention Feller-Lindeberg Theorem that adds an addition hypothesis that makes the Lindeberg condition equivalent to asymptotic normality.

The Lindeberg Theorem above doesn't actually prove weak convergence because of the differentiability assumption on the function φ. Our next step is to use approximation arguments to show that we in fact get weak convergence. The argument has broader applicability than the Central Limit Theorem and is just a validation that proving weak convergence for random vectors only requires use compactly supported smooth test functions.

Lemma 9.3. *Let $\xi, \xi_1, \xi_2, \ldots$ be random vectors in $\mathbb{R}^N$, then $\xi_n \xrightarrow{d} \xi$ if and only if $\lim_{n\to\infty} \boldsymbol{E}[f(\xi_n)] = \boldsymbol{E}[f(\xi)]$ for all $f \in C_c^\infty(\mathbb{R}^N;\mathbb{R})$.*

Proof. Since any $f \in C_c^\infty(\mathbb{R}^N;\mathbb{R})$ is bounded we certainly see that $\xi_n \xrightarrow{d} \xi$ implies $\lim_{n\to\infty} \mathbf{E}[f(\xi_n)] = \mathbf{E}[f(\xi)]$.

In the other direction, take an arbitrary $f \in C_b(\mathbb{R}^N;\mathbb{R})$ and pick $\epsilon > 0$. By Lemma 3.101, we can find $f_n \in C_c^\infty(\mathbb{R}^N;\mathbb{R})$ such that f_n converges uniformly on compact sets and $\|f_n\|_\infty \leq \|f\|_\infty$. The idea of the proof is to note that for any $n, k \geq 0$, we have

$$|\mathbf{E}[f(\xi_n) - f(\xi)]| \leq |\mathbf{E}[f(\xi_n) - f_k(\xi_n)]| + |\mathbf{E}[f_k(\xi_n) - f_k(\xi)]| + |\mathbf{E}[f_k(\xi) - f(\xi)]|$$

and then to bound each term on the right hand side. The second term will be easy to handle because of our hypothesis and the smoothness of f_k. The first and third terms will require that we examine the approximation provided by the uniform convergence of the f_k on all compact sets.

The first task we have is to pick that compact set; it turns out that it suffices to consider closed balls centered at the origin. For any $R \in \mathbb{R}$ with $R > 0$, there exists a $\psi_R \in C_c^\infty(\mathbb{R}^N;\mathbb{R})$ with $\mathbf{1}_{|x|\leq\frac{R}{2}} \leq \psi_R(x) \leq \mathbf{1}_{|x|\leq R}$, therefore

$$\begin{aligned}
\lim_{n\to\infty} \mathbf{P}\{|\xi_n| > R\} &= 1 - \lim_{n\to\infty} \mathbf{E}\left[\mathbf{1}_{|\xi_n|\leq R}\right] \\
&\leq 1 - \lim_{n\to\infty} \mathbf{E}[\psi_R(\xi_n)] \\
&= 1 - \mathbf{E}[\psi_R(\xi)] \\
&\leq 1 - \mathbf{E}\left[\mathbf{1}_{|\xi|\leq\frac{R}{2}}\right] \\
&= \mathbf{P}\{|\xi| > \frac{R}{2}\}
\end{aligned}$$

On the other hand, we know that $\lim_{R\to\infty} \mathbf{1}_{|\xi|\leq\frac{R}{2}} = 0$ a.s. and therefore by Monotone Convergence, $\lim_{R\to\infty} \mathbf{P}\{|\xi| > \frac{R}{2}\} = 0$. Select $R > 0$ such that

$$\mathbf{P}\{|\xi| > R\} \leq \mathbf{P}\{|\xi| > \frac{R}{2}\} \leq \frac{\epsilon}{4\|f\|_\infty}$$

Then we can pick $N_1 > 0$ such that $\mathbf{P}\{|\xi_n| > R\} \leq \frac{\epsilon}{2\|f\|_\infty}$ for all $n > N_1$.

Having picked $R > 0$, we know that f_n converges uniformly to f on $|x| \leq R$ and therefore we can find a $K > 0$ such that for $k > K$ and $|x| \leq R$ we have

$|f_k(x) - f(x)| < \epsilon$. Therefore,

$$\begin{aligned}|\mathbf{E}\left[f_k(\xi) - f(\xi)\right]| &\leq \mathbf{E}\left[|f_k(\xi) - f(\xi)|\,; |\xi| \leq R\right] + \mathbf{E}\left[|f_k(\xi) - f(\xi)|\,; |\xi| > R\right] \\ &\leq \epsilon \mathbf{P}\{|\xi| \leq R\} + 2\|\xi\|_\infty \mathbf{P}\{|\xi| > R\} \\ &\leq \epsilon + \frac{\epsilon}{2} < 2\epsilon\end{aligned}$$

and via the same calculation, for $n > N_1$

$$|\mathbf{E}\left[f_k(\xi_n) - f(\xi_n)\right]| \leq \epsilon + 2\|\xi_n\|_\infty \mathbf{P}\{|\xi_n| > R\} \leq 2\epsilon$$

To finish the proof, pick a single $k > K$ and then we can find $N_2 > 0$ such that for all $n > N_2$, we have $|\mathbf{E}\left[f_k(\xi_n) - f_k(\xi)\right]| < \epsilon$. Putting these three estimates together, we have for $n > \max(N_1, N_2)$,

$$|\mathbf{E}\left[f(\xi_n) - f(\xi)\right]| \leq 5\epsilon$$

□

We are not going to prove the following but we should talk about it:

Theorem 9.4. *Let $\xi, \xi_1, \xi_2, \ldots$ be i.i.d with $\boldsymbol{E}\left[|\xi|^3\right] < \infty$. Let $\Phi(x)$ be the cdf of standard normal and let $G(x) = \boldsymbol{P}\{\frac{S_n - \mu}{\sigma\sqrt{n}} \leq x\}$ be the empirical cdf. Then there exists a constant $C > 0$ such that*

$$\sup_x |G(x) - \Phi(x)| \leq \frac{C\boldsymbol{E}\left[|\xi|^3\right]}{\sigma^3\sqrt{n}}$$

Note the upper bound of the constant C has been reduced to about 0.5600.

10. Characteristic Functions And Central Limit Theorem

In this section we study the weak convergence of random vectors more carefully. Our first goal is to develop just enough of the theory of characteristic functions in order to prove the classical Central Limit Theorem. After that we delve more deeply into theory of characteristic functions.

The motivation for the theory we are about to develop is the intuition that most of the behavior of a probability distribution on $\mathbb{R}$ is captured by its moments. If one could put the information about all of the distribution's moments into a single package simulataneously then the resulting package might characterize the probability distribution in a useful way. A initial naive approach might be to use a *generating function* methodology. For example, one might try to define a function $f(t) = \sum_{n=0}^{\infty} M_n t^n$ where M_n denotes the n^{th} moment. Alas, such a approach fails rather miserably as it is a very rare thing for moments to decrease quickly enough for the formal power series for $f(t)$ to ever converge and make a useful function object. A better approach is to scale the moments to give the series a chance to converge. For example, being a bit sloppy we could write

$$f(t) = \int e^{tx} dP = \sum_{n=0}^{\infty} \frac{M_n}{n!} t^n$$

This idea has a lot more merit and can be used effectively but it has the distinct disadvantage that it only works for distributions that have moments of all orders.

The wonderful idea that we will be exploring in this chapter is that by passing into the domain of complex numbers we get a characterization of the distribution

that is always defined and (at least conceptually) captures all moments in a generating function. Specifically, we define

$$f(t) = \int e^{itx} dP$$

which is the *Fourier Transform* of the probability distribution and we get an object that uniquely determines the distribution and can often be much easier to work with. In particular we we will see that convergence in distribution is described as pointwise convergence of characteristic functions and through that connection we will get another proof of the Central Limit Theorem.

In this section we start to make use of integrals of complex valued measurable functions. Let's establish the basic definitions and facts that we require.

Definition 10.1. A function $f : (\Omega, \mathcal{A}, \mu) \to \mathbb{C}$ is measurable if and only $f = h + ig$ where $h, g : (\Omega, \mathcal{A}, \mu) \to \mathbb{R}$ are measurable. Equivalently, $\mathbb{C}$ is given the Borel σ-algebra.

(i) If $\mu(A) < \infty$, then $\left|\int f \, d\mu\right| \leq \int |f| \, d\mu$.

Proof. By the triangle inequality for the complex norm, we know that given any two $z, w \in \mathbb{C}$ and $t \in [0, 1]$, $|(1-t)z + tw| \leq (1-t)|z| + t|w|$ and therefore the complex norm is convex. Then by Jensen's Inequality (Theorem 6.17, $\left|\int f \, d\mu\right| \leq \int |f| \, d\mu$. □

Definition 10.2. Let μ be a probability measure on $\mathbb{R}^n$. Its *Fourier Transform* is denoted $\hat{\mu}$ and is the complex function on $\mathbb{R}^n$ defined by

$$\hat{\mu}(u) = \int e^{i\langle u,x\rangle} d\mu(x) = \int \cos(\langle u, x\rangle) \, d\mu(x) + i \int \sin(\langle u, x\rangle) \, d\mu(x)$$

The first order of business is to establish the basic properties of the Fourier Transform of a probability measure including the fact that the definition makes sense.

Theorem 10.3. *Let μ be a probability measure, then $\hat{\mu}$ exists and is a bounded uniformly continuous function with $\hat{\mu}(0) = 1$.*

Proof. To see that $\hat{\mu}$ exists, use the representation

$$\hat{\mu}(u) = \int \cos(\langle u, x\rangle) \, d\mu(x) + i \int \sin(\langle u, x\rangle) \, d\mu(x)$$

and use the facts that $|\cos\theta| \leq 1$ and $|\sin\theta| \leq 1$ to conclude that both integrals are bounded.

To see that $\hat{\mu}(0) = 1$, simply calculate

$$\hat{\mu}(0) = \int \cos(\langle 0, x\rangle) \, d\mu(x) + i \int \sin(\langle 0, x\rangle) \, d\mu(x) = \int d\mu(x) = 1$$

In a similar way, boundedness is a simple calculation

$$|\hat{\mu}(u)| \leq \int \left|e^{i\langle u,x\rangle}\right| \, d\mu(x) = \int \, d\mu(x) = 1$$

Lastly, to prove uniform continuity, first note that for any $u, v \in \mathbb{R}^n$, we have

$$\begin{aligned}\left|e^{i\langle u,x\rangle} - e^{i\langle v,x\rangle}\right|^2 &= \left|e^{i\langle u-v,x\rangle} - 1\right|^2 \\ &= (\cos(\langle u-v,x\rangle) - 1)^2 + \sin^2(\langle u-v,x\rangle) \\ &= 2(1-\cos(\langle u-v,x\rangle) \\ &\leq \langle u-v,x\rangle^2 && \text{by Lemma 5.1} \\ &\leq \|u-v\|_2^2\|x\|_2^2 && \text{by Cauchy Schwartz}\end{aligned}$$

On the other hand, it is clear from the triangle inequality that

$$\left|e^{i\langle u,x\rangle} - e^{i\langle u,x\rangle}\right| \leq \left|e^{i\langle u,x\rangle}\right| + \left|e^{i\langle u,x\rangle}\right| \leq 2$$

and therefore we have the bound $\left|e^{i\langle u,x\rangle} - e^{i\langle u,x\rangle}\right| \leq \|u-v\|_2\|x\|_2 \wedge 2$. Note that pointwise in $x \in \mathbb{R}^n$, $\lim_{n\to\infty} \frac{1}{n}\|x\|_2 \wedge 2 = 0$ and trivially $\frac{1}{n}\|x\|_2 \wedge 2 \leq 2$ so Dominated Convergence shows that $\lim_{n\to\infty} \int \frac{1}{n}\|x\|_2 \wedge 2\, d\mu(x) = 0$. Given an $\epsilon > 0$, pick $N > 0$ such that $\int \frac{1}{N}\|x\|_2 \wedge 2\, d\mu(x) < \epsilon$ then for $\|u-v\|_2 \leq \frac{1}{N}$,

$$\begin{aligned}|\hat{\mu}(u) - \hat{\mu}(v)| &\leq \int \left|e^{i\langle u,x\rangle} - e^{i\langle u,x\rangle}\right| d\mu(x) \\ &\leq \int \|u-v\|_2\|x\|_2 \wedge 2\, d\mu(x) \\ &\leq \int \frac{1}{N}\|x\|_2 \wedge 2\, d\mu(x) < \epsilon\end{aligned}$$

proving uniform continuity. □

Definition 10.4. Let ξ be an $\mathbb{R}^n$-valued random variable. Its characteristic function is denoted φ_ξ and is the complex valued function on $\mathbb{R}^n$ defined by

$$\begin{aligned}\varphi_\xi(u) &= \mathbf{E}\left[e^{i\langle u,\xi\rangle}\right] \\ &= \int e^{i\langle u,x\rangle}\mathbf{P}^\xi(dx) = \hat{\mathbf{P}^\xi}(u)\end{aligned}$$

We motivated the definition of the characteristic function by considering how we might encode information about the moments of a probability measure. To make sure that we've succeeded we need to show how to extract moments from the characteristic function. To see what we should expect, let's specialize to $\mathbb{R}$ and suppose that we can write out a power series:

$$\hat{\mu}(t) = \int e^{itx}\, d\mu = \sum_{n=0}^{\infty} \frac{i^n M_n}{n!} t^n$$

Still working formally, we see that we can differentiate the series with respect to t to isolate each individual moment M_n

$$\frac{d^n}{dt^n}\hat{\mu}(0) = i^n M_n$$

The above computation was rather formal and we won't try to make the entire thing rigorous (specifically we won't consider the series expansions). What we make rigorous in the next Theorem is the connection between moments of μ and derivatives of the characteristic function.

Theorem 10.5. *Let μ be a probability measure on $\mathbb{R}^n$ such that $f(x) = |x|^m$ is integrable with respect to μ. Then $\hat{\mu}$ is has continuous partial derivatives up to order m and*

$$\frac{\partial^m \hat{\mu}}{\partial x_{j_1} \dots \partial x_{j_m}}(u) = i^m \int x_{j_1} \dots x_{j_m} e^{i\langle u,x\rangle} \mu(dx)$$

Proof. First we proceed with $m = 1$. Pick $1 \leq j \leq n$ and let $v \in \mathbb{R}^n$ be the vector with $v_j = 1$ and $v_i = 0$ for $i \neq j$. Then for $u \in \mathbb{R}^n$ and $t > 0$,

$$\begin{aligned}\frac{\hat{\mu}(u + tv_j) - \hat{\mu}(u)}{t} &= \frac{1}{t} \int e^{i\langle u+tv_j,x\rangle} - e^{i\langle u,x\rangle} d\mu(x) \\ &= \frac{1}{t} \int e^{i\langle u,x\rangle} \left(e^{itx_j} - 1\right) d\mu(x)\end{aligned}$$

But note that

$$\begin{aligned}\left|\frac{1}{t} e^{i\langle u,x\rangle} \left(e^{itx_j} - 1\right)\right|^2 &= \left|\frac{e^{itx_j} - 1}{t}\right|^2 \\ &= \frac{\cos^2(tx_j) - 2\cos(tx_j) + 1 + \sin^2(tx_j)}{t^2} \\ &= 2\left(\frac{1 - \cos(tx_j)}{t^2}\right) \\ &\leq x_j^2 \qquad \text{by Lemma 5.1}\end{aligned}$$

But $|x_j|$ is assumed to be integrable hence we can apply the Dominated Convergence Theorem to see

$$\begin{aligned}\frac{\partial}{\partial x_j} \int e^{i\langle u,x\rangle} d\mu(x) &= \lim_{t\to 0} \frac{1}{t} \int e^{i\langle u+tv_j,x\rangle} - e^{i\langle u,x\rangle} d\mu(x) \\ &= \int \lim_{t\to 0} \frac{e^{i\langle u+tv_j,x\rangle} - e^{i\langle u,x\rangle}}{t} d\mu(x) \\ &= i \int x_j e^{i\langle u,x\rangle} d\mu(x)\end{aligned}$$

Continuity of the derivative follows from the formula we just proved. Suppose that $u_n \to u \in \mathbb{R}^n$. Then we have shown that

$$\frac{\partial}{\partial x_j} \hat{\mu}(u_n) = i \int x_j e^{i\langle u_n,x\rangle} d\mu(x)$$

and we have the bound on the integrands $\left|x_j e^{i\langle u_n,x\rangle}\right| < |x_j|$ with $|x_j|$ integrable by assumption. We apply Dominated Convergence to see that

$$\begin{aligned}\lim_{n\to\infty} \frac{\partial}{\partial x_j} \hat{\mu}(u_n) &= i \int \lim_{n\to\infty} x_j e^{i\langle u_n,x\rangle} d\mu(x) \\ &= i \int x_j e^{i\langle u,x\rangle} d\mu(x) \\ &= \frac{\partial}{\partial x_j} \hat{\mu}(u)\end{aligned}$$

TODO: Fill in the details of the induction step (it is pretty obvious that argument above IS the induction step). □

The key in unlocking the relationship between weak convergence and characteristic functions is a basic property of Fourier Transforms that is often called the Plancherel Theorem. In our particular case the Plancherel Theorem shows that one may evaluate integrals of continuous functions against probability measures equally well using Fourier Transforms; in this way we'll see that the characteristic function of a probability measure is a faithful representation of the measure when viewed as a functional (the point of view implicit in the definition of weak convergence).

Theorem 10.6. *Let*

$$\rho_\epsilon(x) = \frac{1}{\epsilon\sqrt{2\pi}} e^{-\frac{x^2}{2\epsilon^2}}$$

be the Gaussian density with variance ϵ^2. Given a Borel probability measure $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu)$ and an integrable $f : \mathbb{R} \to \mathbb{R}$, then for any $\epsilon > 0$,

$$\int_{-\infty}^{\infty} f * \rho_\epsilon(x)\, d\mu(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\frac{\epsilon^2 u^2}{2}} \hat{f}(u)\overline{\hat{\mu}(u)}\, du$$

If in addition, $f \in C_b(\mathbb{R})$ and $\hat{f}(u)$ is integrable then

$$\int_{-\infty}^{\infty} f\, d\mu = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{f}(u)\overline{\hat{\mu}(u)}\, du$$

Proof. This is a calculation using Fubini's Theorem (Theorem 3.81) to the triple integral

$$\int\int\int e^{-\frac{\epsilon^2 u^2}{2}} f(x) e^{iux} e^{-iuy}\, d\mu(y)\, dx\, du$$

Note that by Tonneli's Theorem,

$$\begin{aligned}\int\int\int \left| e^{-\frac{\epsilon^2 u^2}{2}} f(x) e^{iux} e^{-iuy} \right| d\mu(y)\, dx\, du &= \int\int\int e^{-\frac{\epsilon^2 u^2}{2}} |f(x)|\, d\mu(y)\, dx\, du \\ &= \int |f(x)|\, dx \int e^{-\frac{\epsilon^2 u^2}{2}}\, du < \infty\end{aligned}$$

and therefore we are justified in using using Fubini's Theorem to calculate via iterated integrals

$$\begin{aligned}\frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\frac{\epsilon^2 u^2}{2}} \hat{f}(u)\overline{\hat{\mu}(u)}\, du &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\frac{\epsilon^2 u^2}{2}} \left(\int_{-\infty}^{\infty} f(x) e^{iux}\, dx \right) \left(\int_{-\infty}^{\infty} e^{-iuy}\, d\mu(y) \right) du \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} f(x) \left(\int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} e^{iu(x-y)} e^{-\frac{\epsilon^2 u^2}{2}}\, du \right) d\mu(y) \right) dx\end{aligned}$$

Now the inner integral is just the Fourier Transform of a Gaussian with mean 0 and variance $\frac{1}{\epsilon^2}$ which we have calculated in Exercise 10.10, so we have by that calculation, another application of Fubini's Theorem and the definition of convolution,

$$\begin{aligned}
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} f(x)\left(\int_{-\infty}^{\infty} \frac{\sqrt{2\pi}}{\epsilon} e^{-(x-y)^2/2\epsilon^2}\, d\mu(y)\right) dx \\
&= \int_{-\infty}^{\infty} f(x)\left(\int_{-\infty}^{\infty} \rho_\epsilon(x-y)\, d\mu(y)\right) dx \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} f(x)\rho_\epsilon(x-y)\, dx\right) d\mu(y) \\
&= \int_{-\infty}^{\infty} f * \rho_\epsilon(y)\, d\mu(y)
\end{aligned}$$

The second part of the theorem is just an application of Lemma 3.103 and the first part of the Theorem. By the Lemma, we know that for any $f \in C_c(\mathbb{R};\mathbb{R})$, we have $\lim_{\epsilon\to 0}\sup_x |f*\rho_\epsilon(x) - f(x)| = 0$. So we have,

$$\lim_{\epsilon\to 0}\left|\int_{-\infty}^{\infty} f - f*\rho_\epsilon\, d\mu\right| \leq \lim_{\epsilon\to 0}\int_{-\infty}^{\infty} |f - f*\rho_\epsilon|\, d\mu \quad \leq \lim_{\epsilon\to 0}\sup_x |f - f*\rho_\epsilon| = 0$$

and by integrability of $\hat{f}(u)$, the fact that $|\hat{\mu}| \leq 1$ (Lemma 10.3) we may use Dominated Convergence to see that

$$\begin{aligned}
\lim_{\epsilon\to 0}\int_{-\infty}^{\infty} f*\rho_\epsilon\, d\mu &= \frac{1}{2\pi}\lim_{\epsilon\to 0}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\epsilon^2 u^2}\hat{f}(u)\overline{\hat{\mu}(u)}\, du \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} \lim_{\epsilon\to 0} e^{-\frac{1}{2}\epsilon^2 u^2}\hat{f}(u)\overline{\hat{\mu}(u)}\, du \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(u)\overline{\hat{\mu}(u)}\, du
\end{aligned}$$

and therefore we have the result. □

As it turns out, we'll get a lot more mileage our the first statement of the Theorem above. We won't really ever be in a position in which we have the required integrability of the Fourier Transform $\hat{f}(t)$ to use the second part. However, the technique used in the proof of the second part of the Theorem will be replayed several times. First we show that the characteristic function completely characterizes probability measures.

Theorem 10.7. *Let μ and ν be a probability measures on $\mathbb{R}^n$ such that $\hat{\mu} = \hat{\nu}$, then $\mu = \nu$.*

Proof. Let $f \in C_c(\mathbb{R})$, then we know by Lemma 3.103 that $\lim_{\epsilon\to 0}\|\rho_\epsilon * f - f\|_\infty = 0$. Then for each $\epsilon > 0$, and using the Plancherel Theorem

$$\begin{aligned}\left|\int f\,d\mu - \int f\,d\nu\right| &\leq \left|\int \rho_\epsilon * f\,d\mu - \int \rho_\epsilon * f\,d\nu\right| + \int |\rho_\epsilon * f - f|\,d\mu + \int |\rho_\epsilon * f - f|\,d\nu \\ &\leq \left|\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\frac{\epsilon^2 u^2}{2}}\hat{f}(u)(\overline{\hat{\mu}(u)} - \overline{\hat{\nu}(u)})\,du\right| + 2\|\rho_\epsilon * f - f\|_\infty \\ &= 2\|\rho_\epsilon * f - f\|_\infty\end{aligned}$$

Taking the limit as ϵ goes to 0, we see that $\int f\,d\mu = \int f\,d\nu$ for all $f \in C_c(\mathbb{R})$.

Now, take a finite interval $[a,b]$ and approximate $\mathbf{1}_{[a,b]}$ by the compactly supported continuous functions

$$f_n(x) = \begin{cases} 1 & \text{for } a \leq x \leq b \\ 0 & \text{for } x < a - \frac{1}{n} \text{ or } x > b + \frac{1}{n} \\ n(x-a)+1 & \text{for } a - \frac{1}{n} \leq x < a \\ 1 - n(x-b) & \text{for } b < x \leq b + \frac{1}{n} \end{cases}$$

It is clear that $f_n(x)$ is decreasing in n and $\lim_{n\to\infty} f_n(x) = \mathbf{1}_{[a,b]}$ so by Monotone Convergence

$$\mu([a,b]) = \lim_{n\to\infty}\int f_n\,d\mu = \lim_{n\to\infty}\int f_n\,d\nu = \nu([a,b])$$

Since the Borel σ-algebra is generated by the closed intervals, we see that $\mu = \nu$. □

Theorem 10.8. *Let $\xi = (\xi_1, \ldots, \xi_n)$ be an $\mathbb{R}^n$-valued random variable. Then the $\mathbb{R}$-valued random variables ξ_i are independent if and only if*

$$\varphi_\xi(u_1, \ldots, u_n) = \prod_{j=1}^{n} \varphi_{\xi_j}(u_j)$$

Proof. TODO: This is a simple corollary that follows by calculating the characteristic function of the product and then using the fact that the characteristic function uniquely defines the distribution. First suppose that the ξ_i and independent. Then we calculate

$$\varphi_\xi(u) = \mathbf{E}\left[e^{i\langle u,\xi\rangle}\right] = \mathbf{E}\left[\prod_{k=1}^{n} e^{iu_k\xi_k}\right] = \prod_{k=1}^{n}\mathbf{E}\left[e^{iu_k\xi_k}\right] = \prod_{k=1}^{n}\varphi_{\xi_k}(u_k)$$

Note that here we have used Lemma 7.16 on a bounded complex valued function. TODO: Do the simple validation that the Lemma extends to this situation.

On the other hand, if we assume that $\varphi_\xi(u_1, \ldots, u_n) = \prod_{j=1}^{n}\varphi_{\xi_j}(u_j)$, then we know that if we pick independent random variables η_j where each η_j has the same distribution as ξ_j then by the above calculation $\varphi_\xi(u) = \varphi_\eta(u)$. By Theorem 10.7 we know that ξ and η have the same distribution. Thus the ξ_j are also independent by Lemma 7.5 and the equality of the distributions of each ξ_j and η_j. □

Lemma 10.9. *Let ξ and η be independent random vectors in $\mathbb{R}^n$. Then $\varphi_{\xi+\eta}(u) = \varphi_\xi(u)\varphi_\eta(u)$.*

Proof. This follows from the calculation

$$\begin{aligned}\varphi_{\xi+\eta}(u) &= \mathbf{E}\left[e^{i\langle u,\xi+\eta\rangle}\right] = \mathbf{E}\left[e^{i\langle u,\xi\rangle}e^{i\langle u,\eta\rangle}\right] \\ &= \mathbf{E}\left[e^{i\langle u,\xi\rangle}\right]\mathbf{E}\left[e^{i\langle u,\eta\rangle}\right] = \varphi_\xi(u)\varphi_\eta(u) \qquad \text{by Lemma 7.16}\end{aligned}$$

□

Example 10.10. Let ξ be an $N(0,1)$ random variable. Then $\varphi_\xi(u) = e^{\frac{-u^2}{2}}$. The least technical way of seeing this requires a bit of a trick. First note that because $\sin ux$ is an odd function we have

$$\begin{aligned}\varphi_\xi(u) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{iux}e^{\frac{-x^2}{2}}\,dx \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\frac{-x^2}{2}}\cos ux\,dx + \frac{i}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\frac{-x^2}{2}}\sin ux\,dx \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\frac{-x^2}{2}}\cos ux\,dx\end{aligned}$$

On the other hand by Lemma 10.5 and the fact that $x\cos ux$ is an odd function we have

$$\begin{aligned}\frac{d\varphi_\xi(u)}{du} &= \frac{i}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{iux}e^{\frac{-x^2}{2}}\,dx \\ &= \frac{i}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{\frac{-x^2}{2}}\cos ux\,dx - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{\frac{-x^2}{2}}\sin ux\,dx \\ &= \frac{-1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xe^{\frac{-x^2}{2}}\sin ux\,dx\end{aligned}$$

This last integral can be integrated by parts (let $df = xe^{\frac{-x^2}{2}}dx$ and $g = \sin ux$, hence $f = -e^{\frac{-x^2}{2}}$ and $dg = u\cos ux$) to yield

$$\frac{d\varphi_\xi(u)}{du} = \frac{-u}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{\frac{-x^2}{2}}\cos ux\,dx$$

and therefore we have shown that characteristic function satisfies the simple first order differential equation $\frac{d\varphi_\xi(u)}{du} = -u\varphi_\xi(u)$ which has the general solution $\varphi_\xi(u) = Ce^{\frac{-u^2}{2}}$ for some constant C. To determine the constant, we use Lemma 10.3 to see that $\varphi_\xi(0) = C = 1$ and we are done.

To extend the previous example to arbitrary normal distributions, we prove the following result that has independent interest.

Lemma 10.11. *Let ξ be a random vector in $\mathbb{R}^N$ then for $a \in \mathbb{R}^M$ and A an $M \times N$ matrix, we have*

$$\varphi_{a+A\xi}(u) = e^{i\langle a,u\rangle}\varphi_\xi(A^*u)$$

where A^* denotes the transpose of A.

Proof. This is a simple calculation

$$\varphi_{a+A\xi}(u) = \mathbf{E}\left[e^{i\langle u,a+A\xi\rangle}\right] = \mathbf{E}\left[e^{i\langle u,a\rangle}e^{i\langle u,A\xi\rangle}\right] = e^{i\langle u,a\rangle}\mathbf{E}\left[e^{i\langle A^*u,\xi\rangle}\right] = e^{i\langle a,u\rangle}\varphi_\xi(A^*u)$$

where we have used the elementary fact from linear algebra that

$$\langle u, Av\rangle = u^*Av = (u^*Av)^* = v^*A^*u = \langle A^*u, v\rangle$$

□

Example 10.12. Let ξ be an $N(\mu, \sigma^2)$ random variable. Then $\varphi_\xi(u) = e^{iu\mu - \frac{1}{2}u^2\sigma^2}$. We know that if η is an $N(0,1)$ random variable then $\mu + \sigma\eta$ is $N(\mu, \sigma^2)$, so by the previous Lemma 10.11 and Example 10.10

$$\varphi_\xi(u) = e^{iu\mu}\varphi_\eta(\sigma u) = e^{iu\mu - \frac{1}{2}u^2\sigma^2}$$

The last piece of the puzzle that we need to put into place before proving the Central Limit Theorem is a result that shows we can test convergence in distribution by looking at pointwise convergence of associated characteristic functions.

Theorem 10.13 (Glivenko-Levy Continuity Theorem)**.** *If $\mu, \mu_1, \mu_2, \ldots$ are probability measures on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, then μ_n converge weakly to μ if and only if $\hat{\mu}_n(u)$ converge to $\hat{\mu}(u)$ pointwise.*

Proof. By Theorem 9.3 it suffices to show that for every $f \in C_c^\infty(\mathbb{R}^n, \mathbb{R})$, we have $\lim_{n\to\infty} \int f\, d\mu_n = \int f\, d\mu$. By 3.103 we know that $\lim_{\epsilon\to 0}\|\rho_\epsilon * f - f\|_\infty = 0$. Pick $\delta > 0$ and find $\epsilon > 0$ such that $\|\rho_\epsilon * f - f\|_\infty < \delta$. Now,

$$\begin{aligned}\left|\int f\, d\mu_n - \int f\, d\mu\right| &\le \left|\int (f - \rho_\epsilon * f)\, d\mu_n\right| + \left|\int \rho_\epsilon * f\, d\mu_n - \int \rho_\epsilon * f\, d\mu\right| + \left|\int (\rho_\epsilon * f - f)\, d\mu\right| \\ &\le \delta + \frac{1}{2\pi}\left|\int \hat{f}(t) e^{-\frac{1}{2}\epsilon^2 t^2}(\hat{\mu}_n(t) - \hat{\mu}(t))\, dt\right| + \delta\end{aligned}$$

where we have used the Plancherel Theorem (Theorem 10.6) and the uniform approximation of f by $\rho_\epsilon * f$ in going from the first to the second line.

Because f is compactly supported, we know that $\hat{f}(t) \le \|f\|_\infty$ and together with Lemma 10.3 we see that

$$\left|\hat{f}(t) e^{-\frac{1}{2}\epsilon^2 t^2}(\hat{\mu}_n(t) - \hat{\mu}(t))\right| \le 2\|f\|_\infty e^{-\frac{1}{2}\epsilon^2 t^2}$$

where the upper bound is an integrable function of t. Therefore by Dominated Convergence we see that $\limsup_{n\to\infty}\left|\int f\, d\mu_n - \int f\, d\mu\right| \le 2\delta$. Since $\delta > 0$ was arbitrary, we have $\int f\, d\mu_n = \int f\, d\mu$. □

Note that part of the hypothesis in the above theorem is the fact that the pointwise limit of the characteristic functions is assumed to be the characteristic function of a probability measure. There is a stronger form of the above theorem that characterizes when a pointwise limit of charactersistic functions is in fact the characteristic function of a probability measure. That stronger result is not needed to prove the Central Limit Theorem so we postpone its statement and proof until later.

Theorem 10.14 (Central Limit Theorem)**.** *Let $\xi, \xi_1, \xi_2, \ldots$ be i.i.d. random variables with $\mu = \boldsymbol{E}[\xi]$ and $\sigma = \boldsymbol{Var}(\xi_n) < \infty$, then*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \xi_i - \mu\right) \xrightarrow{d} N(0, \sigma^2)$$

Proof. The first thing to note is that by using the Theorem on $\frac{\xi_i - \mu}{\sigma}$, it suffices to assume that $\mu = 0$ and $\sigma = 1$. Thus we only have to show that $\frac{1}{\sqrt{n}} \sum_{k=1}^{n} \xi_k \xrightarrow{d} N(0,1)$.

Define $S_n = \sum_{k=1}^{n} \xi_k$. By Theorem 10.13 it suffices to show that

$$\lim_{n\to\infty} \mathbf{E}\left[e^{itS_n/\sqrt{n}}\right] = e^{t^2/2}$$

To calculate the limit, first note that by independence and i.i.d. we have

$$\mathbf{E}\left[e^{itS_n/\sqrt{n}}\right] = \prod_{k=1}^{n} \mathbf{E}\left[e^{it\xi_k/\sqrt{n}}\right] = \left[\mathbf{E}\left[e^{it\xi/\sqrt{n}}\right]\right]^n$$

In order to evaluate the limit, we take the Taylor expansion of the exponential $e^{ix} = 1 + ix - \frac{1}{2}x^2 + R(x)$ where by Lagrange form of the remainder and the fact that $\left|\frac{d}{dx}e^{ix}\right| \leq 1$, we see that $|R(x)| \leq \frac{1}{6}|x|^3$. Note that this estimate isn't very good for large $|x|$ but it is easy to do better for $|x| > 1$ just using the triangle inequality

$$\left|e^{ix} - 1 - ix + \frac{1}{2}x^2\right| \leq 2 + |x| + \frac{1}{2}x^2 \leq \frac{7}{2}x^2$$

Therefore we have the bound $|R(x)| \leq \frac{7}{2}(|x|^3 \wedge x^2)$. Applying the Taylor expansion and using the zero mean and unit variance assumption, we get

$$\mathbf{E}\left[e^{itS_n/\sqrt{n}}\right] = \left(1 - \frac{t^2}{2n} + \mathbf{E}\left[R(\frac{t\xi}{\sqrt{n}})\right]\right)^n$$

By our estimate on the remainder term, we can see that

$$\begin{aligned} n\left|\mathbf{E}\left[R(\frac{t\xi}{\sqrt{n}})\right]\right| &\leq \frac{7}{2}\mathbf{E}\left[\frac{t^3|\xi|^3}{\sqrt{n}} \wedge t^2\xi^2\right] \\ &\leq \frac{7}{2}\mathbf{E}\left[t^2\xi^2\right] = \frac{7t^2}{2} \end{aligned}$$

By the above inequalities and Dominated Convergence we can conclude that

$$\lim_{n\to\infty} n\left|\mathbf{E}\left[R(\frac{t\xi}{\sqrt{n}})\right]\right| = 0$$

so if we define $\epsilon_n = \frac{2n}{t^2}\left|\mathbf{E}\left[R(\frac{t\xi}{\sqrt{n}})\right]\right|$ then we have $\lim_{n\to\infty}\epsilon_n = 0$ and

$$\lim_{n\to\infty} \mathbf{E}\left[e^{itS_n/\sqrt{n}}\right] = \lim_{n\to\infty}\left(1 - \frac{t^2}{2n}(1+\epsilon_n)\right)^n = \lim_{n\to\infty} e^{n\log(1-\frac{t^2}{2n}(1+\epsilon_n))} = e^{-t^2/2}$$

□

Theorem 10.15 (Prokhorov's Theorem, special case)**.** *Let μ_n be a tight sequence of measures on $\mathbb{R}^n$. Then there is a subsequence of that converges in distribution.*

Proof. TODO □

TODO: Do the full Levy Continuity Theorem (and Prokhorov's Theorem) that shows a characteristic function that is continuous at 0 is the characterisitic function of a probability measure (the basic point is that the pointwise limit of characteristic functions of probability measures is almost the characteristic function of a probability measure; the associated distribution function may not have the correct limits

at $\pm\infty$. If we assume continuity at 0, then we can prove tightness which shows that the limits are $0, 1$ as required of a distribution function.

10.1. **Gaussian Random Vectors and the Multidimensional Central Limit Theorem.** There is a version of the Central Limit Theorem for random vectors in $\mathbb{R}^N$ in which Gaussian distributions also occur. The nature of Gaussians in this context is a bit more subtle than in the one dimensional case. We lead with a definition

Definition 10.16. A random vector ξ in $\mathbb{R}^N$ is said to be a *Gaussian random vector* is for every $a \in \mathbb{R}^N$, the random variable $\{a, \xi\}$ is a univariate normal or is almost surely 0 (which we take as the degenerate univariate normal $N(0, 0)$).

The first theorem that we prove gives an alternative characterization of the property in terms of characteristic functions. This result is sometimes used as the definition of a Gaussian random vector; the only real benefit to the definition we've given is that it is more elementary.

Theorem 10.17. *A random vector ξ in $\mathbb{R}^N$ is Gaussian if and only if there is a $\mu \in \mathbb{R}^N$ and a symmetric nonnegative semi-definite matrix $Q \in \mathbb{R}^{N\times N}$ such that*

$$\varphi_\xi(u) = e^{i\langle u,\mu\rangle - \frac{1}{2}\langle u,Qu\rangle}$$

For ξ with characteristic function of this form, $\mu = \boldsymbol{E}[\xi]$ and $Q = \boldsymbol{Cov}(\xi)$; we say that ξ is $N(\mu, Q)$.

Proof. First we assume that we have a characteristic function of the above form. Let $a \in \mathbb{R}^N$ and consider the random variable $\langle a, \xi\rangle$. Notice that $\langle a, \xi\rangle = a^*\xi$ is a special case of an affine transformation so we can apply Lemma 10.11 to calculate

$$\varphi_{\langle a,\xi\rangle}(u) = \varphi_\xi(au) = e^{iu\langle a,\mu\rangle - \frac{1}{2}\langle a,Qa\rangle u^2}$$

Now, by Example 10.12 we see that $\langle a, \xi\rangle$ is $N(\langle a, \mu\rangle, \langle a, Qa\rangle)$. Since a was arbitrary, this shows that ξ is Gaussian.

Now we assume that ξ is Gaussian. Let $\mu = (\mu_1, \ldots, \mu_N) = \mathbf{E}[\xi]$ and let $Q = \mathbf{Cov}(\xi)$. Pick $a \in \mathbb{R}^N$ and note that

$$\begin{aligned}
\mathbf{E}[\langle a, \xi\rangle] &= \langle a, \mu\rangle \\
\mathbf{Var}(\langle a, \xi\rangle) &= \mathbf{E}\left[(\langle a, \xi\rangle - \mathbf{E}[\langle a, \xi\rangle])^2\right] \\
&= \mathbf{E}\left[(\langle a, \xi - \mu\rangle)^2\right] \\
&= \mathbf{E}[a^*(\xi - \mu)(\xi - \mu)^* a] \\
&= a^*\mathbf{E}[(\xi - \mu)(\xi - \mu)^*]\, a = \langle a, Qa\rangle
\end{aligned}$$

Now we know by our assumption and the expectation and variance calculation above that $\langle a, \xi\rangle$ is $N(\langle a, \mu\rangle, \langle a, Qa\rangle)$ and by Example 10.12, we have

$$\varphi_{\langle a,\xi\rangle}(u) = e^{iu\langle a,\mu\rangle - \frac{1}{2}\langle a,Qa\rangle u^2}$$

As above we can apply Lemma 10.11 to see

$$\varphi_\xi(a) = \varphi_{\langle a,\xi\rangle}(1) = e^{i\langle a,\mu\rangle - \frac{1}{2}\langle a,Qa\rangle}$$

Together with the fact two measures with the same characteristic function must be equal (Theorem 10.7), this also proves the last part of the Theorem since we have shown by construction that $\mu = \mathbf{E}[\xi]$ and $Q = \mathbf{Cov}(\xi)$. □

Example 10.18. Let $\xi_1, \ldots, \xi_N$ be independent random variables with ξ_i being normal $N(\mu_i, \sigma_i^2)$. Then $\xi = (\xi_1, \ldots, \xi_N)$ is a Gaussian random vector. In fact, if we let $\mu = (\mu_1, \ldots, \mu_N)$ and

$$Q = \text{Diag}(\sigma_1^2, \ldots, \sigma_N^2)$$

then $\xi = N(\mu, Q)$.

The characterization of Gaussian random vectors using characteristic functions allows us to see that limits of Gaussian random vectors are Gaussian random vectors. We will need this result when we construct Brownian motion later on.

Lemma 10.19. *Let $\xi_1, \xi_2, \ldots$ be a sequence of random vectors in $\mathbb{R}^N$ with ξ_n an $N(\mu_n, C_n)$ Gaussian random vector. Suppose that ξ is a random vector such that ξ_n converges to ξ almost surely. If $\lim_{n\to\infty} \boldsymbol{E}[\xi_n] = \mu$ and $\lim_{n\to\infty} \boldsymbol{Cov}(() \xi_n) = C$ then ξ is a $N(\mu, C)$ Gaussian random vector.*

Proof. Since ξ_n converges almost surely to ξ then it converges in distribution. We know from Lemma 10.17 and the Glivenko-Levy Continuity Theorem (Theorem 10.13) we see

$$\varphi_\xi(u) = \lim_{n\to\infty} \varphi_{\xi_n}(u) = \lim_{n\to\infty} e^{i\langle u, \mu_n\rangle - \frac{1}{2}\langle u, C_n y\rangle} = e^{i\langle <u, \mu\rangle - \frac{1}{2}\langle u, Cy\rangle}$$

where we have used continuity of $e^i x$. Thus, using Lemma 10.17 again shows that ξ is $N(\mu, C)$. □

TODO: Gaussian Random Variables in $\mathbb{R}^n$ and the multidimensional CLT. TODO: Show that a given two independent Gaussian random variables their sum and difference are independent Gaussian (that probably doesn't require Gaussian random vectors). Not sure we really need to call this out as a Lemma.

11. Conditioning

11.1. L^p **Spaces.** Prior to discussing the general formulation of the notion of conditional probabilities we shall need to lay down some techniques of functional analysis pertaining to spaces of measurable (and integrable) random variables.

Definition 11.1. Given a measure space $(\Omega, \mathcal{A}, \mu)$ and $p \geq 1$ we let $L^p(\Omega, \mathcal{A}, \mu)$ be the space of equivalence classes of measurable functions such that $\int |f|^p \, d\mu < \infty$ under the equivalence relation of almost everywhere equality. For any element $f \in L^p(\Omega, \mathcal{A}, \mu)$ we define

$$\|f\|_p = \left(\int |f|^p \, d\mu \right)^{\frac{1}{p}}$$

It is clear that the spaces $L^p(\Omega, \mathcal{A}, \mu)$ but our first goal is to establish that each is a complete normed vector space (a.k.a. Banach space). As our first step in that direction we need to prove the triangle inequality

Lemma 11.2 (Minkowski Inequality)**.** *Given $f, g \in L^p(\Omega, \mathcal{A}, \mu)$ then $f + g \in L^p(\Omega, \mathcal{A}, \mu)$ and $\|f + g\|_p \leq \|f\|_p + \|g\|_p$.*

Proof. Note that it suffices to assume that $f \geq 0$ and $g \geq 0$ since if we have the inequality for positive elements then it follows for all elements by applying the ordinary triangle inequality on $\mathbb{R}$ and using the fact that x^p is increasing to see

$$\|f + g\|_p \leq \||f| + |g|\|_p \leq \||f|\|_p + \||g|\|_p = \|f\|_p + \|g\|_p$$

The case $p = 1$ follows immediately from linearity of integral (in fact we have equality).

For $1 < p < \infty$, first use the following crude pointwise bound to see that $f + g \in L^p(\Omega, \mathcal{A}, \mu)$:

$$(f+g)^p \leq (f \vee g + f \vee g)^p = 2^p f^p \vee g^p \leq 2^p (f^p + g^p)$$

and therefore $\|f+g\|_p^p \leq 2^p(\|f\|_p^p + \|g\|_p^p) < \infty$. To see the triangle inequality, note that we can assume that $\|f+g\|_p > 0$ for otherwise the triangle inequality follows by positivity of the norm. Write

$$\|f+g\|_p^p = \int (f+g)^p \, d\mu = \int f(f+g)^{p-1} \, d\mu + \int g(f+g)^{p-1} \, d\mu$$

Now we can apply the Hölder Inequality (Lemma 6.11) to each of the terms on the right hand side and use the fact that $\frac{1}{p} + \frac{1}{q}$ is equivalent to $q = (p-1)q$ to see

$$\int f(f+g)^{p-1} \, d\mu \leq \left(\int f^p \, d\mu \right)^{\frac{1}{p}} \left(\int (f+g)^{(p-1)q} \, d\mu \right)^{\frac{1}{q}} = \|f\|_p \|f+g\|_q^{p/q}$$

Applying this argument to the term $\int g(f+g)^{p-1} \, d\mu$ as well we get

$$\|f+g\|_p^p \leq (\|f\|_p + \|g\|_p) \cdot \|f+g\|_q^{p/q}$$

and dividing through by $\|f+g\|_p^{p/q}$ and using $p - \frac{p}{q} = 1$ we get $\|f+g\|_p \leq \|f\|_p + \|g\|_p$. □

Lemma 11.3. *For $p \geq 1$ the normed vector space $L^p(\Omega, \mathcal{A}, \mu)$ is complete.*

Proof. Let f_n be a Cauchy sequence in $L^p(\Omega, \mathcal{A}, \mu)$. The first step of the proof is to show that there is a subsequence of f_n that converges almost everywhere to an element $f \in L^p(\Omega, \mathcal{A}, \mu)$.

By the Cauchy property, for each $j \in \mathbb{N}$ we can find an $n_j > 0$ such that $\|f_m - f_{n_j}\|_p \leq \frac{1}{2^j}$ for all $m > n_j$. In this way we get a subsequence f_{n_j} such that $\|f_{n_{j+1}} - f_{n_j}\|_p \leq \frac{1}{2^j}$ for all $j \in \mathbb{N}$. Now by applying Monotone Convergence and the triangle inequality we have

$$\begin{aligned} \| \sum_{j=1}^{\infty} \left| f_{n_{j+1}} - f_{n_j} \right| \|_p &= \lim_{N \to \infty} \| \sum_{j=1}^{N} \left| f_{n_{j+1}} - f_{n_j} \right| \|_p \\ &\leq \lim_{N \to \infty} \sum_{j=1}^{N} \| f_{n_{j+1}} - f_{n_j} \|_p \\ &\leq \lim_{N \to \infty} \sum_{j=1}^{N} \frac{1}{2^j} < \infty \end{aligned}$$

and therefore we know that $\sum_{j=1}^{\infty} \left| f_{n_{j+1}} - f_{n_j} \right|$ is almost surely finite. Anywhere this sum is finite it follows that f_{n_j} is a Cauchy sequence in $\mathbb{R}$. To see this, suppose we are given $\epsilon > 0$ we pick $N > 0$ such that $\sum_{j=N}^{\infty} \left| f_{n_{j+1}} - f_{n_j} \right| < \epsilon$, then for any

$k \geq j \geq N$ we have

$$\left|f_{n_k} - f_{n_j}\right| = \left|\sum_{m=j}^{k} (f_{n_{m+1}} - f_{n_m})\right| \leq \sum_{m=j}^{k} \left|f_{n_{m+1}} - f_{n_m}\right| < \epsilon$$

We know that the set where f_{n_j} converges is measurable (TODO: Where is this?) so we can define f to be the limit of the Cauchy sequence f_{n_j} where valid and define it to be zero elsewhere (a set of measure zero).

To see that $f \in L^p(\Omega, \mathcal{A}, \mu)$ and to show that f_n converges to f, suppose $\epsilon > 0$ is given and pick $N \in \mathbb{N}$ such that for all $m, n \geq N$ we have $\|f_m - f_n\|_p < \epsilon$. Now we can use Fatou's Lemma (Theorem 3.42) to see for any $n \geq N$,

$$\int |f - f_n|^p \, d\mu \leq \liminf_{j\to\infty} \int \left|f_{n_j} - f_n\right|^p \, d\mu \leq \sup_{m\geq n} \int |f_m - f_n|^p \, d\mu < \epsilon^p$$

Therefore by the Minkowski Inequality, we see that $f = f_n + (f - f_n)$ is in $L^p(\Omega, \mathcal{A}, \mu)$ and $f_n \xrightarrow{L^p} f$. □

We know that measurable functions can be approximated by simple functions (Lemma 3.17) with pointwise convergence. It is useful to extend this approximation to L^p spaces.

Lemma 11.4. *Simple functions are dense in $L^p(\Omega, \mathcal{A}, \mu)$.*

Proof. Let's assume for the moment that μ is a finite measure. In that case, every simple function is in L^p. Pick a positive function $f \in L^p(\Omega, \mathcal{A}, \mu)$ and sequence of simple functions such that $f_n \uparrow f$. Then it is also true that $f_n^p \uparrow f^p$ and Monotone Convergence tells us that $\lim_{n\to\infty} \|f_n\|_p = \|f\|_p$. By Lemma 8.52 we conclude that $f_n \xrightarrow{L^p} f$.

To finish the proof, take an arbitrary f and write it as $f = f_+ - f_-$ and use linearity.

TODO: What about non-finite measures? This argument clearly extends to σ-finite by restricting each f_n to a finite subset. In the general case we need to handle the fact that all simple functions are not in L^p. □

Note that for any σ-algebra $\mathcal{F} \subset \mathcal{A}$ we can also consider the space $L^p(\Omega, \mathcal{F}, \mu)$. As we shall soon see, it will become important to understand a bit about these spaces as $\mathcal{F}$ vary. The first thing to note is that for $\mathcal{G} \subset \mathcal{F}$, $L^p(\Omega, \mathcal{G}, \mu)$ is a closed linear subspace of $L^p(\Omega, \mathcal{F}, \mu)$. The inclusion is trivial since any $\mathcal{G}$-measurable function is also $\mathcal{F}$-measurable; closure follows from the completeness of the space $L^p(\Omega, \mathcal{G}, \mu)$ (Lemma 11.3).

The following approximation result will be used only occasionaly.

Lemma 11.5. *$\cup_n L^p(\Omega, \mathcal{F}_n, \mu)$ is dense in $L^p(\Omega, \bigvee_n \mathcal{F}_n, \mu)$*

Proof. The first thing to show the result for indicator functions. A general fact, suppose V is a closed linear subspace of L^p and let $\mathcal{C} = \{A \mid \mathbf{1}_A \in V\}$. We claim that $\mathcal{C}$ is a λ-system. Given $A, B \in \mathcal{C}$ with $A \subset B$, we have $B \setminus A \in \mathcal{C}$ since $\mathbf{1}_{B\setminus A} = \mathbf{1}_B - \mathbf{1}_A$ and V is a linear space. Now assume that $A_1 \subset A_2 \subset \cdots \in \mathcal{C}$. We have that $\mathbf{1}_{A_n} \uparrow \mathbf{1}_A$ and continuity of measure (Lemma 3.27) tells us that $\lim_{n\to\infty} \|\mathbf{1}_{A_n}\|_p = \|\mathbf{1}_A\|_p$ so Lemma 8.52 implies $\mathbf{1}_{A_n} \xrightarrow{L^p} \mathbf{1}_A$. Since V is closed we know $\mathbf{1}_A \in V$. □

TODO: Develop inner product and projection for L^2 spaces.

11.2. **Conditional Expectation.** Before getting into the technical details we want to get set the intuition for the problem and the form that solutions will take. Given a random element ξ in S and a random variable η, we want to formulate the notion of the expected value of η given a value of ξ. The immediate way to think of representing such an object is as a map from S to $\mathbb{R}$. In practice the representation is expressed in a different but equivalent way. Recall from Lemma 3.20 that any random variable γ that is ξ-measurable can be factored as $f \circ \xi$ for some measurable $f : S \to \mathbb{R}$. In this way the conditional expectation may equally be considered as ξ-measurable random variable. It is this latter representation that is most convenient for working with (and constructing) conditional expectations. To remove matters a little further from the initial intuition, one often makes use of the fact that the conditional expection winds up only depending on the σ-field induced by ξ and discusses conditioning with respect to arbitrary sub σ-fields.

TODO: Elaborate on the three faces of conditional expectation: projection, density/Radon-Nikodym derivative and disintegration.

Existence via Radon-Nikodym. The Radon-Nikodym theorem (Theorem 3.93) can be given a martingale proof (hence derived in some sense from the existence of conditional expectations). However, the standard proof for Radon-Nikodym using Hahn Decomposition does not depend on the existence of conditional expection and in fact, the Radon-Nikodym theorem can easily be used to prove the existence of conditional expectations. Given $\xi \geq 0$ and $\mathcal{F} \subset \mathcal{A}$, then define the probability measure $\nu(A) = \mathbf{E}[\xi \mathbf{1}_A]$. Note that ν is absolutely continuous with respect to μ on $\mathcal{F}$. Therefore, the Radon-Nikodym derivative with respect to $(\Omega, \mathcal{F})$ exists and satisfies

$$\nu(A) = \mathbf{E}[\xi \mathbf{1}_A] = \mathbf{E}\left[\frac{d\nu}{d\mu}\mathbf{1}_A\right]$$

for all $A \in \mathcal{F}$. This equality shows that $\frac{d\nu}{d\mu}$ is a conditional expectation of ξ. For general ξ, write $\xi = \xi_+ - \xi_-$ and proceed as above.

TODO: Make sure we have covered the following: Definition of L^p spaces, completeness of L^p spaces, definition of Hilbert space, orthogonal projections in Hilbert spaces. Density of L^2 in L^1. Unique extension of a bounded linear operator from a dense subspace of a complete normed linear space.

On the other hand, there is very appealing construction of conditional expectation using function spaces that we provide here. Recall that for a measurable space $(\Omega, \mathcal{A}, \mu)$ we have associated Banach spaces of p-integrable functions $L^p(\Omega, \mathcal{A}, \mu)$ with norm $\|f\|_p = \left(\int |f|^p \, d\mu\right)^{\frac{1}{p}}$. In the special case $p = 2$ we actually have a Hilbert space $L^2(\Omega, \mathcal{A}, \mu)$ with inner product $< f, g >= \int fg \, d\mu$. Suppose we have a sub σ-algebra $\mathcal{F} \subset \mathcal{A}$ and we have a canonical inclusion $L^p(\Omega, \mathcal{F}, \mu) \subset L^p(\Omega, \mathcal{A}, \mu)$ as a subspace. In fact by the completeness of $L^p(\Omega, \mathcal{F}, \mu)$, we know that this is a *closed* subspace. Therefore if we specialize to the case of $L^2(\mathcal{F}) \subset L^2(\mathcal{A})$ then we have the orthogonal projection onto $L^2(\mathcal{F})$. For square integrable random variables, this orthogonal projection defines the conditional expectation. In the following, we extend this defintion to all integrable random variables and prove the basic properties.

TODO: Elaborate on the "a.s. uniqueness" in the definition.

Theorem 11.6 (Conditional Expectation)**.** *For any $\mathcal{F} \subset \mathcal{A}$ there exists a unique linear operator $\boldsymbol{E}^{\mathcal{F}} : L^1 \to L^1(\mathcal{F})$ such that*

(i) $\boldsymbol{E}\left[\boldsymbol{E}^{\mathcal{F}}\xi; A\right] = \boldsymbol{E}[\xi; A]$ *for all* $\xi \in L^1$, $A \in \mathcal{F}$

The following properties also hold for $\xi, \eta \in L^1$,

(ii) $\boldsymbol{E}\left[\left|\boldsymbol{E}^{\mathcal{F}}\xi\right|\right] \leq \boldsymbol{E}[|\xi|]$ *a.s.*
(iii) $\xi \geq 0$ *implies* $\boldsymbol{E}^{\mathcal{F}}\xi \geq 0$ *a.s.*
(iv) $0 \leq \xi_n \uparrow \xi$ *implies* $\boldsymbol{E}^{\mathcal{F}}\xi_n \uparrow \boldsymbol{E}^{\mathcal{F}}\xi$ *a.s.*
(v) $\boldsymbol{E}^{\mathcal{F}}\xi\eta = \xi\boldsymbol{E}^{\mathcal{F}}\eta$ *if ξ is $\mathcal{F}$-measurable and* $\xi\eta, \xi\boldsymbol{E}^{\mathcal{F}}\eta \in L^1$
(vi) $\boldsymbol{E}\left[\boldsymbol{E}^{\mathcal{F}}\xi \cdot \boldsymbol{E}^{\mathcal{F}}\eta\right] = \boldsymbol{E}\left[\xi \cdot \boldsymbol{E}^{\mathcal{F}}\eta\right] = \boldsymbol{E}\left[\boldsymbol{E}^{\mathcal{F}}\xi \cdot \eta\right]$
(vii) $\boldsymbol{E}^{\mathcal{F}}\boldsymbol{E}^{\mathcal{G}}\xi = \boldsymbol{E}^{\mathcal{F}}\xi$ *a.s. for all* $\mathcal{F} \subset \mathcal{G}$.

Proof. Begin by defining $\mathbf{E}^{\mathcal{F}} : L^2 \to L^2(\mathcal{F})$ as orthogonal projection. If we pick $A \in \mathcal{F}$, then $\mathbf{1}_A \in L^2(\mathcal{F})$ and therefore, $\xi - \mathbf{E}^{\mathcal{F}}\xi \perp \mathbf{1}_A$ which shows

$$\mathbf{E}[\xi; A] =< \xi, \mathbf{1}_A >=< \mathbf{E}^{\mathcal{F}}\xi, \mathbf{1}_A >= \mathbf{E}\left[\mathbf{E}^{\mathcal{F}}\xi; A\right]$$

If we define $A = \{\mathbf{E}^{\mathcal{F}}\xi \geq 0\}$ the above implies

$$\begin{aligned}
\mathbf{E}\left[\left|\mathbf{E}^{\mathcal{F}}\xi\right|\right] &= \mathbf{E}\left[\mathbf{E}^{\mathcal{F}}\xi; A\right] - \mathbf{E}\left[\mathbf{E}^{\mathcal{F}}\xi; A^c\right] && \text{by linearity of expectation} \\
&= \mathbf{E}[\xi; A] - \mathbf{E}[\xi; A^c] && \text{by (i)} \\
&\leq \mathbf{E}[|\xi|; A] + \mathbf{E}[|\xi|; A^c] && \text{since } \xi \leq |\xi| \text{ and } -\xi \leq |\xi| \\
&= \mathbf{E}[|\xi|] && \text{by linearity of expecation}
\end{aligned}$$

This inequality shows us that the linear operator $\mathbf{E}^{\mathcal{F}}$ is bounded in the L^1 norm as well as in the L^2 norm. On the other hand, we know that L^2 is dense in L^1 and L^1 is complete so there is a unique extension of $\mathbf{E}^{\mathcal{F}}$ to a bounded linear operator $L^1 \to L^1\mathcal{F}$. Concretely, for any $\xi \in L^1$, we pick a sequence $\xi_n \in L^2$ such that $\lim_{n\to\infty} \xi_n \to \xi$ in the L^1 norm and define $\mathbf{E}^{\mathcal{F}}\xi = \lim_{n\to\infty} \mathbf{E}^{\mathcal{F}}\xi_n$ where the limit is in the L^1 norm. Since the L^1 closure of $L^2(\mathcal{F})$ is $L^1(\mathcal{F})$, we see that the definition is plausible.

TODO: Show independence, linearity and boundedness of the extension. Perhaps factor this out into a separate Lemma; it is a generic construction.

To see that the condition (i) uniquely defines $\mathbf{E}^{\mathcal{F}}\xi$ a.s., suppose we had two $\mathcal{F}$-measurable random variables η and ρ for which $\mathbf{E}[\eta; A] = \mathbf{E}[\rho; A]$ for all $A \in \mathcal{F}$. Let $A = \{\eta > \rho\}$ which is $\mathcal{F}$-measurable and so we have assumed $\mathbf{E}[\eta - \rho; A] = 0$. If we apply Lemma 3.47 we know that $(\eta - \rho)\mathbf{1}_A = 0$ a.s. which shows that $\mathbf{P}\{A\} = 0$. The same argument shows that $\rho > \eta$ with probability 0, hence $\eta = \rho$ a.s.

To see (iii), let $A = \{\mathbf{E}^{\mathcal{F}}\xi < 0\}$ and observe that

$$0 \leq \mathbf{E}\left[-\mathbf{E}^{\mathcal{F}}\xi; A\right] = \mathbf{E}[-\xi; A] \leq 0$$

and therefore $\mathbf{E}\left[-\mathbf{E}^{\mathcal{F}}\xi; A\right] = 0$ which applying Lemma 3.47 implies $\mathbf{P}\{A\} = 0$.

To see (iv), suppose $0 \leq \xi_n \uparrow \xi$ a.s. Then by Monotone Convergence, $\lim_{n\to\infty} \mathbf{E}[|\xi - \xi_n|] = 0$. Now by (ii) and linearity of conditional expection,

$$0 \leq \lim_{n\to\infty} \mathbf{E}\left[\left|\mathbf{E}^{\mathcal{F}}\xi - \mathbf{E}^{\mathcal{F}}\xi_n\right|\right] \leq \lim_{n\to\infty} \mathbf{E}[|\xi - \xi_n|] = 0$$

which shows that $\mathbf{E}^{\mathcal{F}}\xi_n$ converges to $\mathbf{E}^{\mathcal{F}}\xi$ in L^1. Now by Lemma 8.6 this implies that the converges is in probability and by Lemma 8.9 there is a subsequence that converges a.s. By (iii) we know that $\mathbf{E}^{\mathcal{F}}\xi_n$ is non-decreasing so we know by Lemma 2.15 that that almost sure convergence of the subsequence extends to the almost sure convergence of the entire sequence.

To see (v), note that if ξ is $\mathcal{F}$-measurable then for every $\eta \in L^1$, we know $\xi\mathbf{E}^{\mathcal{F}}\eta$ is $\mathcal{F}$-measurable and by simple calculation

$$\mathbf{E}\left[\xi\mathbf{E}^{\mathcal{F}}\eta; A\right] = \mathbf{E}\left[\xi\eta; A\right]$$

by the apply the extension of the property (i) to the $\mathcal{F}$-measurable function $\xi\mathbf{1}_A$. Now by (v) follows by applying (i) again.

For the property (vi), by symmetry we only have to prove $\mathbf{E}\left[\mathbf{E}^{\mathcal{F}}\xi \cdot \mathbf{E}^{\mathcal{F}}\eta\right] = \mathbf{E}\left[\xi \cdot \mathbf{E}^{\mathcal{F}}\eta\right]$. To prove this first assume that $\xi, \eta \in L^2$. In that case, we know that $\mathbf{E}^{\mathcal{F}}\eta \in L^2(\mathcal{F})$ and $\xi - \mathbf{E}^{\mathcal{F}}\xi \perp L^2(\mathcal{F})$, so

$$\begin{aligned}\mathbf{E}\left[\mathbf{E}^{\mathcal{F}}\xi \cdot \mathbf{E}^{\mathcal{F}}\eta\right] &=< \mathbf{E}^{\mathcal{F}}\xi, \mathbf{E}^{\mathcal{F}}\eta > \\ &=< \mathbf{E}^{\mathcal{F}}\xi - \xi, \mathbf{E}^{\mathcal{F}}\eta > + < \xi, \mathbf{E}^{\mathcal{F}}\eta > \\ &=< \xi, \mathbf{E}^{\mathcal{F}}\eta >= \mathbf{E}\left[\xi \cdot \mathbf{E}^{\mathcal{F}}\eta\right]\end{aligned}$$

Now by the density of $L^2 \subset L^1$, for general $\xi, \eta \in L^1$ we pick $\xi_n \xrightarrow{L^1} \xi$ and $\eta_n \xrightarrow{L^1} \eta$ with $\xi_n, \eta_n \in L^2$. By the above Lastly, we prove (vii). Suppose we are given σ-algebras $\mathcal{F} \subset \mathcal{G}$. Then for $A \in \mathcal{F} \subset \mathcal{G}$,

$$\begin{aligned}\mathbf{E}\left[\mathbf{E}^{\mathcal{G}}\xi; A\right] &= \mathbf{E}\left[\xi; A\right] && \text{by (i) applied to } \mathbf{E}^{\mathcal{G}}\xi \\ &= \mathbf{E}\left[\mathbf{E}^{\mathcal{F}}\xi; A\right] && \text{by (i) applied to } \mathbf{E}^{\mathcal{F}}\xi\end{aligned}$$

where are the equalities are a.s. By definition $\mathbf{E}^{\mathcal{F}}\xi$ is $\mathcal{F}$-measurable which shows by (i) that $\mathbf{E}^{\mathcal{F}}\mathbf{E}^{\mathcal{G}}\xi = \mathbf{E}^{\mathcal{F}}\xi$ a.s. □

When verifying the defining property of conditional expectation it is often useful to observe that it suffices to check indicator functions for sets in a generating π-system.

Lemma 11.7. *Suppose ξ, η are integrable or non-negative random variables and $\mathcal{F}$ is a π-system such that $\Omega \in \mathcal{F}$ and for all $A \in \mathcal{F}$, we have $\boldsymbol{E}[\xi; A] = \boldsymbol{E}[\eta; A]$. Then we have $\boldsymbol{E}[\xi; A] = \boldsymbol{E}[\eta; A]$ for all $A \in \sigma(\mathcal{F})$.*

Proof. We first let $\mathcal{G}$ be the set of all A such that $\mathbf{E}[\xi; A] = \mathbf{E}[\eta; A]$ and show that it is a λ-system. If $A, B \in \mathcal{G}$ and $B \supset A$ then

$$\mathbf{E}[\xi; B \setminus A] = \mathbf{E}[\xi; B] - \mathbf{E}[\xi; A] = \mathbf{E}[\eta; B] - \mathbf{E}[\eta; A] = \mathbf{E}[\eta; B \setminus A]$$

Now suppose that we have $A_1 \subset A_2 \subset \cdots \in \mathcal{G}$. We claim that $\lim_{n\to\infty} \mathbf{E}[\xi; A_n] = \mathbf{E}[\xi; \cup_n A_n]$ and similarly with η. In the case that we assume ξ is integrable then

we have $|\xi \mathbf{1}_{A_n}| \leq |\xi|$, so we may use Dominated Convergence whereas in the case that ξ is non-negative we may use Monotone Convergence. In either case,

$$\mathbf{E}\left[\xi; \cup_n A_n\right] = \lim_{n\to\infty} \mathbf{E}\left[\xi; A_n\right] = \lim_{n\to\infty} \mathbf{E}\left[\eta; A_n\right] = \mathbf{E}\left[\eta; \cup_n A_n\right]$$

We have assumed that $\Omega \in \mathcal{G}$ therefore we have shown $\mathcal{G}$ is a λ-system and our assumption is that $\mathcal{F} \subset \mathcal{G}$ so we apply the π-λ Theorem (Theorem 3.24) to get the result. □

Occasionally it can be useful to extend the defining property of conditional expectation beyond indicator functions.

Lemma 11.8. *Let $\xi \in L^1$ then for a σ-algebra $\mathcal{F}$ and for any $\eta \in L^1(\mathcal{F})$ such that $\eta\xi$ and $\eta \boldsymbol{E}^{\mathcal{F}}\xi$ are both integrable, $\boldsymbol{E}\left[\boldsymbol{E}^{\mathcal{F}}\xi \cdot \eta\right] = \boldsymbol{E}[\xi \cdot \eta]$.*

Proof. This is a simple application of the standard machinery. Property (i) is exactly this statement for $\mathcal{F}$-measurable indicator functions. Linearity of expectation shows that the statement then holds for $\mathcal{F}$-measurable simple functions. For $\mathcal{F}$-measurable $\eta \geq 0$ satisfying the requirements of the Lemma, we pick an increasing approximation by simple functions $\eta_n \uparrow \eta$. Now we can apply Dominated Convergence to the sequences $\mathbf{E}^{\mathcal{F}}\xi \cdot \eta_n$ and $\xi \cdot \eta_n$,

$$\begin{aligned}
\mathbf{E}\left[\xi \cdot \eta\right] &= \lim_{n\to\infty} \mathbf{E}\left[\xi \cdot \eta_n\right] && \text{by Dominated Convergence} \\
&= \lim_{n\to\infty} \mathbf{E}\left[\mathbf{E}^{\mathcal{F}}\xi \cdot \eta_n\right] \\
&= \mathbf{E}\left[\mathbf{E}^{\mathcal{F}}\xi \cdot \eta\right] && \text{by Dominated Convergence}
\end{aligned}$$

For general integrable η split into its positive and negative parts $\eta = \eta_+ - \eta_-$ and use linearity of expectation. □

TODO: Provide an example of conditional expectation and a dyadic σ-algebra.

A last observation is that conditional expectations depend only "local" information in both the random variable and the σ-algebra. This has an intuitive appeal as one can think of the σ-algebra against which the conditional expectation is taken as a specifying a coarser resolution of the random variable and this coarsening is obtained by averaging/integration. So long as the domains over which we integrate are contained entirely inside of a set we are interested in, the conditional expectation should only depend on the σ-algebra restricted to that set and the values of the random variable on that set. We proceed to make this idea more formal and give a proper proof.

Definition 11.9. Given σ-algebras $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{A}$ with $\mathcal{F} \subset \mathcal{A}$ and $\mathcal{G} \subset \mathcal{A}$ and a set $A \in \mathcal{F} \cap \mathcal{G}$, we way that $\mathcal{F}$ and $\mathcal{G}$ *agree on* A if for every $B \subset A$, $B \in \mathcal{F}$ if and only if $B \in \mathcal{G}$.

Lemma 11.10. *Given σ-algebras $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{A}$ with $\mathcal{F} \subset \mathcal{A}$ and $\mathcal{G} \subset \mathcal{A}$ and a set $A \in \mathcal{F} \cap \mathcal{G}$ such that $\mathcal{F}$ and $\mathcal{G}$ agree on A and random variables ξ and η such that ξ and η agree almost surely on A then*

$$\boldsymbol{E}[\xi \mid \mathcal{F}] = \boldsymbol{E}[\eta \mid \mathcal{G}] \ \text{ a.s. on } A$$

Proof. First note that because $A \in \mathcal{F} \cap \mathcal{G}$ and because $\mathbf{1}_A \xi = \mathbf{1}_A \eta$ a.s., we have

$$\mathbf{E}[\xi \mid \mathcal{F}] = \mathbf{E}[\mathbf{1}_A \xi \mid \mathcal{F}] = \mathbf{E}[\mathbf{1}_A \eta \mid \mathcal{F}] = \mathbf{E}[\eta \mid \mathcal{F}]$$

so it suffices to show that $\mathbf{E}[\eta \mid \mathcal{F}] = \mathbf{E}[\eta \mid \mathcal{G}]$ a.s. □

The definition of conditional expectation as given is rather abstract but in the case of random variables with densities, we can make the concept more concrete.

TODO: Where to put this?

Lemma 11.11. *Let (ξ, η) be a random vector in $\mathbb{R}^2$. Suppose that (ξ, η) has a density f, then*

(i) *Both ξ and η have a densities given by the formulas*

$$f_\xi(y) = \int_{-\infty}^{\infty} f(y,z)\,dz \qquad\qquad f_\eta(z) = \int_{-\infty}^{\infty} f(y,z)\,dy$$

(ii) *ξ and η are independent if and only if $f(y,z) = f_\xi(y) f_\eta(z)$.*

(iii) *For any $y \in \mathbb{R}$ such that $f_\xi(y) \neq 0$, we have the density*

$$f_{\xi=y}(z) = \frac{f(y,z)}{f_\xi(y)}$$

(iv) *If we define $h_\eta(y) = \int_{-\infty}^{\infty} z f_{\xi=y}(z)\,dz$ then for every measurable $g : \mathbb{R} \to \mathbb{R}$ such that $g(\xi)$ is integrable, we have*

$$\boldsymbol{E}[g(\xi) \cdot h_\eta(\xi)] = \boldsymbol{E}[\xi \cdot \eta]$$

If we consider η a random element in some $(T, \mathcal{T})$, ξ an integrable random variable then we usually write $\mathbf{E}[\xi \mid \sigma(\eta)] = \mathbf{E}[\xi \mid \eta]$ and speak of the *conditional expectation of ξ with respect to η.*

Lemma 11.12. *There exists a measurable function $f : T \to \mathbb{R}$ such that $\boldsymbol{E}[\xi \mid \eta] = f(\eta)$, furthermore such an f is unique almost surely $P \circ \eta^{-1}$. If we are given another pair $\tilde{\xi}$ and $\tilde{\eta}$ such that $(\xi, \eta) \overset{d}{=} (\tilde{\xi}, \tilde{\eta})$ then $\boldsymbol{E}\left[\tilde{\xi} \mid \tilde{\eta}\right] = f(\tilde{\eta})$.*

Proof. This is a simple corollary of Lemma 3.20 and the almost sure uniqueness of conditional expectations. □

Having defined $\mathbf{E}[\xi \mid \eta]$ in terms of conditional expectation of ξ with respect the σ-algebra $\sigma(\eta)$ is natural to think of the latter as being the more general case. However note that if we are given $\mathcal{F}$ and define $\eta : (\Omega, \mathcal{A}) \to (\Omega, \mathcal{F})$ to be identity function then in fact we see the two notions are equivalent. In some cases, authors (Kallenberg in particular) will refer to conditional expectation with respect to a σ-algebra as the special case. We'll try to avoid making statements about the relative level of generality of the two ideas but will try to avoid using the notation $\mathbf{E}[\xi \mid \eta]$ when we know that η is an identity map.

11.3. **Conditional Independence.**

Definition 11.13. Given σ-algebras $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{H}$ we say that $\mathcal{F}$ and $\mathcal{H}$ are *conditionally independent given* $\mathcal{G}$ if for all $F \in \mathcal{F}$ and all $H \in \mathcal{H}$ we have

$$\mathbf{P}\{F \cap H \mid \mathcal{G}\} = \mathbf{P}\{F \mid \mathcal{G}\}\mathbf{P}\{H \mid \mathcal{G}\}$$

We often write $\mathcal{F} \perp\!\!\!\perp_{\mathcal{G}} \mathcal{H}$.

A technical result that can be helpful when trying to prove conditional independence is the following analogue of Lemma 7.12

Lemma 11.14. *Suppose we are given a σ-algebra $\mathcal{G}$ and two π-systems $\mathcal{S}$ and $\mathcal{T}$ in a probability space $(\Omega, \mathcal{A}, P)$ such that $\boldsymbol{P}\{A \cap B \mid \mathcal{G}\} = \boldsymbol{P}\{A \mid \mathcal{G}\}\boldsymbol{P}\{B \mid \mathcal{G}\}$ for all $A \in \mathcal{S}$ and $B \in \mathcal{T}$. Then $\sigma(\mathcal{S})$ and $\sigma(\mathcal{T})$ are conditionally independent given $\mathcal{G}$.*

Proof. TODO: A straightforward extension of the proof of Lemma 7.12. □

Lemma 11.15. *Given σ-algebras $\mathcal{F}$, $\mathcal{G}$ and $\mathcal{H}$, then $\mathcal{F} \perp\!\!\!\perp_{\mathcal{G}} \mathcal{H}$ if and only if for all $H \in \mathcal{H}$, we have $\boldsymbol{P}\{H \mid \mathcal{G}\} = \boldsymbol{P}\{H \mid \mathcal{F}, \mathcal{G}\}$. In particular, $\mathcal{F} \perp\!\!\!\perp_{\mathcal{G}} \mathcal{H}$ if and only if $(\mathcal{F}, \mathcal{G}) \perp\!\!\!\perp_{\mathcal{G}} \mathcal{H}$*

Proof. We first assume that $\mathcal{F} \perp\!\!\!\perp_{\mathcal{G}} \mathcal{H}$. Let $F \in \mathcal{F}$ and $G \in \mathcal{G}$ and calculate

$$\begin{aligned}
\mathbf{E}[\mathbf{1}_F \mathbf{1}_G \mathbf{1}_H] &= \mathbf{E}[\mathbf{E}[\mathbf{1}_F \mathbf{1}_G \mathbf{1}_H \mid \mathcal{G}]] \\
&= \mathbf{E}[\mathbf{1}_G \mathbf{E}[\mathbf{1}_F \mathbf{1}_H \mid \mathcal{G}]] \\
&= \mathbf{E}[\mathbf{1}_G \mathbf{E}[\mathbf{1}_F \mid \mathcal{G}]\, \mathbf{E}[\mathbf{1}_H \mid \mathcal{G}]] \\
&= \mathbf{E}[\mathbf{E}[\mathbf{1}_F \mathbf{1}_G \mid \mathcal{G}]\, \mathbf{E}[\mathbf{1}_H \mid \mathcal{G}]] \\
&= \mathbf{E}[\mathbf{1}_F \mathbf{1}_G \mathbf{E}[\mathbf{1}_H \mid \mathcal{G}]]
\end{aligned}$$

Now note that set of all intersections $F \cap G$ is a π-system that contains Ω and therefore by Lemma 11.7 and the defining property of conditional expectation we have $\mathbf{E}[\mathbf{1}_H \mid \mathcal{G}] = \mathbf{E}[\mathbf{1}_H \mid \mathcal{F}, \mathcal{G}]$.

To show the converse, we take $F \in \mathcal{F}$ and $H \in \mathcal{H}$ and

$$\begin{aligned}
\mathbf{E}[\mathbf{1}_F \mathbf{1}_H \mid \mathcal{G}] &= \mathbf{E}[\mathbf{E}[\mathbf{1}_F \mathbf{1}_H \mid \mathcal{F}, \mathcal{G}] \mid \mathcal{G}] \\
&= \mathbf{E}[\mathbf{1}_F \mathbf{E}[\mathbf{1}_H \mid \mathcal{F}, \mathcal{G}] \mid \mathcal{G}] \\
&= \mathbf{E}[\mathbf{1}_F \mid \mathcal{G}]\, \mathbf{E}[\mathbf{1}_H \mid \mathcal{F}, \mathcal{G}] \\
&= \mathbf{E}[\mathbf{1}_F \mid \mathcal{G}]\, \mathbf{E}[\mathbf{1}_H \mid \mathcal{G}]
\end{aligned}$$

Now the last claim follows simply we have shown both statements are equivalent to the fact that $\mathbf{P}\{H \mid \mathcal{G}\} = \mathbf{P}\{H \mid \mathcal{F}, \mathcal{G}\}$ for all $H \in \mathcal{H}$. □

Lemma 11.16. *Given σ-algebras $\mathcal{G}$, $\mathcal{H}$ and $\mathcal{F}_1, \mathcal{F}_2, \dots$, then $\mathcal{H} \perp\!\!\!\perp_{\mathcal{G}} (\mathcal{F}_1, \mathcal{F}_2, \dots)$ if and only if $\mathcal{H} \perp\!\!\!\perp_{(\mathcal{G}, \mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_n)} \mathcal{F}_{n+1}$ for all $n \geq 0$.*

Proof. If we assume the second property then we can conclude from Lemma 11.15 and an induction on $n \geq 0$ that for every $H \in \mathcal{H}$,

$$\mathbf{P}\{H \mid \mathcal{G}\} = \mathbf{P}\{H \mid \mathcal{G}, \mathcal{F}_1\} = \mathbf{P}\{H \mid \mathcal{G}, \mathcal{F}_1, \mathcal{F}_2\} = \cdots$$

and therefore by another application of Lemma 11.15, we know that $\mathcal{H} \perp\!\!\!\perp_{\mathcal{G}} (\mathcal{F}_1, \dots, \mathcal{F}_n)$ for every $n \geq 1$. Now $\cup_n \sigma(\mathcal{F}_1, \dots, \mathcal{F}_n)$ is a π-system that generates $\sigma(\mathcal{F}_1, \mathcal{F}_2, \dots)$ and therefore application of Lemma 11.14 shows us that $\mathcal{H} \perp\!\!\!\perp_{\mathcal{G}} (\mathcal{F}_1, \mathcal{F}_2, \dots)$.

On the other hand, if we assume $\mathcal{H} \perp\!\!\!\perp_{\mathcal{G}} (\mathcal{F}_1, \mathcal{F}_2, \dots)$ then for any $n \geq 1$, and $H \in \mathcal{H}$, we apply the telescoping rule, Lemma 11.15 and the pull out rule to get

$$\begin{aligned}
\mathbf{P}\{H \mid \mathcal{G}, \mathcal{F}_1, \dots, \mathcal{F}_n\} &= \mathbf{E}[\mathbf{P}\{H \mid \mathcal{G}, \mathcal{F}_1, \mathcal{F}_2, \dots\} \mid \mathcal{G}, \mathcal{F}_1, \dots, \mathcal{F}_n] \\
&= \mathbf{E}[\mathbf{P}\{H \mid \mathcal{G}\} \mid \mathcal{G}, \mathcal{F}_1, \dots, \mathcal{F}_n] \\
&= \mathbf{P}\{H \mid \mathcal{G}\}
\end{aligned}$$

so in particular, for all $n \geq 0$,

$$\mathbf{P}\{H \mid \mathcal{G}, \mathcal{F}_1, \ldots, \mathcal{F}_n\} = \mathbf{P}\{H \mid \mathcal{G}, \mathcal{F}_1, \ldots, \mathcal{F}_{n+1}\}$$

Another application of Lemma 11.15 shows that $\mathcal{H} \perp\!\!\!\perp_{(\mathcal{G}, \mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n)} \mathcal{F}_{n+1}$ for all $n \geq 0$. □

Lemma 11.17. *Suppose $\mathcal{F} \perp\!\!\!\perp_{\mathcal{G}} \mathcal{H}$ and $\mathcal{A} \subset \mathcal{F}$, then $\mathcal{F} \perp\!\!\!\perp_{\mathcal{A}, \mathcal{G}} \mathcal{H}$.*

Proof. By Lemma 11.15, we know for all $H \in \mathcal{H}$, $\mathbf{P}\{H \mid \mathcal{G}\} = \mathbf{P}\{H \mid \mathcal{F}, \mathcal{G}\}$. On the other hand, since $\mathcal{A} \subset \mathcal{F}$ we also have $\mathcal{G} \subset \sigma(\mathcal{A}, \mathcal{G}) \subset \sigma(\mathcal{F}, \mathcal{G})$ and therefore we can conclude $\mathbf{P}\{H \mid \mathcal{F}, \mathcal{G}\} = \mathbf{P}\{H \mid \mathcal{A}, \mathcal{G}\}$. Since $\mathcal{A} \subset \mathcal{F}$ we know that $\sigma(\mathcal{A}, \mathcal{F}, \mathcal{G}) = \sigma(\mathcal{F}, \mathcal{G})$ and we get $\mathbf{P}\{H \mid \mathcal{F}, \mathcal{A}, \mathcal{G}\} = \mathbf{P}\{H \mid \mathcal{A}, \mathcal{G}\}$. Another application of Lemma 11.15 tells us that $\mathcal{F} \perp\!\!\!\perp_{\mathcal{A}, \mathcal{G}} \mathcal{H}$. □

11.4. **Conditional Distributions and Disintegration.** Now for a more subtle concept in conditioning. Consider a random element ξ in a measurable space $(S, \mathcal{S})$ and a random element η in a measurable space $(T, \mathcal{T})$. We'd like to make sense of the conditional distribution of ξ given a value of η. Two things should occur to us. First, such an object sounds like it should a mapping from T to a space of measures on S. Second, we expect that we'll actually define this object in terms of the conditional expectation and that it will likely wind up as an η-measurable random measure on Ω. A third thing might also occur to us: namely these two representations are equivalent. As it turns out, due to the fact that conditional expectations are only defined up to almost sure equivalence, this last supposition is not true and we often must make additional assumptions to arrange for the existence of the mapping of T to the space of measures on S.

11.4.1. *Probability Kernels.* Before jumping into the development of conditional distributions proper we need to step back a bit and make sure we've laid a proper foundation for the discussion. We wrote heuristically above about a mapping to a space of measures. This is a concept that will come up in a variety of contexts from this point on and we glossed over the fact that we want such a mapping to have measurability properties. There are a couple of equivalent ways of formulating the notion of a measurable family of measures; we explore these now. To formalize, we have the following definition

Definition 11.18. Let $(S, \mathcal{S})$ and $(T, \mathcal{T})$ be measurable spaces. A *probability kernel* from S to T is a function $\mu : S \times \mathcal{T} \to [0, 1]$ such that for every fixed $s \in S$, $\mu(s, \cdot) : \mathcal{T} \to [0, 1]$ is a probability measure and for every fixed $A \in \mathcal{T}$, $\mu(\cdot, A) : S \to [0, 1]$ is Borel measurable.

It is useful to have some alternative characterizations of the measurability properites of kernels but before we can state them we need another definition.

Definition 11.19. Given a measurable space $(S, \mathcal{S})$, then $\mathcal{P}(S)$ is the space of probability measures on S with the σ-algebra generated by all sets of the form $\{\mu \mid \mu(A) \in B\}$ for $A \in \mathcal{S}$ and $B \in \mathcal{B}([0, 1])$. Alternatively, for each $A \in \mathcal{S}$, define the evaluation map $\pi_A : \mathcal{P}(S) \to [0, 1]$ by $\pi_A(\mu) = \mu(A)$ and then take the σ-algebra generated by all of the evaluation maps.

Example 11.20. The following special case of a probability kernel is easy to understand and also comes up in the theory of finite Markov chains. Suppose S and T are two finite probability spaces with the power set sigma algebra. In this case

a probability measure on T is just a set of non-negative real numbers p_t for $t \in T$ such $\sum_{t\in T} p_t = 1$. Therefore a probablity kernel from S to T is just a set of such vectors for each $s \in S$. It is customary in the theory of finite Markov chains to view probabilities on T as row vectors and thus view a probability kernel μ as an $S \times T$ matrix $\mu_{s,t}$ such that $\mu_{s,t} \geq 0$ and for each fixed $s \in S$ we have $\sum_{t\in T} \mu_{s,t} = 1$. Such a matrix with row sums equal to 1 is sometimes called a *stochastic matrix*.

TODO: Is there anything to say about the measurability condition here?

As promised, we have the following lemma that gives a couple of alternative characterizations of the measurability condition of a kernel; including the obligatory monotone class argument.

Lemma 11.21. *Let $(S, \mathcal{S})$ and $(T, \mathcal{T})$ be measurable spaces and μ_s be a family of probability measures on T. Then the following are equivalent*

(i) $\mu : S \times \mathcal{T} \to [0, 1]$ *is a probability kernel*
(ii) $\mu : S \to \mathcal{P}(T)$ *is measurable*
(iii) $\mu(s, A) : S \to [0, 1]$ *is Borel measurable for every A belonging to a π-system that generates $\mathcal{S}$.*

Proof. First suppose that μ is a kernel, $A \in \mathcal{T}$ and B is a Borel measurable subset of $[0, 1]$. Then

$$\mu^{-1}(\{\nu \mid \nu(A) \in B\}) = \{s \in S \mid \mu(s, A) \in B\} = \mu(\cdot, A)^{-1}(B)$$

which is measurable by the kernel property. Since sets of the form $\{\nu \mid \nu(A) \in B\}$ generate the σ-algebra on $\mathcal{P}(T)$ we see that μ is measurable by Lemma 3.12.

To see that (ii) implies (i), observe that for a fixed $A \in \mathcal{T}$ and let $\pi_A(\nu) = \nu(A)$ be the evaluation map. By construction the π_A are measurable. For such a fixed A, we see that $\mu(s, A) = \pi_A(\mu)$ therefore as a composition of measurable maps we see that $\mu(s, A)$ is $\mathcal{S}$-measurable (Lemma 3.13).

The implication (i) implies (iii) is immediate. If we assume (iii) then we derive (i) by a monotone class argument. By Theorem 3.24 it suffices to show that $\mathcal{C} = \{A \mid \mu(s, A) : S \to [0, 1] \text{ is measurable}\}$ is a λ-system. If $A \subset B$ with $A, B \in \mathcal{C}$ then $\mu(s, B \setminus A) = \mu(s, B) - \mu(s, A)$ is measurable. If $A_1 \subset A_2 \subset \cdots$ with $A_n \in \mathcal{C}$ then by continuity of measure (Lemma 3.27) applied pointwise in s, we see $\mu(s, \cup_n A_n) = \lim_n \mu(s, A_n)$ which shows measurability by Lemma 3.14. □

A point that shall occasionally come up is the fact that we shall use the previous lemma to shift interpretations of a kernel: sometimes thinking of it as a map $\mu : S \times \mathcal{T} \to [0, 1]$ and sometimes as a map $\mu : S \to \mathcal{P}(T)$. Often we will make such transitions between these perspectives without comment but there are times in which we may use the notation $\mu(s, A)$ when thinking of the first realization and $\mu(s)$ when thinking of the second. It is also the case that the notation for integrals with respect to kernels needs to be considered. Up to this point we have notation $\int f \, d\mu$ for integrals and in those cases in which we wanted to make it clear what the integration variable is we might right $\int f(x) \, d\mu(x)$. In a world with kernels the latter notation is unfortunate as become difficult to construe wheter the x dependence indicated for the measure means an integration variable or whether it may indicate that the measure is a kernel with x dependence. To resolve this issue we shall adopt a different convention when discussing integrals against kernels and write $\int f(x) \, \mu(dx)$ to denote that x is the integration variable. This notation allows us

to capture both integration variables and measure dependence in expressions such as $\int f(x)\,\mu(s,dx)$.

There is a useful generalization of the product measure construction involving kernels. It is a type of "twisted" product construction.

Definition 11.22. Let $\mu : S \times \mathcal{T} \to [0,1]$ be a probability kernel from S to T and $\nu : S \times T \times \mathcal{U} \to [0,1]$ be a probability kernel from $S \times T$ to U, we then define $\mu \otimes \nu : S \times \mathcal{T} \otimes \mathcal{U} \to [0,1]$ by

$$\mu \otimes \nu(s, A) = \iint \mathbf{1}_A(t,u)\, d\nu(s,t,du)\, d\mu(s,dt)$$

The fact that this construction defines a probability kernel is the content of the next Lemma.

Lemma 11.23. *Suppose $\mu : S \times \mathcal{T} \to [0,1]$ is a probability kernel from S to T and $\nu : S \times T \times \mathcal{U} \to [0,1]$ be a probability kernel from $S \times T$ to U. Let $f : S \times T \to \mathbb{R}_+$ and $g : S \times T \times \to U$ be measurable then*

(i) *$\int f(s,t)\, d\mu(s,dt)$ is a measurable function of $s \in S$.*
(ii) *$\mu_s \circ (g(s,\cdot))^{-1}$ is a kernel from S to U.*
(iii) *$\mu \otimes \nu$ is a kernel from S to $T \times U$.*

Proof. To see (i), we apply the standard machinery. First consider $f(s,t) = \mathbf{1}_{A\times B}(s,t)$ for $A \in \mathcal{S}$ and $B \in \mathcal{T}$. In this case,

$$\int \mathbf{1}_{A\times B}(s,t)\, d\mu(s,dt) = \mathbf{1}_A(s) \int \mathbf{1}_B(t)\, d\mu(s,dt) = \mathbf{1}_A(s)\mu(s,B)$$

which is $\mathcal{S}$-measurable by measurability of A and the fact that μ is a kernel. We extend to the case of general characteristic functions by observing that products $A \times B$ are a generating π-system for the σ-algebra $\mathcal{S} \otimes \mathcal{T}$. Additionally we must show that $\mathcal{C} = \{C \in \mathcal{S} \otimes \mathcal{T} \mid \int \mathbf{1}_{A\times B}(s,t)\, d\mu(s,dt) \text{ is measurable}\}$ is a λ-system. To see this first assume that $A \subset B$ with $A, B \in \mathcal{C}$. Then by linearity of integral, $\int \mathbf{1}_{B\setminus A}(s,t)\, d\mu(s,dt) = \int \mathbf{1}_B(s,t)\, d\mu(s,dt) - \int \mathbf{1}_A(s,t)\, d\mu(s,dt)$ which shows $B\setminus A \in \mathcal{C}$. Secondly if $A_1 \subset A_2 \subset \cdots$ is a chain in $\mathcal{C}$ then by Monotone Convergence applied pointwise in s, we have $\int \mathbf{1}_{\cup_n A_n}(s,t)\, d\mu(s,dt) = \lim_{n\to\infty} \int \mathbf{1}_{A_n}(s,t)\, d\mu(s,dt)$ which shows $\cup_n A_n \in \mathcal{C}$ because limits of measurable functions are measurable (Lemma 3.14). Now an application of Theorem 3.24 shows the result.

By $\mathcal{S}$-measurability for characteristic functions and linearity of integral, we see that $\int f(s,t)\, d\mu(s,dt)$ is $\mathcal{S}$-measurable for simple functions and by definition of integral we see that for any positive measurable f with an approximation by simple functions $f_n \uparrow f$ we note that for each fixed s, f_n are simple functions of t alone so $\int f(s,t)\, d\mu(s,dt) = \lim_n \int f_n(s,t)\, d\mu(s,dt)$ showing $\mathcal{S}$-measurability by another application of Lemma 3.14. Lastly extending to general integrable f, write $f = f_+ - f_-$ and use linearity of integral.

Having proven (i) we derive (ii) and (iii) from it. To see (ii) assume that $A \in \mathcal{U}$ and note that for fixed s, if we denote the section of g at s by $g_s : T \to U$ then it is elementary that $\mathbf{1}_{g_s^{-1}(A)}(t) = \mathbf{1}_{g^{-1}(A)}(s,t)$ and thus

$$\mu_s \circ (g(s,\cdot))^{-1}(A) = \mu(s, g^{-1}(s,A)) = \mu(s, g^{-1}(A))$$

which we have shown is $\mathcal{S}$-measurable in (i).

To see (iii), pick $A \in \mathcal{T} \otimes \mathcal{U}$ and recall that by definition

$$\mu \otimes \nu(A)(s) = \iint \mathbf{1}_A(t,u)\, d\nu(s,t,du)\, d\mu(s,dt)$$

We know that $\mathbf{1}_A(t,u)$ is $\mathcal{T}\otimes\mathcal{U}$-measurable hence also $\mathcal{S}\otimes\mathcal{T}\otimes\mathcal{U}$-measurable. Therefore we can apply (i) to conclude that $\int \mathbf{1}_A(t,u)\, d\nu(s,t,du)$ is $\mathcal{S}\otimes\mathcal{T}$-measurable. Now apply (i) again to conclude that $\mu\otimes\nu(A)(s)$ is $\mathcal{S}$-measurable. □

Example 11.24. For finite probability spaces S, T and U a probability kernel $\mu : S \to \mathcal{P}(T)$ is a stochastic matrix $\mu_{s,t}$ and a probability kernel $\nu : S\times T \to \mathcal{P}(U)$ is a $(S\times T)\times U$ stochastic matrix $\nu_{s,t,u}$ where we consider the pair (s,t) to the row index. If we now identify (t,u) as column index in the $S\times(T\times U)$ matrix $\mu\otimes\nu$ then $(\mu\otimes\nu)_{s,t,u} = \mu_s\nu_{s,t,u}$.

Theorem 11.25. *Let $(S,\mathcal{S})$ be a Borel space and $(T,\mathcal{T})$ be an arbitrary measuable space. Let ξ be a random element in S and η be a random element in T. There is exists a probability kernel $\mu : T\times\mathcal{S}\to\mathbb{R}$ such that $\boldsymbol{P}\{\xi\in A \mid \eta\}(\omega) = \mu(\eta(\omega), A)$ for all $A\in\mathcal{S}$ and $\omega\in\Omega$. Furthermore, if $\tilde{\mu}$ is another probability kernel satisfying this property then $\mu = \tilde{\mu}$ almost surely with respect to $\mathcal{L}(\eta)$.*

Proof. TODO: Reduce to the case of $S=\mathbb{R}$ and use density of rationals and properties of distribution functions to create a regular version. □

Theorem 11.26. *Let $(S,\mathcal{S})$ and $(T,\mathcal{T})$ be measurable spaces and let ξ be a random element in S and η be a random element in T. Suppose*

(i) *$\boldsymbol{P}\{\xi\in\cdot \mid \mathcal{F}\}$ has a regular version $\nu : \Omega\times\mathcal{S}\to\mathbb{R}$*
(ii) *η is $\mathcal{F}$-measurable*
(iii) *$f : S\times T\to\mathbb{R}$ is measurable with either $f\geq 0$ or $\boldsymbol{E}[|f(\xi,\eta)|]<\infty$*

Then

$$\boldsymbol{E}[f(\xi,\eta)] = \boldsymbol{E}\left[\int f(s,\eta)\, d\nu(s)\right]$$

and moreover

$$\boldsymbol{E}[f(\xi,\eta)\mid\mathcal{F}] = \int f(s,\eta)\, d\nu(s) \ \text{ a.s.}$$

Proof. The proof is an application of the standard machinery. To start with we assume that $f = \mathbf{1}_{A\times B}$ for $A\in\mathcal{S}$ and $B\in\mathcal{T}$. Then

$$\begin{aligned}
\mathbf{E}[f(\xi,\eta)] &= \mathbf{E}[\mathbf{1}_A(\xi)\mathbf{1}_B(\eta)] \\
&= \mathbf{E}[\mathbf{E}[\mathbf{1}_A(\xi)\mid\mathcal{F}]\,\mathbf{1}_B(\eta)] \\
&= \mathbf{E}[\nu(A)\mathbf{1}_B(\eta)] \\
&= \mathbf{E}\left[\int \mathbf{1}_A(s)\mathbf{1}_B(\eta) d\nu(s)\right] \\
&= \mathbf{E}\left[\int f(s,\eta) d\nu(s)\right]
\end{aligned}$$

Now we extend to the set of all $C\in\mathcal{S}\otimes\mathcal{T}$ by using a Monotone Class Argument (Theorem 3.24). Let $\mathcal{C} = \{C\in\mathcal{S}\otimes\mathcal{T} \mid \mathbf{E}[\mathbf{1}_C(\xi,\eta)] = \mathbf{E}\left[\int \mathbf{1}_C(s,\eta)d\nu(s)\right]\}$ Since

the set of all $A \times B$ is a π-system containing $S \times T$ it suffices to show that $\mathcal{C}$ is a λ-system. Suppose $C, D \in \mathcal{D}$ and $C \subset D$; then we see $D \setminus C \in \mathcal{C}$ by noting $\mathbf{1}_{D\setminus C} = \mathbf{1}_D - \mathbf{1}_C$ and applying linearity of expectation and integral. If we assume $C_1 \subset C_2 \subset \cdots$ with $C_n \in \mathcal{C}$, then $\mathbf{1}_{\cup_n C_n} = \lim_{n\to\infty} \mathbf{1}_{C_n}$ and the Monotone Convergence Theorem implies $\mathbf{E}\left[\mathbf{1}_{\cup_n C_n}(\xi,\eta)\right] = \lim_{n\to\infty} \mathbf{E}\left[\mathbf{1}_{C_n}(\xi,\eta)\right]$. Similarly for fixed $\omega \in \Omega$, $\int \mathbf{1}_{\cup_n C_n}(s,\eta)\, d\nu(s) = \lim_{n\to\infty} \int \mathbf{1}_{C_n}(s,\eta)\, d\nu(s)$, moreover monotonicity of integral implies that $\int \mathbf{1}_{C_n}(s,\eta)\, d\nu(s)$ is increasing in n. Therefore we may apply Monotone Convergence a second time to conclude that

$$\mathbf{E}\left[\int \mathbf{1}_{\cup_n C_n}(s,\eta)\, d\nu(s)\right] = \lim_{n\to\infty} \mathbf{E}\left[\int \mathbf{1}_{C_n}(s,\eta)\, d\nu(s)\right]$$

Therefore we see that $\cup_n C_n \in \mathcal{C}$.

Extending the result to simple functions is trivial since both sides are linear in f.

Now we suppose that $f : S \times T \in \mathbb{R}$ is positive measurable. We pick an approximation of f by an increasing sequence of positive simple functions $0 \leq f_n \uparrow f$. Now $f_n(\xi,\eta)$ is an increasing sequence of positive simple functions with $\lim_{n\to\infty} f_n(\xi,\eta) = f(\xi,\eta)$ and therefore by definition of expectation, $\mathbf{E}\left[f(\xi,\eta)\right] = \lim_{n\to\infty} \mathbf{E}\left[f_n(\xi,\eta)\right]$. Similarly for fixed $\omega \in \Omega$ we have $f_n(s,\eta)$ are positive simple functions increasing to $f(s,\eta)$ and therefore $\int f(s,\eta)\, d\nu(s) = \lim_{n\to\infty} \int f_n(s,\eta)\, d\nu(s)$. Monotonicity of integral shows that the sequence $\int f_n(s,\eta)\, d\nu(s)$ is positive and increasing and therefore we may apply Monotone Convergence and the fact that result holds for the f_n to show that

$$\mathbf{E}\left[\int f(s,\eta)\, d\nu(s)\right] = \lim_{n\to\infty} \mathbf{E}\left[\int f_n(s,\eta)\, d\nu(s)\right] = \lim_{n\to\infty} \mathbf{E}\left[f_n(\xi,\eta)\right] = \mathbf{E}\left[f(\xi,\eta)\right]$$

Therefore the result for positive measurable f.

Lastly for general integrable f, we know by the result for positive f that

$$\mathbf{E}\left[\int |f(s,\eta)|\, d\nu(s)\right] = \mathbf{E}\left[|f(\xi,\eta)|\right] < \infty$$

Which shows us that $\int |f(s,\eta)|\, d\nu(s) < \infty$ almost surely. Then we can write $f = f_+ - f_-$ and use the the result for postive f and linearity.

The last thing to do is to extend the result to the case of conditional expectations. Let $f : S \times T \to \mathbb{R}_+$ be positive and let $A \in \mathcal{F}$. Consider $(\eta, \mathbf{1}_A)$ as a random element of $T \times \{0,1\}$. Note that this random element is $\mathcal{F}$-measurable since η is and $A \in \mathcal{F}$. Therefore we can apply the case just proven to the function $\tilde{f} : S \times T \times \{0,1\} \to \mathbb{R}_+$ given by $\tilde{f}(s,t,u) = uf(s,t)$ and the elements ξ and $(\eta, \mathbf{1}_A)$ to get

$$\mathbf{E}\left[f(\xi,\eta); A\right] = \mathbf{E}\left[\int f(s,\eta)\mathbf{1}_A\, d\nu(s)\right] = \mathbf{E}\left[\int f(s,\eta)\, d\nu(s); A\right]$$

which shows that $\mathbf{E}\left[f(\xi,\eta) \mid \mathcal{F}\right] = \int f(s,\eta)\, d\nu(s)$ a.s. for $f \geq 0$. The case of integrable f follows as usual by taking differences. □

Theorem 11.27 (Jensen's Inequality)**.** *Let ξ be a random vector and $\mathcal{F}$ be a σ-algebra. If φ is a convex function then $\varphi(\boldsymbol{E}[\xi \mid \mathcal{F}]) \leq \boldsymbol{E}[\varphi(\xi) \mid \mathcal{F}]$ a.s. If φ is strictly convex then $\varphi(\boldsymbol{E}[\xi \mid \mathcal{F}]) = \boldsymbol{E}[\varphi(\xi) \mid \mathcal{F}]$ if and only if $\xi = \boldsymbol{E}[\xi \mid \mathcal{F}]$ a.s.*

Proof. Since $\mathbb{R}^n$ is Borel by Theorem 11.25 we know $\mathbf{P}\{\xi \in \cdot \mid \mathcal{F}\}$ has regular version μ. Now by Theorem 11.26 and the ordinary Jensen Inequality (Lemma 6.17) applied pointwise we know that

$$\varphi(\mathbf{E}[\xi \mid \mathcal{F}]) = \varphi\left(\int f(s)\,\mu(ds)\right) \leq \int \varphi(f(s))\,\mu(ds) = \mathbf{E}[\varphi(f(\xi)) \mid \mathcal{F}]$$

TODO: The strictly convex/equality case □

As another application of Theorem 11.26 we give a little result about the interaction between conditional indpendence and conditional expectations.

Corollary 11.28. *Let ξ be a random element in S such that $\boldsymbol{P}\{\xi \in \cdot \mid \mathcal{G}\}$ has a regular version. Then if $\xi \perp\!\!\!\perp_{\mathcal{F}} \mathcal{G}$ and $f : S \to \mathbb{R}$ is measurable then $\boldsymbol{E}[f(\xi) \mid \mathcal{G}] = \boldsymbol{E}[f(\xi) \mid \mathcal{F}, \mathcal{G}]$.*

Proof. Let μ be a regular version of $\mathbf{P}\{\xi \in \cdot \mid \mathcal{G}\}$. By Lemma 11.15 we know that $\mathbf{P}\{\xi \in \cdot \mid \mathcal{G}\} = \mathbf{P}\{\xi \in \cdot \mid \mathcal{F}, \mathcal{G}\}$ and therefore μ is a regular version for $\mathbf{P}\{\xi \in \cdot \mid \mathcal{F}, \mathcal{G}\}$ as well and by Theorem 11.26

$$\mathbf{E}[f(\xi) \mid \mathcal{G}] = \int f(s)\,\mu(ds) = \mathbf{E}[f(\xi) \mid \mathcal{F}, \mathcal{G}] \text{ a.s.}$$

TODO: Is there a proof of this result that doesn't require the existence of regular versions? □

Special case of random vectors with densities. Suppose we are given $\xi : \Omega \to \mathbb{R}^m$ and $\eta : \Omega \to \mathbb{R}^n$ such that (ξ, η) has density f on $\mathbb{R}^{m+n}$. Then ξ and η have densities f_ξ and f_η called the marginal densities and we get a conditional densities $f(x, y)/f_\xi(x)$ and $f(x, y)/f_\eta(y)$. TODO: Tie this back to conditional distributions as defined in the general case (this is an exercise in Kallenberg for example).

We've seen that given a specified distribution we can always find a random variable with that specfied distribution. Moreover, we know that if we allow ourselves to to extend the probability space then we can construct such a random variable to be independent of any existing random elements (or σ-algebras). We now turn our attention to the analogous problem space for conditional distributions. The simplest such result shows that given a random element and a prescribed probability kernel we can always find a second random element whose conditional distribution is the kernel.

Lemma 11.29. *Let $(S, \mathcal{S})$ and $(T, \mathcal{T})$ be measurable spaces, $\mu : T \times \mathcal{S} \to \mathbb{R}$ be a probability kernel and η be a random element in T. There exists an extension $\hat{\Omega}$ and a random element ξ in $\hat{\Omega} \to S$ such that $\boldsymbol{P}\{\xi \in \cdot \mid \eta\} = \mu(\eta, \cdot)$ a.s. and $\xi \perp\!\!\!\perp_\eta \zeta$ for every random element ζ defined on Ω.*

Proof. The appropriate construction is thrust upon us by Theorem 11.26. Note that if we succeed in constructing ξ then that result tells how to compute expectations on $\hat{\Omega}$. Following that lead, define $(\hat{\Omega}, \hat{\mathcal{A}}) = (S \times \Omega, \mathcal{S} \otimes \mathcal{A})$. Define the probability measure

$$\hat{P}(A) = \mathbf{E}\left[\int \mathbf{1}_A(s, \omega)\,d\mu(\eta, s)\right]$$

Note that $\hat{P}$ is an extension since for $A \in \mathcal{A}$,

$$\hat{P}(S \times A) = \mathbf{E}\left[\int \mathbf{1}_S(s)\mathbf{1}_A(\omega)\, d\mu(\eta, s)\right] = \mathbf{E}\left[\mathbf{1}_A(\omega)\right] = P(A)$$

Now define $\xi(s, \omega) = s$ and note that for $A \in \mathcal{S}$ and $B \in \mathcal{A}$,

$$\hat{P}(\xi \in A; B) = \mathbf{E}\left[\int \mathbf{1}_A(s)\mathbf{1}_B(\omega)\, d\mu(\eta, s)\right] = \mathbf{E}\left[\mu(\eta, A); B\right]$$

which shows $\mathbf{P}\{\xi \in A \mid \mathcal{A}\} = \mu(\eta, A)$ a.s. by the defining property of conditional expectation (note that since $\mu(\eta, A)$ and $\mathbf{1}_B$ are both $\mathcal{A}$-measurable, their expectation with respect to P is the same as their expectation with respect to $\hat{P}$). In particular, since we know that $\mu(\eta, A)$ is η-measurable we also know that $\mathbf{P}\{\xi \in A \mid \mathcal{A}\} = \mathbf{P}\{\xi \in A \mid \eta\} = \mu(\eta, A)$.

This last observation also shows $\xi \perp\!\!\!\perp_\eta \mathcal{A}$ by an application of Lemma 11.15. □

The next result is closely related but uses a different construction that shows how one may use a single uniform randomization variable.

Lemma 11.30. *Let $(S, \mathcal{S})$ be a Borel space and $(T, \mathcal{T})$ be a general measurable space. Let ξ be a random element in S and let $\eta, \tilde{\eta}$ be random elements in T. There exists a measurable function $f : T \times [0, 1] \to S$ such that if ϑ is a uniform random variable with $\vartheta \perp\!\!\!\perp \tilde{\eta}$ and we define $\tilde{\xi} = f(\tilde{\eta}, \vartheta)$ then $(\xi, \eta) \overset{d}{=} (\tilde{\xi}, \tilde{\eta})$.*

Proof. TODO: I believe in some applications of this Lemma it can be convenient to assume that ξ, η and $\tilde{\xi}, \tilde{\eta}$ live on different probability spaces. Validate this fact and restate to make it clear that this is true.

First assume $S = \mathbb{R}$. By Theorem 11.25 we have a probability kernel $\mu : T \times \mathcal{S} \to \mathbb{R}$ such that $\mathbf{P}\{\xi \in \cdot \mid \eta\} = \mu(\eta, \cdot)$.

Furthermore, we know by Lemma TODO:??? we can find measurable $f : T \times [0, 1] \to S$ such that the distribution of $f(t, \vartheta)$ is $\mu(t)$. Now define $\tilde{\xi} = f(\tilde{\eta}, \vartheta)$, assume we have a measurable $g : S \times T \to \mathbb{R}_+$ and calculate

$$\begin{aligned}
\mathbf{E}\left[g(\tilde{\xi}, \tilde{\eta})\right] &= \mathbf{E}\left[g(f(\tilde{\eta}, \vartheta), \tilde{\eta})\right] \\
&= \mathbf{E}\left[\int_0^1 g(f(\tilde{\eta}, x), \tilde{\eta})\, dx\right] && \text{by independence of } \tilde{\eta} \text{ and } \vartheta \\
&= \mathbf{E}\left[\int_0^1 g(f(\eta, x), \eta)\, dx\right] && \text{by } \eta \overset{d}{=} \tilde{\eta} \\
&= \mathbf{E}\left[\int g(s, \eta)\, d\mu(\eta, s)\right] && \text{by Lemma 6.7} \\
&= \mathbf{E}\left[g(\xi, \eta)\right] && \text{by Theorem 11.26}
\end{aligned}$$

□

12. Martingales and Optional Times

TODO: First introduce discrete time martingales then do stopping times and lastly extend to continuous time martingales (at least the basics).

We first begin with a very general notion of *stochastic process* which we rather quickly specialize.

Definition 12.1. Suppose one has a measurable space $(S, \mathcal{S})$ and an index set T. We let S^T denote the set of all functions $f : T \to S$. Then $\mathcal{S}^T$ is the σ-algebra generated by all the evaluation maps $\pi_t : S^T \to S$ defined by $\pi_t(f) = f(t)$. That is to say

$$\mathcal{S}^T = \sigma\left(\{\{f \mid f(t) \in U\} \mid t \in T, U \in \mathcal{S}\}\right)$$

Measurability with respect to the σ-algebra $\mathcal{S}^T$ has a useful alternative characterization. First we establish some notation. If we consider a set function $X : \Omega \to S^T$ then can equivalently view this as a set function $\tilde{X} : \Omega \times T \to S$ via the identification $\tilde{X}(\omega, t) = X(\omega)(t)$ (the process of transforming $\tilde{X}$ to X is called *currying* in computer science). We can also curry X on Ω to get an element $\hat{X} : T \to S^\Omega$. It is customary to write $\hat{X}(t)$ as X_t.

Lemma 12.2. *Suppose one has a probability space $(\Omega, \mathcal{A})$, a measurable space $(S, \mathcal{S})$, an index set T and a subset $U \subset S^T$. Then $X : \Omega \to U$ is $U \cap \mathcal{S}^T$-measurable if and only if $X_t : \Omega \to S$ is $\mathcal{S}$-measurable for all $t \in T$.*

Proof. We know by definition of $\mathcal{S}^T$ that every projection $\pi_t : S^T \to S$ is measurable. Moreover, we know that $X_t = \pi_t \circ X$. Therefore if we assume X is $\mathcal{S}^T$-measurable then it follows from Lemma 3.13 that X_t is measurable.

In the opposite direction, assume that each X_t is measurable. Let $A \in \mathcal{S}$ and $t \in T$ and consider the set $\pi_t^{-1}(A) \in \mathcal{S}^T$. By definition we can see that

$$X^{-1}(\pi_t^{-1}(A)) = \{\omega \in \Omega \mid \pi_t(X(\omega)) \in A\} = X_t^{-1}(A)$$

which is measurable by assumption. Since sets of the form $\pi_t^{-1}(A)$ generate $\mathcal{S}^T$ application of Lemma 3.12 shows that X is measurable. □

Definition 12.3. Suppose one has a probability space $(\Omega, \mathcal{A}, \mu)$, a measurable space $(S, \mathcal{S})$, an index set T and a subset $U \subset S^T$. A $U \cap \mathcal{S}^T$-measurable $X : \Omega \to U$ is called a *stochastic process*.

According to this definition, a stochastic process is simply a random element in a path space $(S^T, \mathcal{S}^T)$. As such it has a distribution $\mu \circ X^{-1}$ which is a measure on path space; as usual we will say that two stochastic processes X and Y are equal in distribution when their laws are equal. Because of the nature of the σ-algebra on $\mathcal{S}^T$ there is a simple way to measure whether two processes are equal in distribution.

Lemma 12.4. *Let X be a stochastic process then for every $t_1, \ldots, t_n$ then $(X_{t_1}, \ldots, X_{t_n}) \in S^n$ is $\mathcal{S}^{\otimes n}$-measurable and the measures $\mu \circ (X_{t_1}, \ldots, X_{t_n})^{-1}$ are called the* finite dimensional distributions *of X. If X and Y are two stochastic processes then $X \overset{d}{=} Y$ if and only if their finite dimensional distributions are equal.*

Proof. The $\mathcal{S}^{\otimes n}$-measurability of $(X_{t_1}, \ldots, X_{t_n})$ is consequence of Lemma 12.2 and the fact that a function in a product σ-algebra is measurable if and only if its coordinate projections are.

If we suppose that $X \overset{d}{=} Y$ TODO: Finish □

There are a great many things to be said about stochastic processes in general, however we will wait a bit to travel that road and instead begin to look at a special subclass of stochastic processes.

The first specialization is to assume our index set $T \subset \overline{\mathbb{R}}$ (e.g. $\mathbb{Z}$, $\mathbb{R}$). A good intuition here is that T represents time and that X_t repsesents the dynamics of a time-varying random variable.

Remaining in the land of intuition, we know that as time progress we learn from our experience; more things become known (or at least knowable). If we translate the term "knowable" into the term "measurable" we get a mathematically precise description of the increasing flow of information with time.

Definition 12.5. Suppose we have a probability space $(\Omega, \mathcal{A})$. A collection of σ-algebras $\mathcal{F}_t \subset \mathcal{A}$ for $t \in T$ is called a *filtration* if $\mathcal{F}_s \subset \mathcal{F}_t$ for all $s < t$.

Given a stochastic process one can easily construct a filtration associated with observation of it.

Definition 12.6. Given a probability space $(\Omega, \mathcal{A})$, an index set $T \subset \overline{\mathbb{R}}$ and a stochastic process $X : \Omega \to U$, the filtration *generated by* X is

$$\mathcal{F}_t = \sigma(\{\sigma(X_s) \mid s \leq t\})$$

We then need to tie back the notion of a stochastic process with the notion of a filtration. In particular one wants to call out the case in which a filtration contains enough information to be able to measure the values of the process (i.e. contains at least as much information as the knowledge of the values of the process itself).

Definition 12.7. Given a probability space $(\Omega, \mathcal{A})$, an index set $T \subset \overline{\mathbb{R}}$, a filtration $\mathcal{F}_t$ for $t \in T$ and a stochastic process $X : \Omega \to U$, we say that X is *adapted* to $\mathcal{F}$ if X_t is $\mathcal{F}_t$-measurable for every $t \in T$.

Example 12.8. X is adapted to its generated filtration (and the generated filtration is the smallest filtration adapted to X).

Now we are able to define the special class of stochastic processes with which well spend some time.

Definition 12.9. Given a probability space $(\Omega, \mathcal{A})$, an index set $T \subset \overline{\mathbb{R}}$ and a filtration $\mathcal{F}_t$ for $t \in T$, a stochastic process $M : \Omega \to \mathbb{R}^T$ is called an *$\mathcal{F}$-martingale* if

(i) M_t is integrable for all $t \in T$
(ii) M is adapted to $\mathcal{F}$
(iii) $\mathbf{E}^{\mathcal{F}_s} M_t = M_s$ a.s. for all $s, t \in T$ with $s \leq t$.

If we replace the condition (iii) by the condition $M_s \leq \mathbf{E}^{\mathcal{F}_s} M_t$ a.s., then M is said to be a *submartingale* and if we replace it with $M_s \geq \mathbf{E}^{\mathcal{F}_s} M_t$ a.s. then M is said to be a *supermartingale*.

A entire class of examples of martingales can be constructed via the following Lemma.

Lemma 12.10. *Given a probability space $(\Omega, \mathcal{A})$, an index set $T \subset \overline{\mathbb{R}}$, a filtration $\mathcal{F}_t$ for $t \in T$ and a integrable random variable ξ, the process $M_t = \boldsymbol{E}^{\mathcal{F}_t}\xi$ is an $\mathcal{F}$-martingale.*

Proof. Integrability $\mathcal{F}$-adaptedness of M_t follows from the definition of conditional expectation. Since for $s, t \in T$ with $s \leq t$ we have $\mathcal{F}_s \subset \mathcal{F}_t$, the chain rule for conditional expectation shows

$$\mathbf{E}^{\mathcal{F}_s} M_t = \mathbf{E}^{\mathcal{F}_s} \mathbf{E}^{\mathcal{F}_t} \xi = \mathbf{E}^{\mathcal{F}_s} \xi = M_s$$

□

A martingale that can be expressed in the form given by the Lemma is referred to as a *closed* martingale.

The unbiased random walk provides one of the simplest examples of a martingale.

Example 12.11. Suppose we are given a collection of independent random variables $\xi_1, \xi_2, \dots$ with $\mathbf{E}[\xi_n] = 0$ for all $n > 0$. Define the filtration $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_n = \sigma(\xi_1, \dots, \xi_n)$ for $n > 0$ and define the process $M_0 = 0$ and $M_n = \xi_1 + \cdots + \xi_n$. Then M_n is an $\mathcal{F}$-martingale.

From the point of view of gambling, if we think of each ξ_n as representing the outcome of a fair game based on a bet of one dollar, then M_n represents the wealth at time n of a gambler that places a one dollar bet on every game. The gambling interpretation of martingales doesn't really depend on the random walk structure of the example. Given any martingale we can interpret M_n as the wealth at time n and then use use a telescoping sum

$$M_n = M_0 + \sum_{j=1}^{n} (M_j - M_{j-1})$$

to represent the wealth at time n as the initial wealth M_0 plus the sum of the return $M_j - M_{j-1}$ on the first j bets.

The second example shows how one can make a martinagale out of the variance of a base martinagle.

Example 12.12. Suppose we have the setup of Example 12.11 except that we also assume a constant variance $\mathbf{E}\left[\xi_n^2\right] = \sigma^2$ for all $n > 0$. Then $M_n^2 - n\sigma^2$ is an $\mathcal{F}$-martingale. Integrability and $\mathcal{F}$-adaptedness are immediate from our assumptions. The martingale property requires a small computation

$$\begin{aligned}
\mathbf{E}\left[M_n^2 - n\sigma^2 \mid \mathcal{F}_{n-1}\right] &= \mathbf{E}\left[M_{n-1}^2 + 2M_{n-1}\xi_n + \xi_n^2 - n\sigma^2 \mid \mathcal{F}_{n-1}\right] \\
&= M_{n-1}^2 + 2M_{n-1}\mathbf{E}\left[\xi_n \mid \mathcal{F}_{n-1}\right] + \mathbf{E}\left[\xi_n^2 \mid \mathcal{F}_{n-1}\right] - n\sigma^2 \\
&= M_{n-1}^2 + 2M_{n-1}\mathbf{E}\left[\xi_n\right] + \mathbf{E}\left[\xi_n^2\right] - n\sigma^2 \\
&= M_{n-1}^2 - (n-1)\sigma^2
\end{aligned}$$

Returning to our gambling interpretation of martingales we discussed in Example 12.11, one can ask whether the "unit bet" assumption can be relaxed. That is we think of each increment $M_n - M_{n-1}$ as the return on a game in which one has wagered on dollar. It would be very interesting indeed to know whether there is a betting strategy that could make a fair game into an advantageous game (either for the gambler or the house). As manifested in our view of the world as a wealth process and a returns process, the bet on the n^{th} game is simply a multiplier A_n applied to the return $M_n - M_{n-1}$. Thus the betting strategy is also a stochastic process. To model reality, there is an important constraint on a betting strategy. A bet on the n^{th} game must be made prior to the n^{th} game being played and therefore should only should only be able to make use of information about the outcome of the first $n-1$ games. Thus a betting strategy must not only bet adapted the filtration $\mathcal{F}$ but satisfy the stonger condition of the following definition.

Definition 12.13. Given a filtration $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots$, we say a process A_n is *$\mathcal{F}$-previsible* or *$\mathcal{F}$-non-anticipating* if A_n is $\mathcal{F}_{n-1}$-measurable.

We make the assumption that a betting strategy is previsible and model the strategy as providing the amount that a gambler will bet. Of interest is that we allow the gambler to "short" the bet (i.e. bet a negative amount). It turns out that under reasonable conditions betting strategies alone cannot alter the fairness of a game.

Lemma 12.14. *Let M_n be a martingale and let A_n be an $\mathcal{F}$-previsible process with each A_n bounded and $A_0 = 1$. Define the* martingale transform $\tilde{M}_n = \sum_{j=0}^{n} A_j (M_j - M_{j-1})$ *(we define $M_{-1} = 0$ for simplicity). Then $\tilde{M}_n$ is a martingale.*

Proof. Clearly $\tilde{M}_n$ is $\mathcal{F}_n$-measurable as M_j and A_j are for each $j \leq n$. Integrability of $\tilde{M}_n$ follows from the integrability of the M_n and the boundedness of A_n. The martingale property follows from a simple computation

$$\begin{aligned}
\mathbf{E}\left[\tilde{M}_n \mid \mathcal{F}_{n-1}\right] &= \sum_{j=0}^{n} \mathbf{E}\left[A_j(M_j - M_{j-1}) \mid \mathcal{F}_{n-1}\right] \\
&= A_n \mathbf{E}\left[(M_n - M_{n-1}) \mid \mathcal{F}_{n-1}\right] + \sum_{j=0}^{n-1} A_j(M_j - M_{j-1}) \\
&= \tilde{M}_{n-1}
\end{aligned}$$

□

Lemma 12.15. *Let M_t be a martingale then $\boldsymbol{E}[M_t]$ is constant in $t \in T$.*

Proof. For $s, t \in T$ with $s < t$, by the martingale property and the chain rule of conditional expectations we have $\mathbf{E}[M_s] = \mathbf{E}\left[\mathbf{E}^{\mathcal{F}_s} M_t\right] = \mathbf{E}[M_t]$. □

Definition 12.16. Given a set $T \subset \overline{\mathbb{R}}$, we call a $T \cup \{\sup T\}$-valued random variable a *random time*. A random time is called an *$\mathcal{F}$-optional time* (also called an *$\mathcal{F}$-stopping time*) if and only if $\{\tau \leq t\} \in \mathcal{F}_t$ for all $t \in T$.

An $\mathcal{F}$-optional time τ represents a random decision rule of when to stop a game such that the decision to stop at time $t$can be made based only on information acccumulated up to and including time t (i.e. without seeing the future). Note that we allow a random time to take the value $\sup T$ (think of this as infinity) but the condition of being an optional time does not place a condition on what happens at $\sup T$.

Provided with an optional time there is a σ-algebra of events that is associated with it.

Definition 12.17. Given an optional time τ, we define

$$F_\tau = \{A \in \mathcal{A} \mid A \cap \{\tau \leq t\} \in \mathcal{F}_t \text{ for all } t \in T\}$$

Note that we have not taken the generated σ-algebra in the above definition, because of the following.

Lemma 12.18. *Given an optional time τ, $\mathcal{F}_\tau$ is a σ-algebra. Furthermore, τ is $\mathcal{F}_\tau$-measurable.*

Proof. Since $\Omega \cap \{\tau \leq t\} = \{\tau \leq t\} \in \mathcal{F}_t$ by definition of optional time, we see that $\Omega \in \mathcal{F}_\tau$. If we suppose that $A \in \mathcal{F}_\tau$ then for all $t \in T$, we apply elementary Boolean algebra and σ-algebra properties of $\mathcal{F}_t$ to see $A^c \cap \{\tau \leq t\} = (A \cap \{\tau \leq t\})^c \cap \{\tau \leq t\} \in \mathcal{F}_t$. Lastly, given $A_1, A_2, \cdots \in \mathcal{F}_\tau$, we have $(\cap_{n=1}^\infty A_n) \cap \{\tau \leq t\} = \cap_{n=1}^\infty (A_n \cap \{\tau \leq t\}) \in \mathcal{F}_t$ and thus $\mathcal{F}_\tau$ is a σ-algebra.

For every $s, t \in T$, we have $\{\tau \leq s\} \cap \{\tau \leq t\} = \{\tau \leq s \wedge t\} \in \mathcal{F}_{s \wedge t} \subset \mathcal{F}_t$ which shows every set $\{\tau \leq s\} \in \mathcal{F}_\tau$ for $s \in T$. Now for $s \in \mathbb{R} \setminus T$, $\{\tau \leq s\} = \cup_{t \in T; t<s} \{\tau \leq t\}$; the trick is that this is an uncountable union so we have to be a bit more careful in handling this case. Let $\tilde{s} = \sup\{t \leq s \mid t \in T\}$. The first thing to note is that $\{\tau \leq s\} = \{\tau \leq \tilde{s}\}$. The inclusion $\supset$ is obvious since $s \geq \tilde{s}$. To see the inclusion $\subset$ note that we cannot have $\tilde{s} < \tau(\omega) \leq s$ since $\tau(\omega) \in T$. If $\tilde{s} \in T$ then we have show $\{\tau \leq s\} \in \mathcal{F}_\tau$. Lets assume that $\tilde{s} \notin T$. By definition, we can find an increasing sequence $s_n \leq \tilde{s}$ such that $s_n \in T$ and $\lim_{n \to \infty} s_n = \tilde{s}$. Now we claim that $\cup_n \{\tau \leq s_n\} = \{\tau \leq \tilde{s}\}$. The inclusion $\subset$ follows since $s_n \leq \tilde{s}$. To see the other inclusion, suppose $\tau(\omega) \leq \tilde{s}$. Because we have assumed $\tilde{s} \notin T$ then in fact $\tau(\omega) < \tilde{s}$ and we can find s_n such that $\tau(\omega) < s_n < \tilde{s}$ showing $\omega \in \cup_n \{\tau \leq s_n\}$. Putting the two equalities together

$$\{\tau \leq s\} = \{\tau \leq \tilde{s}\} = \cup_n \{\tau \leq s_n\} \in \mathcal{F}_\tau$$

and we have shown that for all $s \in \mathbb{R}$, $\{\tau \leq s\} \in \mathcal{F}_\tau$. This suffices to show $\mathcal{F}_\tau$-measurability by Lemma 3.12. □

Conceptually, one thinks of the σ-algebra $\mathcal{F}_\tau$ as being events A such that if $\tau \leq t$ then one only needs information available at time t to determine whether A has occurred or not. More suggestively one may say that $\mathcal{F}_\tau$ as being the events that happen before τ.

Lemma 12.19. *Let σ and τ be optional times with $\sigma \leq \tau$, then $\mathcal{F}_\sigma \subset \mathcal{F}_\tau$.*

Proof. Suppose we have an $A \in \mathcal{F}_\sigma$. Because $\sigma \leq \tau$, we know that$\{\tau \leq t\} \subset \{\sigma \leq t\}$ for all $t \in T$. Take a $t \in T$, then $A \cap \{\tau \leq t\} = (A \cap \{\sigma \leq t\}) \cap \{\tau \leq t\} \in \mathcal{F}_t$. □

Lemma 12.20. *Let $T \subset \overline{\mathbb{R}}$ be a countable subset of the extended reals, let $\mathcal{F}_t$ be a filtration and $\tau : \Omega \to T$ be a random time. Then τ is an optional time if and only if $\{\tau = t\} \in \mathcal{F}_t$ for every $t \in T$.*

Proof. Suppose that $\{\tau = t\} \in \mathcal{F}_t$ then we see that

$$\{\tau \leq t\} = \cup_{s \leq t} \{\tau = s\}$$

which is a countable union of sets $\{\tau = s\} \in \mathcal{F}_s \subset \mathcal{F}_t$ hence is in $\mathcal{F}_t$.

Now if τ is $\mathcal{F}$-optional then similarly we may write

$$\{\tau = t\} = \{\tau \leq t\} \cap (\cup_{s<t} \{\tau \leq s\})^c$$

which shows that $\{\tau = t\} \in \mathcal{F}_t$. □

If we think of an optional time as a random stopping rule for a game, then a useful construct is the random stopping element associated with a process and the stopping rule. An interesting aspect of the proof is that it shows stopped processes can be represented as martingale transforms.

Lemma 12.21. *Let τ be an $\mathcal{F}$-optional time on a countable index set $T \subset \overline{\mathbb{R}}$ and let X be a stochastic process on T adapted to $\mathcal{F}$. Then the random element X_τ is $\mathcal{F}_\tau$-measurable.*

Proof. TODO □

12.1. **Discrete Time Martingales.** For the special case of index set $T = \mathbb{Z}_+$, we often call a martingale a *discrete time martingale*. Discrete martingales are well undertsood objects and as it turns out many important results about discrete martingales can be used to prove corresponding results for general martingales via approximation arguments. Thus, we will start our study of martingales by studying discrete martingales.

The first thing to note is a simple observation that the definition for the special case of discrete martingales can be simplified.

Lemma 12.22. *Let $\mathcal{F}_n$ be a filtration and M_n be a sequence of $\mathcal{F}$-adapted integrable random variables. If $\boldsymbol{E}^{\mathcal{F}_{n-1}} M_n = M_{n-1}$ for $n > 0$ then M_n is an $\mathcal{F}$-martingale.*

Proof. We only have to show that $\mathbf{E}^{\mathcal{F}_m} M_n = M_m$ for all $m \leq n$. Because we know M_n is $\mathcal{F}_n$-measurable then we have $\mathbf{E}^{\mathcal{F}_n} M_n = M_n$. If $m < n - 1$, then we proceed by induction assuming the result is true for $m + 1$,

$$\begin{aligned} \mathbf{E}^{\mathcal{F}_m} M_n &= \mathbf{E}^{\mathcal{F}_m} \mathbf{E}^{\mathcal{F}_{m+1}} M_n \\ &= \mathbf{E}^{\mathcal{F}_m} M_{m+1} && \text{by induction hypothesis} \\ &= M_m && \text{by hypothesis} \end{aligned}$$

□

Furthermore in discrete time we have a simple version of a construction of a useful class of optional times.

Definition 12.23. Let $\mathcal{F}$ be a filtration on $\mathbb{Z}_+$ and let X_n be an $\mathcal{F}$-adapted process with values in a meaurable space $(S, \mathcal{S})$. For every $A \in \mathcal{S}$ be can define the *hitting time* by

$$\tau_A = \min\{n \mid X_n \in A\}$$

where by convention we assume the minimum of the empty set is positive infinity.

For the moment the only thing we want to record about hitting times is that they are indeed optional times. They will soon thereafter start to prove their utility.

Lemma 12.24. *A hitting time is an $\mathcal{F}$-optional time.*

Proof. Simply write for every finite n,

$$\{\tau_A \leq n\} = \cup_{0 \leq m \leq n} \{X_m \in A\}$$

and note that by $\mathcal{F}$-adaptedness of X, we have $\{X_m \in A\} \in \mathcal{F}_m \subset \mathcal{F}_n$. □

Lemma 12.25. *Let M_n be a martingale and let τ be an optional time such that $\tau \leq C < \infty$, then $\boldsymbol{E}[M_\tau] = \boldsymbol{E}[M_0]$.*

Proof.

$$\begin{aligned}\mathbf{E}[M_\tau] &= \sum_{n=0}^{C} \mathbf{E}[M_n; \tau = n] \\ &= \sum_{n=0}^{C} \mathbf{E}[\mathbf{E}[M_C \mid \mathcal{F}_n]; \tau = n] \\ &= \sum_{n=0}^{C} \mathbf{E}[M_C; \tau = n] \\ &= \mathbf{E}\left[M_C; \cup_{n=0}^{C} \tau = n\right] = \mathbf{E}[M_C]\end{aligned}$$

Therefore the result follows from the case of a constant deterministic time. This latter case is just a simple induction on n. □

Theorem 12.26 (Optional Stopping Theorem)**.** *Let σ and τ be bounded $\mathcal{F}$-optional times and let X_n be a martingale, then*

$$\boldsymbol{E}[X_\tau \mid \mathcal{F}_\sigma] \geq X_{\sigma \wedge \tau} \text{ a.s.}$$

Proof. We warn the reader that the following proof is a bit longer than many you'll see in the literature. It intentionally avoids any of the tricks that make for short proofs in hopes of making a clearer explanation for why the result is in fact true.

We first begin with a simple special case that captures the essence of the result. Suppose τ is $\mathcal{F}$-optional and there exist constants k, m such that $k \leq \tau \leq m$. We need to prove that $\mathbf{E}[X_\tau \mid \mathcal{F}_k] = X_k$. We do this by induction on $m - k$. For $m - k = 0$, the result is trivial since in this case $X_\tau = X_k$. For the induction step suppose we have $k \leq \tau \leq m$ with $m - k > 0$ and note that we can use the induction hypothesis on the stopping time $k + 1 \leq \tau \vee k + 1 \leq m$. We get

$$\begin{aligned}\mathbf{E}[X_\tau \mid \mathcal{F}_k] &= \mathbf{E}[X_{\tau \vee k+1} \mid \mathcal{F}_k] + \mathbf{E}[(X_k - X_{k+1})\mathbf{1}_{\tau = k} \mid \mathcal{F}_k] \\ &= \mathbf{E}[\mathbf{E}[X_{\tau \vee k+1} \mid \mathcal{F}_{k+1}] \mid \mathcal{F}_k] + \mathbf{1}_{\tau = k}\mathbf{E}[(X_k - X_{k+1}) \mid \mathcal{F}_k] \\ &= \mathbf{E}[M_{k+1} \mid \mathcal{F}_k] + 0 = M_k\end{aligned}$$

To get the general result, we suppose that we are given $\sigma, \tau \leq N < \infty$ and we suppose we are given $A \in \mathcal{F}_\sigma$. Note that we can write $A = \cup_{n=0}^{N} A \cap \{\sigma = n\}$ where

$A \cap \{\sigma = n\} \in \mathcal{F}_n$ for all $0 \leq n \leq N$.

$$\begin{aligned}
\mathbf{E}[X_\tau; A] &= \sum_{n=0}^{N} \sum_{m=0}^{N} \mathbf{E}[X_n \mathbf{1}_{\tau=n} \mathbf{1}_{\sigma=m} \mathbf{1}_A] \\
&= \sum_{n=0}^{N} \left(\sum_{m=n}^{N} \mathbf{E}[X_m \mathbf{1}_{\tau=m} \mathbf{1}_{\sigma=n} \mathbf{1}_A] + \sum_{m=n+1}^{N} \mathbf{E}[X_n \mathbf{1}_{\tau=n} \mathbf{1}_{\sigma=m} \mathbf{1}_A] \right) \\
&= \sum_{n=0}^{N} \mathbf{E}[(X_{\tau \vee n} - X_n \mathbf{1}_{\tau<n}) \mathbf{1}_{\sigma=n} \mathbf{1}_A] + \mathbf{E}[X_n \mathbf{1}_{\tau=n} \mathbf{1}_{\sigma \geq n+1} \mathbf{1}_A] \\
&= \sum_{n=0}^{N} \mathbf{E}[X_n \mathbf{1}_{\tau \geq n} \mathbf{1}_{\sigma=n} \mathbf{1}_A] + \mathbf{E}[X_n \mathbf{1}_{\tau=n} \mathbf{1}_{\sigma \geq n+1} \mathbf{1}_A] \\
&= \sum_{n=0}^{N} \mathbf{E}[X_n \mathbf{1}_{\tau \wedge \sigma = n} \mathbf{1}_A] \\
&= \mathbf{E}[X_{\tau \wedge \sigma}; A]
\end{aligned}$$

and therefore by the defining property of conditional expectations we are done. □

Corollary 12.27. *Let M_n be a martingale and let τ be an optional time, then $M_{\tau \wedge n}$ is a martingale.*

Proof. This is an immediate consequence of Optional Stopping as $\tau \wedge n$ and $n-1$ are both bounded optional times and therefore

$$\mathbf{E}[M_{\tau \wedge n} \mid \mathcal{F}_{n-1}] = M_{\tau \wedge n \wedge (n-1)} = M_{\tau \wedge (n-1)}$$

Note that this can also be proven by a direct computation using the fact that $\{\tau \geq n\} = \{\tau \leq n-1\}^c \in \mathcal{F}_{n-1}$:

$$\begin{aligned}
\mathbf{E}[M_{\tau \wedge n} \mid \mathcal{F}_{n-1}] &= \sum_{m=0}^{n-1} \mathbf{E}[M_m \mathbf{1}_{\tau=m} \mid \mathcal{F}_{n-1}] + \mathbf{E}[M_n \mathbf{1}_{\tau \geq n} \mid \mathcal{F}_{n-1}] \\
&= \sum_{m=0}^{n-1} M_m \mathbf{1}_{\tau=m} + M_{n-1} \mathbf{1}_{\tau \geq n} \\
&= \sum_{m=0}^{n-2} M_m \mathbf{1}_{\tau=m} + M_{n-1} \mathbf{1}_{\tau \geq n-1} = M_{\tau \wedge (n-1)}
\end{aligned}$$

□

Lemma 12.28 (Doob Decomposition)**.** *Let X_n be a submartingale, then there exists a martingale M_n and an almost surely increasing $\mathcal{F}$-previsible process A_n such that $X_n = X_0 + M_n + A_n$.*

Proof. We start with $M_0 = A_0 = 0$ and proceed to define M_n by induction for $n > 0$ in the most natural way possible

$$\begin{aligned}
M_n &= X_n - \mathbf{E}[X_n \mid \mathcal{F}_{n-1}] + M_{n-1} \\
A_n &= X_n - M_n - X_0 = \mathbf{E}[X_n \mid \mathcal{F}_{n-1}] - M_{n-1} + X_0
\end{aligned}$$

a simple induction validating that M_n is $\mathcal{F}_n$-measurable, A_n is $\mathcal{F}_{n-1}$-measurable and M_n is integrable.

The martingale property follows immediately from the definition and the $\mathcal{F}_{n-1}$-measurability of $\mathbf{E}[X_n \mid \mathcal{F}_{n-1}]$ and M_{n-1}:

$$\mathbf{E}[M_n \mid \mathcal{F}_{n-1}] = \mathbf{E}[X_n \mid \mathcal{F}_{n-1}] - \mathbf{E}[\mathbf{E}[X_n \mid \mathcal{F}_{n-1}] \mid \mathcal{F}_{n-1}] + \mathbf{E}[M_{n-1} \mid \mathcal{F}_{n-1}] = M_{n-1}$$

The fact that A_n is increasing follows from

$$A_n = \mathbf{E}[X_n \mid \mathcal{F}_{n-1}] - M_{n-1} = \mathbf{E}[X_n \mid \mathcal{F}_{n-1}] - X_{n-1} + A_{n-1}$$

so that

$$A_n - A_{n-1} = \mathbf{E}[X_n \mid \mathcal{F}_{n-1}] - X_{n-1} \geq 0 \text{ a.s.}$$

by the submartingale property of X_n. □

The Doob Decomposition allows us to carry over the optional stopping theorem to submartingales

Corollary 12.29. *Let X_n be a submartingale and let σ and τ be bounded optional times, then $\boldsymbol{E}[X_\tau \mid \mathcal{F}_\sigma] \geq X_\sigma$ a.s.*

Proof. We write $X_n = M_n + A_n + X_0$ with M_n a martingale and A_n positive increasing previsible. Applying optional stopping (Theorem 12.26) and the Doob Decomposition we get

$$\mathbf{E}[X_\tau \mid \mathcal{F}_\sigma] = \mathbf{E}[M_\tau + A_\tau + X_0 \mid \mathcal{F}_\sigma] = M_{\sigma\wedge\tau} + \mathbf{E}[A_\tau \mid \mathcal{F}_\sigma] + X_0$$

so by a reverse application of the Doob Decomposition we just need to show $\mathbf{E}[A_\tau \mid \mathcal{F}_\sigma] \geq A_{\sigma\wedge\tau}$ a.s.

To see last fact first note that the monotonicity of A_n and the fact that $\sigma\wedge\tau \leq \tau$ shows us that $A_{\sigma\wedge\tau} \leq A_\tau$ a.s. Also we know that $\mathcal{F}_{\sigma\wedge\tau} \subset \mathcal{F}_\sigma$ and therefore the $\mathcal{F}_{\sigma\wedge\tau}$-measurability of $A_{\sigma\wedge\tau}$ implies $\mathcal{F}_\sigma$-measurability. Therefore applying these observations and monotonicity of conditional expectation we get

$$\mathbf{E}[A_\tau \mid \mathcal{F}_\sigma] - A_{\sigma\wedge\tau} = \mathbf{E}[A_\tau - A_{\sigma\wedge\tau} \mid \mathcal{F}_\sigma] \geq 0 \text{ a.s.}$$

and we are done. □

12.1.1. *Martingale Inequalities.* Intuitively one thinks of martingales as being essentially constant and submartingales as essentially increasing. These intuitions can be helpful when thinking of the types of properties that martingales should have. There are several fundamental inequalities that describe these ideas in a precise way. The first result we prove is a maximal inequality that can be viewed as an analogue of Kolmogorov's Maximal Inequality (Lemma 8.16) for a special case of dependent random variables.

Lemma 12.30 (Doob's Maximal Inequality)**.** *Let M_t be a submartingale on a countable index set T, then for every $\lambda > 0$,*

$$\lambda \boldsymbol{P}\{\sup_{s\leq t} M_s \geq \lambda\} \leq \boldsymbol{E}\left[M_t; \sup_{s\leq t} M_s \geq \lambda\right] \leq \boldsymbol{E}\left[M_t^+\right]$$

where $M_t^+ = M_t \vee 0$.

Proof. First we assume that T is a finite set. By reindexing we may as well assume that $T = \{n \mid n \leq N \text{ and } n \in \mathbb{Z}_+\}$ for some $N \geq 0$. Now pick an $n \in T$. The first thing to note is that for any submartingale M_n, $n \geq m$ and $A_m \in \mathcal{F}_m$, $\mathbf{E}[M_n; A_m] = \mathbf{E}[\mathbf{E}[M_n \mid \mathcal{F}_m]; A_m] \geq \mathbf{E}[M_m; A_m]$.

Now the event $\{\sup_{0\leq k\leq n} M_k \geq \lambda\}$ can be nicely reexpressed in terms of optional times. Define

$$\tau = \min\{n \mid M_n \geq \lambda\}$$

where we assume the minimum of the empty set is positive infinity. Note that $\{\sup_{0\leq k\leq n} M_k \geq \lambda\} = \{\tau \leq n\}$. If we consider the stopped process $M_\tau \mathbf{1}_{\tau\leq n} = \sum_{m=0}^n M_m \mathbf{1}_{\tau=m}$, take expectations and use the initial observation, $\mathbf{E}[M_\tau \mathbf{1}_{\tau\leq n}] \leq \sum_{m=0}^n \mathbf{E}[M_n \mathbf{1}_{\tau=m}] = \mathbf{E}[M_n \mathbf{1}_{\tau\leq n}]$. But on the other hand, by defintion of τ, we know that $\mathbf{E}[M_\tau \mathbf{1}_{\tau\leq n}] \geq \lambda \mathbf{E}[\mathbf{1}_{\tau\leq n}] = \lambda \mathbf{P}\{\sup_{0\leq k\leq n} M_k \geq \lambda\}$ which shows the first inequality.

The second inequality is true because nonnegativity of M_n^+ implies

$$0 \leq M_n \mathbf{1}_{\sup_{0\leq k\leq n} M_k \geq \lambda} \leq M_n^+$$

so we can apply monotonicity of expectation.

Now we want to extend the result to martingales on arbitrary countable index sets T. The proof above shows that the result holds for finite subsets of T. Now note that for any finite subsets $T' \subset T''$ such that $t \in T'$ we have

$$\{\sup_{\substack{s\leq t\\ s\in T'}} M_s \geq \lambda\} \subset \{\sup_{\substack{s\leq t\\ s\in T''}} M_s \geq \lambda\}$$

so if we write T as an increasing union of finite sets $T_0 \subset T_1 \subset \cdots$ then by continuity of measure (Lemma 3.27) we have

$$\mathbf{P}\{\sup_{\substack{s\leq t\\ s\in T}} M_s \geq \lambda\} = \lim_{m\to\infty} \mathbf{P}\{\sup_{\substack{s\leq t\\ s\in T_m}} M_s \geq \lambda\}$$

and by Dominated Convergence

$$\mathbf{E}\left[M_t; \sup_{\substack{s\leq t\\ s\in T}} M_s \geq \lambda\right] = \lim_{m\to\infty} \mathbf{E}\left[M_t; \sup_{\substack{s\leq t\\ s\in T_m}} M_s \geq \lambda\right]$$

proving the result for countable T. □

Having proven a tail inequality it is often a good idea to see what it might say about expectations via Lemma 6.8. In this case, with a bit of care we get the following result of Doob that can be interpreted as giving a bound on the extent to which a non-negative submartingale can deviate from being increasing.

Lemma 12.31 (Doob's L^p Inequality)**.** *Let X_t be a non-negative submartingle on a countable index set T, then for all $p > 1$ and $t \in T$,*

$$\|\sup_{s\leq t} X_s\|_p \leq \frac{p}{p-1}\|X_t\|_p$$

Proof. As with the proof of the maximal inequality we begin by assuming that T is finite and by reindexing equal to $\{n \in \mathbb{Z}_+ \mid n \leq N\}$ for some $N \geq 0$. We begin

let us start by assuming that $\mathbf{E}\left[(\sup_{0\le k\le n} X_k)^p\right] < \infty$. With this assumption in place we can apply Lemma 6.8 and Lemma 12.30 to get

$$\begin{aligned}
\mathbf{E}\left[(\sup_{0\le k\le n} X_k)^p\right] &= p\int_0^\infty \lambda^{p-1}\mathbf{P}\{\sup_{0\le k\le n} X_k \ge \lambda\}\,d\lambda \\
&\le p\int_0^\infty \lambda^{p-2}\mathbf{E}\left[X_n;\ \sup_{0\le k\le n} X_k \ge \lambda\right]\,d\lambda \\
&= p\mathbf{E}\left[X_n\int_0^\infty \lambda^{p-2}\mathbf{1}_{\sup_{0\le k\le n} X_k\ge\lambda}\,d\lambda\right] \\
&= \frac{p}{p-1}\mathbf{E}\left[X_n(\sup_{0\le k\le n} X_k)^{p-1}\right] \\
&\le \frac{p}{p-1}\|X_n\|_p\mathbf{E}\left[(\sup_{0\le k\le n} X_k)^p\right]^{\frac{p-1}{p}} \qquad \text{by Hölder's Inequality}
\end{aligned}$$

But now, we can divide both sides by $\mathbf{E}\left[(\sup_{0\le k\le n} X_k)^p\right]^{\frac{p-1}{p}}$ to get the result.

It remains to remove the assumption that $\mathbf{E}\left[(\sup_{0\le k\le n} X_k)^p\right] < \infty$. Obviously if $\|X_n\|_p = \infty$ then the result is trivially true so we may assume that $\|X_n\|_p < \infty$. Now we have for all $k \le n$, by Jensen's Inequality (Theorem 11.27), the submartingale property and the tower rule for conditional expectation

$$\mathbf{E}\left[X_k^p\right] \le \mathbf{E}\left[\mathbf{E}\left[X_n \mid \mathcal{F}_k\right]^p\right] \le \mathbf{E}\left[\mathbf{E}\left[X_n^p \mid \mathcal{F}_k\right]\right] = \mathbf{E}\left[X_n^p\right] < \infty$$

which shows that $\|X_k\|_p < \infty$ for all $0 \le k \le n$. But this implies that $\|\sup_{0\le k\le n} X_k\|_p < \infty$ (e.g. for any $\xi, \eta \in L^p$, write $\xi \vee \eta = \xi\mathbf{1}_{\xi>\eta} + \eta\mathbf{1}_{\xi\le\eta}$ and induct) and so the previous calculation proves the lemma for finite index sets.

Now to extend the result to arbitrary countable index sets T, simply observe if $t \in T' \subset T''$ then

$$\sup_{\substack{s\le t\\ s\in T'}} M_s \le \sup_{\substack{s\le t\\ s\in T''}} M_s$$

so we may take finite sets $T_0 \subset T_1 \subset \cdots$ such that $t \in T_0$ and $\cup_n T_n = T$ and use Monotone Convergence to conclude

$$\mathbf{E}\left[\sup_{\substack{s\le t\\ s\in T}} M_s\right] = \lim_{n\to\infty}\mathbf{E}\left[\sup_{\substack{s\le t\\ s\in T_n}} M_s\right] \le \frac{p}{p-1}\|X_t\|_p$$

□

Conceptually there are two ways that a real valued sequence can fail to converge: either the sequence escapes to infinity or the sequence oscillates. Our next goal is a result that puts explicit bounds on the expected amount of oscillation in any submartingle. More specifically, assume that we have fixed two reals numbers $a < b$; then we can focus in on the oscillations between the values a and b. Alternatively one can measure the number of times the value of the submartingale pass from below the lower bound a to above the upper bound b; each such transition is referred to as an *upcrossing*. To describe upcrossings precisely we first define the times at which pass below a and then the time we pass above b.

Lemma 12.32. *Let $\mathcal{F}_n$, M_n be a discrete $\mathcal{F}$- adapted process and let $a < b$ be real numbers. Let $\tau_0 = 0$ and for each $j \geq 0$ define inductively*

$$\sigma_j = \min\{n \mid n \geq \tau_j \text{ and } M_n \leq a\}$$
$$\tau_{j+1} = \min\{n \mid n \geq \sigma_j \text{ and } M_n \geq b\}$$

then each τ_j and σ_j is an $\mathcal{F}$-optional time. Furthermore if we define

$$U_a^b(n) = \max\{m \mid \tau_m \leq n\}$$

to be the number of upcrossings of X_n before n, then each $U_a^b(n)$ is measurable.

Proof. To see that τ_j and σ_j is an induction. Assume that τ_j is $\mathcal{F}$-optional for $j \leq n$. We write

$$\{\sigma_n = m\} = \bigcup_{0 \leq k < m} \left(\{\tau_n = k\} \cap \bigcap_{k < l < m} \{X_l > a\} \right) \cap \{X_m \leq a\}$$

and by $\mathcal{F}$-adaptedness of X_n and the fact that τ_n is $\mathcal{F}$-optional we see that $\{\sigma_n = m\} \in \mathcal{F}_m$. In a similar way we can express

$$\{\tau_{n+1} = m\} = \bigcup_{0 \leq k < m} \left(\{\sigma_n = k\} \cap \bigcap_{k < l < m} \{X_l < b\} \right) \cap \{X_m \geq b\}$$

and by $\mathcal{F}$-adaptedness of X_n and the just proven fact that σ_n is $\mathcal{F}$-optional we see that $\{\tau_{n+1} = m\} \in \mathcal{F}_m$.

To see measurability of $U_a^b(n)$ we just express

$$\{U_a^b(n) = m\} = \{\tau_m \leq n\} \cap \bigcap_{k > m} \{\tau_m > n\}$$

which is measurable because we have just shown each τ_m is an optional time (in particular is measurable). □

Lemma 12.33 (Doob's Upcrossing Inequality)**.** *Let X_n be a submartingale and let $U_a^b(n)$ be the number of upcrossings up to time n. Then*

$$\boldsymbol{E}\left[U_a^b(n)\right] \leq \frac{\boldsymbol{E}\left[(X_n - a)_+\right]}{b - a}$$

Proof. The first step of the proof is a reduction to a notationally simpler case. As the function $f(x) = (x - a)_+$ is convex and nondecreasing we know that $(X_n - a)_+$ is a positive submartingale. Furthermore $X_n \geq b$ if and only if $(X_n - a)_+ \geq b - a$ and $X_n \geq a$ if and only if $(X_n - a)_+ = 0$ and therefore the number of upcrossings of X_n between a and b is the same as the number of upcrossings of $(X_n - a)_+$ between 0 and $b - a$. Therefore the result is proven if we show that for every positive submartingale X_n and $b > 0$ we have

$$U_0^b(n) \leq \frac{\mathbf{E}\left[X_n\right]}{b}$$

To finish the proof, note that by definition, for any $j > 0$ we have $\sigma_j - \tau_j > 0$ and $\tau_{j+1} - \sigma_j > 0$ and therefore we have $\sigma_n \geq n$ and $\tau_n \geq n$ and we get the following

expression of X_n as the telescoping sum of stopped processes

$$X_n = X_{\tau_0 \wedge n} + \sum_{j=0}^{n} \left(X_{\sigma_j \wedge n} - X_{\tau_j \wedge n} \right) + \sum_{j=0}^{n} \left(X_{\tau_{j+1} \wedge n} - X_{\sigma_j \wedge n} \right)$$

Taking expectations we note that from the positivity of X_n we have $\mathbf{E}\left[X_{\tau_0 \wedge n}\right] \geq 0$ and because $\sigma_j \geq \tau_j$ and the optional stopping theorem for submartingles (Corollary 12.29) we have

$$\begin{aligned} \mathbf{E}\left[X_{\sigma_j \wedge n} - X_{\tau_j \wedge n}\right] &= \mathbf{E}\left[\mathbf{E}\left[X_{\sigma_j \wedge n} - X_{\tau_j \wedge n} \mid \mathcal{F}_{\tau_j}\right]\right] \\ &\geq \mathbf{E}\left[X_{\tau_j \wedge n} - X_{\tau_j \wedge n}\right] = 0 \end{aligned}$$

and similarly $\mathbf{E}\left[X_{\tau_{j+1} \wedge n} - X_{\sigma_j \wedge n}\right] \geq 0$. Thus

$$\begin{aligned} \mathbf{E}\left[X_n\right] &= \mathbf{E}\left[X_{\tau_0 \wedge n}\right] + \sum_{j=0}^{n} \mathbf{E}\left[X_{\sigma_j \wedge n} - X_{\tau_j \wedge n}\right] + \sum_{j=0}^{n} \mathbf{E}\left[X_{\tau_{j+1} \wedge n} - X_{\sigma_j \wedge n}\right] \\ &\geq \sum_{j=0}^{n} \mathbf{E}\left[X_{\tau_{j+1} \wedge n} - X_{\sigma_j \wedge n}\right] \\ &\geq b\mathbf{E}\left[U_0^b(n)\right] \end{aligned}$$

and therefore the result is proved. □

TODO: Add comments about the result that $\mathbf{E}\left[X_{\sigma_j \wedge n} - X_{\tau_j \wedge n}\right] \geq 0$. Given the definition of σ_j and τ_j this result might seem a bit counterintuitive. The explanation for how this result can hold is that in fact is very unlikely that $\sigma_j < n$; with high probability $\sigma_j \geq n$ and $X_{\sigma_j \wedge n} = X_n \geq X_{\tau_j \wedge n}$. This explanation is completely consistent with the conceptual model that submartingales are not oscillating much and is really one of the two main points of the result (the other main point being the fact that a lower bound for the terms $\mathbf{E}\left[X_{\tau_{j+1} \wedge n} - X_{\sigma_j \wedge n}\right]$ is given by $(b-a)U_a^b(n)$).

Theorem 12.34. *Let X_n be a submartingale with $\sup_n \|X_n\|_1 < \infty$ then there exists and integrable random variable X such that $X_n \overset{a.s.}{\to} X$.*

Proof. The first order of business here is leverage the Doob Upcrossing Inequality to show that X_n is not oscillatory and therefore has a limit (possibly infinite) almost surely. To do that for every $a \in \mathbb{R}$, we note the elementary inequality $(x-a)_+ \leq |x|+|a|$ and therefore we can that $\mathbf{E}\left[(X_n - a)_+\right] \leq \sup_n \|X_n\|_1 + |a| < \infty$. Supposing $a, b \in \mathbb{R}$ with $a < b$ and $U_a^b(n)$ be the number of upcrossings of $[a, b]$ before n, we can see that $U_a^b(n)$ is positive and increasing in n and Lemma 12.33 and Monotone Convergence tell us that

$$\mathbf{E}\left[\lim_{n\to\infty} U_a^b(n)\right] = \lim_{n\to\infty} \mathbf{E}\left[U_a^b(n)\right] \leq \lim_{n\to\infty} \frac{\|X_n\|_1 + |a|}{b-a} \leq \frac{\sup_n \|X_n\|_1 + |a|}{b-a} < \infty$$

If we let $U_a^b(\infty) = \lim_{n\to\infty} U_a^b(n)$ be the number of upcrossing on $\mathbb{Z}_+$, then $U_a^b(\infty)$ is integrable and therefore almost surely finite.

Now for each $a, b \in \mathbb{Q}$ with $a < b$ let

$$\Lambda_a^b = \{\liminf_{n\to\infty} X_n < a < b < \limsup_{n\to\infty} X_n\}$$

and note that $\Lambda_a^b \subset \{U_a^b(\infty) = \infty\}$ (we can pick subsequences N and M such that X_n converges to $\liminf_{n\to\infty} X_n$ along N and $\limsup_{n\to\infty} X_n$ along M and in this

way construct an infinite number of upcrossings of $[a,b]$). Thus $\mathbf{P}\{\Lambda_a^b\} = 0$ and taking the countable union

$$\mathbf{P}\{\liminf_{n\to\infty} X_n < \limsup_{n\to\infty} X_n\} = \mathbf{P}\{\bigcup_{\substack{a,b\in\mathbb{Q}\\ a<b}} \Lambda_a^b\} = 0$$

and therefore $\lim_{n\to\infty} X_n$ exists almost surely.

Let $X = \lim_{n\to\infty} X_n$ and our last task is to show that X is integrable (hence almost surely finite as well). This follows from Fatou's Lemma

$$\mathbf{E}\left[|X|\right] \leq \liminf_{n\to\infty} \mathbf{E}\left[|X_n|\right] \leq \sup_n \|X_n\|_1 < \infty$$

and we are done. □

Note that despite the fact that the limit of the submartingale is integrable in the above theorem, it is not necessarily the case that the convergence is L^1. TODO: Provide example of a non-uniformly integrable martingale with almost sure but not L^1 convergence.

In the martingale case we can characterize the conditions under which the convergence to a limit is in L^1. Furthermore in this case, the martingale is closed (see Lemma 12.10 for the definition of closed martingales).

Theorem 12.35. *Let X_n be a martingale then the following are equivalent*

(i) *X_n is uniformly integrable*
(ii) *there exists an integrable X such that $X_n \xrightarrow{L^1} X$*
(iii) *there exists an integrable X such that $X_n = \boldsymbol{E}^{\mathcal{F}_n} X$ almost surely.*

Proof. To see (i) implies (ii) we know from Lemma 8.48 that X_n uniformly integrable implies L^1 boundedness, hence we can apply Theorem 12.34 to conclude the existence of an integrable X such that $X_n \xrightarrow{a.s.} X$. However almost sure convergence implies convergence in probability (Lemma 8.4) which together with uniform integrability implies $X_n \xrightarrow{L^1} X$ (Lemma 8.53).

To that (ii) implies (iii) suppose that $\epsilon > 0$ is given and let $N > 0$ be such that $\|X_n - X\|_1 = \mathbf{E}\left[|X_n - X|\right] < \epsilon$ for all $n \geq N$. Pick an $m \in \mathbb{Z}_+$, $n \geq N \vee m$ and let $A \in \mathcal{F}_m$. We calculate

$$\begin{aligned} |\mathbf{E}\left[X;A\right] - \mathbf{E}\left[X_m;A\right]| &= |\mathbf{E}\left[X;A\right] - \mathbf{E}\left[X_n;A\right]| \qquad \text{since } \mathbf{E}\left[X_n \mid \mathcal{F}_m\right] = X_m \\ &\leq \mathbf{E}\left[|X - X_n|\,;A\right] \\ &\leq \mathbf{E}\left[|X - X_n|\right] < \epsilon \end{aligned}$$

and since ϵ is arbitrary, we conclude $\mathbf{E}\left[X;A\right] = \mathbf{E}\left[X_m;A\right]$ and therefore $\mathbf{E}\left[X \mid \mathcal{F}_m\right] = X_m$ a.s.

To see that (ii) implies (iii), we use Lemma 8.48. First note that by contraction property of conditional expectation, we have $\sup_n \mathbf{E}\left[|\mathbf{E}\left[X \mid \mathcal{F}_n\right]|\right] \leq \mathbf{E}\left[|X|\right]$ so the first condition of the lemma holds. To see the second condition, let $\epsilon > 0$ be fixed and pick $R > 0$ such that $\mathbf{E}\left[|X|\,;|X| > R\right] < \frac{\epsilon}{2}$ and pick A such that $\mathbf{P}\{A\} < \frac{\epsilon}{2R}$.

Now, for every n,

$$\begin{aligned}
|\mathbf{E}\left[\mathbf{E}\left[X \mid \mathcal{F}_n\right]; A\right]| &\leq \mathbf{E}\left[\mathbf{E}\left[|X| \mid \mathcal{F}_n\right]; A\right] \\
&= \mathbf{E}\left[|X| \cdot \mathbf{E}\left[\mathbf{1}_A \mid \mathcal{F}_n\right]\right] \\
&= \mathbf{E}\left[|X| \cdot \mathbf{E}\left[\mathbf{1}_A \mid \mathcal{F}_n\right]; |X| \leq R\right] + \mathbf{E}\left[|X| \cdot \mathbf{E}\left[\mathbf{1}_A \mid \mathcal{F}_n\right]; |X| > R\right] \\
&\leq R\mathbf{E}\left[\mathbf{E}\left[\mathbf{1}_A \mid \mathcal{F}_n\right]\right] + \mathbf{E}\left[|X|; |X| > R\right] \\
&\leq \epsilon
\end{aligned}$$

and therefore we have condition (ii) of Lemma 8.48 satisfied and uniform integrability is shown. □

It should be noted that the proof of (iii) implies (i) in previous argument did not depend on the fact that we were dealing with a filtration; in fact we have following corollary to the proof.

Corollary 12.36. *Suppose ξ is an integrable random variable the collection of random variables $\boldsymbol{E}[\xi \mid \mathcal{F}]$ for all σ-algebras $\mathcal{F}$ is uniformly integrable.*

Proof. For any $\mathcal{F}$ just replay the argument that (iii) implies (i) in the previous result. □

Convergence of martingales in L^p spaces with $p > 1$ is equivalent to boundedness.

Theorem 12.37 (L^p Martingale Convergence)**.** *Given a martingale M_n, then for $p > 1$, there exists an $M \in L^p$ such that $M_n \xrightarrow{L^p} M$ if and only if M_n is L^p bounded.*

Proof. Suppose M_n is an L^p bounded martingles. By L^p boundedness, we know that M_n is uniformly integrable thus by Theorem 12.35 we know there is an integrable M such that $M_n \xrightarrow{a.s.} M$ (thus $|M_n|^p \xrightarrow{a.s.} |M|^p$ and $M_n \xrightarrow{L^1} M$. By Doob's L^p inequality, for every n we have $\|\sup_{0\leq k\leq n} M_k\|_p \leq \frac{p}{p-1}\|M_n\|_p < \frac{p}{p-1}\sup_n\|M_n\|_p < \infty$ therefore by Monotone Convergence we have $\|\sup_{0\leq k\leq\infty} M_k\|_p = \lim_{n\to\infty}\|\sup_{0\leq k\leq n} M_k\|_p < \infty$. Now we clearly have $|M_n|^p \leq (\sup_{0\leq k\leq\infty} M_k)^p$ and Dominated Convergence gives us $M_n \xrightarrow{L^p} M$.

Now if we assume that $M_n \xrightarrow{L^p} M$ then we also know from Corollary 6.12 that $M_n \xrightarrow{L^1} M$ so Theorem 12.35 implies that $M_n = \mathbf{E}\left[M \mid \mathcal{F}_n\right]$ a.s. for every n. Now by Jensen's Inequality (Theorem 11.27)

$$\mathbf{E}\left[|M_n|^p\right] = \mathbf{E}\left[|\mathbf{E}\left[M \mid \mathcal{F}_n\right]|^p\right] \leq \mathbf{E}\left[\mathbf{E}\left[|M|^p \mid \mathcal{F}_n\right]\right] = \|M\|_p^p < \infty$$

Now the L^p boundedness follows from the fact that $M_n \xrightarrow{L^p} M$ since we can find $N > 0$ such that $\|M_n\|_p \leq \|M\|_p + \|M_n - M\|_p \leq \|M\|_p + 1$ for $n > N$ and then $M = \max(\|M_1\|_p, \ldots, \|M_N\|_p, \|M\|_p + 1)$ is a bound for L^p norms of the M_n. □

We now give a result that we'll use in the transition to continuous time.

Theorem 12.38. *Let ξ be an integrable random variable and let $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots$ be filtration, then $\boldsymbol{E}[\xi \mid \mathcal{F}_n]$ converges to $\boldsymbol{E}[\xi \mid \bigvee_n \mathcal{F}_n]$ both almost surely and in L^1.*

Proof. We know from the tower property of conditional expectation and Corollary 12.36 that $\mathbf{E}\left[\xi \mid \mathcal{F}_n\right]$ is a uniformly integrable martingle and is closable and converges both almost surely and in L^1. Let M be the limit and we need to show that $\mathbf{E}\left[\xi \mid \bigvee_n \mathcal{F}_n\right] = M$ almost surely. We know that M is $\bigvee_n \mathcal{F}_n$-measurable since

it is an almost sure limit of M_n each of which is $\bigvee_n \mathcal{F}_n$-measurable. Furthermore by Theorem 12.35 we also know that $\mathbf{E}[M \mid \mathcal{F}_n] = \mathbf{E}[\xi \mid \mathcal{F}_n]$ almost surely. Now suppose that we have $A \in \mathcal{F}_n$ for some n. We have

$$\begin{aligned}\mathbf{E}[M; A] &= \mathbf{E}[\mathbf{E}[M \mid \mathcal{F}_n]; A] \\ &= \mathbf{E}[\mathbf{E}[\xi \mid \mathcal{F}_n]; A] \\ &= \mathbf{E}[\mathbf{E}[\mathbf{E}[\xi \mid \vee_n \mathcal{F}_n] \mid \mathcal{F}_n]; A] \\ &= \mathbf{E}[\mathbf{E}[\xi \mid \vee_n \mathcal{F}_n]; A]\end{aligned}$$

thus $\mathbf{E}[M; A] = \mathbf{E}[\mathbf{E}[\xi \mid \bigvee_n \mathcal{F}_n]; A]$ for all A belonging to the π-system $\cup_n \mathcal{F}_n$. By a monotone class argument (Lemma 11.7) we conclude that $M = \mathbf{E}[\xi \mid \bigvee_n \mathcal{F}_n]$ almost surely. □

TODO: This result can be proven directly without appealing to the martingle convergence theorems (Stroock does this). Is there any point in doing so here? Should we move this result further down and put it in the context of the discussion of approximating continuous optional times by discrete ones? Stroock has some other interesting consequences of this theorem too. Here is the proof that depends only on the Doob Maximal Inequality.

Proof. Before we begin, we can clean up the notation that follows by assuming that $\mathcal{A} = \bigvee_n \mathcal{F}_n$. For if ξ is integrable then we know that $\mathbf{E}[\xi \mid \bigvee_n \mathcal{F}_n]$ is also integrable and convergence in $L^1(\Omega, \bigvee_n \mathcal{F}_n, \mu)$ implies convergence in $L^1(\Omega, \mathcal{A}, \mu)$.

First goal is to validate the following claim:

$$\lambda \mathbf{P}\{\sup_{n \in \mathbb{Z}_+} |\mathbf{E}[\xi \mid \mathcal{F}_n]| \geq \lambda\} \leq \mathbf{E}\left[|\xi|; \sup_{n \in \mathbb{Z}_+} |\mathbf{E}[\xi \mid \mathcal{F}_n]| \geq \lambda\right] \leq \mathbf{E}[|\xi|]$$

Here is where Stroock reduces this to Doob's Maximal Inequality along the way claiming that we may assume $\xi \geq 0$. I don't understand how to validate his claim about the positivity assumption and I am stuck trying to use Doob's Maximal Inequality as we've stated it but it is easy to adapt the proof of the Maximal Inequality to prove the above as you'll see. We first prove the claim for a finite index set. Since we know from Lemma 12.10 that $\mathbf{E}[\xi \mid \mathcal{F}_n]$ is an $\mathcal{F}$-martingale. we know from that $|\mathbf{E}[\xi \mid \mathcal{F}_n]|$ is a submartingale. We let τ be hitting time of the interval $[\lambda, \infty)$ and note that

$$\{\sup_{n \in \mathbb{Z}_+} |\mathbf{E}[\xi \mid \mathcal{F}_n]| \geq \lambda\} = \cup_{0 \leq m \leq n} \{\tau = m\}$$

where the union is disjoint. Since τ is an optional time (Lemma 12.24) we also know that $\{\tau = m\} \in \mathcal{F}_m$ and therefore

$$\mathbf{E}[|\xi|; \tau = m] = \mathbf{E}[\mathbf{E}[|\xi| \mid \mathcal{F}_m]; \tau = m] \geq \mathbf{E}[|\mathbf{E}[\xi \mid \mathcal{F}_m]|; \tau = m] \geq \lambda \mathbf{P}\{\tau = m\}$$

and summing for m from 0 to n yields

$$\lambda \mathbf{P}\{\max_{0 \leq m \leq n} |\mathbf{E}[\xi \mid \mathcal{F}_m]| \geq \lambda\} \leq \mathbf{E}\left[|\xi|; \max_{0 \leq m \leq n} |\mathbf{E}[\xi \mid \mathcal{F}_m]| \geq \lambda\right]$$

The result is completed by taking the limit as n goes to infinity and using continuity of measure (Lemma 3.27) and Montone Convergence.

Here is the result from Stroock We know from Lemma 12.10 that $\mathbf{E}[\xi \mid \mathcal{F}_n]$ is an $\mathcal{F}$-martingale. By Doob's Maximal Inequality (Lemma 12.30), the $\mathcal{F}_n$-measurability

of $\{\sup_{0\leq k\leq n}\mathbf{E}\left[\xi\mid\mathcal{F}_k\right]\geq\lambda\}$ and another application of the tower property we know that

$$\begin{aligned}\lambda\mathbf{P}\{\sup_{0\leq k\leq n}\mathbf{E}\left[\xi\mid\mathcal{F}_k\right]\geq\lambda\}&\leq\mathbf{E}\left[\mathbf{E}\left[\xi\mid\mathcal{F}_n\right];\sup_{0\leq k\leq n}\mathbf{E}\left[\xi\mid\mathcal{F}_k\right]\geq\lambda\right]\\&=\mathbf{E}\left[\xi;\sup_{0\leq k\leq n}\mathbf{E}\left[\xi\mid\mathcal{F}_k\right]\geq\lambda\right]\end{aligned}$$

By continuity of measure (Lemma 3.27) we know that

$$\mathbf{P}\{\sup_k\mathbf{E}\left[\xi\mid\mathcal{F}_k\right]\geq\lambda\}=\lim_{n\to\infty}\mathbf{P}\{\sup_{0\leq k\leq n}\mathbf{E}\left[\xi\mid\mathcal{F}_k\right]\geq\lambda\}$$

and by Dominated Convergence

$$\mathbf{E}\left[\xi;\sup_k\mathbf{E}\left[\xi\mid\mathcal{F}_k\right]\geq\lambda\right]=\lim_{n\to\infty}\mathbf{E}\left[\xi;\sup_{0\leq k\leq n}\mathbf{E}\left[\xi\mid\mathcal{F}_k\right]\geq\lambda\right]$$

so we have shown

$$\lambda\mathbf{P}\{\sup_k\mathbf{E}\left[\xi\mid\mathcal{F}_k\right]\geq\lambda\}\leq\mathbf{E}\left[\xi;\sup_k\mathbf{E}\left[\xi\mid\mathcal{F}_k\right]\geq\lambda\right]$$

TODO: Is this result only used to show uniform integrability? *End result from Stroock*

Since we know that the family $\mathbf{E}\left[\xi\mid\mathcal{F}_n\right]$ is uniformly integrable by Corollary 12.36 it suffices to show that $\mathbf{E}\left[\xi\mid\mathcal{F}_n\right]\overset{a.s.}{\to}\xi$.

To show almost sure convergence, we let $\mathcal{G}$ denote the set of all integrable ξ such that $\mathbf{E}\left[\xi\mid\mathcal{F}_n\right]\overset{a.s.}{\to}\xi$. Note that any $\mathcal{F}_n$-measurable ξ is in $\mathcal{G}$ since the sequence of conditional expectations is eventually almost surely constant and equal to ξ. On the other hand we know that $\cup_n L^1(\Omega,\mathcal{F}_n,\mu)$ is dense in $L^1(\Omega,\bigvee_n\mathcal{F}_n,\mu)=L^1(\Omega,\mathcal{A},\mu)$ (Lemma 11.5) so it suffices to show that $\mathcal{G}$ is closed in L^1. So suppose that ξ_n is a sequence in $\mathcal{G}$ such that $\xi_n\overset{L^1}{\to}\xi$. We show that $\mathbf{E}\left[\xi\mid\mathcal{F}_n\right]\overset{a.s.}{\to}\xi$ by using Lemma 8.3. Suppose $\epsilon>0$ is given, then for every m,n

$$\begin{aligned}\mathbf{P}\{\sup_{k\geq m}|\mathbf{E}\left[\xi\mid\mathcal{F}_k\right]-\xi|>\epsilon\}&\leq\mathbf{P}\{\sup_{k\geq m}|\mathbf{E}\left[\xi-\xi_n\mid\mathcal{F}_k\right]|>\frac{\epsilon}{3}\}+\\&\qquad\mathbf{P}\{\sup_{k\geq m}|\mathbf{E}\left[\xi_n\mid\mathcal{F}_k\right]-\xi_n|>\frac{\epsilon}{3}\}+\mathbf{P}\{|\xi_n-\xi|>\frac{\epsilon}{3}\}\\&\leq\frac{6}{\epsilon}\mathbf{E}\left[|\xi-\xi_n|\right]+\mathbf{P}\{\sup_{k\geq m}|\mathbf{E}\left[\xi_n\mid\mathcal{F}_k\right]-\xi_n|>\frac{\epsilon}{3}\}\end{aligned}$$

where the first term is bounded by our claim at the beginning of proof applied to $\xi_n-\xi$, the second term goes to zero as m goes to infinity by our assumption that $\xi_n\in\mathcal{G}$ and Lemma 8.3 and the third term is bounded by the Markov Inequality (Lemma 14.1).

Taking the limit as m goes to infinity and using our assumption that $\xi_n\in\mathcal{G}$ and Lemma 8.3 we get

$$\lim_{m\to\infty}\mathbf{P}\{\sup_{k\geq m}|\mathbf{E}\left[\xi\mid\mathcal{F}_k\right]-\xi|>\epsilon\}\leq\frac{6}{\epsilon}\mathbf{E}\left[|\xi-\xi_n|\right]$$

and then taking the limit as n goes to infinity we get

$$\mathbf{P}\{\sup_{k\geq m} |\mathbf{E}[\xi \mid \mathcal{F}_k] - \xi| > \epsilon\} = 0$$

so the result it proved. □

12.2. **Continuous Time Martingales and Weakly Optional Times.** Our next goal is to extend the theory we've developed to a continuous time setting. For the most part we proceed by using approximation arguments to reduce results to the discrete time analogues proven in the last section. First we have to come to grips with some subtleties related to filtrations, optional times and measurability in continuous time.

Definition 12.39. A T-valued random variable is called a *weakly $\mathcal{F}$-optional time* (also called a *weak $\mathcal{F}$-stopping time*) if and only if $\{\tau < t\} \in \mathcal{F}_t$ for all $t \in T$.

Just as with optional times, if the filtration $\mathcal{F}$ is clear from context, we'll simply refer to a weakly optional time.

A weakly $\mathcal{F}$-optional time τ is a decision rule to stop at t that requires an arbitrarily small amount of future information to determine that one should stop at t. Alternatively one can characterize it as a decision rule such that $\tau + \epsilon$ is $\mathcal{F}$-optional for all $\epsilon > 0$.

Let $\mathcal{F}^+ = \cup_{s>t}\mathcal{F}_s$ (note that $\mathcal{F} = \mathcal{F}^+$ if and only if $\mathcal{F}$ is right continuous).

One way of defining the σ-algebra associated with a weakly $\mathcal{F}$-optional time is as a limit of the σ-algebras associated the $\mathcal{F}$-optional times $\tau + \epsilon$

$$\mathcal{F}_{\tau^+} = \cup_{\epsilon>0}\mathcal{F}_{\tau+\epsilon}$$

Lemma 12.40. *τ is $\mathcal{F}^+$-optional if and only if τ is weakly $\mathcal{F}$-optional. In this case,*

$$\mathcal{F}_\tau^+ = \mathcal{F}_{\tau^+} = \{A \in \mathcal{A} \mid A \cap \{\tau < t\} \in \mathcal{F}_t \text{ for all } t \in T\}$$

Proof. The first thing is to notice that for any random time τ (not just optional or weakly optional times) we have the equalities

$$\{\tau \leq t\} = \bigcap_{\substack{r\in\mathbb{Q}\\ r>t}} \{\tau < r\} \qquad\qquad \{\tau < t\} = \bigcup_{\substack{r\in\mathbb{Q}\\ r<t}} \{\tau \leq r\}$$

TODO: Justify (but it's kinda obvious by density of $\mathbb{Q}$)

Armed with these facts we proceed to show the equality

$$\mathcal{F}_\tau^+ = \{A \in \mathcal{A} \mid A \cap \{\tau < t\} \in \mathcal{F}_t \text{ for all } t \in T\}$$

for any random time τ.

Suppose $A \cap \{\tau \leq t\} \in \mathcal{F}_t^+ = \cap_{s>t}\mathcal{F}_s$ for every $t \in T$. Then for all $t \in T$,

$$A \cap \{\tau < t\} = A \cap \left(\bigcup_{\substack{r\in\mathbb{Q}\\ r<t}} \{\tau \leq r\}\right) = \bigcup_{\substack{r\in\mathbb{Q}\\ r<t}} (A \cap \{\tau \leq r\}) \in \mathcal{F}_t$$

since for any $r < t$, $\mathcal{F}_r^+ \subset \mathcal{F}_t$.

On the other hand, if $A \cap \{\tau < t\} \in \mathcal{F}_t$ for all $t \in T$, then

$$A \cap \{\tau \leq t\} = A \cap \left(\bigcap_{\substack{r \in \mathbb{Q} \\ r > t}} \{\tau < r\} \right) = \bigcap_{\substack{r \in \mathbb{Q} \\ r > t}} (A \cap \{\tau < r\}) \in \mathcal{F}_t^+$$

where the last inclusion follows from the fact that for any $r < s$, $A \cap \{\tau < r\} \subset A \cap \{\tau < s\}$, so for any $s \in T$ with $s > t$ we in fact have

$$\bigcap_{\substack{r \in \mathbb{Q} \\ r > t}} (A \cap \{\tau < r\}) = \bigcap_{\substack{r \in \mathbb{Q} \\ s \geq r > t}} (A \cap \{\tau < r\}) \in \mathcal{F}_s$$

Now note that by definition, τ is weakly $\mathcal{F}$-optional if and only if $\Omega \in \{A \in \mathcal{A} \mid A \cap \{\tau < t\} \in \mathcal{F}_t \text{ for all } t \in T\}$ and τ is $\mathcal{F}^+$-optional if and only if $\Omega \in \mathcal{F}_\tau^+$. Therefore the equality just shown tells us that τ is weakly $\mathcal{F}$-optional if and only if τ is $\mathcal{F}^+$-optional.

We finish by showing that $\mathcal{F}_\tau^+ = \mathcal{F}_{\tau^+}$. To see this, note that $A \in \mathcal{F}_{\tau^+}$ if and only if $A \in \mathcal{F}_{\tau+\epsilon}$ for all $\epsilon > 0$ which is true if and only if $A \cap \{\tau + \epsilon \leq t\} = A \cap \{\tau \leq t - \epsilon\} \in \mathcal{F}_t$ for all $t \in T$, $\epsilon > 0$ which is true if and only if $A \cap \{\tau \leq t\} \in \mathcal{F}_{t+\epsilon}$ for all $t \in T$, $\epsilon > 0$. This last statement is simply that $A \cap \{\tau \leq t\} \in \mathcal{F}_t^+$ for all $t \in T$ so we are done. □

When passing from discrete time results to continuous time results it is often useful to approximate an optional time on a continuous domain by a discrete one. The following approximation scheme is so useful it deserves to be called out.

Lemma 12.41. *Let τ be a weakly optional time on $\mathbb{R}_+$, then define*

$$\tau_n = \frac{1}{2^n} \lfloor 2^n \tau + 1 \rfloor$$

τ_n is a sequence of optional times with values in a countable index set such that $\tau_n \downarrow \tau$.

Proof. The fact that each τ_n is an optional time follows from the definition and the fact that τ is a weakly optional time:

$$\{\tau_n \leq \frac{k}{2^n}\} = \{\frac{k-1}{2^n} \leq \tau < \frac{k}{2^n}\} = \{\tau < \frac{k-1}{2^n}\}^c \cap \{\tau < \frac{k}{2^n}\} \in \mathcal{F}_{\frac{k}{2^n}}$$

To see the fact that τ_n is decreasing, note $\tau_n = \frac{k}{2^n}$ if and only if $\frac{k-1}{2^n} \leq \tau < \frac{k}{2^n}$ which implies

$$\tau_{n+1} = \begin{cases} \frac{k}{2^n} & \text{if } \frac{2k-1}{2^{n+1}} \leq \tau < \frac{k}{2^n} \\ \frac{2k-1}{2^{n+1}} & \text{if } \frac{k-1}{2^n} \leq \tau < \frac{2k-1}{2^{n+1}} \end{cases}$$

Convergence to τ follows easily since $|\tau - \tau_n| \leq \frac{1}{2^n}$ by definition. □

If we have approximation scheme for an optional time we may also want to understand how the associated σ-algebras behave. For the decreasing approximation of the previous lemma, part (ii) of the following gives us the answer.

Lemma 12.42. *If we have a finite or countable collection of optional times τ_n then $\sup_n \tau_n$ is an optional time. If we have a finite or countable collection of weakly optional times τ_n then $\tau = \inf_n \tau_n$ is a weakly optional time and furthermore*

$$\mathcal{F}_\tau^+ = \cap_n \mathcal{F}_{\tau_n}^+$$

Proof. If τ_n are optional times then if follows from the definition of supremum that $\{\tau \leq t\} = \cap_n \{\tau_n \leq t\}$ and therefore τ is an optional time.

If τ_n are weakly optional times then if follows from the definition of infimum that $\{\tau < t\} = \cup_n \{\tau_n < t\}$ and therefore τ is a weakly optional time. Furthermore because $\tau \leq \tau_n$ for all n we know that $\mathcal{F}_\tau^+ \subset \mathcal{F}_{\tau_n}^+$ for all n. On the other hand by Lemma 12.40, if we know that $A \in \cap_n \mathcal{F}_{\tau_n}^+$ then $A \cap \{\tau_n < t\} \in \mathcal{F}_t$ for all n and t. Therefore we can write $A \cap \{\tau < t\} = \cup_n A \cap \{\tau_n < t\} \in \mathcal{F}_t$ which shows that $A \in \mathcal{F}_\tau^+$ by another application of Lemma 12.40. □

TODO: We need the decreasing/intersection version of Jessen.

TODO: Introduce complete right continuous filtration and existence of a cadlag version of martingales.

Lemma 12.43. *Let X_t be a cadlag submartingale on $\mathbb{R}_+$, then for any t and λ we have*

$$\lambda \boldsymbol{P}\{\sup_{s \leq t} X_s \geq \lambda\} \leq \boldsymbol{E}\left[X_t; \sup_{s \leq t} X_s \geq \lambda\right] \leq \boldsymbol{E}\left[X_t^+\right]$$

Furthermore if X_t is non-negative then for any $p > 1$ we have

$$\boldsymbol{E}\left[\sup_{s \leq t} X_s\right] \leq \frac{p}{p-1} \|X_t\|_p$$

Proof. Claim 1: For any $\omega \in \Omega$ such that $X_t(\omega)$ is cadlag, we have

$$\sup_{\substack{s \leq t \\ s \in \mathbb{Q} \cup \{t\}}} X_s(\omega) = \sup_{\substack{s \leq t \\ s \in \mathbb{R}}} X_s(\omega)$$

To see this note that given any $\epsilon > 0$ we can find $s \leq t$ with $s \in \mathbb{R}$ such that $X_s(\omega) > \sup_{\substack{s \leq t \\ s \in \mathbb{R}}} X_s(\omega) - \frac{\epsilon}{2}$. By right continuity and density of rationals, we can find $r \in \mathbb{Q} \cup \{t\}$ such that $s \leq r \leq t$ and $|X_r(\omega) - X_s(\omega)| < \frac{\epsilon}{2}$ which by the triangle inequality tells us that $X_r(\omega) > \sup_{\substack{s \leq t \\ s \in \mathbb{R}}} X_s(\omega) - \epsilon$. Therefore

$$\sup_{\substack{s \leq t \\ s \in \mathbb{Q} \cup \{t\}}} X_s(\omega) \geq \sup_{\substack{s \leq t \\ s \in \mathbb{R}}} X_s(\omega) - \epsilon$$

Since $\epsilon > 0$ was arbitrary we can set it to zero to get

$$\sup_{\substack{s \leq t \\ s \in \mathbb{Q} \cup \{t\}}} X_s(\omega) \geq \sup_{\substack{s \leq t \\ s \in \mathbb{R}}} X_s(\omega)$$

The opposite inequality is immediate from the definition of supremum so the claim is verified.

By the Claim 1 and the countable index set maximal inequality (Lemma 12.30) we get the first result. By Claim 1 and the countable index set L^p inequality we get the second result. □

12.3. **Progressive Measurability.** For many applications the notion of an adapted process suffices. However when dealing with continuous time processes there are anomalies that can occur with such processes that are inconvenient and it is best to define a stronger notion of measurability. To understand the issue we're trying to address, note that adaptedness only addresses the behavior of $X_t(\omega)$ as a function of ω for fixed t. If we take the sample path point of view and think of $X_t(\omega)$ as a function of t for fixed ω then there little constraint on how horribly it can behave.

In fact the general definition of a process allows T to be an arbitrary set so it isn't even in scope to talk about an type of regularity of sample paths.

For the special case of processes indexed by $\mathbb{R}_+$ we can discuss measurability, continuity and even differentiability of sample paths. For the moment, there is a very mild restriction that we make.

Definition 12.44. A process X is said to be *progressively measurable* or simply *progressive* if for every t, the restriction of X to the time interval $[0,t]$, $X : \Omega \times [0,t] \to S$ is $\mathcal{F}_t \otimes \mathcal{B}([0,t])$ measurable.

Definition 12.45. The set of *progressively measurable sets* is defined as

$$\mathcal{PM} = \{A \subset \Omega \times \mathbb{R}_+ \mid A \cap \Omega \times [0,t] \in \mathcal{F}_t \otimes \mathcal{B}([0,t]) \text{ for all } t \geq 0\}$$

Lemma 12.46. *The set $\mathcal{PM}$ is a sub σ-algebra of $\mathcal{A} \otimes \mathcal{B}(\mathbb{R}_+)$. A process $X : \Omega \times \mathbb{R}_+ \to S$ is progressive if and only if X is $\mathcal{PM}$-measurable.*

Proof. Since for all $t \geq 0$, $\Omega \times \mathbb{R}_+ \cap \Omega \times [0,t] = \Omega \times [0,t] \in \mathcal{F}_t \otimes \mathcal{B}([0,t])$ we have $\Omega \times \mathbb{R}_+ \in \mathcal{PM}$. Suppose $A \in \mathcal{PM}$ and then note by the elementary set theory equality $B^c \cap C = (B \cap C)^c \cap C$ and the fact that $\mathcal{F}_t \otimes \mathcal{B}([0,t])$ is a σ-algebra

$$A^c \cap \Omega \times [0,t] = (A \cap \Omega \times [0,t])^c \cap \Omega \times [0,t] \in \mathcal{F}_t \otimes \mathcal{B}([0,t])$$

thus showing $\mathcal{PM}$ is closed under set complement. Lastly if we assume that $A_1, A_2, \cdots \in \mathcal{PM}$, then clearly for every $t \geq 0$,

$$(\cap_n A_n) \cap \Omega \times [0,t] = \cap_n (A_n \cap \Omega \times [0,t]) \in \mathcal{F}_t \otimes \mathcal{B}([0,t])$$

so we see that $\mathcal{PM}$ is a σ-algebra.

To see that $\mathcal{PM}$ is a sub σ-algebra of $\mathcal{A} \otimes \mathcal{B}(\mathbb{R}_+)$, if for $A \in \mathcal{PM}$ we define $A_n = A \cap \Omega \times [0,n]$ then by definition of $\mathcal{PM}$ we know $A_n \in \mathcal{F}_n \otimes \mathcal{B}([0,n]) \subset \mathcal{A} \otimes \mathcal{B}(\mathbb{R}_+)$. But we can write $A = \cup_n A_n$ thus showing $A \in \mathcal{A} \otimes \mathcal{B}(\mathbb{R}_+)$.

To see the characterization of progressive processes, assume X is a process and that $A \in \mathcal{S}$ and observe

$$\{X \in A\} \cap \Omega \times [0,t] = \{(\omega, s) \in \Omega \times [0,t] \mid X_s(\omega) \in A\}$$

which shows that X is progressive if and only if it is $\mathcal{PM}$-measurable. □

Example 12.47. The following is an example of a measurable adapted process that is not progressively measurable. Take $\Omega = [0,1]$ and $S = \mathbb{R}$ all supplied with the Borel σ-algebra and Lebesgue measure. Let $A \subset [0,1]$ be non-measurable. Define

$$X_t(\omega) = \begin{cases} t + \omega & \text{for } t \in A \\ -t - \omega & \text{for } t \notin A \end{cases}$$

with filtration defined by $\mathcal{F}_t = \mathcal{B}([0,1])$. (Note that for every $t \geq 0$, $\sigma(X_t) = \mathcal{B}([0,1])$ hence this is the filtration induced by X). It is easy to see that this is a process (i.e. is measurable) since for each fixed t, $X_t : [0,1] \to \mathbb{R}$ is continuous hence measurable. However that $\{(\omega, s) \mid X_s(\omega) \geq 0\} = \Omega \times A$ hence is not measurable thus showing that X is not progressively measurable.

There is the simpler example but the current example also provides an example of the type of anomaly that can occur.

Define a random time

$$\tau(\omega) = \inf\{t \mid 2t \geq |X_t(\omega)|\} = \inf\{t \mid 2t \geq t + \omega\} = \inf\{t \mid t \geq \omega\} = \omega$$

which because $\{\tau \le t\} = [0,t] \in \mathcal{B}([0,1])$ is seen to be an optional time. Because $\mathcal{F}_t = \mathcal{B}([0,1])$ we see that for every Borel measurable A, $A \cap \{\tau \le t\} = A \cap [0,t] \in \mathcal{F}_t$ so we also have $\mathcal{F}_\tau = \mathcal{B}([0,1])$. On the other hand, the stopped process

$$X_\tau(\omega) = \begin{cases} 2\omega & \text{if } \omega \in A \\ -2\omega & \text{if } \omega \notin A \end{cases}$$

and again we see that $\{X_\tau > 0\} = A$ is not $\mathcal{F}_\tau$-measurable.

Note that because sections are measurable (Lemma 3.80) a progressively measurable process is adapted.

13. Notes on Lasso

TODO

14. Concentration Inequalities

Lemma 14.1 (Markov Inequality)**.** *Let ξ be a positive integrable random variable. Then $\boldsymbol{P}\{\xi > t\} \le \frac{E(\xi)}{t}$*

Proof. $E(\xi) \le E(\xi \mathbf{1}_{\{\xi > t\}}) \le E(t\mathbf{1}_{\{\xi > t\}}) = t\mathbf{P}\{\xi > t\}$ □

Lemma 14.2 (Chebeshev's Inequality)**.** *Let ξ be a random variable with finite mean μ and finite variance σ. Then $\boldsymbol{P}\{|\xi - \mu| > t\} \le \frac{\sigma^2}{t^2}$*

Proof. $\mathbf{P}\{|\xi - \mu| > t\} = \mathbf{P}\{(\xi - \mu)^2 > t^2\} \le \frac{\mathbf{E}\left[(\xi-\mu)^2\right]}{t^2} = \frac{\sigma^2}{t^2}$ □

Lemma 14.3 (One Sided Chebeshev's Inequality)**.** *Let ξ be a random variable with finite mean μ and finite variance σ. Then $\boldsymbol{P}\{\xi - \mu > \lambda\} \le \frac{\sigma^2}{\sigma^2 + \lambda^2}$*

Proof. First we assume $\mathbf{E}[\xi] = 0$. We prove a family of inequalities for a real parameter $c > 0$.

$$\begin{aligned} \mathbf{P}\{\xi > \lambda\} &= \mathbf{P}\{\xi + c > \lambda + c\} \\ &\le \mathbf{P}\{(\xi + c)^2 > (\lambda + c)^2\} \qquad \text{because } \lambda + c > 0 \\ &\le \frac{\mathbf{E}\left[\xi^2\right] + c^2}{(\lambda + c)^2} \end{aligned}$$

Now we extract the best estimate by finding the minimum of the right hand side with respect to c. Differentiating we get a vanishing first derivative when $(\lambda^2 + c^2)2c = (\mathbf{E}\left[\xi^2\right] + c^2)2(\lambda + c)$. Divide by $2(\lambda + c)$ and subtract c^2 to get the minimum at $c = \mathbf{E}[\xi]/\lambda > 0$. Plug this value in to get the final estimate.

$$\begin{aligned} \frac{\mathbf{E}\left[\xi^2\right] + (\frac{\mathbf{E}[\xi^2]}{\lambda})^2}{(\lambda + \frac{\mathbf{E}[\xi^2]}{\lambda})^2} &= \frac{\mathbf{E}\left[\xi^2\right](1 + \frac{\mathbf{E}[\xi^2]}{\lambda^2})}{\lambda^2(1 + \frac{\mathbf{E}[\xi^2]}{\lambda^2})^2} \\ &= \frac{\mathbf{E}\left[\xi^2\right]}{\lambda^2 + \mathbf{E}\left[\xi^2\right]} \end{aligned}$$

Now apply the above inequality to the centered random variable $\xi - \mu$ to get the general result. □

Definition 14.4. We say that a random variable ξ is *subgaussian* if and only if there exist constants $c, C > 0$ such that $\mathbf{P}\{|\xi| \ge \lambda\} \le Ce^{-c\lambda^2}$ for all $\lambda > 0$.

TODO: Show that any Gaussian is subgaussian (independent of its mean?).

TODO: Show any bounded (or almost surely bounded) random variable is subgaussian.

Example 14.5. Given the nomenclature it isn't surprising that Gaussian random variables are subgaussian. As it turns out it is useful to analyze the case of a $N(0,\sigma^2)$ random variable separately since it has slightly different behavior than the general $N(\mu,\sigma^2)$ case. Let us assume that ξ is a normal random variable with mean 0 and variance σ^2. We have a standard tail estimate for $\lambda \geq \sigma$

$$\mathbf{P}\{\xi \geq \lambda\} = \frac{1}{\sqrt{2\pi}\sigma}\int_\lambda^\infty e^{-x^2/2\sigma^2}\,dx \leq \frac{1}{\sqrt{2\pi}\sigma}\int_\lambda^\infty \frac{x}{\sigma}e^{-x^2/2\sigma^2}\,dx = \frac{1}{\sqrt{2\pi}}e^{-\lambda^2/2\sigma^2}$$

The $0 \leq \lambda \leq \sigma$ case can easily be handled with a constant multiplier but we can actually find the constant that gives a tight bound. Note that $\frac{1}{\sqrt{2\pi}\sigma}\int_0^\infty e^{-x^2/2\sigma^2} = \frac{1}{2}$ so we can't do any better than $\mathbf{P}\{\xi \geq \lambda\} \leq \frac{1}{2}e^{-\lambda^2/2\sigma^2}$; in fact this bound works for all $\lambda \geq 0$. We've already shown this for $\lambda \geq 1$ and $\lambda = 0$. To show the bound on $[0,1]$ we calculate the derivative

$$\frac{d}{d\lambda}\left(\frac{1}{2}e^{-\lambda^2/2\sigma^2} - \frac{1}{\sqrt{2\pi}\sigma}\int_\lambda^\infty e^{-x^2/2\sigma^2}\,dx\right) = \left(-\frac{\lambda}{2\sigma^2} + \frac{1}{\sqrt{2\pi}\sigma}\right)e^{-\lambda^2/2}$$

from which we conclude there is a unique maximum of the function at $\lambda = \sigma\sqrt{\frac{2}{\pi}} \in (0,\sigma)$. We have already validated that the function is nonnegative at the endpoints of $[0,\sigma]$ so it must be nonnegative on the entire interval. Now by symmetry of ξ, the calculation also shows that $\mathbf{P}\{\xi \leq -\lambda\} \leq \frac{1}{2}e^{-\lambda^2/2\sigma^2}$ and therefore $\mathbf{P}\{|\xi| \geq \lambda\} \leq e^{-\lambda^2/2\sigma^2}$.

Now for a general $N(\mu,\sigma)$ normal random variable ξ we have by change of variables

$$\mathbf{P}\{\xi \geq \lambda\} = \frac{1}{\sqrt{2\pi}\sigma}\int_\lambda^\infty e^{-(x-\mu)^2/2\sigma^2}\,dx = \frac{1}{\sqrt{2\pi}}\int_{(\lambda-\mu)/\sigma}^\infty e^{-x^2/2}\,dx \leq \frac{1}{\sqrt{2\pi}}e^{-(\lambda-\mu)^2/2\sigma^2}$$

TODO: Finish

Lemma 14.6. *Let $\{\xi_i\}_{i=1}^m$ be jointly independent subgaussian random variables. Then $\boldsymbol{E}\left[e^{\sum_{i=1}^m \xi}\right] = \prod_{i=1}^m \boldsymbol{E}\left[e^{\xi_i}\right]$.*

Proof. First show that for a subgaussian ξ, we have by dominated convergence the Taylor expansion

$$\mathbf{E}\left[e^{t\xi}\right] = 1 + \sum_{k=1}^\infty \frac{t^k}{k!}\mathbf{E}\left[\xi^k\right]$$

The proof of this fact is to exhibit an integrable function that dominates the sequence of partial sums $1 + \sum_{k=1}^n \frac{t^k\xi^k}{k!}$. This is obvious if ξ is almost surely bounded but it's not obvious to me that this should be true for a subgaussian ξ. TODO: Perhaps we need to use uniform integrability or something like that in the subgaussian/subexponential case.

In any case, assuming the validity of the above identity for each ξ, we turn to the case of the sum. □

Lemma 14.7. *ξ is subgaussian if and only if there exists C such that $\boldsymbol{E}\left[e^{t\xi}\right] \leq Ce^{Ct^2}$ and if and only if there exists C such that $\boldsymbol{E}\left[|\xi|^k\right] \leq (Ck)^{\frac{k}{2}}$ for all $t \in \mathbb{R}$.*

Proof. Suppose ξ is subgaussian and calculate:

$$\begin{aligned}
\mathbf{E}\left[e^{t\xi}\right] &= \int_0^\infty \mathbf{P}\{e^{t\xi} \geq \lambda\} d\lambda \\
&= \int_{-\infty}^\infty \mathbf{P}\{e^{t\xi} \geq e^{t\eta}\} t e^{t\eta} d\eta \\
&= \int_{-\infty}^\infty \mathbf{P}\{\xi \geq \eta\} t e^{t\eta} d\eta \\
&\leq \int_{-\infty}^\infty C t e^{t\eta - c\eta^2} d\eta \\
&= C t e^{\frac{t^2}{4c}} \int_{-\infty}^\infty e^{-\left(\sqrt{c}\eta - \frac{t}{2\sqrt{c}}\right)^2} d\eta \\
&= C' t e^{\frac{t^2}{4c}} \\
&\leq C' e^{\frac{5ct^2}{4c}}
\end{aligned}$$

Now assume that we have $\mathbf{E}\left[e^{t\xi}\right] \leq Ce^{Ct^2}$ for all t. Pick an arbitrary $t > 0$ to be chosen later and proceed by using first order moment method:

$$\begin{aligned}
\mathbf{P}\{\xi \geq \lambda\} &= \mathbf{P}\{e^{t\xi} \geq e^{t\lambda}\} \\
&\leq \frac{\mathbf{E}\left[e^{t\xi}\right]}{e^{t\lambda}} \\
&\leq Ce^{Ct^2 - t\lambda}
\end{aligned}$$

Now we pick t to minimize the upper bound derived above; simple calculus shows this occcurs at $t = \frac{\lambda}{2C}$. Subtituting yields the bound

$$\mathbf{P}\{\xi \geq \lambda\} \leq Ce^{-\frac{\lambda^2}{4C}}$$

For the other tail, we note that our assumption holds equally well for $-\xi$. Thus we can use the same method to bound

$$\mathbf{P}\{\xi \leq -\lambda\} = \mathbf{P}\{-\xi \geq \lambda\} \leq Ce^{-\frac{\lambda^2}{4C}}$$

therefore taking the union bound we get

$$\mathbf{P}\{|\xi| \geq \lambda\} \leq 2Ce^{-\frac{\lambda^2}{4C}}$$

Now consider absolute moments of subgaussian variables. We can assume that $\xi \geq 0$ and calculate as before:

$$\begin{aligned}
\mathbf{E}\left[\xi^k\right] &= \int_0^\infty \mathbf{P}\{\xi^k \geq x\}dx \\
&= k\int_0^\infty \mathbf{P}\{\xi^k \geq y^k\}y^{k-1}dy \\
&= kC\int_0^\infty y^{k-1}e^{-cy^2}dy \\
&= kC\frac{c^{k-3}}{2}\int_0^\infty x^{\frac{k}{2}-1}e^{-x}dx \\
&= kC\frac{c^{k-3}}{2}\Gamma\left(\frac{k}{2}\right) \\
&\leq kC\frac{c^{k-3}}{2}\left(\frac{k}{2}\right)^{\frac{k}{2}}
\end{aligned}$$

To go the other direction, assume $\mathbf{E}\left[|\xi|^k\right] \leq (Ck)^{\frac{k}{2}}$ and pick a constant $0 < c < \frac{e}{2C}$

$$\begin{aligned}
\mathbf{E}\left[e^{K\xi^2}\right] &= 1 + \sum_{k=1}^\infty \frac{t^k\mathbf{E}\left[\xi^{2k}\right]}{k!} \\
&\leq 1 + \sum_{k=1}^\infty \frac{(2tCk)^k}{k!} \\
&\leq 1 + \sum_{k=1}^\infty \left(\frac{2tC}{e}\right)^k < \infty
\end{aligned}$$

Now use the elementatry bound $ab \leq \frac{(a^2+b^2)}{2}$ so see

$$\mathbf{E}\left[e^{t\xi}\right] \leq$$

□

The definition of subgaussian random variables differs in a minor way from another in common use in the literature. In particular, in some descriptions a random variable ξ is called subgaussian if and only if $\mathbf{E}\left[e^{t\xi}\right] \leq e^{\frac{c^2t^2}{2}}$ for all $t \in \mathbb{R}$. The important difference here compared with the characterization in Lemma 14.7 is that the constant on the right hand side is 1. With this defintion, we must add the hypothesis $\mathbf{E}[\xi] = 0$ to get equivalence with the other definition.

Lemma 14.8. *Suppose ξ is a random variable such that there exists $c > 0$ for which*

$$\boldsymbol{E}\left[e^{t\xi}\right] \leq e^{\frac{c^2t^2}{2}} \text{ for all } t \in \mathbb{R}$$

then $\boldsymbol{E}[\xi] = 0$ and $\boldsymbol{E}\left[\xi^2\right] \leq c^2$.

Proof. By Dominated Convergence and the hypothesis we get

$$\sum_{n=0}^{\infty} \frac{t^n}{n!} \mathbf{E}\left[\xi^n\right] = \mathbf{E}\left[e^{t\xi}\right] \leq e^{\frac{c^2t^2}{2}} = \sum_{n=0}^{\infty} \frac{c^{2n}}{2^n n!} t^{2n}$$

so in particular by taking only terms up to order t^2 and using the fact that the constant term in on both sides is 1, we have

$$t\mathbf{E}\left[\xi\right] + \frac{t^2}{2}\mathbf{E}\left[\xi^2\right] = \frac{c^2t^2}{2} + o(t^2) \text{ as } t \to 0$$

If we divide both sides by $t > 0$ and take the limit as $t \to 0^+$ then we get $\mathbf{E}\left[\xi\right] \leq 0$. If we divide by $t < 0$ and take the limit as $t \to 0^-$ then we get $\mathbf{E}\left[\xi\right] \geq 0$. Thus we can conclude $\mathbf{E}\left[\xi\right] = 0$. If we plug that in and divide by t^2 and take the limit as $t \to 0$ then see $\mathbf{E}\left[\xi^2\right] \leq c^2$. □

Note that the argument in the proof above doesn't even get off the ground unless the constant of the bounding exponential is assumed to be 1.

The following lemma is useful for the second moment method for deriving tail bounds.

Lemma 14.9. *Let $\{\xi_i\}_{i=1}^m$ be pairwise independent random variables and c_i be scalars. Then $\boldsymbol{Var}(\sum_{i=1}^m c_i\xi) = \sum_{i=1}^m |c_i|^2 \boldsymbol{Var}(\xi_i)$.*

Proof. TODO □

Lemma 14.10 (Bennett's Inequality)**.** *Let $\{\xi_i\}_{i=1}^m$ be independent random variables with means μ_i and variances σ_i. Set $\Sigma^2 = \sum_{i=1}^m \sigma_i^2$. If for every i, $|\xi_i - \mu_i| \leq M$ almost everywhere then for every $\lambda > 0$ we have*

$$\boldsymbol{P}\{\sum_{i=1}^m [\xi_i - \mu_i] > \lambda\} \leq e^{-\frac{\lambda}{M}\{(1+\frac{\Sigma^2}{M\lambda})\log(1+\frac{M\lambda}{\Sigma^2})-1\}}$$

Proof. First it is easy to see that by subtracting means we may assume that $\mu_i = 0$. Then we have $\sigma_i = \mathbf{E}\left[\xi_i^2\right]$. We use the exponential moment method. We show a family of inequalities depending on a real parameter $c > 0$ which we will pick later. First we have

$$\begin{aligned}
\mathbf{P}\{\sum_{i=1}^m \xi_i > \lambda\} &= \mathbf{P}\{c\sum_{i=1}^m \xi_i > c\lambda\} && \text{since } c > 0. \\
&= \mathbf{P}\{e^{c\sum_{i=1}^m \xi_i} > e^{c\lambda}\} && \text{since } e^x \text{ is increasing} \\
&\leq e^{-c\lambda}\mathbf{E}\left[e^{c\sum_{i=1}^m \xi_i}\right] && \text{by Markov's Inequality(14.1)} \\
&= e^{-c\lambda}\prod_{i=1}^m \mathbf{E}\left[e^{c\xi_i}\right] && \text{by independence and boundedness. TODO: do we really need boundedne}
\end{aligned}$$

Now we consider an individual term $\mathbf{E}\left[e^{c\xi_i}\right]$ for an almost surely bounded ξ_i with zero mean.

$$\begin{aligned}
\mathbf{E}\left[e^{c\xi_i}\right] &= \mathbf{E}\left[\sum_{k=0}^{\infty}\frac{c^k\xi_i^k}{k!}\right] = \sum_{k=0}^{\infty}\frac{c^k}{k!}\mathbf{E}\left[\xi_i^k\right] && \text{by dominated convergence}\\
&= 1+\sum_{k=2}^{\infty}\frac{c^k}{k!}\mathbf{E}\left[\xi_i^k\right] && \text{by mean zero}\\
&\le 1+\sum_{k=2}^{\infty}\frac{c^k M^{k-2}\sigma_i^2}{k!} && \text{by boundedness and definition of variance}\\
&\le e^{\sum_{k=2}^{\infty}\frac{c^k M^{k-2}\sigma_i^2}{k!}} && \text{since } 1+x\le e^x (5.1)\\
&= e^{\frac{(e^{cM}-1-cM)\sigma_i^2}{M^2}}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbf{P}\{\sum_{i=1}^{m}\xi_i > \lambda\} &\le e^{-c\lambda}\prod_{i=1}^{m} e^{\frac{(e^{cM}-1-cM)\sigma_i^2}{M^2}}\\
&= e^{\frac{(e^{cM}-1-cM)\Sigma^2}{M^2}}
\end{aligned}$$

Now we pick $c>0$ to minimize the bound above ($e^{cM}-1=\frac{M\lambda}{\Sigma^2}$ or equivalently $c=\frac{1}{M}\ln(1+\frac{M\lambda}{\Sigma^2})$). Substituting yields the final bound

$$\begin{aligned}
\mathbf{P}\{\sum_{i=1}^{m}\xi_i > \lambda\} &\le e^{-(\lambda+\frac{\Sigma^2}{M})\frac{1}{M}\ln(1+\frac{M\lambda}{\Sigma^2})+\frac{\lambda}{M}}\\
&= e^{-\frac{\lambda}{M}\{(1+\frac{\Sigma^2}{\lambda M})\ln(1+\frac{M\lambda}{\Sigma^2})-1\}}
\end{aligned}$$

□

Lemma 14.11 (Bernstein's or Chernoff's Inequality)**.** *Let $\{\xi_i\}_{i=1}^m$ be independent random variables with means μ_i and variances σ_i. Set $\Sigma^2=\sum_{i=1}^m\sigma_i^2$. If for every i, $|\xi_i-\mu_i|\le M$ almost everywhere then for every $\lambda>0$ we have*

$$\boldsymbol{P}\{\sum_{i=1}^{m}[\xi_i-\mu_i]>\lambda\} \le e^{-\{\frac{\lambda^2}{2(\Sigma^2+\frac{1}{3}M\lambda)}\}}$$

Proof. TODO □

The next inequality has a pleasing form because the resulting bound is of the form of a Gaussian random variable. Such bounds are interesting enough that they warrant the following definition.

Definition 14.12. Let ξ be a real valued random variable with mean μ. We say that ξ has a *subgaussian upper tail* if there exists a constants $C>0$ and $c>0$ such that for all $\lambda>0$,

$$\mathbf{P}\{[\xi-\mu]>\lambda\}\le Ce^{-c\lambda^2}.$$

We say that ξ has a *subgaussian tail up to* λ_0 if the above bound holds for $\lambda<\lambda_0$. We say that ξ has a *subgaussian tail* if both ξ and $-\xi$ have subgaussian upper tails (or equivalently if $|\xi|$ has a subgaussian tail.

The boundedness assumption on the individual random variables in the above sums can be relaxed to an assumption that the individual random variables has subgaussian tails. Moreover, one can generalize the sum of random variables to an arbitrary linear combination of random variables on the unit sphere.

Lemma 14.13. *Let $\{\xi_i\}_{i=1}^m$ be independent random variables with $E[\xi_i] = 0$ and $E[\xi_i^2] = 1$ and uniform subgaussian tails. Let $\{\alpha_i\}_{i=1}^m$ be real coefficients satisfying $\sum_{i=1}^m \alpha_i^2 = 1$. The then random variable $\eta = \sum_{i=1}^m \alpha_i \xi_i$ has $E[\eta] = 0$, $E[\eta^2] = 1$ and a subgaussian tail.*

Proof. TODO □

Lemma 14.14 (Exercise 7 Lugosi)**.** *Let $\{\xi_i\}_{i=1}^n$ be independent random variables with values in $[0,1]$. Let $S_n = \sum_{i=1}^n \xi_i$ and let $\mu = \boldsymbol{E}[S_n]$. Show that for any $\lambda \geq \mu$,*

$$\boldsymbol{P}\{S_n \geq \lambda\} \leq \left(\frac{\mu}{\lambda}\right)^{\lambda} \left(\frac{n-\mu}{n-\lambda}\right)^{n-\lambda}.$$

Proof. Use Chernoff bounding. Looking at the solution, we can pattern match that we may want to use the convexity of e^x since the solution seems to reference the endpoints of the interval $[0, n]$; indeed that is the way to proceed. TODO: convert the argument below for $n = 1$ to cover general n. To estimate $\mathbf{E}\left[e^{s\xi_i}\right]$ we first use convexity of $e^s x$ on the interval $x \in [0,1]$,

$$e^{sx} \leq xe^s + (1-x)$$

Substituting ξ_i and taking expectations we get

$$\mathbf{E}\left[e^{s\xi_i}\right] \leq \mu_i e^s + (1-\mu_i).$$

So now we minimize the Chernoff bound by using elementary calculus

$$\frac{d}{ds}\mu_i e^{s(1-\lambda)} + (1-\mu_i)e^{-s\lambda} = \mu_i(1-\lambda)e^{s(1-\lambda)} + \lambda(1-\mu_i)e^{-s\lambda}$$

which equals 0 when $s = \ln\left(\frac{\lambda(1-\mu_i)}{\mu_i(1-\lambda)}\right)$. This value is positive when $\lambda \geq \mu$. Back-substituing this value and doing some algebra shows

$$e^{-s\lambda}\mathbf{E}\left[e^{s\xi_i}\right] \leq \left(\frac{\mu_i}{\lambda}\right)^{\lambda} \left(\frac{1-\mu_i}{1-\lambda}\right)^{1-\lambda}$$

Note also an argument for a related estimate (Exercise 8) that uses bounds similar to those in Bennett can be made as follows. Since $\xi_i \in [0,1]$, we have that $\xi_i^k \leq \xi_i$. With this observation,

$$\begin{aligned}
\mathbf{E}\left[e^{s\xi_i}\right] &= 1 + \sum_{k=1}^{\infty} \frac{s^k \mathbf{E}\left[\xi_i^k\right]}{k!} \\
&\leq 1 + \sum_{k=1}^{\infty} \frac{s^k \mu_i}{k!} \\
&= 1 + \mu_i(e^s - 1) \\
&\leq e^{\mu_i(e^s-1)}
\end{aligned}$$

Now we select s to minimize the Chernoff bound $e^{\mu_i(e^s-1)-s\lambda}$ which simple calculus shows happens at $s = \ln\left(\frac{\lambda}{\mu_i}\right)$; the location of the minimum being positive precisely when $\lambda \geq \mu_i$. Backsubstituting yields a bound $\left(\frac{\mu_i}{\lambda}\right)^{\lambda} e^{\lambda-\mu_i}$. □

15. Likelihood Theory

TODO:

(i) Definition of Likelihood function
(ii) Definition of Maximum Likelihood estimate
(iii) Fisher information: regularity conditions (FI and Le Cam), score function and information matrix; information matrix as Riemannian metric on manifold of parameters
(iv) Cramer-Rao Lower Bound
(v) Asymptotic distribution/Asymptotic Normality : Delta Method and Second Order Delta Method
(vi) Asymptotic consistency of MLEs
(vii) Asymptotic efficiency of MLEs
(viii) Hypothesis testing with MLE: Likelihood Ratio Tests Wilks Theorem(Schervish Thm 7.125, van der Vaart 16.9), Wald Tests and Score Tests
(ix) Problems with boundaries lack of regularity
(x) M-estimators
(xi) Observed information matrix...

As a quick motivation for where maximum likelihood estimation comes from, consider the following measure of distance between two probability distributions that was motivated by information theory.

Definition 15.1. Suppose μ and ν such that $\mu << \nu$. The the *Kullback-Liebler divergence* or *relative entropy* of μ and ν is defined as

$$D(\mu \mid\mid \nu) = \mathbf{E}_\mu[\log \frac{d\mu}{d\nu}]$$

If μ is not absolutely continuous with respect to ν then by convention $D(\mu \mid\mid \nu) = \infty$.

Example 15.2. Suppose μ and ν are probability measures that are both absolutely continuous with respect to a third measure λ and furthermore $\mu << \nu$. Then we may write $\mu = f \cdot \lambda$ and $\nu = g \cdot \lambda$ where we assume that λ- almost surely $g = 0$ implies $f = 0$ (otherwise the event $A = \{g = 0; f > 0\}$ satisfies $\nu(A) = 0$ but $\mu(A) \neq 0$). In this case we can make sense of the ratio $\frac{f}{g}$ if we agree that $\frac{0}{0} = 0$ and then $\frac{d\mu}{d\nu} = \frac{f}{g}$.

In this case we get the formula

$$D(\mu \mid\mid \nu) = \int \log(\frac{f}{g}) f \, d\lambda$$

that the user may have encountered before.

Example 15.3. One interpretation of relative entropy is that is the number of bits of information that one gains updating ones that belief that a probability distribution is ν to a belief that a probability distribtuion is μ. The following simple example illustrates the point. In what follows we interpret log to be the base 2 logarithm as opposed to the standard assumption that it represents the natural logarithm. Suppose you believe that a coin is fair. In this case you believe that the distribution is $\nu(H) = \nu(T) = 1/2$. If someone tells you that the coin is a trick coin that only lands with heads up then you change belief to $\mu(H) = 1$ and

$\mu(T) = 0$. It is easy to see that $\mu << \nu$ and using the formula for relative entropy in terms of densities in the previous example we compute

$$D(\mu \mid\mid \nu) = \log(\frac{1}{1/2}) \cdot 1 + \log(\frac{0}{1/2}) \cdot 0 = \log 2 = 1$$

Thus one has gained 1 bit of information; which is intuitively correct because on updating one's view of the probability distribution one has learned the outcome of a single binary trial.

It is also instructive to consider the example with the roles of μ and ν reversed. In this case $\mu(T) = 0$ but $\nu(T) \neq 0$ hence ν is not absolutely continuous with respect μ and therefore we have agreed that the relative entropy is infinite. The convention is corroborated by the heuristic calculation

$$D(\nu \mid\mid \mu) = \log(\frac{1/2}{1}) \cdot \frac{1}{2} + \log(\frac{1/2}{0}) \cdot \frac{1}{2} = \infty$$

The intuition here is that in going from μ to ν we are learning that something that was formerly thought to be impossible is in fact possible and that the information gained from this is infinitely large. Along the lines of this example one will often hear the relative entropy referred to as *information gain* : particularly in the machine learning literature.

Lemma 15.4 (Gibbs Inequality)**.** *For all probability distributions μ and ν, $D(\mu \mid\mid \nu) \geq 0$ with equality if and only if μ and ν agree except on a set of measure zero with respect to ν.*

Proof. It suffices to handle the case in which $\nu << \mu$. In this case we can simply use the strict convexity of $x \log x$ and apply Jensen's inequality and the definition of the Radon-Nikodym derivative to see

$$D(\mu \mid\mid \nu) = \mathbf{E}_\mu[\log \frac{d\mu}{d\nu}] = \mathbf{E}_\nu[\frac{d\mu}{d\nu} \log \frac{d\mu}{d\nu}] \geq \mathbf{E}_\nu[\frac{d\mu}{d\nu}] \log \mathbf{E}_\nu[\frac{d\mu}{d\mu}] = \mathbf{E}_\mu[1] \log \mathbf{E}_\mu[1] = 0$$

By strict convexity of $x \log x$, we have equality if and only if $\frac{d\mu}{d\nu}$ is almost surely (with respect to ν) a constant. This constant must be 1 because μ and ν are both probability measures. □

Example 15.5. Continuing the previous example we specialize to case in which we consider a family of densities indexed by a set Θ. Specifically for each $\theta \in \Theta$, we suppose we have a density $f(x \mid \theta)$ with respect to a base measure λ. The problems of (parametric) statistical estimation generally start with such an assumption and and assume there is distinguished *true* value θ_0 from among the elements of the set Θ. Lemma 15.4 suggests a potential path. We know from the previous example that

$$D(\theta_0 \mid\mid \theta) = \mathbf{E}_{\theta_0}[\log(\frac{f(x \mid \theta_0)}{f(x \mid \theta)})] = \mathbf{E}_{\theta_0}[\log(f(x \mid \theta_0))] - \mathbf{E}_{\theta_0}[\log(f(x \mid \theta))] \geq 0$$

with equality if an only $f(x \mid \theta_0)$ and $f(x \mid \theta)$ give the same measure (which we generally assume to imply that $\theta_0 = \theta$; a condition referred to as *identifiability*). So this means that $\mathbf{E}_{\theta_0}[\log(f(x \mid \theta))]$ has a unique maximum at the value θ_0. Now this isn't of much use directly since it assume knowledge of the density $f(x \mid \theta_0)$ in order to compute the expectations, but it suggests that we should consider using an approximation of the measure defined by the density such as one defined by sampling and consider contexts in which we maximize the function $f(x \mid \theta)$ considered

as a function of θ. This insight leads to the method of maximum likelihood which we shall study in some detail in the following chapter.

Now we apply this idea in the context of parametric estimation. If we suppose that we are given a parametric family of densities $f(x;\theta)$ relative to some measure ν.

TODO: To be continued...

15.1. **The Delta Method.**

Definition 15.6. Given a metric space (S,d) and arbitrary index set A, a set of random elements ξ_α in S with $\alpha \in A$ is said to be *tight* if for every $\epsilon > 0$ there exists a compact set $K \subset S$ such that $\sup_\alpha \mathbf{P}\{\xi_\alpha \notin K\} < \epsilon$. In the case in which ξ_α are random vectors in some $\mathbb{R}^n$ it is also common to that a tight set of random vectors is *bounded in probability*.

Just as with convergence in distribution, note that tightness is really a property of the law of the random elements ξ_α. We will eventually see that tightness is a type of sequential compactness; if one goes a bit farther than we intend to go, one can in fact show that there is a metric on the space of measures (the Levy-Prohorov metric which metrizes convergence in distribution) and that tight sets are compact sets of measures in the corresponding metric space (are all compact sets tight???).

The first thing that we shall see about tightness is the fact that sequence that converge in distribution are tight.

Lemma 15.7. *Suppose $\xi_n \xrightarrow{d} \xi$ with $\xi, \xi_1, \xi_2, \ldots$ random vectors, then ξ_n is a tight sequence.*

Proof. TODO: Can we use Portmanteau and clean up the argument by making the continuous approximation unnecessary? Answer is certainly yes but it's not clear how much simpler it makes the argument.

Suppose we are given an $\epsilon > 0$. First since ξ is almost surely finite, continuity of measure shows that $\lim_{M\to\infty} \mathbf{P}\{|\xi| > M\} = 0$ and therefore we can find $M_1 > 0$ such $\mathbf{P}\{|\xi| > M_1\} < \frac{\epsilon}{2}$. Now pick an arbitrary $M_2 > M_1$ and let f be a bounded continuous function such that $\mathbf{1}_{|x|>M_2} \leq f \leq \mathbf{1}_{|x|>M_1}$. Then we have

$$\mathbf{P}\{|\xi_n| > M_2\} \leq \mathbf{E}\left[f(\xi_n)\right]$$

and

$$\mathbf{E}\left[f(\xi)\right] \leq \mathbf{P}\{|\xi| > M_1\} \leq \frac{\epsilon}{2}$$

but also we can find $N > 0$ such that $|\mathbf{E}\left[f(\xi_n)\right] - \mathbf{E}\left[f(\xi)\right]| < \frac{\epsilon}{2}$ for all $n \geq N$. Putting the pieces together we have for all $n \geq N$,

$$\mathbf{P}\{|\xi_n| > M_2\} \leq \mathbf{E}\left[f(\xi_n)\right] \leq \mathbf{E}\left[f(\xi)\right] + |\mathbf{E}\left[f(\xi_n)\right] - \mathbf{E}\left[f(\xi)\right]| < \epsilon$$

Now for each $0 \leq n \leq N$, we can find M_n' such that $\mathbf{P}\{|\xi_n| > M_n'\} < \epsilon$, so if we take $M = \max(M_2, M_1', \ldots, M_{n-1}')$ then we get $\sup_n \mathbf{P}\{|\xi_n| > M\} < \epsilon$ and tightness is shown. □

Lemma 15.8. *Suppose r_n is a sequence of real numbers such that $\lim_{n\to\infty} |r_n| = \infty$ and $\eta, \xi, \xi_1, \xi_2, \ldots$ is a sequence of random vectors such that $r_n(\xi_n - \xi) \xrightarrow{d} \eta$. Then $\xi_n \xrightarrow{P} \xi$.*

Proof. The proof only relies on the fact that $r_n(\xi_n - \xi)$ is a tight sequence (Lemma 15.7). Suppose we are given $\epsilon, \delta > 0$. By tightness, we can pick $M > 0$ such that

$$\sup_n \mathbf{P}\{|r_n(\xi_n - \xi)| > M\} = \sup_n \mathbf{P}\{|\xi_n - \xi| > \frac{M}{|r_n|}\} < \delta$$

Because $\lim_n |r_n| = \infty$ we pick $N > 0$ such that $\frac{M}{|r_n|} \leq \epsilon$ for $n \geq N$. Then

$$\mathbf{P}\{|r_n(\xi_n - \xi)| > \epsilon\} \leq \sup_n \mathbf{P}\{|\xi_n - \xi| > \frac{M}{|r_n|}\} < \delta$$

for $n \geq N$ and we have show $\xi_n \xrightarrow{P} \xi$. □

In this result we have restricted ourselves to random vectors in $\mathbb{R}^n$ because it is an important special case (especially in parametric statistics) and because it is a trivial matter to show that all random vectors are tight. Generalization to arbitrary metric spaces is subtle because it is no longer the case that an arbitrary random element is tight. One can repair the argument above by adding the assumption that the elements of the sequence are tight random elements or one can explore what conditions on a metric space guarantee that all random elements are tight. Though we don't go into it at the moment, it turns out separability and completeness (i.e. Polishness) are sufficient to guarantee tightness of arbitrary random elements and there is also a more subtle necessary and sufficient condition that has been identified (universal measurability see Dudley's RAP).

Part of the importance of tightness is lies in its role as a compactness property (that is to say the fact that it implies weak convergence of a subsequence). On the other hand, in some cases one uses only the boundedness aspect. This is particularly true in asymptotic statistics. TODO: Introduce the $O_P(r_n)$ and $o_p(r_n)$ notation.

Lemma 15.9. *Let $\xi_1, \xi_2, \ldots$ and $\eta_1, \eta_2, \ldots$ be sequences of random vectors.*

(i) *If $\xi_n \xrightarrow{P} 0$ then ξ_n is tight. ($o_p(1) = O_P(1)$).*
(ii) *If $\xi_n \xrightarrow{P} 0$ and $\eta_n \xrightarrow{P} 0$ then $\xi_n + \eta_n \xrightarrow{P} 0$. ($o_P(1) + o_P(1) = o_P(1)$).*
(iii) *If ξ_n is tight and $\eta_n \xrightarrow{P} 0$ then $\xi_n + \eta_n$ is tight. ($O_P(1) + o_P(1) = O_P(1)$).*
(iv) *If ξ_n is tight and $\eta_n \xrightarrow{P} 0$ then $\xi_n * \eta_n \xrightarrow{P} 0$ (this is true for many kinds of multiplication; scalar multiplication, dot product, matrix multiplication). ($O_P(1)o_P(1) = o_P(1)$).*
(v) *If η_n is tight sequence of random variables and $\xi_n \eta_n \xrightarrow{P} 0$ then $\xi_n \xrightarrow{P} 0$. ($o_P(O_P(1)) = o_P(1)$).*

Proof. To prove (i) simply note that $\xi_n \xrightarrow{P} 0$ implies $\xi_n \xrightarrow{d} 0$ (Lemma 8.28) the therefore we know ξ_n is tight by Lemma 15.7.

The statement of (ii) is a corollary to the Continuous Mapping Theorem (Corollary 8.13).

TODO: Finish... □

Here is a slightly more involved fact that we shall use in the sequel.

Lemma 15.10. *Let Ψ_n be a sequence of random matrices such that $\Psi_n \xrightarrow{P} \Psi$ with Ψ almost surely equal to a constant nonsingular matrix. Suppose ξ_n is a sequence of random vectors such that $\Psi_n \xi_n$ is tight, then ξ_n is tight.*

Proof. Recall that because convergence in probability only depends on the underlying topology induced by a metric (Corollary 8.10) and that all norms on a finite dimensional vector space are equivalent; this means that we are free to choose the operator norm when dealing with the convergence of the matrices Ψ_n.

We remind the reader of some basic facts about the operator norm. In any normed vector space of linear operators with the operator norm we have Neumann series for inverting perturbations of the identity operator. Specifically for any A with $\|A\| < 1$, we have

$$\begin{aligned}
(1-A)^{-1} &= \sum_{n=0}^{\infty} A^n && \text{converges absolutely} \\
\|(1-A)^{-1}\| &\leq \sum_{n=0}^{\infty} \|A^n\| \leq \sum_{n=0}^{\infty} \|A\|^n = (1-\|A\|)^{-1} \\
(1-A)(1-A)^{-1} &= \sum_{n=0}^{\infty} A^n - \sum_{n=1}^{\infty} A^n = 1 \\
(1-A)^{-1}(1-A) &= \sum_{n=0}^{\infty} A^n - \sum_{n=1}^{\infty} A^n = 1
\end{aligned}$$

which shows that $(1-A)$ is invertible with inverse $(1-A)^{-1}$ defined by the Neumann series. We now extend this argument to show there is an open neighborhood of any invertible operator in the space of invertible operators. Suppose T is invertible and let $\|T-A\| < \frac{1}{\|T^{-1}\|}$. Then we can write $T - A = T(1 - T^{-1}A)$ where $\|T^{-1}A\| \leq \|T^{-1}\|\|A\| < 1$ so that $(1-T^{-1}A)$ is invertible. This shows $T-A$ is product of invertible operators hence is itself invertible. Moreover we have the norm bound

$$\|(T-A)^{-1}\| \leq \|T\|\|(1-T^{-1}A)^{-1}\| \leq \frac{\|T\|}{1-\|T^{-1}A\|} \leq \frac{\|T\|}{1-\|T^{-1}\|\|A\|}$$

With that little piece of operator theory out of the way we can return statistics proper. We have assumed $\Psi_n \xrightarrow{P} \Psi$ with Ψ an invertible a.s. constant matrix. Pick $\delta > 0$ and $0 < \epsilon < \frac{1}{2\|\Psi^{-1}\|}$, then we know that there exists an $N > 0$ such that $\mathbf{P}\{\|\Psi_n - \Psi\| \leq \epsilon\} \geq 1 - \frac{\delta}{2}$ for all $n > N$. By the preceeding discussion we know that whenever $\|\Psi_n - \Psi\| \leq \epsilon$, Ψ_n is invertible and $\|\Psi_n^{-1}\| < 2\|\Psi^{-1}\|$. By tightness of $\Psi_n \xi_n$ we can find $M > 0$ such that

$$\sup_n \mathbf{P}\{\|\Psi_n \xi_n\| > M\} < \frac{\delta}{2}$$

Therefore by applying the inverse of Ψ_n and using its operator norm bound we get

$$\sup_{n>N} \mathbf{P}\{\|\xi_n\| > 2M\|\Psi^{-1}\|\} < \delta$$

Because random vectors in $\mathbb{R}^n$ are tight, we know that there is an M' such that $\mathbf{P}\{\|\xi_n\| > M'\} < \delta$ for all $0 < n \leq N$ and therefore ξ_n is tight. □

Definition 15.11. Given an open set $U \subset \mathbb{R}^m$ and function $\phi : U \to \mathbb{R}^n$ we say that ϕ is *Frechet differentiable* at a point $x \in U$ if there is a linear map $A : \mathbb{R}^m \to \mathbb{R}^n$

such that for every sequence $h_n \in \mathbb{R}^m$ such that $\lim_{n\to\infty} |h_n| = 0$ we have

$$\lim_{n\to\infty} \frac{\phi(x+h_n) - \phi(x) - Ah_n}{|h_n|} = 0$$

The linear map A is called the *Frechet derivative* of ϕ at x is usually written $D\phi(x)$.

Theorem 15.12 (Delta Method)**.** *Let $\phi : D \subset \mathbb{R}^k \to \mathbb{R}^m$ be Frechet differentiable at $\theta \in D$. Let $\xi, \xi_1, \xi_2, \ldots$ be random vectors with values in D and r_n be a sequence of real numbers such that $\lim_{n\to\infty} r_n = \infty$ and $r_n(\xi_n - \theta) \xrightarrow{d} \xi$. Then*

$$r_n(\phi(\xi_n) - \phi(\theta)) \xrightarrow{d} D\phi(\theta)\xi$$

and moreover

$$|r_n(\phi(\xi_n) - \phi(\xi)) - D\phi(\theta) r_n(\xi_n - \theta)| \xrightarrow{P} 0$$

Proof. By Lemma 15.8 be know that $\xi_n - \theta \xrightarrow{P} 0$. By differentiability of ϕ we know that for every sequence $h_n \to 0$,

$$\lim_n \frac{\phi(\theta + h_n) - \phi(\theta) - D\phi(\theta)h_n}{|h_n|} = 0$$

The first thing to show is that we can extend this fact to random sequences. We state this as a general fact. Suppose $\psi(x)$ is a function such that for every $h_n \to 0$ we have $\frac{\psi(h_n)}{|h_n|} \to 0$. We claim that if we are given random vectors η_n such that $\eta_n \xrightarrow{P} 0$ then $\frac{\psi(\eta_n)}{|\eta_n|} \xrightarrow{P} 0$. To see this define a new function by

$$f(x) = \begin{cases} \frac{\psi(x)}{|x|} & \text{for } x \neq 0 \\ 0 & \text{for } x = 0 \end{cases}$$

and note that by assumption f is continuous at 0. Now by the Continuous Mapping Theorem (Theorem 8.43) we know that $f(\eta_n) \xrightarrow{P} f(0) = 0$.

Having shown the above fact, we can use $\xi_n - \theta \xrightarrow{P} 0$ to conclude

$$\frac{\phi(\xi_n) - \phi(\theta) - D\phi(\theta)(\xi_n - \theta)}{|\xi_n - \theta|} \xrightarrow{P} 0$$

and if we multiply top and bottom by r_n and use linearity of the Frechet derivative we get

$$\frac{r_n(\phi(\xi_n) - \phi(\theta)) - D\phi(\theta) r_n(\xi_n - \theta)}{|r_n(\xi_n - \theta)|} \xrightarrow{P} 0$$

Tightness of $r_n(\xi_n - \theta)$ allows us to conclude that

$$r_n(\phi(\xi_n) - \phi(\theta)) - D\phi(\theta) r_n(\xi_n - \theta) \xrightarrow{P} 0$$

which gives us the second conclusion of the Theorem.

To prove this last fact suppose ξ_n, η_n are random vectors such that $\frac{\xi_n}{|\eta_n|} \xrightarrow{P} 0$ and η_n is tight. Suppose we are given $\epsilon, \delta > 0$. Use tightness to pick an $M > 0$ such that $\sup_n \mathbf{P}\{|\eta_n| > M\} < \frac{\delta}{2}$ and use $\frac{\xi_n}{|\eta_n|} \xrightarrow{P} 0$ to pick an N such that $\mathbf{P}\{\left|\frac{\xi_n}{\eta_n}\right| > \frac{\epsilon}{M}\} < \frac{\delta}{2}$

for all $n \geq N$. Then

$$\begin{aligned}\mathbf{P}\{|\xi_n| > \epsilon\} &= \mathbf{P}\{|\xi_n| > \epsilon; |\eta_n| > M\} + \mathbf{P}\{|\xi_n| > \epsilon; |\eta_n| \leq M\} \\ &\leq \mathbf{P}\{|\eta_n| > M\} + \mathbf{P}\{\frac{|\xi_n|}{|\eta_n|} > \frac{\epsilon}{M}\} \\ &< \delta\end{aligned}$$

for all $n \geq N$ which shows $\xi_n \xrightarrow{P} 0$. TODO: Is it better to think of this as $O_P(1)o_P(1) = o_P(1)$; probably better to think of this as $o_P(O_P(1)) = o_P(1)$?

To get the first conclusion we simply use the fact that matrix multiplication is continuous and the Continuous Mapping Theorem (Theorem 8.43) to see that $D\phi(\theta)r_n(\xi_n - \theta) \xrightarrow{d} D\phi(\theta)\xi$ and Slutsky's Lemma (Lemma 8.44)) and the part of this Theorem just proven to conclude $r_n(\phi(\xi_n) - \phi(\theta)) \xrightarrow{d} D\phi(\theta)\xi$. □

Example 15.13. One of the most common problems in statistics is the comparison of binomial populations. For example, to estimate treatment effectiveness one might want to compare the proportion of postive responses between a treated group and a control group. One common way to estimate the difference in proportions between two independent populations is the *risk ratio*

$$\hat{RR} = \frac{\hat{p}_1}{\hat{p}_2}$$

where $\hat{p}_i$ denotes the sample proportion. Here we calculate the asymptotic distribution of the risk ratio by using the Delta method.

The trick is to apply a logarithm to convert the division into subtraction. First we consider a single sample proportion $\hat{p}$. Since $\hat{p} = \frac{1}{n}\sum_n \xi_i$ for ξ_i a Bernoulli random variable with rate p, we can apply the Central Limit Theorem to conclude that

$$\sqrt{n}(\hat{p} - p) \xrightarrow{d} N(0, p(1-p))$$

Assuming $p \neq 0$, the Delta Method (Theorem 15.12) yields

$$\sqrt{n}(\ln(\hat{p}) - \ln(p)) \xrightarrow{d} \frac{1}{p}N(0, p(1-p)) = N(0, \frac{1-p}{p})$$

Therefore is we apply this reasoning to the risk ratio and use the fact that a sum of independent normal random variables is normal, we see that

$$\sqrt{n}(\ln(\hat{RR}) - \ln(RR)) \xrightarrow{d} N(0, \frac{1-p_1}{p_1} + \frac{1-p_2}{p_2})$$

This result can then be used to create asymptotic confidence intervals for the estimation of risk ratio

$$\ln(\hat{p}_1/\hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{1-\hat{p}_1}{n_1\hat{p}_1} + \frac{1-\hat{p}_2}{n_2\hat{p}_2}}$$

TODO: Discuss the implications of substituting the variance estimate into this formula.

TODO: Lay down the conceptual framework in which parametric statistics is modeled. Basic problem statement is this. Assume that one has a probability space $(\Omega, \mathcal{A}, P)$ and a family of random elements ξ_θ in a measure space $(X, \mathcal{X}, \mu)$ with $\theta \in \Theta$ an unknown parameter that determines the distribution of ξ_θ. Assume

we make observations of the value of ξ (or more properly observations of generally independent random variables with the same distribution as ξ), we want to find an estimate of the value (or the distribution) of θ.

There is the subtlety around the notion of having a random variable ξ with *conditional density* $f(x \mid \theta)$. The question is how rigorously one needs to think about the parameter θ. In the simplest form, one can just think of having a family of random variables ξ_θ for $\theta \in \Theta$ and not concern oneself with measurability in θ. This seems to be sufficient when discussing frequentist methods for example. Note also that the notation $f(x \mid \theta)$ seems to hedge on how we want to think of the functional dependence on θ. We'll see that understanding the dependence on θ is important but doesn't map nicely to standard probabilisitc or measure theoretic notions and has its own somewhat idiosyncratic notions of regularity. In the Bayesian formulation it appears that one wants to view θ as a random quantity as well and one assumes the existence of a random element θ in Θ and a random element ξ in X and take the conditional distribution $P_\theta = \mathbf{P}\{\xi \in \cdot \mid \theta\}$. Then one assumes that the conditional distributions are all absolutely continuous with respect to μ and thereby get the conditional densities $f(x \mid \theta)$ such that $P_\theta = f(x \mid \theta) \cdot \mu$. It is not yet clear to me at what point one is forced to take the latter approach.

Here is one account of the FI regularity conditions.

Definition 15.14. Suppose we are given a measure space $(X, \mathcal{X}, \mu)$ and a family of probability measures P_θ with $\theta \in \Theta \subset \mathbb{R}^n$ for some $n > 0$. Suppose that such that there exist densities $f(x \mid \theta)$ for each P_θ with respect so μ. The $f(x \mid \theta)$ are said to satisfy the *FI regularity constraints* if the following are true:

(i) $\Theta \subset \mathbb{R}^n$ is convex and contains and open set. There exists a set $B \in \mathcal{X}$ with $\mu(B^c) = 0$ such that $\frac{\partial}{\partial \theta_i} f(x \mid \theta)$ exists for every $i = 1, \ldots, n$, every $\theta \in \Theta$ and every $x \in B$.

(ii) For every $k = 1, \ldots, n$,

$$\frac{\partial}{\partial \theta_i} \int f(x \mid \theta)\, d\mu(x) = \int \frac{\partial}{\partial \theta_i} f(x \mid \theta)\, d\mu(x)$$

(iii) The set $C = \{x \in X \mid f(x \mid \theta) > 0$ does not depend on θ.

Definition 15.15. Let ξ be a random element in the measure space $(X, \mathcal{X}, \mu)$ with conditional density $f(x \mid \theta)$ with respect to μ. Suppose that $f(x \mid \theta)$ satisfy the FI regularity constraints. Then the random vector

$$U(\xi \mid \theta) = \left(\frac{\partial}{\partial \theta_1} \log f(\xi \mid \theta), \ldots, \frac{\partial}{\partial \theta_n} \log f(\xi \mid \theta) \right)$$

is called the *score function*.

The basic calculation with the score function is that if we assume that ξ is a random element with density $f(x \mid \theta)$ then

$$\begin{aligned} \mathbf{E}_\theta[\frac{\partial}{\partial \theta_i} \log f(\xi \mid \theta)] &= \int \frac{\frac{\partial}{\partial \theta_i} f(x \mid \theta)}{f(x \mid \theta)} f(x \mid \theta)\, d\mu(x) \\ &= \int \frac{\partial}{\partial \theta_i} f(x \mid \theta)\, d\mu(x) \\ &= \frac{\partial}{\partial \theta_i} \int f(x \mid \theta)\, d\mu(x) = \frac{\partial}{\partial \theta_i} 1 = 0 \end{aligned}$$

and therefore $\mathbf{E}_\theta[U(\xi \mid \theta)] = 0$ under the FI regularity constraints.

If we differentiate both side of this latter equality

$$\begin{aligned} 0 &= \frac{\partial}{\partial \theta_j} \int \frac{\partial}{\partial \theta_i} \log f(x \mid \theta) f(x \mid \theta)\, d\mu(x) \\ &= \int \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(x \mid \theta) f(x \mid \theta) + \frac{\partial}{\partial \theta_i} \log f(x \mid \theta) \frac{\partial}{\partial \theta_j} f(x \mid \theta)\, d\mu(x) \\ &= \int (\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(x \mid \theta) + \frac{\partial}{\partial \theta_i} \log f(x \mid \theta) \frac{\partial}{\partial \theta_j} \log f(x \mid \theta)) f(x \mid \theta)\, d\mu(x) \end{aligned}$$

which shows that when ξ has density $f(x \mid \theta)$, we have the identity

$$-\mathbf{E}_\theta[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(\xi \mid \theta)] = \mathbf{E}_\theta[\frac{\partial}{\partial \theta_i} \log f(\xi \mid \theta) \frac{\partial}{\partial \theta_j} \log f(\xi \mid \theta)]$$

This quantity is called the *Fisher information matrix*. TODO: The Fisher information as a Riemannian metric on Θ.

TODO: What kind of object is the score function (i.e. what domain and range). More specifically, how does one think of the θ dependence in the score function? In the Bayesian formulation everything is fine because ξ is an honest random element and we are just composing it with a deterministic function. In the formulation in which we don't think of θ as being random, then are we thinking of ξ as having θ-dependence when we plug it in? The answer to this is YES.

Example 15.16. Let ξ be a parameteric Gaussian family with $\theta = (\mu, \sigma)$. Then $f(x \mid \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$ and $U(\xi|\theta) = \frac{\xi - \mu}{\sigma^2}$.

Definition 15.17. Let ξ be a random element with conditional density $f(x \mid \theta)$ with respect to a measure space $(X, \mathcal{X}, \mu)$. For every $x \in X$, the function

$$L(\theta) = f(x \mid \theta)$$

is called the *likelihood function*.

Any random element $\hat{\theta}$ in Θ that satisfies

$$\max_{\theta \in \Theta} f(\xi \mid \theta) = f(\xi \mid \hat{\theta})$$

is called a *maximum likelihood estimator* of θ.

It is important to note that in most statistical applications the random element ξ whose likelihood we are investigating is a random vector that corresponds to sampling from a population. This is to say that is some underlying distribution of interest that corresponds to some random element ξ and that we model repeated sampling as a random element $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ in a product space $\mathcal{X}^n$. In all cases we shall be concerned about for the moment, we assume that the samples are i.i.d. hence the joint density of the sample is just the product of the density of ξ. In some cases it may be convenient to emphasize that the likelihood function is of such a form; in those cases we may choose to write $L_n(\theta)$ for the sample likelihood.

The fact that likelihood functions for independent samples are products is leveraged constantly in what follows and is in large part responsible for the nice asymptotic properties of maximum likelihood estimators. To release the power of this fact

we simply convert the product into a sum by taking log and create the log likelihood. Note that because the log is monotonic, one can perform maximum likelihood estimation equally well by taking maxima of the log likelihood. We shall usually write $\ell(x \mid \theta)$ to denote a log likelihood and the case of i.i.d. samples we shall use a subscript to emphasize the dependence on sample size $\ell_n(\boldsymbol{\xi} \mid \theta) = \sum_{i=1}^n \log f(\xi_i \mid \theta)$. The maximum likelihood estimator associated with i.i.d. samples of size n is denoted:

$$\hat{\theta}_n = \max_{\theta \in \Theta} \sum_{i=1}^n \log f(\xi_i \mid \theta)$$

and it is the estimator that we shall spend some time studying. The motivation behind this mechanism is that we know from the Gibbs Inequality (Lemma 15.4) that the true parameter θ_0 is characterized as the maximum of $\mathbf{E}_{\theta_0}[\log f(x \mid \theta)]$. Now we can view $\hat{\theta}_n$ as the result of substituting the (random) empirical measure in the expectation. To the extent that the empirical measure converges we may hope that the estimator converges as well. Less abstractly, we know from the Strong Law of Large Numbers that $\frac{1}{n}\sum_{i=1}^n \log f(\xi_i \mid \theta) \overset{a.s.}{\to} \mathbf{E}_{\theta_0}[\log f(x \mid \theta)]$ so thinking of this as convergence of functions of θ we may hope that the convergence is strong enough so that the maxima converge.

Note that the definition of the maximum likelihood estimator is using the max and not the sup; this means that in the case the supremum is not actually attained on the set Θ (e.g. Θ is open and the supremum is attained on the boundary) then MLE may not exist. In some accounts of the theory, the maximum is taken over the closure of the parameter domain (should we do this?)

Example 15.18. Consider the case parameter estimation in a normal distribution $\frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$. If we consider μ unknown and σ known the the MLE for the mean is given by setting the derivative with respect to μ to be zero

$$\frac{\partial}{\partial \mu} \sum_{i=1}^n \log \frac{1}{\sqrt{2\pi}\sigma} e^{-(\xi_i-\mu)^2/2\sigma^2} = -\frac{1}{\sigma^2} \sum_{i=1}^n (\xi_i - \mu) = 0$$

which implies it is the sample mean $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \xi_n$.

If we assume that μ is known and σ is unknown the finding the maximum by differentiation we get

$$\frac{\partial}{\partial \sigma} \sum_{i=1}^n \log \frac{1}{\sqrt{2\pi}\sigma} e^{-(\xi_i-\mu)^2/2\sigma^2} = \frac{1}{\sigma^3} \sum_{i=1}^n (\xi_i - \mu) - n\frac{1}{\sigma} = 0$$

and therefore the biased estimate of standard deviation $\hat{\sigma}_n = \frac{1}{n}\sum_{i=1}^n (\xi_n - \mu)^2$.

TODO: Example of estimating the rate of a Bernoulli r.v. Note the boundary behavior.

TODO: Example of ξ as a random vector of independent observations (factoring the likelihood function).

Note that we have allowed an MLE to be an arbitrary random element in Θ. It makes intuitive sense however that the estimator should depend on the value of ξ. That is indeed the case in many cases of interest and one of our goals shall be to understand the conditions under which that dependence holds.

Theorem 15.19. *If there is a sufficient statistic and the MLE exists, then the MLE is a function of the sufficient statistic.*

Proof. TODO: Apply the factorization theorem. □

TODO: Bring up the notion of *identifiability*; clearly if the likelihood function attains its maximum value for multiple values of θ then it is subtle to describe what consistency means (which is the correct value of θ).

As we've seen in Example 15.18 we cannot expect that maximum likelihood estimators will be consistent. However it is often the case that they will be asymptotically consistent. TODO: Define weakly and strongly asymptotically consistent. The following theorem provides a set of sufficient conditions under which a maximum likelihood estimator is strongly asymptotically consistent.

Theorem 15.20 (Asymptotic Consistency of MLE)**.** *Let $\xi, \xi_1, \xi_2, \ldots$ be i.i.d. parametric family with distribution $f(x \mid \theta)\, d\mu$ with respect to measure space $(X, \mathcal{X}, \mu)$. Assume that θ_0 is fixed and define*

$$Z(M, x) = \inf_{\theta \in M} \log \frac{f(x \mid \theta_0)}{f(x \mid \theta)}$$

Assume that for all $\theta \neq \theta_0$ there is an open neighborhood U_θ such that $\theta \in U_\theta$ and $\boldsymbol{E}_{\theta_0}[Z(U_\theta, \xi)] > 0$.

If Θ is not compact, assume that there is a compact $K \subset \Theta$ such that $\theta_0 \in K$ and $\boldsymbol{E}_{\theta_0}[Z(\Theta \setminus K, \xi)] > 0$. Then

$$\lim_{n \to \infty} \hat{\theta}_n = \theta_0$$

almost surely with respect to P_{θ_0}.

Before starting in on the proof make sure to understand the nature of the hypotheses. Given the observation x we have $Z(U, x) < 0$ if there is a $\theta \in U$ such that a θ this more likely than θ_0, whereas $Z(U, x) > 0$ tells us that θ_0 is more likely than any $\theta \in U$. Thus the conditions $\mathbf{E}_{\theta_0}[Z(U_\theta, \xi)] > 0$ are statements that on average there is no better explanation than θ_0. One thing that is interesting about the result is that it is only required that θ_0 be the best average estimate locally in Θ (admittedly the weakening to a local property is only allowed over a compact set).

Proof. By Lemma 8.3, the Theorem is proven if we can show that $\mathbf{P}_{\theta_0}\{d(\hat{\theta}_n, \theta_0) \geq \epsilon \text{ i.o.}\} = 0$ for every $\epsilon > 0$. So assume that we have fixed $\epsilon > 0$ and let $B(\theta_0, \epsilon)$ be the ϵ-ball around θ_0. Since $K \setminus B(\theta_0, \epsilon)$ is compact and U_θ is a cover, we can find an finite subcover $U_1, \ldots, U_{m-1}$ of $K \setminus B(\theta_0, \epsilon)$ such that each U_j satisfies $\mathbf{E}_{\theta_0}[Z(U_j, \xi)] > 0$. If we define $U_m = \Theta \setminus K$ then we by hypothesis have a finite cover $U_1, \ldots, U_m$ of $\Theta \setminus B(\theta_0, \epsilon)$ with each U_j satisfying the same property.

Now on each U_j we can apply the Strong Law of Large Numbers to conclude that for each j, $\frac{1}{n} \sum_{i=1}^{n} Z(U_j, \xi_i) \overset{a.s.}{\to} \mathbf{E}_{\theta_0}[Z(U_j, \xi)] > 0$ a.s. The key point from this point on is to understand that if we assume that $\hat{\theta}_n \in U_j$ infinitely often it would

force the expectation $\mathbf{E}_{\theta_0}[Z(U_j, \xi)]$ to be nonpositive. Precisely,

$$
\begin{aligned}
&\mathbf{P}_{\theta_0}\{\hat{\theta}_n \notin B(\theta_0, \epsilon) \text{ i.o.}\} \\
&\leq \mathbf{P}_{\theta_0}\{\hat{\theta}_n \in \cup_{j=1}^m U_j \text{ i.o.}\} && \text{since } B^c \subset \cup_{j=1}^m U_j \\
&= \mathbf{P}_{\theta_0}\{\cup_{j=1}^m \{\hat{\theta}_n \in U_j \text{ i.o.}\}\} && \text{by finiteness of } n \\
&\leq \sum_{j=1}^m \mathbf{P}_{\theta_0}\{\hat{\theta}_n \in U_j \text{ i.o.}\} && \text{by subadditivity} \\
&\leq \sum_{j=1}^m \mathbf{P}_{\theta_0}\{\inf_{\theta \in U_j} \sum_{i=1}^n \log \frac{f(\xi_i, \theta_0)}{f(\xi_i, \theta)} \leq 0 \text{ i.o.}\} && \text{because } \sum_{i=1}^n \log \frac{f(\xi_i, \theta_0)}{f(\xi_i, \hat{\theta}_n)} \leq 0 \\
&\leq \sum_{j=1}^m \mathbf{P}_{\theta_0}\{\sum_{i=1}^n \inf_{\theta \in U_j} \log \frac{f(\xi_i, \theta_0)}{f(\xi_i, \theta)} \leq 0 \text{ i.o.}\} \\
&= \sum_{j=1}^m \mathbf{P}_{\theta_0}\{\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \inf_{\theta \in U_j} \log \frac{f(\xi_i, \theta_0)}{f(\xi_i, \theta)} \leq 0\} \\
&= \sum_{j=1}^m \mathbf{P}_{\theta_0}\{\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n Z(U_j, \xi_i) \leq 0\} \\
&= 0
\end{aligned}
$$

since as noted the last equality follows from fact that the Strong Law of Large Numbers tells us that almost surely for all $1 \leq j \leq m$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n Z(U_j, \xi_i) = \mathbf{P}_{\theta_0}\{Z(U_j, \xi)\} > 0$$

□

Note that the proof above has a gap in it from the outset. The functions $Z(M, x)$ for a fixed $M \subset \Theta$ are defined as an infimum of an uncountable collection of random variables hence we do not know that they are measurable. On the other hand we clearly need them to be in order to take expectations. TODO: How do we get around these issues? I suspect there are two paths to explore: 1) take a countable dense subset and show that the infimum can be reduced to a countable one or 2) abandon measurability and see if we can make due with outer expectations (a la empirical process theory).

Example 15.21. Consider the problem of estimating the parameter $\theta \in [0, \infty)$ in the family $U(0, \theta)$. Assume that θ_0 is the true parameter and we want to show consistency of the maximum likelihood estimator. The likelihood function in this case is

$$f(x \mid \theta) = \begin{cases} \frac{1}{\theta} & \text{if } 0 \leq x \leq \theta \\ 0 & \text{if } x < 0 \text{ or } x > \theta \end{cases}$$

Note that you should be thinking of $f(x \mid \theta)$ as a function of θ with x fixed. To apply Theorem 15.20 we need to show $\mathbf{E}_{\theta_0}[Z(U, \theta)] > 0$ for appropriately chosen $U \subset [0, \infty)$. Since $\mathbf{P}_{\theta_0}\{x < 0\} = \mathbf{P}_{\theta_0}\{x > \theta_0\} = 0$ for purposes of computing the

expectations we may assume that $0 \leq x \leq \theta_0$. With this in mind, for such an x, we have the likelihood ratio

$$\log \frac{f(x \mid \theta_0)}{f(x \mid \theta)} = \begin{cases} +\infty & \text{if } 0 \leq \theta < x \\ \log \frac{\theta}{\theta_0} & \text{if } x \leq \theta \end{cases}$$

So now we find our neighborhoods. Pick $\theta > \theta_0$ and define $U_\theta = (\frac{\theta+\theta_0}{2}, \infty)$ (any left hand endpoint between θ_0 and θ would suffice). In this case,

$$Z(U_\theta, x) = \inf_{\psi > \frac{\theta+\theta_0}{2}} \frac{f(x \mid \theta_0)}{f(x \mid \theta)} = \inf_{\psi > \frac{\theta+\theta_0}{2}} \log \frac{\psi}{\theta_0} = \log \frac{\theta + \theta_0}{2\theta_0} > 0$$

therefore $\mathbf{E}_{\theta_0}[Z(U_\theta, x)] > 0$.

If we pick $\theta < \theta_0$ then pick $U_\theta = (\theta/2, \frac{\theta+\theta_0}{2})$ and note that

$$Z(U_\theta, x) = \begin{cases} \log \frac{\theta}{2\theta_0} & \text{if } x \leq \frac{\theta}{2} \\ \log \frac{x}{\theta_0} & \text{if } \frac{\theta}{2} < x < \frac{\theta+\theta_0}{2} \\ +\infty & \text{if } \frac{\theta+\theta_0}{2} \leq x \leq \theta_0 \end{cases}$$

and therefore $\mathbf{E}_{\theta_0}[Z(U_\theta, x)] = +\infty$.

Lastly we have to find a compact set K such that $\mathbf{E}_{\theta_0}[Z(\mathbb{R}_+ \setminus K, x)] > 0$. Pick $a > 1$ and consider the interval $[\theta_0/a, a\theta_0]$. Note that

$$Z(\mathbb{R}_+ \setminus [\theta_0/a, a\theta_0], x) = \begin{cases} \log \frac{x}{\theta_0} & \text{if } x < \frac{\theta_0}{a} \\ \log a & \text{if } \frac{\theta_0}{a} \leq x \leq \theta_0 \end{cases}$$

so integrating,

$$\begin{aligned} \mathbf{E}_{\theta_0}[Z(\mathbb{R}_+ \setminus [\theta_0/a, a\theta_0], x)] &= \frac{1}{\theta_0} \int_0^{\frac{\theta_0}{a}} \log \frac{x}{\theta_0}\, dx + \frac{\theta_0 - \frac{\theta_0}{a}}{\theta_0} \log a \\ &= (\frac{1}{a} \log \frac{1}{a} - \frac{\theta_0}{a}) + \frac{\theta_0 - \frac{\theta_0}{a}}{\theta_0} \log a \end{aligned}$$

Note that the first term goes to 0 as a goes to ∞ and the second term goes to ∞ as a goes to ∞ and therefore for sufficiently large a we have $\mathbf{E}_{\theta_0}[Z(\mathbb{R}_+ \setminus [\theta_0/a, a\theta_0], x)] > 0$.

Note also that as a approaches 1 the expectation approaches $-\theta_0 \leq 0$. In this specific sense if we allow ourselves to consider regions of parameter space like $(\theta, \theta_0 + \epsilon)$ for $\epsilon > 0$ small, then under sampling we expect the there is an estimate that is better (more likely) than the true parameter value. TODO: Think more carefully about this fact and how to interpret it; should this disturb us? Perhaps this shouldn't disturb us because the thing that allows us to create these regions on which $\mathbf{E}_{\theta_0}[Z(U, x)] < 0$ is precisely the fact that we are allowing ourselves to include $\theta_0 \in U$; without allowing that we can't create such a set.

The basic phenomenon in this example can be summarized as:

(i) Given a single observation x then the MLE is x with likelihood $1/x$; any $\theta > x$ has strictly smaller likelihood $1/\theta$ while any $\theta < x$ has likelihood 0.

(ii) For any $\theta \geq \theta_0$ we know that θ_0 is always a better estimator since we can only observe $x \leq \theta_0$ and for these observations θ_0 is always better.

(iii) For any $\theta < \theta_0$ for any observations $x \leq \theta$ we know that θ is a better estimator than θ_0 by a finite factor, however for $\theta < x \leq \theta_0$ then θ_0 is a infinitely better estimator than θ.

Example 15.22. This example illustrates the difficulties that can arise in applying the above results to conclude that an MLE is consistent when the parameter space is not compact. Consider a normal family with parameter $\Theta = \{(\mu, \sigma) \mid \sigma > 0\}$ given by $\frac{1}{\sqrt{2\pi\sigma}} e^{-(x-\mu)^2/2\sigma}$. We show that for any compact $K \subset \mathbb{R}^2$ we have $\mathbf{E}_{\theta_0}[Z(K^c, x)] = -\infty$ and therefore Theorem 15.20 does not apply. In fact we show that for any compact K we have $Z(K^c, x) = -\infty$. This follows by noting that any compact K is bounded hence there exists a value of μ such that $\{(\mu, \sigma) \mid \sigma > 0\} \subset K^c$. Now we see that for such a μ,

$$\lim_{\sigma \to 0^+} \frac{f(x \mid \mu_0, \sigma_0)}{f(x \mid \mu, \sigma)} = \lim_{\sigma \to 0^+} \left(\log \sigma - \log \sigma_0 - \frac{(x-\mu_0)^2}{2\sigma_0} + \frac{(x-\mu)^2}{2\sigma} \right) \neq -\infty$$

TODO: Fix this argument; it is broken. The limit is only negative infinity when x is large enough so that $\{(x, \sigma) \mid \sigma > 0\} \subset \Theta \setminus K$. That should be enough if we can show that the integral over the rest of the domain is not $+\infty$.

On the other hand, one can compute the MLE explicitly in this case and verify that it is asymptotically consistent so we have shown that conditions of the theorem are sufficient but not necessary.

TODO: The following Theorem only requires upper semi-continuity.

Theorem 15.23. *Let $\xi, \xi_1, \xi_2, \ldots$ be i.i.d. parametric family with distribution $f(x \mid \theta)\, d\mu$ with respect to measure space $(X, \mathcal{X}, \mu)$. Assume that θ_0 is fixed and define*

$$Z(M, x) = \inf_{\theta \in M} \log \frac{f(x \mid \theta_0)}{f(x \mid \theta)}$$

Assume that for all $\theta \neq \theta_0$ there is an open neighborhood U_θ such that $\theta \in U_\theta$ and $\boldsymbol{E}_{\theta_0}[Z(U_\theta, \xi)] > -\infty$. Assume $f(x \mid \theta)$ is a continuous function of θ for almost all x with respect to P_{θ_0}

If Θ is not compact, assume that there is a compact $K \subset \Theta$ such that $\theta_0 \in K$ and $\boldsymbol{E}_{\theta_0}[Z(\Theta \setminus K, \xi)] > 0$. Then

$$\lim_{n \to \infty} \hat{\theta}_n = \theta_0$$

almost surely with respect to P_{θ_0}.

Proof. We show that for all $\theta \neq \theta_0$ there exists a neighborhood $\theta \in U_\theta$ such that $\mathbf{E}_{\theta_0}[Z(U_\theta, \xi)] > 0$ and then apply the previous Theorem 15.20.

Pick $\theta \neq \theta_0$ and assume that we have an open neighborhood U_θ with $\theta \in U_\theta$ and $\mathbf{E}_{\theta_0}[Z(U_\theta, \xi)] > -\infty$. If $\mathbf{E}_{\theta_0}[Z(U_\theta, \xi)] > 0$ then have found a suitable neighborhood so we may assume $\mathbf{E}_{\theta_0}[Z(U_\theta, \xi)] \leq 0$ as well (we really just need to assume that the value if finite a bit later in the proof). Now for each $n \in \mathbb{N}$ pick a closed ball $U_\theta^n = B(\theta, r_n) \subset U_\theta$ such that $r_n \leq \frac{1}{n}$ and r_n are non-increasing. Furthermore because $U_\theta^{n+1} \subset U_\theta^n$ we have for fixed x, $Z(U_\theta^n, x)$ is increasing in n.

Now assume that we have an x such that $f(x \mid \theta)$ is continuous. This implies $\log \frac{f(x|\theta_0)}{f(x|\theta)}$ is continuous as well. This continuity coupled with the compactness of U_θ^n implies that there exists a $\theta_n(x) \in U_\theta^n$ such that $Z(U_\theta^n, x) = \log \frac{f(x|\theta_0)}{f(x|\theta_n(x))}$. Clearly

we have $\cap_n U_\theta^n = \{\theta\}$ and this implies $\lim_{n\to\infty} \theta_n(x) = \theta$. Again by continuity we get

$$\lim_{n\to\infty} Z(U_\theta^n, x) = \lim_{n\to\infty} \log \frac{f(x \mid \theta_0)}{f(x \mid \theta_n(x))} = \log \frac{f(x \mid \theta_0)}{f(x \mid \theta)}$$

Now because $U_\theta^n \subset U_\theta$ we have $Z(U_\theta^n, x) \geq Z(U_\theta, x)$ and $\mathbf{E}_{\theta_0}[Z(U_\theta, \xi)]$ is finite, we may apply Fatou's Lemma (Theorem 3.42)

$$\begin{aligned}
\liminf_{n\to\infty} \mathbf{E}_{\theta_0}[Z(U_\theta^n, x)] - \mathbf{E}_{\theta_0}[Z(U_\theta, x)] &= \liminf_{n\to\infty} \mathbf{E}_{\theta_0}[Z(U_\theta^n, x) - Z(U_\theta, x)] \\
&\geq \mathbf{E}_{\theta_0}[\lim_{n\to\infty} (Z(U_\theta^n, x) - Z(U_\theta, x))] \\
&= \mathbf{E}_{\theta_0}[\log \frac{f(x \mid \theta_0)}{f(x \mid \theta)}] - \mathbf{E}_{\theta_0}[Z(U_\theta, x)]
\end{aligned}$$

Cancelling the (finite) common term $\mathbf{E}_{\theta_0}[Z(U_\theta, x)]$ we get

$$\liminf_{n\to\infty} \mathbf{E}_{\theta_0}[Z(U_\theta^n, x)] \geq \mathbf{E}_{\theta_0}[\log \frac{f(x \mid \theta_0)}{f(x \mid \theta)}] > 0$$

where the last inequality follows from the positivity of relative entropy (Lemma 15.4). Now by this inequality we can find an $N > 0$ such that $\mathbf{E}_{\theta_0}[Z(U_\theta^n, x)] > 0$ for all $n \geq N$, but in particular there is a single neighborhood U_θ^N with this property. □

The technical conditions above are sufficient to prove asymptotic efficient of MLEs but it is certainly not necessary.

TODO: Example showing consistency without conditions.

TODO: Note a different condition that suffices (Martingale proof: Schervish Lemma 7.83)

Maximum likelihood estimators are asymptotically normal under certain circumstances. If is unfortunate that any precise statement of those circumstances is technical and verbose. It is also unfortunate that there is no definitive characterization of asymptotic normality as a set of necessary and sufficient conditions. Instead there are a number of sufficient conditions available with different levels of generality and sophistication. TODO: This is equally true about asymptotic consistency and asymptotic results in general; move this comment to an appropriate place and generalize.

Before stating a rather classical version of such a result let's consider the case of a scalar parameter in a somewhat heuristic fashion. If we assume that we have a consistent MLE such that $\hat{\theta}_n \xrightarrow{a.s.} \theta_0$ and we want to prove that $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \sigma^2)$ for an appropriate σ. We assume that $f(x \mid \theta)$ is twice continuously differentiable as a function of θ; under these conditions the maximum of the likelihood implies a vanishing derivative

$$\frac{\partial}{\partial\theta} \ell_n(\xi \mid \hat{\theta}_n) = 0$$

If we apply the mean value theorem to the function $\frac{\partial}{\partial\theta}\ell_n(\xi \mid \theta)$ to conclude that there is a value θ_n^* that lies between $\hat{\theta}_n$ and θ_0 such that

$$\frac{\frac{\partial}{\partial\theta}\ell_n(\xi \mid \hat{\theta}_n) - \frac{\partial}{\partial\theta}\ell_n(\xi \mid \theta_0)}{\hat{\theta}_n - \theta_0} = \frac{\partial^2}{\partial\theta^2}\ell_n(\xi \mid \theta_n^*)$$

or rearranging terms to set up ourselves up to take advantage of the Central Limit Theorem (ignore the possiblity that the denominator vanishes):

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\frac{\sqrt{n}\frac{\partial}{\partial\theta}\ell_n(\xi \mid \theta_0)}{\frac{\partial^2}{\partial\theta^2}\ell_n(\xi \mid \theta_n^*)}$$

Now consider the numerator $\mu = \mathbf{E}_{\theta_0}[\log f(\xi \mid \theta_0)] = 0$ and variance $i(\theta_0) = \mathbf{E}_{\theta_0}[\log^2 f(\xi \mid \theta_0)]$ and we can apply the Central Limit Theorem to see

$$\frac{1}{\sqrt{n}}\ell_n'(\xi \mid \theta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(\xi_i \mid \theta_0) \xrightarrow{d} N(0, i(\theta_0))$$

This looks quite promising but there is a factor of $\frac{1}{\sqrt{n}}$ that was added that will have to be addressed.

Now if we consider the denominator things don't look so good; however a small modification seems amenable to analysis. If we consider $\frac{\partial^2}{\partial\theta^2}\ell_n(\xi \mid \theta_0)$, then we see that the Weak Law Of Large Numbers tells us that

$$-\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\ell_n(\xi \mid \theta_0) = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\ell_n(\xi_i \mid \theta_0) \xrightarrow{P} \mathbf{E}_{\theta_0}[-\frac{\partial^2}{\partial\theta^2}\log f(\xi \mid \theta_0)] = i(\theta_0)$$

Moreover, the factor of $\frac{1}{n}$ that we needed here to apply the Law of Large Numbers cancelled exactly with our use of $\frac{1}{\sqrt{n}}$ in the Central Limit Theorem application so that our Taylor expansion can be written as

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\frac{\frac{\partial}{\partial\theta}\ell_n(\xi \mid \theta_0)}{\sqrt{n}} \cdot \frac{n}{\frac{\partial^2}{\partial\theta^2}\ell_n(\xi \mid \theta_0)} \cdot \frac{\frac{\partial^2}{\partial\theta^2}\ell_n(\xi \mid \theta_0)}{\frac{\partial^2}{\partial\theta^2}\ell_n(\xi \mid \theta_n^*)}$$

and we are in position to use Slutsky's Lemma to extend the asymptotic normality of the first factor to $\sqrt{n}(\hat{\theta}_n - \theta_0)$. The rub is that we have a term

$$\frac{\frac{\partial^2}{\partial\theta^2}\ell_n(\xi \mid \theta_0)}{\frac{\partial^2}{\partial\theta^2}\ell_n(\xi \mid \theta_n^*)}$$

to understand. By consistency of the estimator we know that $\theta_n^* \xrightarrow{a.s.} \theta_0$ we might hope that this term converges to 1 (at least in probability). In fact additional smoothness assumptions on f are sufficient to guarantee that this is the case; the expression of these smoothness constraints is what provides the complexity to statements of asymptotic normality of MLEs. When that is shown, then keeping track of the factors of $i(\theta_0)$ we see that Slutsky's Lemma will tell us that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, i(\theta_0)^{-1})$$

In the following Theorem we capture all the varied assumptions that are required to make an argument like the above rigorous; the result is also stated for multivariate parameters. The details of the proof are organized a bit differently than the outline of the scalar case given above (e.g. dealing with boundaries in parameter space) but the main points of the proof remain the same:

1) Taylor expand the likelihood function around θ_0
2) Use the Central Limit Theorem to prove convergence of the first derivative term at θ_0
3) Use the Weak Law of Large Numbers to prove convergence of the second derivative term at θ_0

4) Use asymptotic consistency of $\hat{\theta}_n$ and bounds on the variation of the second derivative to conclude that the difference between the second derivatives at θ_0 and $\hat{\theta}_n$ go to zero in probability.
5) Use Slutsky's Lemma to glue all the pieces together.

Theorem 15.24. *Let $\xi, \xi_1, \xi_2, \ldots$ be i.i.d. parametric family with distribution $f(x \mid \theta)\, d\mu$ with respect to measure space $(X, \mathcal{X}, \mu)$ with $\Theta \subset \mathbb{R}^k$ for some $k > 0$. Assume*

(i) *$\hat{\theta}_n \xrightarrow{P} \theta_0$ in P_{θ_0} for every $\theta_0 \in \Theta$.*
(ii) *$f(x \mid \theta)$ has continuous second partial derivatives with respect to θ and that differentiation can be passed under the integral sign*
(iii) *there exists $H_r(x, \theta)$ such that for each $\theta_0 \in int(\Theta)$ and each k, j,*

$$\sup_{\|\theta - \theta_0\| \leq r} \left| \frac{\partial^2}{\partial \theta_k \partial \theta_j} \log f(x \mid \theta_0) - \frac{\partial^2}{\partial \theta_k \partial \theta_j} \log f(x \mid \theta) \right| \leq H_r(x, \theta_0)$$

with $\lim_{r \to 0} \boldsymbol{E}_{\theta_0}[H_r(\xi, \theta_0)] = 0$.
(iv) *the Fisher information matrix $\mathcal{I}_\xi(\theta_0)$ is finite and nonsingular.*

Then

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N(0, \mathcal{I}_\xi^{-1}(\theta_0))$$

Proof. We start with

Claim 1: $\frac{1}{\sqrt{n}} D_{\hat{\theta}_n} \ell_n(\boldsymbol{\xi} \mid \theta) \xrightarrow{P} 0$

One might jump to the conclusion that $D_{\hat{\theta}_n} \ell_n(\boldsymbol{\xi} \mid \theta) = 0$ everywhere because $\hat{\theta}_n$ is a maximum, however there are some details about handling the issue of boundaries on Θ. One does know that $D_{\hat{\theta}_n} \ell_n(\boldsymbol{\xi} \mid \theta) = 0$ when $\hat{\theta}_n \in \text{int}(\Theta)$ but there is the possibility that some $\hat{\theta}_n$ lies on the boundary of Θ and the derivative might not vanish in this case. To handle the boundary effects, first we know that $\theta_0 \in \text{int}(\Theta)$ and therefore there is an open neighborhood $\theta_0 \in U \subset \text{int}(\Theta)$. By the vanishing of the derivative at any maximum in the interior, we know

$$\begin{aligned} \frac{1}{\sqrt{n}} D_{\hat{\theta}_n} \ell_n(\boldsymbol{\xi} \mid \theta) &= \frac{1}{\sqrt{n}} D_{\hat{\theta}_n} \ell_n(\boldsymbol{\xi} \mid \theta) \mathbf{1}_{\hat{\theta}_n \in U} + \frac{1}{\sqrt{n}} D_{\hat{\theta}_n} \ell_n(\boldsymbol{\xi} \mid \theta) \mathbf{1}_{\hat{\theta}_n \notin U} \\ &= \frac{1}{\sqrt{n}} D_{\hat{\theta}_n} \ell_n(\boldsymbol{\xi} \mid \theta) \mathbf{1}_{\hat{\theta}_n \notin U} \end{aligned}$$

Using the fact that $\hat{\theta}_n \xrightarrow{P} \theta_0$ allows us to conclude that

$$\lim_{n \to \infty} \mathbf{P}_{\theta_0}\{\hat{\theta}_n \notin U\} = 0$$

so in particular,

$$\lim_{n \to \infty} \mathbf{P}_{\theta_0}\{\frac{1}{\sqrt{n}} D_{\hat{\theta}_n} \ell_n(\boldsymbol{\xi} \mid \theta) \mathbf{1}_{\hat{\theta}_n \notin U} = 0\} = 0$$

Putting these two pieces of information together we see

$$\frac{1}{\sqrt{n}} D_{\hat{\theta}_n} \ell_n(\boldsymbol{\xi} \mid \theta) = \frac{1}{\sqrt{n}} D_{\hat{\theta}_n} \ell_n(\boldsymbol{\xi} \mid \theta) \mathbf{1}_{\hat{\theta}_n \notin U} \xrightarrow{P} 0$$

Now we derive a quadratic approximation to the likelihood by using a Taylor expansion (actually just the Mean Value Theorem) of $D_\theta \ell_n(\boldsymbol{\xi} \mid \theta)$ around θ_0. Once again there is the issue of boundaries but moreover the domain Θ is not convex so the Taylor series only applies cleanly when $\hat{\theta}_n$ belongs to a ball around θ_0. To

handle this, pick an $R > 0$ such that we have $B(\theta_0; R) \subset \text{int}(\Theta)$. In this case, when $\|\hat{\theta}_n - \theta_0\| < R$ then we know there exists a θ_n^* between θ_0 and $\hat{\theta}_n$ such that

$$D_{\hat{\theta}_n} \ell_n(\boldsymbol{\xi} \mid \theta) - D_{\theta_0} \ell_n(\boldsymbol{\xi} \mid \theta) = D^2_{\theta_n^*} \ell_n(\boldsymbol{\xi} \mid \theta) \cdot (\hat{\theta}_n - \theta_0)$$

As it turns out what happens when $\|\hat{\theta}_n - \theta_0\| \geq R$ won't matter since it is an event that occurs with vanishingly small probability as n grows. Accordingly, we define

$$\Delta_n = \begin{cases} D^2_{\theta_n^*} \ell_n(\boldsymbol{\xi} \mid \theta) & \text{when } \|\hat{\theta}_n - \theta_0\| < R \\ 0 & \text{when } \|\hat{\theta}_n - \theta_0\| \geq R \end{cases}$$

TODO: Do we need to justify measurability here...

Claim 2: $\frac{1}{\sqrt{n}}(D_{\theta_0} \ell_n(\boldsymbol{\xi} \mid \theta) + \Delta_n \cdot (\hat{\theta}_n - \theta_0)) \xrightarrow{P} 0$

Pick an $\epsilon > 0$. From the definition of Δ_n we have

$$\frac{1}{\sqrt{n}}(D_{\theta_0} \ell_n(\boldsymbol{\xi} \mid \theta) + \Delta_n \cdot (\hat{\theta}_n - \theta_0)) = \begin{cases} \frac{1}{\sqrt{n}}(D_{\hat{\theta}_n} \ell_n(\boldsymbol{\xi} \mid \theta) & \text{when } \|\hat{\theta}_n - \theta_0\| < R \\ \frac{1}{\sqrt{n}}(D_{\theta_0} \ell_n(\boldsymbol{\xi} \mid \theta) & \text{when } \|\hat{\theta}_n - \theta_0\| \geq R \end{cases}$$

and therefore

$$\begin{aligned} &\lim_{n \to \infty} \mathbf{P}_{\theta_0}\{\frac{1}{\sqrt{n}}(D_{\theta_0} \ell_n(\boldsymbol{\xi} \mid \theta) + \Delta_n \cdot (\hat{\theta}_n - \theta_0)) > \epsilon\} \\ &= \lim_{n \to \infty} \mathbf{P}_{\theta_0}\{\frac{1}{\sqrt{n}} D_{\hat{\theta}_n} \ell_n(\boldsymbol{\xi} \mid \theta) > \epsilon; \|\hat{\theta}_n - \theta_0\| < R\} \\ &+ \lim_{n \to \infty} \mathbf{P}_{\theta_0}\{\frac{1}{\sqrt{n}} D_{\theta_0} \ell_n(\boldsymbol{\xi} \mid \theta) > \epsilon; \|\hat{\theta}_n - \theta_0\| \geq R\} \\ &\leq \lim_{n \to \infty} \mathbf{P}_{\theta_0}\{\frac{1}{\sqrt{n}} D_{\hat{\theta}_n} \ell_n(\boldsymbol{\xi} \mid \theta) > \epsilon\} + \lim_{n \to \infty} \mathbf{P}_{\theta_0}\{\|\hat{\theta}_n - \theta_0\| \geq R\} = 0 \end{aligned}$$

where we have used Claim 1 and the weak consistency of the estimator $\hat{\theta}_0$.

Claim 3: $\frac{1}{n}\Delta_n \xrightarrow{P} -\mathcal{I}_\xi(\theta_0)$

Write

$$\begin{aligned} \frac{1}{n}\Delta_n = &\frac{1}{n} D^2_{\theta_0} \ell_n(\boldsymbol{\xi} \mid \theta) \mathbf{1}_{\|\hat{\theta}_n - \theta_0\| < R} \\ &+ (D^2_{\theta_n^*} \ell_n(\boldsymbol{\xi} \mid \theta) - D^2_{\theta_0} \ell_n(\boldsymbol{\xi} \mid \theta)) \mathbf{1}_{\|\hat{\theta}_n - \theta_0\| < R} \end{aligned}$$

and we address the convergence of each of the summands. First note that by weak consistency of the estimator $\hat{\theta}_n$ we have $\mathbf{1}_{\|\hat{\theta}_n - \theta_0\| < R} \xrightarrow{P} 1$. By the Weak Law of Large Numbers and the fact we can exchange derivatives and expectations we have

$$\frac{1}{n} D^2_{\theta_0} \ell_n(\boldsymbol{\xi} \mid \theta) = \frac{1}{n} \sum_{i=1}^n D^2_{\theta_0} \log f(\xi_i \mid \theta) \xrightarrow{P} \mathbf{E}_{\theta_0}[D^2_{\theta_0} \log f(\xi \mid \theta)] = -\mathcal{I}_\xi(\theta_0)$$

and therefore by Corollary 8.13 to the Continuous Mapping Theorem we can combine these facts to conclude

$$\frac{1}{n} D^2_{\theta_0} \ell_n(\boldsymbol{\xi} \mid \theta) \mathbf{1}_{\|\hat{\theta}_n - \theta_0\| < R} \xrightarrow{P} -\mathcal{I}_\xi(\theta_0)$$

We turn attention to the error term which we show is $o_P(1)$. Let $\epsilon > 0$ be given. Pick any $0 < r \leq R$ such that $\mathbf{E}_{\theta_0}[H_r(\xi, \theta_0)] < \frac{\epsilon}{2}$. Again applying the Weak Law

of Large Numbers

$$\frac{1}{n}\sum_{i=1}^{n} H_r(\xi_i,\theta_0) \xrightarrow{P} \mathbf{E}_{\theta_0}[H_r(\xi,\theta_0)] < \frac{\epsilon}{2}$$

and therefore

$$\lim_{n\to\infty} \mathbf{P}_{\theta_0}\{\frac{1}{n}\sum_{i=1}^{n} H_r(\xi_i,\theta_0) < \epsilon\} \leq \lim_{n\to\infty} \mathbf{P}_{\theta_0}\{\left|\frac{1}{n}\sum_{i=1}^{n} H_r(\xi_i,\theta_0) - \mathbf{E}_{\theta_0}[H_r(\xi,\theta_0)]\right| < \frac{\epsilon}{2}\} = 0$$

Now apply this fact to get a bound on each entry of the Hessian matrix

$$\begin{aligned}
&\lim_{n\to\infty} \mathbf{P}_{\theta_0}\{\frac{1}{n}\left|D^2_{\theta_n^*,j,k}\ell_n(\boldsymbol{\xi}\mid\theta) - D^2_{\theta_0,j,k}\ell_n(\boldsymbol{\xi}\mid\theta)\right| \mathbf{1}_{\|\theta_n^*-\theta_0\|<R} < \epsilon\}\\
&\lim_{n\to\infty} \mathbf{P}_{\theta_0}\{\frac{1}{n}\left|D^2_{\theta_n^*,j,k}\ell_n(\boldsymbol{\xi}\mid\theta) - D^2_{\theta_0,j,k}\ell_n(\boldsymbol{\xi}\mid\theta)\right| \mathbf{1}_{\|\theta_n^*-\theta_0\|<r} < \epsilon\}\\
&+ \lim_{n\to\infty} \mathbf{P}_{\theta_0}\{\frac{1}{n}\left|D^2_{\theta_n^*,j,k}\ell_n(\boldsymbol{\xi}\mid\theta) - D^2_{\theta_0,j,k}\ell_n(\boldsymbol{\xi}\mid\theta)\right| \mathbf{1}_{r\leq\|\theta_n^*-\theta_0\|<R} < \epsilon\}\\
&\leq \lim_{n\to\infty} \mathbf{P}_{\theta_0}\{\frac{1}{n}\sum_{i=1}^{n} H_r(\xi_i,\theta_0) < \epsilon\} + \lim_{n\to\infty} \mathbf{P}_{\theta_0}\{\mathbf{1}_{r\leq\|\theta_n^*-\theta_0\|<R}\}\\
&= 0
\end{aligned}$$

and therefore we have shown $\frac{1}{n}(D^2_{\theta_n^*,j,k}\ell_n(\boldsymbol{\xi}\mid\theta) - D^2_{\theta_0,j,k}\ell_n(\boldsymbol{\xi}\mid\theta))\mathbf{1}_{\|\theta_n^*-\theta_0\|<R} \xrightarrow{P} 0$.

Claim 4: $\frac{1}{\sqrt{n}} D_{\theta_0}\ell_n(\boldsymbol{\xi}\mid\theta) \xrightarrow{d} N(0,\mathcal{I}_\xi(\theta_0))$

First note that

$$\frac{1}{n} D_{\theta_0}\ell_n(\boldsymbol{\xi}\mid\theta) = \frac{1}{n}\sum_{i=1}^{n} D_{\theta_0}\log f(\xi_i\mid\theta) \xrightarrow{P} \mathbf{E}_{\theta_0}[D_{\theta_0}\log f(\xi\mid\theta)]$$

since we have an i.i.d. sum and we can apply the Weak Law of Large Numbers. Because we assume we can exchange expectations and derivatives for any partial derivative

$$\mathbf{E}_{\theta_0}[\frac{\partial}{\partial\theta_i}\log f(\xi_i\mid\theta)] = \int \frac{\partial}{\partial\theta_i}\log f(x\mid\theta) f(x\mid\theta_0)dx = \int \frac{\partial}{\partial\theta_i} f(x\mid\theta_0)\,dx = \frac{\partial}{\partial\theta_i}\int f(x\mid\theta_0)\,dx = 0$$

and thus we conclude $\frac{1}{n} D_{\theta_0}\ell_n(\boldsymbol{\xi}\mid\theta) \xrightarrow{P} 0$. We can also calculate the covariance matrix of the random variable $D_{\theta_0}\log f(\xi\mid\theta)$ as $\mathcal{I}_\xi(\theta_0)$.

Now we simply apply the multivariate Central Limit Theorem and the Claim is proven.

Claim 5: $\frac{1}{\sqrt{n}} D_{\theta_0}\ell_n(\boldsymbol{\xi}\mid\theta) - \sqrt{n}\mathcal{I}_\xi(\theta_0)\cdot(\hat{\theta}_n-\theta_0) \xrightarrow{P} 0$

We already know from Claim 2 that $\frac{1}{\sqrt{n}}(D_{\theta_0}\ell_n(\boldsymbol{\xi}\mid\theta) + \Delta_n\cdot(\hat{\theta}_n-\theta_0)) \xrightarrow{P} 0$ so it suffices to show that $\frac{1}{\sqrt{n}}\Delta_n\cdot(\hat{\theta}_n-\theta_0) + \sqrt{n}\mathcal{I}_\xi(\theta_0)\cdot(\hat{\theta}_n-\theta_0) \xrightarrow{P} 0$ as well.

By Claim 4 and Lemma 15.7, we know that $\frac{1}{\sqrt{n}} D_{\theta_0}\ell_n(\boldsymbol{\xi}\mid\theta)$ is tight. Together with Claim 2 this tells us that $\frac{1}{\sqrt{n}}\Delta_n\cdot(\hat{\theta}_n-\theta_0)$ is $o_P(1)+O_P(1)$ hence is tight as well (Lemma 15.9). Claim 3 and the invertibility of $\mathcal{I}_\xi(\theta_0)$ allows us to apply Lemma 15.10 to conclude that $\frac{1}{\sqrt{n}}\Delta_n\cdot(\hat{\theta}_n-\theta_0)$ is tight. Now by Claim 3 and Lemma 15.9 we can conclude that $(\frac{1}{n}\Delta_n + \mathcal{I}_\xi(\theta_0))\cdot\sqrt{n}(\hat{\theta}_n-\theta_0) \xrightarrow{P} 0$ as required.

Now when we combine Claim 4 and Claim 5 with Slutsky's Lemma (Theorem 8.44) we conclude that $\mathcal{I}_\xi(\theta_0)\cdot\sqrt{n}(\hat{\theta}_n-\theta_0) \xrightarrow{d} N(0,\mathcal{I}_\xi(\theta_0)^{-1})$. Because $\mathcal{I}_\xi(\theta_0)$ is

invertible and matrix multiplication is continuous, the Continuous Mapping Theorem allows us to conclude $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{I}_\xi(\theta_0)^{-1} N(0, \mathcal{I}_\xi(\theta_0)) = N(0, \mathcal{I}_\xi(\theta_0)^{-1})$. and we are done. □

As a side effect of having shown that an MLE may be asymptotically normal we computed its asymptotic variance. Now it is intuitively clear that given two estimators that are equal in every other way the one with a smaller variance is to be preferred. So a natural question to ask is whether a variance of $\mathcal{I}_\xi(\theta)^{-1}$ is a good by some objective standard. It is in fact optimal.

Theorem 15.25 (Cramer-Rao Lower Bound)**.** *blah blah*

TODO: Binomial estimation Ideas: Frequentist vs. Bayesian. Two sampling approaches: sample fixed n vs. sequentially sample till n successes. Same means but different variances in frequentist approaches (failure of the likelihood principle) but same in Bayesian. The normal approximation and confidence intervals. Discuss issues with coverage. Ratio of binomial (e.g. Koopman and the Bayesian approach).

TODO: Maybe a good idea to cover logistic regression as an application of MLE. Expressing regression as an MLE: requires a distribution assumption on the residual and then regression becomes a location scale family. I don't see that the standard proofs of consistency and normality work in these cases though (since the observations now are independent but have differing distributions..) I think this is an accurate state of affairs; there are direct proofs of MLE asymptotic properties for GLMs (and I suppose GAMs). See also Hjort and Pollard, "Asymptotics for minimisers of convex processes" As for intuition about why i.i.d. should not be necessary to prove asymptotic results recall that the Weak Law of Large Numbers doesn't require i.i.d. but only uniform integrability and that the Lindeberg C.L.T. applies without full blown i.i.d. It'll be an interesting exercise to see how the asymptotic theory of logistic regression unfolds.

15.2. **Logistic Regression.** To motivate the logistic regression, assume that we have a binomial random variable $y \sim B(n, p)$ and consider the maximum likelihood estimate of the parameter p. Introduce the log odds $\theta = logit(p) = \ln(p/(1-p))$ rewrite the binomial distribution in terms of θ.

$$\binom{n}{m} p^m (1-p)^{n-m} = e^{\ln(\binom{n}{m})} e^{\ln(p^m)} e^{\ln((1-p)^{n-m}} \tag{5}$$

$$= e^{\ln(\binom{n}{m}) + m\ln(p/(1-p)) + n\ln(1-p)} \tag{6}$$

$$= e^{\ln(\binom{n}{m}) + m\ln(p/(1-p)) - n\ln(1+p/(1-p))} \tag{7}$$

$$= e^{\ln(\binom{n}{m}) + m\theta - n\ln(1+e^\theta)} \tag{8}$$

This allows us to write the loglikelihood function in terms of the parameter θ as:

$$l(\theta; y) = y\theta - n\ln(1+e^\theta) + \ln\binom{n}{y}$$

and then it is easy to get the score and information functions

$$s(\theta; y) = \frac{\partial}{\partial\theta} l(\theta; y) = y - \frac{ne^\theta}{1+e^\theta} = y - np \tag{9}$$

$$i(\theta; y) = -\frac{\partial}{\partial\theta} s(\theta; y) = np(1-p) \tag{10}$$

16. Brownian Motion

We begin by studying the one dimensional version of Brownian motion.

Definition 16.1. A real-valued stochastic process B_t on $[0,\infty)$ is said to be a *Brownian motion* at $x \in \mathbb{R}$ if

(i) $B(0) = x$
(ii) For all times $0 \leq t_1 \leq t_2 \leq \cdots \leq t_n$ the increments $B_{t_2} - B_{t_1}, B_{t_3} - B_{t_2}, \ldots, B_{t_n} - B_{t_{n-1}}$ are independent random variables
(iii) For all $0 \leq s < t$, the increment $B_t - B_s$ is normally distributed with expectation zero and variance $t - s$.
(iv) Almost surely the sample path $B(t)$ is continuous.

The existence of Brownian motion is a non-trivial fact that was first proved by Norbert Weiner. Here we present a construction by Paul Levy whose details are worth understanding because many properties of Brownian motion follow from them.

Theorem 16.2. *Standard Brownian motion exists.*

Proof. Before we construct Brownian motion on the entire real line, we construct it on the interval $[0,1]$ (that is to say we only construct the values $B(t)$ for $t \in [0,1]$). To motivate the construction of Brownian motion, we take as our driving goals the fact that we have to construct a continuous random path $B(x)$ for which the distribution of $B(x)$ for fixed $x \in [0,1]$ is $N(0,x)$. The approach to the construction is to proceed iteratively such that at stage n of the iteration we have a piecewise linear approximation $B_n(x)$ with the distribution of $B_n(x)$ being $N(0,x)$ at the points $x = 0, 1/2^n, \ldots, 1$. The set of rational numbers of the form $\frac{k}{2^n}$ for $n \geq 0$ and $0 \leq k \leq 2^n$ is known as the *dyadic rationals* in $[0,1]$. We will sometime have need for the notation

$$\mathcal{D}_n = \{\frac{k}{2^n} \mid 0 \leq k \leq 2^n\}$$

and $\mathcal{D} = \cup_{n=0}^{\infty} \mathcal{D}_n$ when discussing the dyadic rationals. To support the construction, we need a probability space which we assume to be $([0,1], \mathcal{B}([0,1]), \lambda)$. As a concrete source of randomness, for each $d \in \mathcal{D}$ let Z_d be an $N(0,1)$ random variable with the Z_d independent (we may do this by Lemma 7.32).

It is worth walking through the first couple of iterations in rather gory detail to reinforce the idea and to convince the reader that the construction really is determined by the vague prescription given above. So our first goal is to construct a random piecewise linear path that is constant at $x = 0$ and has distribution $N(0,1)$ at $x = 1$. The simplest idea turns out to be the right one to get started: define $B_0(x) = xZ_1$. Then $\mathbf{Var}(B_0(x)) = x^2$ which is correct for $x \in \{0,1\}$ but nowhere in between. The critical point is the $x^2 < x$ for all $x \in (0,1)$ so we have *too little* variance. Getting a bit more variance is easy whereas we'd be rather doomed if we already had too much.

So recall the next step was to get the correct variance at the points $\{0, 1/2, 1\}$ not just at the points $\{0,1\}$. By the above, $\mathbf{Var}(B_0(1/2)) = 1/4$ but we require that $B_1(1/2) = 1/2$ so we need to add a random variable with distribution $N(0,1/4)$ at $x = 1/2$ satisfy our goal. But since we had the correct variance at $0, 1$ we have make sure not to add any more at either of those points. This motivates the introduction

of the function

$$\Delta(x) = \begin{cases} 2x & \text{for } 0 \le x \le \frac{1}{2} \\ 2 - 2x & \text{for } \frac{1}{2} < x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Now if we define $B_1(x) = B_0(x) + \frac{1}{2}\Delta(x)Z_{1/2}$ the we see that $B_1(1/2)$ is a sum of two $N(0, 1/4)$ random variables hence is $N(0, 1/2)$ as desired. Because $\Delta(0) = \Delta(1) = 0$, we have $B_1(0) = B_0(0)$ and $B_1(1) = B_0(1)$ so these two are still in good shape.

TODO: Make the following into an exercise. Just to turn the crank one more time, by the definition of $B_1(x)$ we can easily see that since in general $B_1(x)$ is an $N(0, x^2 + \frac{1}{2}\Delta_{0,0}(x))$ random variable,

$$\begin{aligned} \mathbf{Var}\left(B_1(1/4)\right) &= \frac{1}{16} + \frac{1}{16} = 1/8 = 1/4 - 1/8 \\ \mathbf{Var}\left(B_1(3/4)\right) &= \frac{9}{16} + \frac{1}{16} = 5/8 = 3/4 - 1/8 \end{aligned}$$

so in both cases we need to add a variance of $1/8$ at the points $\{1/4, 3/4\}$ without changing things at $\{0, 1/2, 1\}$. Mimicing what we have already done, we now need a "double sawtooth" to modify $B_1(x)$ into $B_2(x)$. For reasons that we'll explain later we actually break the modification into two pieces: one for the interval $(0, 1/2)$ and one for the interval $(1/2, 1)$. So define,

$$\Delta_{1,0}(x) = \Delta(2x) = \begin{cases} 4x & \text{for } 0 \le x \le \frac{1}{4} \\ 2 - 4x & \text{for } \frac{1}{4} < x \le \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

and

$$\Delta_{1,1}(x) = \Delta(2x - 1) = \begin{cases} 4x - 2 & \text{for } \frac{1}{2} \le x \le \frac{3}{4} \\ 4 - 4x & \text{for } \frac{3}{4} < x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Now if we define $B_2(x) = B_1(x) + \frac{1}{\sqrt{8}}(\Delta_{1,0}(x)Z_{/1/4} + \Delta_{1,1}(x)Z_{3/4})$, then we have added the appropriate variance of $1/8$ at $x = 1/4$ and $x = 3/4$.

To state the general construction, we first generalize the definition of our sawtooth functions. For $n > 0$ and $k = 0, \cdots, 2^n - 1$, we define

$$\Delta_{n,k}(x) = \Delta(2^n x - k) = \begin{cases} 2^{n+1}x - 2k & \text{for } \frac{2k}{2^{n+1}} \le x \le \frac{2k+1}{2^{n+1}} \\ 2k + 2 - 2^{n+1}x & \text{for } \frac{2k+1}{2^{n+1}} < x \le \frac{2k+2}{2^{n+1}} \\ 0 & \text{otherwise} \end{cases}$$

With the definition we can complete the induction definition. So our definition of $B_n(x)$ can be completed. We point out that $\Delta_{0,0}(x) = \Delta(x)$ so the definition below

is compatible with our definition of $B_1(x)$ and $B_2(x)$ above:

$$\begin{aligned} B_0(x) &= xZ_1 \\ B_n(x) &= B_{n-1}(x) + \frac{1}{\sqrt{2^{n+1}}} \sum_{k=0}^{2^{n-1}-1} \Delta_{n-1,k}(x) Z_{\frac{2k+1}{2^n}} \\ &= B_0(x) + \sum_{j=0}^{n-1} \frac{1}{\sqrt{2^{j+2}}} \sum_{k=0}^{2^j-1} \Delta_{j,k}(x) Z_{\frac{2k+1}{2^{j+1}}} \qquad \text{for } n > 0 \end{aligned}$$

We will sometimes find it convenient to use the definition

$$F_n(x) = \frac{1}{\sqrt{2^{n+2}}} \sum_{k=0}^{2^n-1} \Delta_{n,k}(x) Z_{\frac{2k+1}{2^{n+1}}}$$

so that we may write

$$B_n(x) = B_0(x) + \sum_{j=0}^{n-1} F_n(x)$$

$$B(x) = B_0(x) + \sum_{n=0}^{\infty} F_n(x)$$

There are host of important facts about the $B_n(x)$ and $B(x)$ that proceed to prove. No individual fact is difficult to prove but there are many of them to keep track of.

Lemma 16.3. *The following are true:*

(i) *$B_n(x)$ is linear on every interval $[\frac{k}{2^n}, \frac{k+1}{2^n}]$ for $k = 0, \ldots, 2^n - 1$.*

(ii) *For every $n \geq 0$, and $0 < 2k + 1 < 2^n$,*

$$B(\frac{2k+1}{2^n}) = \frac{1}{2}(B(\frac{2k}{2^n}) + B(\frac{2k+2}{2^n})) + \frac{1}{\sqrt{2^{n+1}}} Z_{\frac{2k+1}{2^n}}$$

(iii) *For every $n \geq 0$ and every pair $0 \leq j < k \leq 2^n$, $B(k/2^n) - B(j/2^n)$ is an $N(0, (k-j)/2^n)$ random variable. Furthermore for $0 \leq j < k \leq l < m \leq 2^n$, the increments $B(k/2^n) - B(j/2^n)$ and $B(m/2^n) - B(l/2^n)$ are independent.*

Proof. FIrst we prove (i). This follows from a simple induction. It is clear for $B_0(x)$. For $B_{n+1}(x)$ we are adding multiples of the functions $\Delta_{n,k}(x)$ each of which is linear on intervals of the form $[\frac{k}{2^{n+1}}, \frac{k+1}{2^{n+1}}]$.

Next we prove (ii). This follows from the fact that $B(\frac{2k+1}{2^n}) = B_n(\frac{2k+1}{2^n})$, the definition of $B_n(x)$ and the linearity of $B_{n-1}(x)$ on the interval $[\frac{k}{2^{n-1}}, \frac{k+1}{2^{n-1}}]$.

To see (iii) first note that it suffices to prove this for increments $j + 1 = k$ and $l + 1 = m$. For if we have proven that then we can write a general increment as a sum of independent increments of the former form. We proceed by induction on n. The case $n = 0$ is trivial because the only non-trivial increment is the $N(0, 1)$ random variable $B(1) - B(0) = Z_1$. Now consider the case for $n > 0$. To see this first we consider "adjacent" increments of the form $B((2k + 1)/2^n) - B(2k/2^n)$ and $B((2k + 2)/2^n) - B((2k + 1)/2^n)$. Here we use the formula $B((2k + 1)/2^n) =$

$\frac{B((2k+2)/2^n)+B(2k/2^n)}{2} + \frac{1}{\sqrt{2^{n+1}}} Z_{(2k+1)/2^n}$ to see

$$B((2k+1)/2^n) - B(2k/2^n) = \frac{B((2k+2)/2^n) - B(2k/2^n)}{2} + \frac{1}{\sqrt{2^{n+1}}} Z_{(2k+1)/2^n}$$
$$B((2k+2)/2^n) - B((2k+1)/2^n) = \frac{B((2k+2)/2^n) - B(2k/2^n)}{2} - \frac{1}{\sqrt{2^{n+1}}} Z_{(2k+1)/2^n}$$

The random variables $B((2k+2)/2^n)$ and $B(2k/2^n)$ only depend on the Z_d for $d \in \mathcal{D}_{n-1}$ and therefore $Z_{(2k+1)/2^n}$ is independent of both. The induction hypothesis is that $B((2k+2)/2^n) - B(2k/2^n)$ is an $N(0, \frac{1}{2^{n-1}})$ random variable therefore $\frac{B((2k+2)/2^n)-B(2k/2^n)}{2}$ is $N(0, \frac{1}{2^{n+1}})$. But both $\pm\frac{1}{\sqrt{2^{n+1}}} Z_{(2k+1)/2^n}$ are also $N(0, \frac{1}{2^{n+1}})$ so we've expressed the increments as a sum of two independent $N(0, \frac{1}{2^{n+1}})$ random variable proving that each is $N(0, \frac{1}{2^n})$. Furthermore the increments are independent. Because we know they are normal it suffices to show they are uncorrelated which is a simple computation using the formulae above and the induction hypothesis

$$\begin{aligned}
&\mathbf{E}\left[(B((2k+1)/2^n) - B(2k/2^n))(B((2k+2)/2^n) - B((2k+1)/2^n))\right] \\
&= \mathbf{E}\left[(\frac{B((2k+2)/2^n) - B(2k/2^n)}{2} + \frac{1}{\sqrt{2^{n+1}}} Z_{(2k+1)/2^n})(\frac{B((2k+2)/2^n) - B(2k/2^n)}{2} - \frac{1}{\sqrt{2^{n+1}}} Z_{(2k+1)/2^n})\right] \\
&= \frac{1}{4}\mathbf{E}\left[(B((2k+2)/2^n) - B(2k/2^n))^2\right] - \frac{1}{2^{n+1}} \\
&= \frac{1}{4}\frac{1}{2^{n-1}} - \frac{1}{2^{n+1}} = 0
\end{aligned}$$

It remains to show the independence of increments $B((k+1)/2^n) - B(k/2^n)$ and $B((j+1)/2^n) - B(j/2^n)$ with $0 \le j < k \le 2^n$. In a similar way to the case above we know that by using the result (ii) we can see that for $0 \le k < 2^n$,

$$B((k+1)/2^n) - B(k/2^n) = \begin{cases} \frac{B((k+1)/2^n)-B((k-1)/2^n)}{2} - \frac{1}{\sqrt{2^{n+1}}} Z_{k/2^n} & k \text{ is odd} \\ \frac{B((k+2)/2^n)-B(k/2^n)}{2} + \frac{1}{\sqrt{2^{n+1}}} Z_{(k+1)/2^n} & k \text{ is even} \end{cases}$$

If we assume that we are not in the case already proven then we are either assuming that $j+1 \neq k$ or k is even. The upshot is that we can write each increment of length $\frac{1}{2^n}$ as a sum of an increment of length $\frac{1}{2^{n-1}}$ and an independent $N(0, \frac{1}{2^{n+1}})$ random variable. The increments of length $\frac{1}{2^{n-1}}$ are independent by the induction hypothesis and therefore the original increments are seen to be independent. TODO: Make this more precise. □

We make the following claim about $B_n(x)$: for $\frac{k}{2^n} \le x \le \frac{k+1}{2^n}$ and $0 \le k < 2^n$, we have $\mathbf{Var}(B_n(x)) = 2^n(x - \frac{k}{2^n})^2 + \frac{k}{2^n}$. We use an induction to prove the claim. Note that the claim is easily seen to be true for $n = 0$ (it reduces to earlier observation that $\mathbf{Var}(B_0(x)) = x^2$). Now assuming that it is true for n we extend to $n+1$. Pick an interval $[\frac{k}{2^n}, \frac{k+1}{2^n}]$ and consider passing from $B_n(x)$ to $B_{n+1}(x)$ on the interval. There are two subcases corresponding to the subinteval $[\frac{k}{2^n}, \frac{2k+1}{2^{n+1}}]$ and the subinterval $[\frac{2k+1}{2^{n+1}}, \frac{k+1}{2^n}]$.

On the first subinterval, by the definition of $B_{n+1}(x)$ we are adding to $B_n(x)$ a normal random variable with variance $\left(\frac{1}{\sqrt{2^{n+2}}}\Delta_{n,k}(x)\right)^2 = 2^n(x - \frac{k}{2^n})^2$. So at such

an x, $B_{n+1}(x)$ is normal with variance

$$\begin{aligned}\mathbf{Var}\left(B_{n+1}(x)\right) &= \mathbf{Var}\left(B_n(x)\right) + 2^n(x - \frac{k}{2^n})^2 \\ &= 2^n(x - \frac{k}{2^n})^2 + \frac{k}{2^n} + 2^n(x - \frac{k}{2^n})^2 \\ &= 2^{n+1}(x - \frac{k}{2^n})^2 + \frac{k}{2^n}\end{aligned}$$

On the second subinterval, by the definition of $B_{n+1}(x)$ we are adding to $B_n(x)$ a normal random variable with variance $2^n(x - \frac{k+1}{2^n})^2$. So at such an x, $B_{n+1}(x)$ is normal with variance

$$\begin{aligned}\mathbf{Var}\left(B_{n+1}(x)\right) &= \mathbf{Var}\left(B_n(x)\right) + 2^n(x - \frac{k+1}{2^n})^2 \\ &= 2^n(x - \frac{k}{2^n})^2 + \frac{k}{2^n} + 2^n(x - \frac{k+1}{2^n})^2 \\ &= 2^n\left[(x - \frac{2k+1}{2^{n+1}})^2 + \frac{1}{2^{n+1}}(x - \frac{2k+1}{2^{n+1}}) + \frac{1}{2^{2n+2}}\right] + \\ &2^n\left[(x - \frac{2k+1}{2^{n+1}})^2 - \frac{1}{2^{n+1}}(x - \frac{2k+1}{2^{n+1}}) + \frac{1}{2^{2n+2}}\right] + \frac{k}{2^n} \\ &= 2^{n+1}(x - \frac{2k+1}{2^{n+1}})^2 + \frac{2k+1}{2^{n+1}}\end{aligned}$$

which verifies the claim.

We reiterate the importance of this fact is that the approximate path $B_n(x)$ has the variance x (the "correct" variance for a Brownian path) at all $x = 0, \frac{1}{2^n}, \ldots, 1$, so that as n increases $B_n(x)$ has the correct variance on an increasing fine grid in $[0, 1]$. In between the points of the grid, the variance of $B_n(x)$ is a quadratic function of x that is strictly less than x.

Having defined the series expansion of our candidate Brownian motion, the first order of business is to validate that it converges almost surely. To show convergence we need to make sure that the increments we add at each n get small fast enough; these increments are multiples of independent standard normal random variables. Convergence will follow if we can get an appropriate almost sure bound on a random sample from a sequence of independent standard normals.

To see this we start with a tail bound for an $N(0, 1)$ distribution.

$$\begin{aligned}\mathbf{P}\{|Z_d| \geq \lambda\} &= \frac{2}{\sqrt{2\pi}} \int_\lambda^\infty e^{\frac{-u^2}{2}}\, du \\ &\leq \frac{2}{\sqrt{2\pi}} \int_\lambda^\infty \frac{u}{\lambda} e^{\frac{-u^2}{2}}\, du \\ &= \frac{1}{\lambda\sqrt{2\pi}} e^{\frac{-\lambda^2}{2}}\end{aligned}$$

so if we pick any constant $c > 1$ and $n > 0$, then

$$\mathbf{P}\{|Z_d| \geq c\sqrt{n}\} \leq \frac{1}{c\sqrt{2\pi n}} e^{\frac{-c^2 n}{2}} \leq e^{\frac{-c^2 n}{2}}$$

Now using this bound, we see that

$$\begin{aligned}\sum_{n=0}^{\infty} \mathbf{P}\{\text{there exists } d \in \mathcal{D}_n \text{ such that } |Z_d| \geq c\sqrt{n}\} &\leq \sum_{n=0}^{\infty} \sum_{d \in \mathcal{D}_n} \mathbf{P}\{|Z_d| \geq c\sqrt{n}\} \\ &\leq \sum_{n=0}^{\infty} 2^n e^{\frac{-c^2 n}{2}} \\ &= \sum_{n=0}^{\infty} e^{-n(c^2 - 2\ln 2)/2}\end{aligned}$$

which converges if $c > \sqrt{2 \ln 2}$. Picking such a c, we apply the Borel Cantelli Theorem to conclude that

$$\mathbf{P}\{\text{there exists } d \in \mathcal{D}_n \text{ such that } |Z_d| \geq c\sqrt{n} \text{ i.o.}\} = 0$$

and therefore for almost all $\omega \in \Omega$ there exists $N_\omega > 0$ such that $|Z_d| < c\sqrt{n}$ for all $n > N_\omega$ and $d \in \mathcal{D}_n$ and by definition of $F_n(x)$, we have $\|F_n\|_\infty \leq 2^{-(n+2)/2} c\sqrt{n}$ which shows that $\sum_{n=0}^{\infty} F_n(x)$ converges absolutely and uniformly in x. Because each $F_n(x)$ is a continuous function, uniform convergence of the series implies $B(x) = B_0(x) + \sum_{n=0}^{\infty} F_n(x)$ is continuous as well (Theorem 2.36).

TODO: Show that for every $x \in [0,1]$, $B(x)$ is integrable and has finite variance. Not sure we need this because we'll prove a stronger statement below.

The next step is to validate that $B(x)$ has independent Gaussian increments. TODO: Show that we have Gaussian increments, independent increments, zero mean and proper variance/covariance. The first step is to note that we have already proven that increments at dyadic rational numbers are independent and Gaussian. But we have also shown that $B(x)$ is almost surely continuous so we may approximate arbitrary increments by those at dyadic rationals.

Suppose we are given $0 \leq x_1 < x_2 < \cdots < x_n \leq 1$. By the density of the dyadic rationals we can find sequences $x_{j,m}$ of dyadic rationals with $x_{j-1} < x_{j,m} \leq x_j$ such that $\lim_{m\to\infty} x_{j,m} = x_j$ (in the case $j = 1$, we only require $0 \leq x_{1,m} \leq x_1$). By almost sure continuity of $B(x)$ we know that $B(x_{j,m}) - B(x_{j-1,m})$ converges to $B(x_j) - B(x_{j-1})$ for $1 < j \leq n$. Moreover we know that

$$\lim_{m\to\infty} \mathbf{E}\left[B(x_{j,m}) - B(x_{j-1,m})\right] = 0$$

and

$$\begin{aligned}\lim_{m\to\infty} \mathbf{Cov}\left(B(x_{j,m}) - B(x_{j-1,m}), B(x_{i,m}) - B(x_{i-1,m})\right) &= \delta_{i,j} \lim_{m\to\infty} (x_{i,m} - x_{i-1,m}) \\ &= \delta_{i,j}(x_i - x_{i-1})\end{aligned}$$

and therefore by Lemma 10.19 we know that the $B(x_j) - B(x_{j-1})$ are independent $N(0, x_j - x_{j-1})$ random variables and we are done. □

TODO: Note the connection of the construction to wavelets. What we are doing here is expressing the Brownian motion as a linear combination of integrals of the Haar wavelet basis (in some sense we are integrating "white noise" which is called an *isonormal process* in the mathematical literature these days). Note that the such a form for a Brownian motion can be anticipated by examining the covariance of Brownian motion (see Steele).

TODO: Modulus of continuity of Brownian paths; Holder continuity and nowhere differentiability.

TODO: Some of these proofs use the specifics of the Levy construction of Brownian motion and not just the defining properties of Brownian motion. In what way is this justified; i.e. to what extent is the Levy construction unique? The answer to this question is that Wiener measure on $C[0,\infty)$ is uniquely defined by its finite dimensional distributions.

Theorem 16.4 (Markov Property of Brownian motion)**.** *Let B_t be a Brownian motion starting at x and let $s \geq 0$. Then $B_{t+s} - B_s$ is a Brownian motion starting at 0 that is independent of B_t for $0 \leq t \leq s$.*

Proof. The fact that $B_{t+s} - B_s$ is a Brownian motion follows from the fact that increments of the translated process are increments of the original Brownian motion. More precisely if we select $t_1 \leq \cdots \leq t_n$ then each $(B_{t_{i+1}+s} - B_s) - (B_{t_i+s} - B_s) = B_{t_{i+1}+s} - B_{t_i+s}$ and therefore they are we can conclude they are jointly independent Gaussian with variance $(t_{i+1} - s) - (t_i - s) = t_{i+1} - t_i$.

The independence of the Brownian motion $B_{t+s} - B_s$ and B_t for $0 \leq t \leq s$ follows from the property of independent increments. Specifically, by the montone class argument of Lemma 7.15 we know that it is sufficient to show independence for finite sets $\{B_{t_1+s} - B_s, \ldots, B_{t_n+s} - B_s\}$ and $\{B_{s_1}, \ldots, B_{s_m}\}$ for all finite sequence of times $s_1 \leq \cdots \leq s_m \leq s$ and $0 \leq t_1 \leq \cdots \leq t_n$. Observe that for any measurable random vectors $\xi_1, \ldots, \xi_n$ we have $\sigma(\xi_1, \xi_2 - \xi_1, \ldots, \xi_n - \xi_1) = \sigma(\xi_1, \xi_2 - \xi_1, \ldots, \xi_n - \xi_{n-1})$ (to see this note that every term on the left is a sum of terms on the right and vice versa). In particular by independence of increments and Lemma 7.13 we know that $\sigma(B_{t_1+s} - B_s, \ldots, B_{t_n+s} - B_{t_{n-1}})$ and $\sigma(B_{s_1} - B_0, \ldots, B_{s_m} - B_{s_{m-1}})$ are independent which establishes the result by applying the previous observation. □

16.1. **Skorohod Embedding and Donsker's Theorem.** TODO: Clarify what we mean when we say a Brownian motion is independent of a σ-algebra.

TODO: Introduce the right continuous filtration $\mathcal{F}_t^+$

TODO: Strong Markov Property

Theorem 16.5 (Markov Property)**.** *Let B_t be a Brownian motion then for any $s \geq 0$ the process $B_t^* = B_{t+s} - B_s$ is a standard Brownian motion independent of $\{B_t \mid 0 \leq t \leq s\}$.*

Proof. We simply walk through the defining properties of Brownian motion:

(i) Clearly $B_0^* = B_s - B_s = 0$.

(ii) For any $0 \leq t_1 \leq \cdots \leq t_n$ the increment $B_{t_j}^* - B_{t_{j-1}}^* = B_{s+t_j}^* - B_{s+t_{j-1}}^*$ therefore the independence of the increments $B_{t_2}^* - B_{t_1}^*, \ldots, B_{t_n}^* - B_{t_{n-1}}^*$ follows from the fact that B_t is a Brownian motion

(iii) By the same argument as in (ii), for any $t_1 < t_2$ we have $B_{t_2}^* - B_{t_1}^* = B_{s+t_2} - B_{s+t_1}$ is normallly distributed with mean 0 and variance $(s + t_2) - (s + t_1) = t_2 - t_1$.

(iv) The paths $B_t^* = B_{s+t}$ are almost surely continuous because B_t is a Brownian motion

To see the independence statement pick $0 \leq t_1 \leq \cdots \leq t_n$ and $0 \leq s_1 \leq \cdots \leq s_m \leq s$

TODO: Finish □

Theorem 16.6 (Strong Markov Property)**.** *Let B_t be a Brownian motion and let τ be an almost surely finite $\mathcal{F}^+$-optional time, then $B_t^* = B_{\tau+t} - B_\tau$ is a standard Brownian motion independent of $\mathcal{F}_\tau^+$.*

Proof. □

The following corollary of the strong Markov property turns out to be a very useful tool in calculating the distributions of various functions of Brownian motion. It is called the reflection principle because it shows that if one runs a Brownian motion up to an optional time τ and then reverses the sign of all subsequent increments (reflecting the graph of the Brownian motion with respect to the line $y = \tau$) then the resulting process has same distribution. TODO: Draw a picture illustrating the geometry of reflection.

Lemma 16.7 (Reflection Principle)**.** *Let B_t be a Brownian motion and let τ be an optional time then*

$$B'_t = B_{\tau \wedge t} - (B_t - B_{\tau \wedge t}) = \begin{cases} B_t & \text{when } t \leq \tau \\ 2B_\tau - B_t & \text{when } t > \tau \end{cases}$$

is a Brownian motion with the same distribution as B_t.

Proof. □

TODO: Skorohod Embedding

Donsker's Theorem states roughly that Brownian motion can be approximated in distribution by a suitably rescaled random walk. Moreover it states that essentially all possible random walks that one might expect could approximate Brownian motion in fact do. This fact shows that Brownian motion is analogous to standard normal distributions and Donsker's Theorem is often referred to as the Functional Central Limit Theorem.

Theorem 16.8 (Donsker's Invariance Principle)**.** *Suppose we are given an i.i.d. sequence of random variables $\xi_1, \xi_2, \ldots$ such that $\boldsymbol{E}[\xi_n] = 0$ and $\boldsymbol{Var}(\xi_n) = 1$ for all $n \in \mathbb{N}$. Define the random walk*

$$S_n = \sum_{j=1}^{n} \xi_j$$

its linear interpolation

$$S(t) = S_{\lfloor t \rfloor} + (t - \lfloor t \rfloor)(S_{\lfloor t \rfloor + 1} - S_{\lfloor t \rfloor})$$

and its rescaling from the interval $[0, n]$ to $[0, 1]$

$$S^*_n(t) = \frac{1}{\sqrt{n}} S(nt) \qquad \text{for } t \in [0, 1]$$

On the space $C[0, 1]$ with the uniform norm, the sequence $S^*_n(t)$ converges in distribution to the standard Brownian motion.

Proof. TODO □

TODO: Extension of Donsker's Theorem to convergence of errors of empirical distributions to Brownian bridge. This may be harder because the convergence takes place not in the separable space $C[0, 1]$ but rather the space of cadlag functions (which is only separable under the Skorohod topology). The alternative here is presumably to use the generalized form of weak convergence from empirical process theory.

17. Markov Processes

TODO: Thinking about Markov processes as dynamical/deterministic systems with noise.

17.1. **Markov Processes.**

Definition 17.1. Let X be a process in $(S, \mathcal{S})$ with time scale T which is adapted to a filtration $\mathcal{F}_t$. We say that X has the *Markov property* if $\mathcal{F}_s \perp\!\!\!\perp_{X_s} X_t$ for all $s \leq t \in T$.

Given any process that satisfies the Markov property it is not hard to show using properties of conditional independence that it automatically satisfies a seemingly stronger condition

Lemma 17.2 (Extended Markov Property)**.** *Let X be a process that satsifies the Markov property then $\mathcal{F}_t \perp\!\!\!\perp_{X_t} \sigma(\bigvee_{u \geq t} X_u)$ for all $t \in T$.*

Proof. Let $t_0 \leq t_1 \leq \cdots$ with $t_i \in T$. By the Markov property we know for each $0 \leq n$ that $\mathcal{F}_{t_n} \perp\!\!\!\perp_{X_{t_n}} X_{t_{n+1}}$. Because X is adapted to $\mathcal{F}$, we know that X_{t_m} is $\mathcal{F}_{t_n}$-measurable for $m \leq n$ and therefore $\sigma(X_{t_0}, \ldots, X_{t_{n-1}}, \mathcal{F}_{t_n}) \perp\!\!\!\perp_{X_{t_n}} X_{t_{n+1}}$. By Lemma 11.16 we conclude that $\mathcal{F}_{t_n} \perp\!\!\!\perp_{X_{t_0}, \ldots, X_{t_n}} X_{t_{n+1}}$ for all $n \geq 0$; because $\mathcal{F}_{t_0} \subset \mathcal{F}_{t_n}$ we get $\mathcal{F}_{t_0} \perp\!\!\!\perp_{X_{t_0}, \ldots, X_{t_n}} X_{t_{n+1}}$ for all $n \geq 0$. Another application of Lemma 11.16 shows that $\mathcal{F}_{t_0} \perp\!\!\!\perp_{X_{t_0}} \sigma(X_{t_1}, X_{t_2}, \ldots)$.

Since the union of the σ-algebras $\sigma(X_{t_1}, X_{t_2}, \ldots)$ for all $t_0 \leq t_1 \leq \cdots$ is clearly a π-system that generates $\sigma(\bigvee_{u \geq t_0} X_u)$, the result follows by montone classes (specifically Lemma 11.14). □

TODO: Introduce the example of Markov Chains here as it is quite a bit simpler and helps the understanding of the abstract case quite a bit.

We know make a regularity assumption that for each pair $s, t \in T$ with $s \leq t$, we have a probability kernel $\mu_{s,t} : S \times \mathcal{S} \to \mathbb{R}$ such that for every $A \in \mathcal{S}$

$$\mu_{s,t}(X_s, A) = \mathbf{P}\{X_t \in A \mid X_s\} = \mathbf{P}\{X_t \in A \mid \mathcal{F}_s\} \text{ a.s.}$$

(e.g. if S is a Borel space then this is true by Theorem 11.25). We let ν_t denote the distribution of X_t. These conditional distributions characterize the distribution of the process X itself. In particular we have the following nice formula for finite dimensional distributions of the process.

Lemma 17.3. *Let X be a stochastic process on a time scale $T \subset \mathbb{R}_+$ that has the Markov property, one dimensional distributions ν_t and transition kernels $\mu_{s,t}$. Then for all $t_0 \leq \cdots \leq t_n$ and $A \in \mathcal{S}^{\otimes n}$ we have*

$$\boldsymbol{P}\{(X_{t_1}, \ldots, X_{t_n}) \in A\} = \nu_{t_1} \otimes \mu_{t_1,t_2} \otimes \cdots \otimes \mu_{t_{n-1},t_n}(A)$$
$$\boldsymbol{P}\{(X_{t_1}, \ldots, X_{t_n}) \in A \mid \mathcal{F}_{t_0}\}(\omega) = \mu_{t_0,t_1} \otimes \cdots \otimes \mu_{t_{n-1},t_n}(X_{t_0}(\omega), A)$$

Proof. We begin by proving the first equality via induction. The case $n = 0$ is true by definition. The induction step is really just a specific case of Theorem 11.26

applied to the Markov transition kernels. Let $A \in \otimes_{i=0}^{n} \mathcal{S}$ then

$$\begin{aligned}
&\mathbf{P}\{(X_{t_0}, \cdots, X_{t_n}) \in A\} \\
&= \mathbf{E}\left[\mathbf{1}_A(X_{t_0}, \cdots, X_{t_n})\right] \\
&= \mathbf{E}\left[\int \mathbf{1}_A(X_{t_0}, \cdots, X_{t_{n-1}}, s)\, \mu_{t_{n-1},t_n}(X_{n-1}, ds)\right] \\
&= \int \left[\int \mathbf{1}_A(u_0, \cdots, u_{n-1}, s)\, \mu_{t_{n-1},t_n}(X_{n-1}, ds)\right] \nu_{t_0} \otimes \cdots \otimes \mu_{t_{n-2},t_{n-1}}(du_0, \ldots, du_{n-1}) \\
&= \nu_{t_0} \otimes \cdots \otimes \mu_{t_{n-1},t_n}(A)
\end{aligned}$$

The second equality is derived from the first. Suppose we have $A \in \mathcal{S}$ and $B \in \mathcal{S}^{\otimes n}$. Then we can compute

$$\begin{aligned}
&\mathbf{E}\left[\mathbf{1}_A(X_{t_0})\mathbf{1}_B(X_{t_1}, \ldots, X_{t_n})\right] \\
&= \nu_{t_0} \otimes \mu_{t_0,t_1} \otimes \cdots \otimes \mu_{t_{n-1},t_n}(A \times B) \\
&= \int \left[\int \mathbf{1}_B(u_1, \ldots, u_n)\mu_{t_0,t_1} \otimes \cdots \otimes \mu_{t_{n-1},t_n}(u_0, du_1, \ldots, du_n)\right] \mathbf{1}_A(u_0)\nu_{t_0}(u_0) \\
&= \mathbf{E}\left[\mu_{t_0,t_1} \otimes \cdots \otimes \mu_{t_{n-1},t_n}(X_0, B)\mathbf{1}_A(X_0)\right]
\end{aligned}$$

Now the $\sigma(X_{t_0})$-measurability of $\mu_{t_0,t_1} \otimes \cdots \otimes \mu_{t_{n-1},t_n}(X_0, B)$ tells us that

$$\mathbf{P}\{(X_{t_1}, \ldots, X_{t_n}) \in B \mid X_{t_0}\} = \mu_{t_0,t_1} \otimes \cdots \otimes \mu_{t_{n-1},t_n}(X_0, B)$$

The last thing is to show that $\mathbf{P}\{(X_{t_1}, \ldots, X_{t_n}) \in B \mid X_{t_0}\} = \mathbf{P}\{(X_{t_1}, \ldots, X_{t_n}) \in B \mid \mathcal{F}_{t_0}\}$ a.s. This follows from Lemma 17.2 since by the tower property of conditional expectations and that result for any $A \in \mathcal{S}^{\otimes n}$ and $B \in \mathcal{F}_{t_0}$

$$\begin{aligned}
\mathbf{P}\{(X_{t_1}, \ldots, X_{t_n}) \in A; B\} &= \mathbf{E}\left[\mathbf{P}\{(X_{t_1}, \ldots, X_{t_n}) \in A; B \mid X_{t_0}\}\right] \\
&= \mathbf{E}\left[\mathbf{P}\{(X_{t_1}, \ldots, X_{t_n}) \in A \mid X_{t_0}\}\mathbf{P}\{B \mid X_{t_0}\}\right] \\
&= \mathbf{E}\left[\mathbf{P}\{(X_{t_1}, \ldots, X_{t_n}) \in A \mid X_{t_0}\}\mathbf{1}_B\right]
\end{aligned}$$

so the $\mathcal{F}_{t_0}$-measurability of $\mathbf{P}\{(X_{t_1}, \ldots, X_{t_n}) \in A \mid X_{t_0}\}$ gives the result by the defining property of conditional expectations. □

A special case of the relations above should be called out as it motivates a property that will assume as part of the definition of a Markov process. But first we need a definition.

Definition 17.4. Let μ and ν be probability kernels from S to S. Then we define the probability kernel $\mu\nu$ from S to S by

$$\mu\nu(s, A) = (\mu \otimes \nu)(s, A \times S)$$

for all $s \in S$ and $A \in \mathcal{S}$.

Example 17.5. Let S be a finite set and view μ and ν as $S \times S$ matrices. Then $\mu\nu$ is just matrix multiplication:

$$\mu\nu(s, \{t\}) = \iint \mathbf{1}_{\{t\} \times S}(u, v)\nu(u, dv)\mu(s, du) = \int \nu(u, \{t\})\mu(s, du) = \sum_{u \in S} \nu_{u,t}\mu_{s,u}$$

Corollary 17.6 (Chapman-Kolmogorov Relations)**.** *Let X be a stochastic process on a time scale $T \subset \mathbb{R}$ with values in Borel space $(S, \mathcal{S})$ and suppose that X has the Markov property. Then for every $s, t, u \in T$ with $s \leq t \leq u$ we have*

$$\mu_{s,t}\mu_{t,u} = \mu_{s,u} \text{ a.s. } \nu_s$$

Proof. Since we have assume S is a Borel space we know from Theorem 11.25 that regular versions $\mu_{s,t}$ exist. By definition of $\mu_{s,t}\mu_{t,u}$, Lemma 17.3 and the uniqueness clause of Theorem 11.25

$$\begin{aligned}
\mu_{s,t}\mu_{t,u}(x, A) &= (\mu_{s,t} \otimes \mu_{t,u})(x, S \times A) \\
&= \mathbf{P}\{(X_t, X_u) \in S \times A \mid \mathcal{F}_s\} \\
&= \mathbf{P}\{X_u \in A \mid \mathcal{F}_s\} \\
&= \mathbf{P}\{X_u \in A \mid X_s\} \\
&= \mu_{s,u}(x, A) \text{ a.s. } \nu_s
\end{aligned}$$

□

The ability to derive the almost sure version of the Chapman-Kolmogorov relations is really just motivational for our purposes. In fact we will want to assume they hold identically in what follows. Absent a workable set of conditions from which we can derive this fact, we build into our definitions. Collecting all of the conditional independence and regularity properties we've identified we finally make the formal definition of a Markov process.

Definition 17.7. A *Markov Process* is a stochastic process X_t on a time scale $T \subset \mathbb{R}_+$ and a state space $(S, \mathcal{S})$ such that

(i) $\mathcal{F}_s \perp\!\!\!\perp_{X_s} X_t$ for all $s \leq t$
(ii) there exists a regular version $\mu_{s,t} : S \times \mathcal{S} \to [0, 1]$ of $\mathbf{P}\{X_t \in \cdot \mid \mathcal{F}_s\}$ for each $s \leq t$.
(iii) $\mu_{s,t}\mu_{t,u} = \mu_{s,u}$ everywhere on S for each $s \leq t \leq u$.

In lieu of general techinique for proving that a process is Markov from general principle, we give a result that shows that we can construct them from a set of transition kernels that obey the Chapman-Kolmogorov relations.

Theorem 17.8. *Suppose we are given*

(i) *a time scale starting at 0, $T \subset \mathbb{R}_+$*
(ii) *a Borel space $(S, \mathcal{S})$*
(iii) *a probability distribution ν on $(S, \mathcal{S})$*
(iv) *probability kernels $\mu_{s,t} : S \times \mathcal{S} \to [0, 1]$ for each $s \leq t \in T$ such that*

$$\mu_{s,t}\mu_{t,u} = \mu_{s,u} \text{ for all } s \leq t \leq u \in T$$

then there exists a Markov process X_t with initial distribution ν and transition kernels $\mu_{s,t}$.

Proof. TODO: Define FDDs, show consistency and use Kolmogorov extension theorem □

Definition 17.9. Suppose that a family of transition kernels $\mu_{s,t}$ is given. For a distribution ν on $(S, \mathcal{S})$, let $\mathbf{P}_\nu$ denote the distribution on S^T of the Markov process with initial distribution ν. If $\nu = \delta_x$ for some $x \in S$ then it is customary to write $\mathbf{P}_x$ instead of $\mathbf{P}_{\delta_x}$.

Lemma 17.10. *The family $\boldsymbol{P}_x$ is a kernel from S to S. Futhermore, given an initial distribution ν*

$$\boldsymbol{P}_\nu\{A\} = \int \boldsymbol{P}_x\{A\}\, d\nu(x)$$

Proof. First assume that $A = (\pi_{t_1}, \dots, \pi_{t_n})^{-1}(B)$ for some $B \in \mathcal{S}^{\otimes n}$. We can use Lemma 17.3 to compute for any ν,

$$\begin{aligned}\mathbf{P}_\nu\{A\} &= \mathbf{P}_x\{(\pi_0, \pi_{t_1}, \dots, \pi_{t_n})^{-1}(S \times B)\} \\ &= \nu \otimes \mu_{0,t_1} \otimes \cdots \mu_{t_{n-1},t_n}(S \times B) \\ &= \int \mu_{0,t_1} \otimes \cdots \mu_{t_{n-1},t_n}(x, B)\, d\nu(x)\end{aligned}$$

In particular, for $\nu = \delta_x$ we get

$$\mathbf{P}_x\{A\} = \mu_{0,t_1} \otimes \cdots \mu_{t_{n-1},t_n}(x, B)$$

which shows both that $\mathbf{P}_x\{A\}$ is measurable and that $\mathbf{P}_\nu\{A\} = \int \mathbf{P}_x\{A\}\, d\nu(x)$.

To extend to general measurable sets, we note that the set of A of the form given above is a π-system therefore we can apply Lemma 11.21 to conclude $\mathbf{P}_x$ is a kernel. Similarly we may conclude that $\mathbf{P}_\nu\{A\} = \int \mathbf{P}_x\{A\}\, d\nu(x)$ for arbitrary measurable A by the fact that probability measures are uniquely determined by their values on a generating π-system (Lemma 3.65). □

17.2. **Homogeneous Markov Processes.** We have described a relatively general version of Markov processes compared to what it needed in many applications and the goal of this section is to define the assumptions that lead to useful simplifications and to understand how to look at these simplifying assumptions from a couple of points of view.

Definition 17.11. Suppose $(S, \mathcal{S})$ is a measurable Abelian group and $\mu : S \times \mathcal{S} \to [0, 1]$ is a kernel. We say μ is *homogeneous* if for every $s \in S$ and $A \in \mathcal{S}$ we have $\mu(0, A) = \mu(s, A + s)$.

A useful observation for computing conditional expectations is that integrals are invariant under certain changes of variables.

Lemma 17.12. *Let $(S, \mathcal{S})$ be a measurable Abelian group with a homogeneous kernel $\mu : S \times \mathcal{S} \to [0, 1]$, then for each $y, z \in S$ and integrable $f : S \to \mathbb{R}$,*

$$\int f(x + y)\, \mu(z, dx) = \int f(x)\, \mu(y + z, dx)$$

Proof. For $y \in S$, let $t_y : S \to S$ be translation by y: $t_y(x) = x + y$. Thinking of the kernel as a measurable measure valued map (which we denote $\mu(z)$) we compute the pushforward of $\mu(z)$ under t_y using homogeneity

$$\mu(z) \circ t_y^{-1}(A) = \mu(z, t_y^{-1}(A)) = \mu(z, A - y) = \mu(z + y, A)$$

thus showing $\mu(z) \circ t_y^{-1} = \mu(y + z)$. Now we can apply the Expectation Rule (Lemma 6.7) to see that

$$\int f(x + y)\, \mu(z, dx) = \int f(x)\, d\left[\mu(z) \circ t_y^{-1}\right] = \int f(x)\, \mu(y + z, dx)$$

□

A Markov process with homogeneous kernels is said to be *space-homogeneous*; intuitively the probability of starting out in a set A at time s and winding up in set B at time t only depends on the relative positions of A and B (under translations).

Definition 17.13. Suppose $(S, \mathcal{S})$ is a measurable Abelian group and let X_t be a Markov process with transition kernels $\mu_{s,t}$. Then X_t is *space-homogeneous* if and only if $\mu_{s,t}$ is homogeneous for every $s \leq t$.

Lemma 17.14. *Let $\mu_{s,t}$ be a family of space homogeneous transition kernels on a measurable Abelian group, then for every $A \in \mathcal{S}^T$ and $x \in S$, $\boldsymbol{P}_x\{A\} = \boldsymbol{P}_0\{A - x\}$.*

Proof. TODO: This proof only seems to require space homogeneity of the kernels $\mu_{0,t}$; is this a mistake (or does Chapman Kolmogorov imply the rest of the kernels are space homogeneous as well...)

We begin by establishing the result for sets of the form $\{(X_{t_1}, \dots, X_{t_n}) \in A\}$ for $A \in \mathcal{S}^{\otimes n}$ and $t_1 \leq \cdots \leq t_n$. The key point is that we know from the proof of Lemma 17.10 that $\mathbf{P}_x\{(X_{t_1}, \dots, X_{t_n}) \in A\} = \mu_{0,t_1} \otimes \ldots \otimes \mu_{t_{n-1},t_n}(x, A)$, so in particular the case $n = 1$ follows directly from the assumption that each $\mu_{0,t}$ is homogeneous. To see the result for $n > 1$ we calculate using Lemma 17.12

$$
\begin{aligned}
&\mathbf{P}_x\{(X_{t_1}, \dots, X_{t_n}) \in A\} \\
&= \mu_{0,t_1} \otimes \ldots \otimes \mu_{t_{n-1},t_n}(x, A) \\
&= \int \int \mathbf{1}_A(x_1, x_2, \dots, x_n) \mu_{t_1,t_2} \otimes \cdots \otimes \mu_{t_{n-1},t_n}(x_1, dx_2, \dots, dx_n)\, \mu_{0,t_1}(x, dy) \\
&= \int \int \mathbf{1}_A(x_1 + x, x_2, \dots, x_n) \mu_{t_1,t_2} \otimes \cdots \otimes \mu_{t_{n-1},t_n}(x_1, dx_2, \dots, dx_n)\, \mu_{0,t_1}(0, dy) \\
&= \int \int \mathbf{1}_{A-x}(x_1, x_2, \dots, x_n) \mu_{t_1,t_2} \otimes \cdots \otimes \mu_{t_{n-1},t_n}(x_1, dx_2, \dots, dx_n)\, \mu_{0,t_1}(0, dy) \\
&= \mu_{0,t_1} \otimes \ldots \otimes \mu_{t_{n-1},t_n}(0, A - x) \\
&= \mathbf{P}_0\{(X_{t_1}, \dots, X_{t_n}) \in A - x\}
\end{aligned}
$$

Now we complete the result by a monotone class argument. We know that sets of the form $\{(X_{t_1}, \dots, X_{t_n}) \in A\}$ are a generating π-system so by the π-λ Theorem (Theorem 3.24) it suffices to show that $\mathcal{C} = \{A \mid \mathbf{P}_x\{A\} = \mathbf{P}_0\{A - x\}\}$ is a λ-system. If $A, B \in \mathcal{C}$ with $A \subset B$ then

$$\mathbf{P}_x\{B \setminus A\} = \mathbf{P}_x\{B\} - \mathbf{P}_x\{A\} = \mathbf{P}_0\{B - x\} - \mathbf{P}_0\{A - x\} = \mathbf{P}_0\{B \setminus A - x\}$$

where we have used the elementary fact that $B \setminus A - x = (B - x) \setminus (A - x)$ (let $y \in B$ and $y \notin A$ then clearly $y - x \in B - x$ and $y - x \notin A - x$). Similarly if $A_n \in \mathcal{C}$ for $n \in \mathbb{N}$ with $A_1 \subset A_2 \subset \cdots$ then it is also true that $A_1 - x \subset A_2 - x \subset \cdots$ and continuity of measure (Lemma 3.27) shows

$$\mathbf{P}_x\{\cup_n A_n\} = \lim_{n \to \infty} \mathbf{P}_x\{A_n\} = \lim_{n \to \infty} \mathbf{P}_0\{A_n - x\} = \mathbf{P}_0\{\cup_n A_n - x\}$$

□

There is another way of thinking about the space-homogeneous Markov processes. We know that for any $s \leq t$, given the value of X_s the probability distribution of X_t is independent of the history of X up to s. Space homogeneity tells us that moreover that the probability distribution X_t only depends on the *increment*

$X_t - X_s$. Putting these two observations together we should expect that $X_t - X_s$ is independent (not just conditionally independent) of the history of X up to s. In fact this provides an equivalent characterisation of space homogeneous Markov processes as we prove in the following result.

Definition 17.15. Let $(S, \mathcal{S})$ be a measurable Abelian group with a time scale $T \subset \mathbb{R}_+$, a filtration $\mathcal{F}_t$ and an S-valued $\mathcal{F}$-adapted process X_t. We say that X_t has $\mathcal{F}$-independent increments if and only if $X_t - X_s$ is independent of $\mathcal{F}_s$ for all $s \leq t$.

Lemma 17.16. *Let $(S, \mathcal{S})$ be a measurable Abelian group with a time scale $T \subset \mathbb{R}_+$, a filtration $\mathcal{F}_t$ and an S-valued $\mathcal{F}$-adapted process X_t. The X_t has $\mathcal{F}$-independent increments if and only if X_t is a space-homogeneous Markov process. In this case the transition kernels of X_t are given by*

$$\mu_{s,t}(x, A) = \boldsymbol{P}\{X_t - X_s \in A - x\} \text{ for } x \in S,\ A \in \mathcal{S} \text{ and } s \leq t \in T$$

TODO: The proof actually requires regular versions of $\boldsymbol{P}\{X_t \mid \mathcal{F}_s\}$; do we need to assume that G is Borel or something? Also we've defined a Markov process as satisfying the Chapman Kolmogorov relations identically; can that be derived?

Proof. Suppose that X_t is a space homogeneous Markov Process with transition kernels $\mu_{s,t}$. Then for every $s \leq t$ and $A \in \mathcal{S}$,

$$\begin{aligned}
\mathbf{P}\{X_t - X_s \in A \mid \mathcal{F}_s\} &= \int \mathbf{1}_A(x - X_s)\, \mu_{s,t}(X_s, dx) && \text{by Theorem 11.26} \\
&= \int \mathbf{1}_A(x)\, \mu_{s,t}(0, dx) && \text{by Lemma 17.12} \\
&= \mu_{s,t}(0, A)
\end{aligned}$$

which shows that $\mathbf{P}\{X_t - X_s \in A \mid \mathcal{F}_s\}$ is almost surely constant hence $X_t - X_s \perp\!\!\!\perp \mathcal{F}_s$. Moreover by the tower rule we also know that $\mathbf{P}\{X_t - X_s \in A\} = \mathbf{P}\{X_t - X_s \in A \mid \mathcal{F}_s\} = \mu_{s,t}(0, A)$ and therefore by another application of space homogeneity, $\mu_{s,t}(x, A) = \mu_{s,t}(0, A - x) = \mathbf{P}\{X_t - X_s \in A\}$.

Suppose that X_t has independent increments. The key point is that this property determines the conditional distributions

$$\mu_{s,t}(x, A) = \mathbf{P}\{X_t - X_s \in A - x\}$$

and moreover this form is a regular version. First note that $\mathbf{P}\{X_t - X_s \in A - x\}$ is a probability kernel since for fixed A it is measurable in x by Lemma 3.80 and for fixed x it is just the distribution of the measurable random element $X_t - X_s - x$.

Showing that $\mathbf{P}\{X_t - X_s \in A - x\}$ is a version of $\mathbf{P}\{X_t \in A \mid \mathcal{F}_s\}$ is not hard but requires a bit of care because the random element X_s plays two different roles in the calculation and it is worth making this fact explicit. We start by defining $\tilde{\mu}_{s,t}(x, A) = \mathbf{P}\{X_t - X_s \in A\}$ and observing that because $X_t - X_s \perp\!\!\!\perp \mathcal{F}_s$, $\tilde{\mu}_{s,t}$ is a kernel for $\mathbf{P}\{X_t - X_s \in \cdot \mid \mathcal{F}_s\}$. With this fact and the $\mathcal{F}$-adaptedness of X, we can apply Theorem 11.26 (using the function $f(x, y) = \mathbf{1}_{A-y}(x)$ evaluated at

$(X_t - X_s, X_s)$) to conclude

$$\begin{aligned}\mathbf{P}\{X_t \in A \mid \mathcal{F}_s\} &= \mathbf{P}\{X_t - X_s \in A - X_s \mid \mathcal{F}_s\} \\ &= \int \mathbf{1}_{A-X_s}(x)\, \tilde{\mu}_{s,t}(dx) \\ &= \tilde{\mu}_{s,t}(A - X_s) \\ &= \mu_{s,t}(X_s, A)\end{aligned}$$

Now note that $\mu_{s,t}(X_s, A)$ is X_s-measurable hence we have $\mathbf{P}\{X_t \in A \mid \mathcal{F}_s\} = \mathbf{P}\{X_t \in A \mid X_s\}$ for all $A \in \mathcal{S}$ thus the Markov property holds by Lemma 11.15. Using the explicit form of the kernel we calculate

$$\mu_{s,t}(x, A) = \mathbf{P}\{X_t - X_s \in A - x\} = \mu_{s,t}(0, A - x)$$

demonstrating space homogeneity. □

Here is what the proof that space homogeneous Markov implies independent increments looks like in elementary probability theory (discrete time countable state space).

Proof. Space homogeneity means that $\mathbf{P}\{X_n = x \mid X_{n-1} = y\} = \mathbf{P}\{X_n = x - y \mid X_{n-1} = 0\}$. This implies that for any $y \in S$ we have $\mathbf{P}\{X_n - X_{n-1} = z\} = \mathbf{P}\{X_n = z + y \mid X_{n-1} = y\}$:

$$\begin{aligned}\mathbf{P}\{X_n - X_{n-1} = z\} &= \sum_x \mathbf{P}\{X_n - X_{n-1} = z; X_{n-1} = x\} \\ &= \sum_x \mathbf{P}\{X_n - X_{n-1} = z \mid X_{n-1} = x\}\mathbf{P}\{X_{n-1} = x\} \\ &= \sum_x \mathbf{P}\{X_n = z + x \mid X_{n-1} = x\}\mathbf{P}\{X_{n-1} = x\} \\ &= \mathbf{P}\{X_n = z + y \mid X_{n-1} = y\} \sum_x \mathbf{P}\{X_{n-1} = x\} \\ &= \mathbf{P}\{X_n = z + y \mid X_{n-1} = y\}\end{aligned}$$

Now we use this fact along with the Markov property to see

$$\begin{aligned}&\mathbf{P}\{X_n - X_{n-1} = z; X_1 = x_1; \cdots; X_{n-1} = x_{n-1}\} \\ &= \mathbf{P}\{X_n = z + x_{n-1}; X_1 = x_1; \cdots; X_{n-1} = x_{n-1}\} \\ &= \mathbf{P}\{X_n = z + x_{n-1} \mid X_1 = x_1; \cdots; X_{n-1} = x_{n-1}\}\mathbf{P}\{X_1 = x_1; \cdots; X_{n-1} = x_{n-1}\} \\ &= \mathbf{P}\{X_n = z + x_{n-1} \mid X_{n-1} = x_{n-1}\}\mathbf{P}\{X_1 = x_1; \cdots; X_{n-1} = x_{n-1}\} \\ &= \mathbf{P}\{X_n = z\}\mathbf{P}\{X_1 = x_1; \cdots; X_{n-1} = x_{n-1}\}\end{aligned}$$

□

TODO: Motivate time homogeneity by thinking about discrete time and the fact that you can generate everything from the unit time transitions. Time homogeneity is the property that all of these transition kernels are the same and therefore the Markov process is determined by a single kernel (and the initial distribution).

Definition 17.17. A time homogenous Markov process ... TODO

TODO: Show that a time homogenous Markov process is a dynamical system with noise.

17.3. **Strong Markov Property.**

Lemma 17.18. *Let X be a time homogeneous Markov process and let τ be an optional time with at most countably many values. Then for every measurable $A \subset S^T$,*

$$\boldsymbol{P}\{\theta_\tau X \in A \mid \mathcal{F}_\tau\}(\omega) = \boldsymbol{P}_{X_\tau(\omega)}\{A\} \text{ for all } \omega \text{ such that } \tau(\omega) < \infty$$

Proof. We first prove the result for deterministic times and extend to countably valued optional times. Note that the content of result is vacuous for an infinite deterministic time, so pick a finite deterministic time t and let $t_1 \leq \cdots \leq t_n$ and $A \in \mathcal{S}^{\otimes n}$ and calculate using Lemma 17.3, time homogeneity and Lemma 17.10

$$\begin{aligned}
\mathbf{P}\{((\theta_t X)_{t_1}, \ldots, (\theta_t X)_{t_n}) \in A \mid \mathcal{F}_t\} &= \mathbf{P}\{(X_{t+t_1}, \ldots, X_{t+t_n}) \in A \mid \mathcal{F}_t\} \\
&= \mu_{t,t+t_1} \otimes \cdots \otimes \mu_{t+t_{n-1},t+t_n}(X_t, A) \\
&= \mu_{0,t_1} \otimes \cdots \otimes \mu_{t_{n-1},t_n}(X_t, A) \\
&= \mathbf{P}_{X_t}\{A\}
\end{aligned}$$

Now we know that sets of the form $\{((\theta_t X)_{t_1}, \ldots, (\theta_t X)_{t_n}) \in A\}$ are a generating π-system for the σ-algebra $\mathcal{S}^T$ and the full result for deterministic times t follows from a monotone class argument.

To see this we simply show that the set of A such that $\mathbf{P}\{\theta_t X \in A \mid \mathcal{F}_t\} = \mathbf{P}_{X_t}\{A\}$ a.s. is a λ-system. The case for $B \setminus A$ follows from linearity of conditional expectation and finite additivity of measure and the case $A_1 \subset A_2 \subset \cdots$ follows from monotone convergence for conditional expectations and continuity of measure.

Now we extend to the case of countably valued optional times. Let $A \in \mathcal{S}^T$ and $B \in \mathcal{F}_\tau$ and calculate

$$\begin{aligned}
\mathbf{E}\left[\mathbf{1}_A(\theta_\tau X); B\right] &= \sum_t \mathbf{E}\left[\mathbf{1}_A(\theta_t X); \{\tau = t\} \cap B\right] \\
&= \sum_t \mathbf{E}\left[\mathbf{P}_{X_t}\{A\}; \{\tau = t\} \cap B\right] \\
&= \mathbf{E}\left[\mathbf{P}_{X_\tau}\{A\}; B\right]
\end{aligned}$$

so the result follows by the definition of conditional expectation.

TODO: What about the $\mathcal{F}_\tau$-measurability of $\mathbf{P}_{X_\tau}\{A\}$? Note that this is a consequence of result since we haven't assumed X is progressive(see Lemma 17.19 below where we make this implication explicit). Double check that we don't assume it in the proof above.

TODO: Some comments about the well-definedness of the expressions on $\{\tau < \infty\}$; this boils down to the locality of conditional expectation. □

In the case of a space homogeneous Markov process the strong Markov property can be expressed more concisely as an extension of the independent increments characterization of Lemma 17.16 to optional times. In many scenarios it is more convenient to use these properties. Note that the Lemma does not require the countable range assumption.

Lemma 17.19. *Let S be a measurable Abelian group with a filtration $\mathcal{F}$, X be a time homogeneous and space homogeneous S- valued Markov process and τ be an almost surely finite optional time. Then*

$$\boldsymbol{P}\{\theta_\tau X \in A \mid \mathcal{F}_\tau\} = \boldsymbol{P}_{X_\tau}\{A\}$$

if and only if X_τ is $\mathcal{F}_\tau$-measurable, $\theta_\tau X - X_\tau \perp\!\!\!\perp \mathcal{F}_\tau$ and $X - X_0 \stackrel{d}{=} \theta_\tau X - X_\tau$

Proof. Assume that X satisfies $\mathbf{P}\{\theta_\tau X \in A \mid \mathcal{F}_\tau\} = \mathbf{P}_{X_\tau}\{A\}$ for all $A \in \mathcal{S}^T$. To see that X_τ is $\mathcal{F}_\tau$-measurable observe that if we let $\pi_0 : S^T \to S$ be evaluation at time 0, then for any $B \in \mathcal{S}$ and $x \in S$,

$$\mathbf{P}_x\{\pi_0^{-1}B\} = \begin{cases} 1 & \text{if } x \in B \\ 0 & \text{if } x \notin B \end{cases}$$

therefore we have

$$\mathbf{1}_{X_\tau \in B} = \mathbf{P}_{X_\tau}\{\pi_0^{-1}B\} = \mathbf{P}\{\theta_\tau X \in \pi_0^{-1}B \mid \mathcal{F}_\tau\}$$

which shows that $\{X_\tau \in B\} \in \mathcal{F}_\tau$.

Having established $\mathcal{F}_\tau$-measurability of X_τ we know that P_{X_τ} is a not just a *regular* version for $\mathbf{P}\{\theta_\tau X \in \cdot \mid \mathcal{F}_\tau\}$ and we can apply Theorem 11.26 and space homogeneity of $\mathbf{P}_x$ (Lemma 17.14) to calculate for $A \in \mathcal{S}^T$ (using $f : S^T \times S \to \mathbb{R}_+$ given by $f(x, y) = \mathbf{1}_{A+y}(x)$ in the disintegration)

$$\mathbf{P}\{\theta_\tau X - X_\tau \in A \mid \mathcal{F}_\tau\} = \int \mathbf{1}_{A+X_\tau}(x)\, \mathbf{P}_{X_\tau}(dx) = \mathbf{P}_{X_\tau}\{A + X_\tau\} = \mathbf{P}_0\{A\} \text{ a.s.}$$

which is almost surely constant and therefore independence is proven. This also shows that the distribution of $\theta_\tau X - X_\tau$ is equal to $\mathbf{P}_0$ and letting $\tau = 0$ shows $\theta_\tau X - X_\tau \stackrel{d}{=} X - X_0$.

To prove the converse, note that $X - X_0$ is has initial distribution δ_0 hence using our independence and equidistribution assumptions and the definition of the measure $\mathbf{P}_0$ we get for any $A \in \mathcal{S}^T$,

$$\mathbf{P}\{\theta_\tau X - X_\tau \in A \mid \mathcal{F}_\tau\} = \mathbf{P}\{\theta_\tau X - X_\tau \in A\} = \mathbf{P}\{X - X_0 \in A\} = \mathbf{P}_0\{A\}$$

which provides us with a regular version for $\mathbf{P}\{\theta_\tau X - X_\tau \in \cdot \mid \mathcal{F}_\tau\}$. Now by the $\mathcal{F}_\tau$-measurability of X_τ we can apply Theorem 11.26 and Lemma 17.14 to get

$$\begin{aligned} \mathbf{P}\{\theta_\tau X \in A \mid \mathcal{F}_\tau\} &= \mathbf{P}\{\theta_\tau X - X_\tau \in A - X_\tau \mid \mathcal{F}_\tau\} \\ &= \int \mathbf{1}_{A-X_\tau}(x)\, \mathbf{P}_0(dx) \\ &= \mathbf{P}_{A-X_\tau}\{0\} \\ &= \mathbf{P}_A\{X_\tau\} \end{aligned}$$

and we are done. □

18. More Real Analysis

Holding area for more advanced topics in real analysis that are eventually required (and in some cases there may be some topics that I am just interested in).

18.1. **Topological Spaces.**

Lemma 18.1. *A set $U \subset X$ is open if and only if for every $x \in U$ there is an open set $V \subset U$ such that $x \in V$.*

Proof. Suppose U is open and $x \in U$, then let $V = U$.

Suppose for every $x \in U$ there exist an open set V_x such that $x \in V_x \subset U$. Note that $\cup_x V_x \subset U$ because each $V_x \subset U$ and on the other hand $\cup_x V_x \supset U$ since every $x \in U$ satisfies $x \in V_x$. Thus $U = \cup_x V_x$ which shows that U is open. □

Definition 18.2. A mapping $f : X \to Y$ between topological spaces is said to be *continuous* if and only if $f^{-1}(V)$ is open in X for every V open in Y.

Definition 18.3. A mapping $f : X \to Y$ between topological spaces is said to be *continuous at x* if and only if for every V open in Y such that $f(x) \in V$, there exists an open set U in X with $x \in U$ and $f(U) \subset V$.

Lemma 18.4. *A mapping $f : X \to Y$ between topological spaces is continuous if and only if it is continuous at x for every $x \in X$.*

Proof. Suppose f is continuous and let $x \in X$ and V be open in Y with $f(x) \in V$. By continuity of f, we know that $f^{-1}(V)$ is open in X and $x \in f^{-1}(V)$. By Lemma 18.1 we can pick an open set U such that $x \in U$ and $U \subset f^{-1}(V)$. It follows that $f(U) \subset V$.

Now suppose f is continuous at every $x \in X$ and let V be open in Y. If $x \in f^{-1}(V)$ then f is continuous at x hence there exists and open U such that $x \in U$ and $f(U) \subset V$. It follows that $U \subset f^{-1}(V)$ and by Lemma 18.1 we have shown that $f^{-1}(V)$ is open. □

Definition 18.5.
(i) A topological space is said to be *separable* if and only if it has a countable dense subset.
(ii) A topological space is said to be *first countable* if and only if every point has a countable local base.
(ii) A topological space is said to be *second countable* if and only if every the topology has a countable base.

Lemma 18.6. *A metric space is separable if and only if it is second countable.*

Proof. TODO: outline of proof is to pick a countable dense subset $\{x_n\}$ and then pick the open balls $B(x_n; \frac{1}{m})$ for $m \in \mathbb{N}$. Show this is a base of the topology. □

TODO: The goal of the next set of results is to show that separable complete metric spaces are Borel.

The following appears in Royden as Theorem 8.11 (with proof delgated to exercises)

Lemma 18.7. *Let X be a Hausdorff topological space, Y be a complete metric space and $Z \subset X$ be a dense subset. If $f : Z \to X$ is a homeomorphism then Z is a countable intersection of open sets.*

Proof. For each n let

$$O_n = \{x \in X \mid \text{there exists } U \text{ open with } x \in U \text{ and } \operatorname{diam}(f(U \cap Z)) < \frac{1}{n}\}$$

Note that O_n is open because for any $x \in O_n$ by definition we have the open set U that provides the evidence that $x \in O_n$; U also provides the evidence that proves

that every $y \in U$ belongs to O_n. Also note that $Z \subset O_n$ since for any n, by continuity of f at $x \in Z$ and Lemma 18.1 we can find an open $U \subset X$ such that $x \in U \cap Z$ and $f(U \cap Z) \subset B(f(x), \frac{1}{2n})$ (sets of the form $U \cap Z$ being precisely the open sets in Z).

Now define $E = \cap_n O_n$. As noted we know $Z \subset E$ so we will be done if we can show $E \subset Z$ as well. Let $x \in E$; we will construct $z \in Z$ such that $x = z$. For each n pick U_n such $x \in U_n$ and $\operatorname{diam}(f(U_n \cap Z)) < \frac{1}{n}$ and let x_n be an arbitrary point in $\cap_{j=1}^n U_j \cap Z$ (the intersection is non-empty because Z is dense in X). For every n and $m \geq n$ we have by construction that $x_n \in U_n$ and $x_m \in U_n$ hence $d(f(x_n), f(x_m)) < \frac{1}{n}$. Therefore $f(x_n)$ is Cauchy in Y and by completeness of Y we know that $f(x_n)$ converges to a value $y \in Y$ with $d(y, f(x_n)) \leq \frac{1}{n}$. Because f is a homeomorphism we know that there is a unique $z \in Z$ such that $f(z) = y$; we claim that $x = z$. Suppose that $x \neq z$, then by the Hausdorff property on X we can pick open sets U and V such that $U \cap V = \emptyset$, $x \in U$ and $z \in V$. Since f is a homeomorphism, we know $f(Z \cap V)$ is open and contains $f(z)$ hence for sufficiently large n, $f^{-1}(B(f(z), \frac{1}{n})) \subset Z \cap V \subset V$. On the other hand, by the definition of x we have U_{2n} open such that $x \in U_{2n}$ and $\operatorname{diam}(f(Z \cap U_{2n})) < \frac{1}{2n}$. By openness of $U \cap U_{2n}$ and density of Z we know there is a $w \in U \cap U_{2n} \cap Z$. Putting these observations together we have

$$d(f(w), f(z)) \leq d(f(w), f(x_{2n})) + d(f(x_{2n}), f(z)) < \frac{1}{2n} + \frac{1}{2n} = \frac{1}{n}$$

which implies $w \in V$ providing a contradiction of $U \cap V = \emptyset$ hence we conclude $x = z$. □

Definition 18.8. Given a topological space $(X, \mathcal{T})$ the Baire σ-algebra is smallest σ-algebra for which all bounded continuous functions are measurable. Equivalently

$$Ba(X, \mathcal{T}) = \sigma(\{f^{-1}(U) \mid U \subset \mathbb{R} \text{ is open; } f \in C_b(X, \mathbb{R})\})$$

Lemma 18.9. *For every topological space $(X, \mathcal{T})$, $Ba(X) \subset \mathcal{B}(X)$. For a metric space (S, d), $Ba(S) = \mathcal{B}(S)$.*

Proof. To see the inclusion $Ba(X) \subset \mathcal{B}(X)$, note that by continuity of $f \in C_b(X; \mathbb{R})$, every set $f^{-1}(U)$ is open.

Now suppose (S, d) is a metric space. To show $\mathcal{B}(S) \subset Ba(S)$, it suffices if we show every closed set $F \subset S$ can be written as $f^{-1}(G)$ where $G \subset \mathbb{R}$ is closed and $f \in C_b(S; \mathbb{R})$. By the triangle inequality (see e.g. Lemma 8.39) we know that $g(x) = d(x, F)$ is continuous (in fact Lipschitz) and by Lemma 8.40 we know that $f(x) = d(x, F) \wedge 1$ is also Lipschitz and therefore $f(x) \in C_b(S; \mathbb{R})$. Because F is closed we also know that $F = f^{-1}(\{0\})$ and we are done. □

TODO: How much this stuff on regularity can be extended to outer measures???? I want to understand the overlap with the results in Evans and Gariepy.

Lemma 18.10. *Let X be a Hausdorff topological space, $\mathcal{A}$ a σ-algebra on X and μ a finite tight measure. Then*

$$\mathcal{R} = \{A \in \mathcal{A} \mid A \text{ and } A^c \text{ are } \mu\text{-inner regular}\}$$

is a σ-algebra. The same is true if the condition is replaced by sets that are μ-closed inner regular (without the requirement that μ is tight).

Proof. By definition, $\mathcal{R}$ is closed under complement. By assumption that μ is tight we have $X \in \mathcal{R}$ so all that needs to be shown is closure under countable union.

Assume $A_1, A_2, \cdots \in \mathcal{R}$ and let $\epsilon > 0$ be given. By finiteness of μ, $\mu(\cup_{n=1}^{\infty} A_n) < \infty$ and continuity of measure (Lemma 3.27) there exists $M > 0$ such that $\mu(\cup_{n=1}^{M} A_n) > \mu(\cup_{n=1}^{\infty} A_n) - \epsilon$. By assumption that $A_n \in \mathcal{R}$ and finiteness of μ, for each A_n there exists a compact K_n such that $\mu(A_n \setminus K_n) < \frac{\epsilon}{2^n}$ and there exists compact L_n such that $\mu(A_n^c \setminus L_n) < \frac{\epsilon}{2^n}$. Let

$$K = \cup_{n=1}^{M} K_n$$
$$L = \cap_{n=1}^{\infty} L_n$$

and note that both K and L are compact (in the latter case, because X is Hausdorff we know that each L is closed hence the intersection is a closed subset of a compact set hence compact). Furthermore we can compute

$$\begin{aligned}
\mu(\cup_{n=1}^{\infty} A_n \setminus K) &= \mu(\cup_{n=1}^{\infty} A_n \setminus \cup_{n=1}^{M} K_n) \\
&= \mu(\cup_{n=1}^{M} A_n \setminus \cup_{n=1}^{M} K_n) + \mu(\cup_{n=1}^{\infty} A_n \setminus \cup_{n=1}^{M} A_n \setminus \cup_{n=1}^{M} K_n) \\
&\leq \mu(\cup_{n=1}^{M} A_n \setminus K_n) + \mu(\cup_{n=1}^{\infty} A_n \setminus \cup_{n=1}^{M} A_n) \\
&\leq \sum_{n=1}^{M} (A_n \setminus K_n) + \epsilon \\
&\leq 3\epsilon
\end{aligned}$$

and

$$\begin{aligned}
\mu((\cup_{n=1}^{\infty} A_n)^c \setminus L) &= \mu(\cap_{n=1}^{\infty} A_n^c \setminus \cap_{n=1}^{\infty} L_n) \\
&= \mu(\cap_{n=1}^{\infty} A_n^c \cap \cup_{n=1}^{\infty} L_n^c) \\
&= \mu(\cup_{n=1}^{\infty} \cap_{m=1}^{\infty} A_m^c \cap L_n^c) \\
&\leq \mu(\cup_{n=1}^{\infty} A_n^c \cap L_n^c) \\
&\leq \sum_{n=1}^{\infty} (A_n^c \setminus L_n) \\
&\leq 2\epsilon
\end{aligned}$$

TODO: The closed inner regular case... □

TODO: In metric space, tightness is equivalent to inner regularity. Then Ulam's Theorem that finite measures on separable metric spaces are automatically inner regular. Also finite measures on arbitrary metric spaces are closed inner regular as well as outer regular.

Lemma 18.11. *Let (S, d) be a metric space and μ be a Borel measure on $(S, \mathcal{B}(S))$, the μ is closed inner regular. If in addition μ is a finite measure then it is outer regular.*

Proof. Let U be an open set in S. Then U^c is closed and the function $f(x) = d(x, U^c)$ is continuous. If we define

$$F_n = f^{-1}([1/n, \infty))$$

then each F_n is closed, $F_1 \subset F_2 \subset \cdots$ and $\cup_{n=1}^{\infty} F_n = U$. By continuity of measure (Lemma 3.27) we know that $\lim_{n\to\infty} \mu(F_n) = \mu(U)$. So this shows that every open

set is inner closed regular. Furthermore it is trivial to note that U^c is inner closed regular because it is closed.

By Lemma 18.10 we know know that

$$\mathcal{B}(S) \subset \mathcal{R} = \{A \subset S \mid A \text{ and } A^c \text{ are inner closed regular}\}$$

Outer regularity follows from taking complements and using the finiteness of μ. □

If we add the criterion that the metric space is separable, then we can upgrade the regularity to inner regularity.

Lemma 18.12. *Let (S, d) be a separable metric space and μ be a Borel measure on $(S, \mathcal{B}(S))$, then μ is inner regular if and only if it is tight.*

Proof. Clearly inner regularity implies tightness (which is just inner regularity of the set S), so it suffices to show that tightness implies inner regularity.

Suppose that μ is a tight measure. By Lemma 18.10 it suffices to show that both open and closed sets are inner regular.

Pick $\epsilon > 0$ and select $K \subset S$ a compact set such that $\mu(S \setminus K) < \frac{\epsilon}{2}$. By Lemma 18.11 we know that for any Borel set B there exists a closed set $F \subset B$ such that $\mu(B \setminus F) < \frac{\epsilon}{2}$. Note that $F \cap K$ is compact. We have

$$\mu(B \setminus (F \cap K)) \leq \mu(B \cap F^c) + \mu(B \cap K^c) \leq \mu(B \cap F^c) + \mu(S \cap K^c) < \epsilon$$

□

Theorem 18.13 (Ulam's Theorem)**.** *Let (S, d) be a separable metric space and μ be a Borel measure on $(S, \mathcal{B}(S))$, then μ is inner regular.*

Proof. By Lemma 18.12 it suffices to show that μ is tight. Pick $\epsilon > 0$ and we construct a compact set $K \subset S$ such that $\mu(S \setminus K) < \epsilon$. Let $\overline{B}(x, r)$ denote the closed ball of radius r around $x \in S$. Pick a countable dense subset $x_1, x_2, \ldots \in S$. For each $m \in \mathbb{N}$, by density of $\{x_n\}$, we know $\cap_{n=1}^{\infty} \left(S \setminus \cup_{j=1}^{n} \overline{B}(x_j, \frac{1}{m})\right) = \emptyset$, thus by continuity of measure (Lemma 3.27) there exists $N_m > 0$ such that $\mu(S \setminus \cup_{j=1}^{n} \overline{B}(x_j, \frac{1}{m}) < \frac{\epsilon}{2^m}$ for all $n \geq N_m$. If we define

$$K = \cap_{m=1}^{\infty} \cup_{j=1}^{N_m} \overline{B}(x_j, \frac{1}{m})$$

we claim that K is compact. Note that K is easily seen to be closed as it is an intersection of a finite union of closed balls. Since S is complete this implies that K is also complete. Also it is easy to see that K is totally bounded since by construction we have demonstrated a cover by a finite number of balls of radius $\frac{1}{m}$ for each $m \in \mathbb{N}$. So by Theorem 2.27 we know K is compact.

To finish the result we claim $\mu(S \setminus K) < \epsilon$:

$$\begin{aligned}
\mu(S \setminus K) &= \mu(S \cap \left(\cap_{m=1}^{\infty} \cup_{j=1}^{N_m} \overline{B}(x_j, \frac{1}{m}) \right)^c) \\
&= \mu(S \cap \cup_{m=1}^{\infty} \left(\cup_{j=1}^{N_m} \overline{B}(x_j, \frac{1}{m}) \right)^c) \\
&= \mu(\cup_{m=1}^{\infty} S \setminus \cup_{j=1}^{N_m} \overline{B}(x_j, \frac{1}{m})) \\
&\leq \sum_{m=1}^{\infty} \mu(S \setminus \cup_{j=1}^{N_m} \overline{B}(x_j, \frac{1}{m})) \\
&< \epsilon
\end{aligned}$$

□

18.2. **Riesz Representation.**

Definition 18.14. Let μ be a measure on the Borel σ-algebra of a topological space S.

(i) A Borel set B is *inner regular* if for $\mu(B) = \sup_{K \subset B} \mu(K)$ where K is compact. μ is inner regular if every Borel set is inner regular.
(ii) A Borel set B is *outer regular* if $\mu(B) = \inf_{U \supset B} \mu(U)$ where U is open. A measure μ is outer regular if every Borel set B is outer regular.
(iii) μ is *locally finite* if every $x \in S$ has an open neighborhood $x \in U$ such that $\mu(U) < \infty$.
(iv) μ is a *Radon measure* it is inner regular and locally finite.
(v) μ is a *Borel measure* when????? In some cases I've seen it required that $\mu(B) < \infty$ for all Borel sets B (reference?) and in other cases just that the Borel sets are measurable.
(vi) A Borel set B is *closed regular* if $\mu(B) = \inf_{F \subset B} \mu(F)$ where F is closed (e.g. Dudley pg. 224). A measure μ is closed regular if every Borel set B is closed regular.
(vii) If μ is finite, then we say *tight* if and only if X is inner regular (e.g. Dudley pg. 224).

Definition 18.15. Let μ be a Borel measure on a Hausdorff topological space. A set measurable set A is called *regular* if

(i) $\mu(A) = \inf_{U \supset A} \mu(A)$ where U are open
(ii) $\mu(A) = \sup_{F \subset A} \mu(A)$ where F are closed

TODO: Alternative def assumes that F are compact (see inner regularity above). If every measurable set is regular then μ is said to be regular. Note that if we assume the definition of regularity uses compact inner approximations then regular measures are inner and outer regular (although inner and outer regularity refer to only Borel sets; is that a meaningful distinction?)

TODO: Regularity of outer measures and the relationship to regularity of measures as defined above (see Evans and Gariepy). Note that regularity of outer measure implies that if we take an outer measure μ and the measure on the μ-measurable sets and then take the induced outer measure we get μ back if and only μ is a regular outer measure. Evans and Gariepy show that Radon outer measures on $\mathbb{R}^n$ are inner regular as measures on the μ-measurable sets. Note that inner

regular is part of the most common definition of Radon measure so their result can be taken as showing a weaker definition of Radon measure holds on $\mathbb{R}^n$ (but also they phrase everything in terms of outer measures...).

Theorem 18.16. *Let μ be a finite Borel measure on a metric space S, then μ is closed regular. If μ is tight then μ is regular.*

TODO: Specialize the definition of Radon measure in the presence of more assumptions on X (in particular local compactness, σ-compactness, second countability).

TODO: Are Radon measures automatically outer regular?

Tao proves Riesz representation under assumption of local compactness and σ-compactness.

Kallenberg proves Riesz representation under assumption of local compactness and second countability (this is more general than the Tao result as σ-compactness implies second countability (I think)).

Evans and Gareipy prove Riesz representation only on $\mathbb{R}^n$.

Dudley proves Riesz representation of compact Hausdorff spaces (in which cases the dual measures are Baire measures instead of Radon measures). Dudley does not really discuss Radon measures.

18.3. **Covering Theorems in $\mathbb{R}^n$.** Since our purposes have been to understand probability theory we have hitherto avoided making assumptions that we are dealing with $\mathbb{R}^n$. While this decision has benefits, it has drawbacks as well. Among them we lose sight of some history but also some of the beautiful and deep understanding of the measure theory of the reals. TODO: Vitali and Besicovich.

18.4. **Hausdorff Measure.**

18.4.1. *Introduction.* In this section we discuss the construction of a family of outer measures on $\mathbb{R}^n$ called *Hausdorff measures*. Note the construction can be generalized to metric spaces. The following is motivation why a tool like Hausdorff measure may be useful. Suppose very specifically that we are in $\mathbb{R}^3$, then the Lebesgue product measure essentially corresponds to a notion of volume. What about the surface area of a 2-dimensional object or the length of a 1-dimensional object? As you may have learned in advanced calculus these ideas can indeed be describe in great generality by the notion of differential forms. However, the formalism of forms usually has some notion of smoothness associated with it (hence the adjective differential); a natural question to ask is whether one can fine a purely measure theoretic approach to the problem. Hausdorff measures provide one answer to this question. The broad form of the theory is perhaps a bit more general than one might expect; for any space there is a Hausdorff outer measure for every real number s. The case of integers $s = 1$ corresponds to arclength, $s = 2$ surface area, $s = 3$ volume and so on. Measures with s non-integral are *fractal*. On $\mathbb{R}^n$, the Hausdorff measure with $s = n$ is equal to Lebesgue measure and any Hausdorff measure with $s > n$ is trivial (gives 0 measure to all sets). We'll prove all of this and more in what follows.

18.4.2. *Construction of Hausdorff Measure.* The following technical Lemma is useful (we'll use it when discussing Hausdorff outer measures). If the reader is in a hurry, no harm will come from skipping over this result and returning to it when the need arises. Note that if the user is only interested in probability theory this result may never come up.

Lemma 18.17 (Caratheodory Criterion)**.** *Let (S, d) be a metric space with an outer measure μ^*. Then μ^* is a Borel outer measure (i.e. all Borel sets are μ^*-measurable) if and only if $\mu^*(A \cup B) = \mu^*(A) + \mu^*(B)$ for all A, B such that $d(A, B) > 0$.*

Proof. We begin with the only if direction. Let A be a closed set in S and let $B \subset S$. To show A is μ^*-measurable it suffices to show $\mu^*(B) \geq \mu^*(A \cap B) + \mu^*(A^c \cap B)$. Since the inequality is trivially satisfied when $\mu^*(B) = \infty$ we assume that $\mu^*(B) < \infty$. For every $n \in \mathbb{N}$, let $A_n = \{x \in S \mid d(x, A) \leq \frac{1}{n}\}$. By definition of A_n, we have $d(A, A_n^c) > \frac{1}{n} > 0$ and therefore $d(A \cap B, A_n^c \cap B) > \frac{1}{n} > 0$. Now by our assumption, we can conclude $\mu^*((A \cap B) \cup (A_n^c \cap B)) = \mu^*(A \cap B) + \mu^*(A_n^c \cap B)$.

We claim that $\lim_{n\to\infty} \mu^*(A_n^c \cap B) = \mu^*(A^c \cap B)$. Note that if we prove the claim the Lemma is proven because then we have

$$\begin{aligned} \mu^*(B) &\geq \mu^*((A \cap B) \cup (A_n^c \cap B)) && \text{by monotonicity} \\ &= \mu^*(A \cap B) + \mu^*(A_n^c \cap B) \end{aligned}$$

and taking limits we have

$$\mu^*(B) \geq \lim_{n\to\infty} \mu^*(A \cap B) + \mu^*(A_n^c \cap B) = \mu^*(A \cap B) + \mu^*(A^c \cap B)$$

To prove the claim we observe that monotonicity of outer measure implies that $\lim_{n\to\infty} \mu^*(A_n^c \cap B) \leq \mu^*(A^c \cap B)$ so we just need to work on the opposite inequality. To see it first define the rings around A

$$R_n = \{x \mid \frac{1}{n+1} < d(x, A) \leq \frac{1}{n}\}$$

and note that because A is closed, for each n,

$$\begin{aligned} A^c &= \{x \in S \mid d(x, A) > 0\} \\ &= \{x \in S \mid d(x, A) > n\} \cup \bigcup_{m=n}^{\infty} \{x \in S \mid \frac{1}{m+1} < d(x, A) \leq \frac{1}{m}\} \\ &= A_n^c \cup \bigcup_{m=n}^{\infty} R_m \end{aligned}$$

It follows that $A^c \cap B = A_n^c \cap B \cup \cup_{m=n}^{\infty} R_m \cap B$ and therefore by subadditivity of outer measure

$$\mu^*(A^c \cap B) \leq \mu^*(A_n^c \cap B) + \sum_{m=n}^{\infty} \mu^*(R_m \cap B)$$

The claim will follow if we can show $\lim_{n\to\infty} \sum_{m=n}^{\infty} \mu^*(R_m \cap B) = 0$ which in turn will follow if we can show that $\sum_{m=1}^{\infty} \mu^*(R_m \cap B)$ converges. By construction, $d(R_{2m}, R_{2n}) > 0$ and therefore $d(R_{2m} \cap B, R_{2n} \cap B) > 0$ for any $m \neq n$. So if we consider only the even terms of the series we can use our hypothesis to show that for any n

$$\sum_{m=1}^{n} \mu^*(R_{2m} \cap B) = \mu^*(\cup_{m=1}^{n} R_{2m} \cap B) \leq \mu^*(B) < \infty$$

and by taking limits $\sum_{m=1}^{\infty} \mu^*(R_{2m} \cap B) \leq \mu^*(B)$ The same argument applies to the odd indexed terms and we get

$$\sum_{m=1}^{\infty} \mu^*(R_m \cap B) \leq 2\mu^*(B) < \infty$$

The claim and the Lemma follow. □

TODO: Here I am taking the path of Evans and Gariepy and normalizing Hausdorff measure so that $\mathcal{H}^n = \lambda_n$. I am not sure if this winds up being inconvenient when one considers Hausdorff measure in arbitrary metric spaces (nor do I know whether we'll bother considering Hausdorff measures in metric spaces).

Lemma 18.18. *Let λ_n be Lebesgue measure on $\mathbb{R}^n$, then $\lambda_n(B(0,1)) = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}$.*

Proof. TODO □

Definition 18.19. Let (S,d) be a metric space and $A \subset S$, the *diameter* of A is

$$\operatorname{diam}(A) = \sup\{d(x,y) \mid x,y \in A\}$$

Definition 18.20. Let (S,d) be a metric space, $0 \leq s < \infty$ and $0 < \delta$. Then for $A \subset S$,

$$\mathcal{H}^s_\delta(A) = \inf\{\sum_{n=1}^{\infty} \alpha(s) \left(\frac{\operatorname{diam}(C_n)}{2}\right)^s \mid A \subset \cup_{n=1}^{\infty} C_n \text{ where } \operatorname{diam}(C_n) \leq \delta \text{ for all } n\}$$

where

$$\alpha(s) = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}$$

For A and s as above define

$$\mathcal{H}^s(A) = \lim_{\delta \to 0} \mathcal{H}^s_\delta(A) = \sup_{\delta > 0} \mathcal{H}^s_\delta(A)$$

19. Exercises

Exercise 19.1. Let $f(x)$ be a Lebesgue integrable function on $\mathbb{R}$. Show that there exists a measurable $a(x)$ with $\lim_{x\to\infty} a(x) = \infty$ such that $a(x)f(x)$ remains integrable.

Proof. It suffices to assume that $f(x) \geq 0$ and $\int f(x)\,dx = 1$. We know from Fundamental Theorem of Calculus that $g(y) = \int_{-\infty}^{y} f(x)\,dx$ is almost everywhere differentiable (and montone) and $g'(y) = f(y)$. By definition $\lim_{y\to\infty} g(y) = 1$. Now define $h(z) = 1 - \sqrt{1-z}$ and note that by the Chain Rule (TODO: Show that the Chain Rule is still valid for functions that are merely absolutely continuous)

$$\frac{d}{dy} h(g(y)) = \frac{f(y)}{2\sqrt{1-g(y)}}$$

Now by the Fundamental Theorem of Calculus again, if we define $a(x) = \frac{1}{2\sqrt{1-g(x)}}$ then

$$\int a(x)f(x)\,dx = \lim_{y\to\infty} h(g(y)) = h(1) = 1$$

but

$$\lim_{x\to\infty} a(x) = \lim_{x\to\infty} \frac{1}{2\sqrt{1-g(x)}} = \infty$$

□

Exercise 19.2. Let ξ be a random variable, show that for all $\lambda > 0$,

$$\min_k \mathbf{E}\left[\xi^q\right]\lambda^{-k} \leq \inf_{s>0} \mathbf{E}\left[e^{s(\xi-\lambda)}\right]$$

Note that this shows that the best moment bound for a tail probability is always better than the best Chernoff bound.

Proof. Let $q = \arg\min_k \mathbf{E}\left[\xi^k\right]\lambda^{-k}$. Now expand as a series

$$\begin{aligned}\mathbf{E}\left[e^{s(\xi-\lambda)}\right] &= e^{-s\lambda}\sum_{k=0}^{\infty}\frac{s^k\mathbf{E}\left[\xi^k\right]}{k!}\\ &\geq e^{-s\lambda}\mathbf{E}\left[\xi^q\right]\lambda^{-q}\sum_{k=0}^{\infty}\frac{s^k\lambda^k}{k!} = \mathbf{E}\left[\xi^q\right]\lambda^{-q}\end{aligned}$$

□

Exercise 19.3. Let ξ be a nonnegative integer valued random variable. Show $\mathbf{P}\{\xi \neq 0\} \leq \mathbf{E}\left[\xi\right]$ and

$$\mathbf{P}\{\xi = 0\} \leq \frac{\mathbf{Var}\left(\xi\right)}{\mathbf{Var}\left(\xi\right) + \left(\mathbf{E}\left[\xi\right]\right)^2}$$

Proof. For the first inequality,

$$\mathbf{P}\{\xi \neq 0\} = \sum_{k=1}^{\infty}\mathbf{P}\{\xi = k\} \leq \sum_{k=1}^{\infty} k\mathbf{P}\{\xi = k\} = \mathbf{E}\left[\xi\right]$$

For the second inequality, use Cauchy-Schwartz

$$\begin{aligned}\left(\mathbf{E}\left[\xi\right]\right)^2 &\leq \left(\mathbf{E}\left[\mathbf{1}_{\xi>0}\xi\right]\right)^2\\ &\leq \mathbf{E}\left[\xi^2\right]\mathbf{P}\{\xi > 0\}\end{aligned}$$

Now use $\mathbf{P}\{\xi > 0\} = 1 - \mathbf{P}\{\xi = 0\}$ and $\mathbf{Var}\left(\xi\right) = \mathbf{E}\left[\xi^2\right] - \left(\mathbf{E}\left[\xi\right]\right)^2$ and rearrangement of terms to get the result. □

Exercise 19.4. Let $f : S \to T$ be function. If $\mathcal{T}$ is a σ-algebra on T then $\mathcal{T} \subset f_* f^{-1}(\mathcal{T})$. If $\mathcal{S}$ is a σ-algebra on S, then $f^{-1}f_*(\mathcal{S}) \subset \mathcal{S}$. Find examples where the inclusions are strict.

Proof. To see the inclusions just unwind the definitions. For the first inclusion

$$\begin{aligned}f_* f^{-1}(\mathcal{T}) &= \{A \subset T \mid f^{-1}(A) \in f^{-1}(\mathcal{T})\}\\ &= \{A \subset T \mid f^{-1}(A) = f^{-1}(B) \text{ for some } B \in \mathcal{T}\}\\ &\supset \mathcal{T}\end{aligned}$$

and for the second

$$\begin{aligned}f^{-1}f_*(\mathcal{S}) &= \{f^{-1}(A) \mid A \in f_*(\mathcal{S})\}\\ &= \{f^{-1}(A) \mid A \subset T \text{ and } f^{-1}(A) \in \mathcal{S}\}\\ &\subset \mathcal{S}\end{aligned}$$

TODO: Find the examples of strict inclusion. □

Exercise 19.5. Let $f : S \to T$ be a set function and let $\mathcal{C} \subset 2^T$ then $f^{-1}(\sigma(\mathcal{C})) = \sigma(f^{-1}(\mathcal{C}))$.

Proof. We know that $f^{-1}(\sigma(\mathcal{C}))$ is a σ-algebra and clearly $f^{-1}(\mathcal{C}) \subset f^{-1}(\sigma(\mathcal{C}))$ therefore showing $\sigma(f^{-1}(\mathcal{C})) \subset f^{-1}(\sigma(\mathcal{C}))$.

To see the reverse inclusion we know that

$$f_*(\sigma(f^{-1}(\mathcal{C}))) = \{A \subset T \mid f^{-1}(A) \in \sigma(f^{-1}(\mathcal{C}))\}$$

is a σ-algebra and clearly $\mathcal{C} \subset f_*(\sigma(f^{-1}(\mathcal{C})))$. This implies $\sigma(\mathcal{C}) \subset f_*(\sigma(f^{-1}(\mathcal{C})))$ and thus by the result of the previous exercise

$$f^{-1}(\sigma(\mathcal{C})) \subset f^{-1}(f_*(\sigma(f^{-1}(\mathcal{C})))) \subset \sigma(f^{-1}(\mathcal{C}))$$

□

Exercise 19.6. Let $f(x) = |x|$. Show that $f_*(\mathcal{B}(\mathbb{R}))$ is a strict σ-subalgebra of $\mathcal{B}(\mathbb{R})$.

Exercise 19.7. Let $f : S \to T$ be a function, $\mathcal{C} \in 2^S$ and define $f_*(\mathcal{C}) = \{A \subset T \mid f^{-1}(A) \in \mathcal{C}\}$. Show by counterexample that $\sigma(f_*(\mathcal{C})) \neq f_*(\sigma(\mathcal{C}))$.

Exercise 19.8. Let A_n be a sequence of events. Show that

$$\mathbf{P}\{A_n \text{ i.o.}\} \geq \limsup_{n\to\infty} \mathbf{P}\{A_n\}$$

Proof. Note that we know that for every $k \geq n$, $A_k \subset \cup_{k=n}^{\infty} A_k$ and therefore monotonicity of measure implies $\mathbf{P}\{A_k\} \leq \mathbf{P}\{\cup_{k=n}^{\infty} A_k\}$ for $k \geq n$. Therefore we know $\sup_{k\geq n} \mathbf{P}\{A_k\} \leq \mathbf{P}\{\cup_{k=n}^{\infty} A_k\}$.

By definition and continuity of measure and applying the above,

$$\begin{aligned}
\mathbf{P}\{A_n \text{ i.o.}\} &= \mathbf{P}\{\cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k\} \\
&= \lim_{n\to\infty} \mathbf{P}\{\cup_{k=n}^{\infty} A_k\} \\
&\geq \lim_{n\to\infty} \sup_{k\geq n} \mathbf{P}\{A_k\} = \limsup_{n\to\infty} \mathbf{P}\{A_n\}
\end{aligned}$$

□

Exercise 19.9. Suppose we toss a coin repeatedly and the probability of heads is $0 < p < 1$ (i.e. the coin may be unfair but not pathological). Without using the Strong Law of Large Numbers show that the probability of flipping only a finite number heads is 0.

Proof. Let $A_n = \{\text{heads is flipped on the } n^{th} \text{ toss}\}$. We know that $\mathbf{P}\{A_n\} = p > 0$, therefore $\sum_{n=1}^{\infty} \mathbf{P}\{A_n\} = \infty$. We also know that A_n are independent events, therefore the converse of the Borel-Cantelli Theorem (Theorem 7.21) tells us that $\mathbf{P}\{A_n \text{ i.o.}\} = 1$. The probability of tossing only a finite number of heads is $1 - \mathbf{P}\{A_n \text{ i.o.}\} = 0$. □

Exercise 19.10. A sequence of random variables $\xi_1, \xi_2, \ldots$ is said to be *completely convergent* to ξ if for every $\epsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbf{P}\{|\xi_n - \xi| > \epsilon\} < \infty$$

Show that if ξ_n are independent then complete convergence is equivalent to almost sure convergence.

Proof. First assume that $\xi = 0$.

We first assume complete convergence. If for a given $\epsilon > 0$, we know $\sum_{n=1}^{\infty} \mathbf{P}\{|\xi_n| > \epsilon\} < \infty$ then we can apply Borel Cantelli to conclude that $\mathbf{P}\{\xi_n > \epsilon \text{ i.o.}\} = 0$. Thus there exists a set A_ϵ of measure zero such that for all $\omega \notin A_\epsilon$, we can find $N > 0$ such that $\xi_n(\omega) \leq \epsilon$. Define $A = \cup_{m=1}^{\infty} A_{\frac{1}{m}}$, note that $\mathbf{P}\{A\} = 0$ and that for every $\omega \notin A$, and every $\epsilon > 0$ we can pick $\frac{1}{m} < \epsilon$ and then we know $N > 0$ such that $\xi_n(\omega) \leq \frac{1}{m} \leq \epsilon$

Then if $\xi_n \overset{a.s.}{\to} 0$, then there exists an event A with $\mathbf{P}\{A\} = 1$ and such that for any $\omega \in A$, $\epsilon > 0$ we can find $N > 0$ such that $|\xi_n| < \epsilon$, thus $\mathbf{P}\{|\xi_n| > \epsilon \text{ i.o.}\} \leq 1 - \mathbf{P}\{A\} = 0$. By independence of ξ_n and Borel Cantelli we conclude that $\sum_{n=1}^{\infty} \mathbf{P}\{|\xi_n| > \epsilon\} < \infty$.

Now in the case in which $\xi \neq 0$ we can reduce to the case in which $\xi = 0$. Note that by Corollary 7.27 to the Kolmogorov 0-1 Theorem, we know that ξ is almost surely a constant c. Then we can define $\xi_n - c$ and note that $\xi_n - c$ are independent by Lemma 7.14. □

Exercise 19.11. Suppose $\eta, \xi_1, \xi_2, \ldots$ are random variables with $|\xi_n| \leq \eta$ a.s. for all $n > 0$. Show that $\sup_n |\xi_n| \leq \eta$ a.s.

Proof. Let $A_n = \{\xi_n \leq \eta\}$ and $A = \cup_n A_n$. By assumption, $\mathbf{P}\{A_n\} = 0$ and therefore by countable subadditivity of measure, $\mathbf{P}\{A\} = 0$. For all $\omega \notin A$, we know for all $n > 0$, $\xi_n(\omega) \leq \eta(\omega)$ and therefore $\sup_n \xi_n(\omega) \leq \eta(\omega)$. □

Exercise 19.12. Suppose $\xi, \xi_1, \xi_2, \ldots$ are random variables with $\xi_n \overset{a.s.}{\to} \xi$ and $\xi < \infty$ a.s. Let $\eta = \sup_n |\xi_n|$ and show that $\eta < \infty$ a.s.

Proof. TODO □

Exercise 19.13 (Kallenberg Ex 3.6)**.** Let $\mathcal{F}_{t,n}$ with $t \in T$ and $n \in \mathbb{N}$ be σ-algebras such that for a fixed t they are nondecreasing in n and for a fixed n they are independent in t. Show that the σ-algebras $\bigvee_n \mathcal{F}_{t,n}$ are independent.

Proof. Because for fixed $t \in T$, we have $\mathcal{F}_{t,0} \subset \mathcal{F}_{t,1} \subset \cdots$ we can see that $\bigcup_n \mathcal{F}_{t,n}$ is a π-system. Since by definition $\bigcup_n \mathcal{F}_{t,n}$ generates $\bigvee_n \mathcal{F}_{t,n}$ by Lemma 7.12 it suffices to show that $\bigcup_n \mathcal{F}_{t,n}$ are independent.

Pick $A_{t_1} \in \mathcal{F}_{t_1,n_1}, \ldots, A_{t_m} \in \mathcal{F}_{t_m,n_m}$. Let $n = n_1 \vee \cdots \vee n_m$ and use the nondecreasing property of $\mathcal{F}_{t,n}$ to observe that $A_{t_1} \in \mathcal{F}_{t_1,n}, \ldots, A_{t_m} \in \mathcal{F}_{t_m,n}$. By the assumption that each of $\mathcal{F}_{t_j,n}$ is independent therefore $\mathbf{P}\{A_1 \cup \cdots \cup A_m\} = \mathbf{P}\{A_1\} \cdots \mathbf{P}\{A_m\}$ and we are done. □

Exercise 19.14 (Kallenberg Ex 3.7)**.** Let T be an arbitrary index set and let $(S_t, \mathcal{B}(S_t))$ be metric spaces with Borel σ-algebras. For each $t \in T$ suppose have random elements random elements $\xi^t, \xi_n^t \in S_t$ for $n \in \mathbb{N}$ such that $\xi_n^t \overset{a.s.}{\to} \xi^t$. If for each fixed $n \in \mathbb{N}$ the ξ_n^t are independent show that ξ^t are independent.

Proof. Pick a finite subset $\{t_1, \ldots, t_m\} \subset T$ and assume we are given bounded continuous functions $f_j : S_{t_j} \to \mathbb{R}$ for $j = 1, \ldots, m$. By Lemma 7.16 and the independence of the $\xi_n^{t_j}$ we have $\mathbf{E}\left[f_1(\xi_n^{t_1}) \cdots f(\xi_n^{t_m})\right] = \mathbf{E}\left[f_1(\xi_n^{t_1})\right] \cdots \mathbf{E}\left[f(\xi_n^{t_m})\right]$ for

each $n \in \mathbb{N}$. But now we can use the boundedness and continuity of the f_j

$$
\begin{aligned}
&\mathbf{E}\left[f_1(\xi^{t_1}) \cdots f_m(\xi^{t_m})\right] \\
&= \mathbf{E}\left[\lim_{n\to\infty} f_1(\xi_n^{t_1}) \cdots f_m(\xi_n^{t_m})\right] && \text{by continuity} \\
&= \lim_{n\to\infty} \mathbf{E}\left[f_1(\xi_n^{t_1}) \cdots f_m(\xi_n^{t_m})\right] && \text{boundedness of } f_j \text{ and Dominated Convergence} \\
&= \lim_{n\to\infty} \mathbf{E}\left[f_1(\xi_n^{t_1})\right] \cdots \mathbf{E}\left[f_m(\xi_n^{t_m})\right] && \text{independence} \\
&= \mathbf{E}\left[f_1(\xi^{t_1})\right] \cdots \mathbf{E}\left[f_m(\xi^{t_m})\right] && \text{continuity and Dominated Convergence}
\end{aligned}
$$

We now prove a slight extension of Lemma 7.16 that shows this is sufficient to see that ξ^t are independent. Let (S, d) be a metric space and let $U \subset S$ be open. We show how to approximate the indicator function $\mathbf{1}_U$ be bounded continuous functions. Let $d(x, U^c) = \inf\{d(x, y) \mid y \not\in U\}$. Note that $d(x, U^c)$ is continuous (see proof Lemma 8.39). Let $f_n(x) = 1 \wedge n d(x, U^c)$ and observe that $f_n \uparrow \mathbf{1}_U$. Now suppose $U_j \subset S_{t_j}$ are open sets for $j = 1, \ldots, m$ and use the construction just presented to create bounded continuous functions $f_n^j \uparrow \mathbf{1}_{U_j}$. Then it is also true that $f_n^1 \cdots f_n^m \uparrow \mathbf{1}_{U_1} \cdots \mathbf{1}_{U_m}$ and so we can apply Montone convergence to see

$$
\begin{aligned}
\mathbf{P}\{\xi^{t_1} \in U_1 \cap \cdots \cap \xi^{t_m} \in U_m\} &= \lim_{n\to\infty} \mathbf{E}\left[f_n^1(\xi^{t_1}) \cdots f_n^m(\xi^{t_m})\right] \\
&= \lim_{n\to\infty} \mathbf{E}\left[f_n^1(\xi^{t_1})\right] \cdots \mathbf{E}\left[f_n^m(\xi^{t_m})\right] \\
&= \mathbf{P}\{\xi^{t_1} \in U_1\} \cdots \mathbf{P}\{\xi^{t_m} \in U_m\}
\end{aligned}
$$

Now it suffices to note that the open sets in a metric space are a π-system that generates all of the Borel sets so by Lemma 7.12 it suffices to show independence on open sets. □

A simpler subcase of the above

Exercise 19.15. Let ξ, ξ_n be random elements in a metric space S such that $\xi_n \xrightarrow{P} \xi$ and each ξ_n is $\mathcal{F}_n$-measurable. Furthermore suppose $\mathcal{G}$ is a σ-algebra such that $\mathcal{F}_n \perp\!\!\!\perp \mathcal{G}$ for all $n \in \mathbb{N}$, then show ξ is independent of $\mathcal{G}$. TODO: In the proof we mention that $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$. Is that really required? If not provide a counter example.

Proof. Since $\xi_n \xrightarrow{P} \xi$ we know there is a subsequence that converges almost surely. Note that all of the hypotheses restrict cleanly to the subsequence so we might as well assume that $\xi_n \xrightarrow{a.s.} \xi$. By the $\mathcal{F}_n$ measurability of ξ_n we see that each ξ_n is $\bigvee_n \mathcal{F}_n$-measurable and therefore ξ is almost surely equal to a $\bigvee_n \mathcal{F}_n$-measurable function. It therefore suffices to show that $\bigvee_n \mathcal{F}_n \perp\!\!\!\perp \mathcal{G}$ (TODO: show this simple fact; if $\xi = \eta$ a.s. and $\xi \perp\!\!\!\perp \mathcal{G}$ then $\eta \perp\!\!\!\perp \mathcal{G}$). This follows from the fact that the nestedness of the $\mathcal{F}_n$ implies $\bigcup_n \mathcal{F}_n$ is a π-system. Since by definition it generates $\bigvee_n \mathcal{F}_n$ we get the result from Lemma 7.12. □

Exercise 19.16. Let $\xi_1, \xi_2, \ldots$ be independent random variables with values in $[0, 1]$. Show that $\mathbf{E}\left[\prod_{n=1}^{\infty} \xi_n\right] = \prod_{n=1}^{\infty} \mathbf{E}\left[\xi_n\right]$. In particular, for independent events A_n we have $\mathbf{P}\{\cup_{n=1}^{\infty} A_n\} = \prod_{n=1}^{\infty} \mathbf{P}\{A_n\}$.

Proof. Note that because ξ_n have values in $[0,1]$, the partial products $\prod_{k=1}^{n} \xi_k \leq 1$ and therefore by Dominated Convergence and Lemma 7.16, we have

$$\mathbf{E}\left[\prod_{k=1}^{\infty} \xi_k\right] = \lim_{n\to\infty} \mathbf{E}\left[\prod_{k=1}^{n} \xi_k\right] = \lim_{n\to\infty} \prod_{k=1}^{n} \mathbf{E}\left[\xi_k\right] = \prod_{k=1}^{\infty} \mathbf{E}\left[\xi_k\right]$$

□

Exercise 19.17. Provide an example of uncorrelated but non-independent random variables.

Proof. See Example 7.19. □

Exercise 19.18. Let $\xi_1, \xi_2, \dots$ be random variables. Show that there exist constants $c_1 > 0, c_2 > 0, \cdots$ such that $\sum_{n=1}^{\infty} c_n \xi_n$ converges almost surely.

Proof. First note that we can make a few assumptions about ξ_n without loss of generality. First, we can assume that $\xi_n \geq 0$ for all n; knowing that that will show absolute convergence for all series. Next, note that by a comparison test argument, we may further assume that $\xi_n > 0$ for all n (e.g. for a random variable ξ that takes 0 as a value we can always create the modification $\xi + \mathbf{1}_{\xi^{-1}(0)}$ which is nonzero and dominates ξ).

The idea here is to leverage freshman calculus and use the ratio test. We first verify the following almost sure version of the ratio test: Let ξ_n be positive random variables such that there exists a $0 < C < 1$ such that $\sum_{n=1}^{\infty} \mathbf{P}\{\frac{|\xi_{n+1}|}{|\xi_n|} > C\} < \infty$, then $\sum_{n=1}^{\infty} \xi_n$ converges almost surely.

To verify the claim, we apply Borel Cantelli to conclude that $\mathbf{P}\{\frac{|\xi_{n+1}|}{|\xi_n|} > C \text{ i.o.}\} = 0$. Unwinding the definitions in this statement, we see that for almost every $\omega \in \Omega$, there exists an $N > 0$ such that $\frac{|\xi_{n+1}(\omega)|}{|\xi_n(\omega)|} \leq C$ for all $n > N$. The ratio test tells us $\sum_{n=1}^{\infty} \xi_n(\omega)$ converges and the almost sure convergence is verified.

Now we apply the claim in our case by choosing $C = \frac{1}{2}$ and inductively defining c_n so that we guarantee $\mathbf{P}\{\frac{c_{n+1}\xi_{n+1}}{c_n \xi_n} > \frac{1}{2}\} < \frac{1}{n^2}$. To see that this is possible, suppose we've defined c_n and note that because $\xi_n > 0$, we know that $0 < \frac{\xi_{n+1}}{c_n \xi_n} < \infty$. This tells us that $\lim_{N\to\infty} \mathbf{P}\{\frac{\xi_{n+1}}{c_n \xi_n} > N\} = 0$ and therefore we can find $M > 0$ such that $\mathbf{P}\{\frac{\xi_{n+1}}{c_n \xi_n} > N\} < \frac{1}{n^2}$ for all $N \geq M$. Pick $c_{n+1} = \frac{1}{2M}$ and we are done.

Here is some things that I tried that proved to be a dead end. Is there a learning opportunity in looking at this? Note that almost sure convergence of $\sum_{n=1}^{\infty} c_n \xi_n$ is equivalent to $\mathbf{P}\{|\sum_{n=1}^{\infty} c_n \xi_n| \geq N \text{ i.o.}\}$. The idea was to try to find c_n so that we could provide bounds on $\mathbf{P}\{c_n |\xi_n| \geq N\}$ and leverage those to show bounds on the series. The problem I had with this approach is that to go from a bound on $c_n |\xi_n|$ to convergence of the series meant that $c_n |\xi_n|$ had to decay fast enough to get convergence. If we assume a finite moment then Markov could provide a rate of decay but in the absence of that one has to deal with the fact that tails of ξ_n can decay increasingly slowly. I tried a truncation argument but fact that ξ_n are not related meant that I couldn't figure out how to control the residuals of the truncations. Maybe this line of reasoning could be made to work but I got stuck.

Guolong asks a good follow on question: either prove this or (more likely) provide a counterexample on general (non-finite) measure spaces (e.g. Lebesgue measure on $\mathbb{R}$). □

Exercise 19.19. Let $\xi_1, \xi_2, \dots$ be positive independent random variables, then $\sum_{n=1}^{\infty} \xi_n$ converges almost surely if and only if $\sum_{n=1}^{\infty} \mathbf{E}[\xi_n \wedge 1] < \infty$. TODO: Provide hints

Proof. One direction is easy and doesn't require the assumption of independence; namely assume that $\sum_{n=1}^{\infty} \mathbf{E}[\xi_n \wedge 1] < \infty$. Apply Tonelli's Theorem (Corollary 3.41) to conclude $\mathbf{E}[\sum_{n=1}^{\infty} \xi_n \wedge 1] < \infty$ which implies that $\sum_{n=1}^{\infty} \xi_n \wedge 1 < \infty$ almost surely. For any $\omega \in \Omega$ such that $\sum_{n=1}^{\infty} \xi_n(\omega) \wedge 1 < \infty$ this implies $lim_{n\to\infty} \xi_n(\omega) \wedge 1 = 0$ so there exists an $N_\omega > 0$ such that $\xi_n(\omega) \wedge 1 = \xi_n(\omega)$ for all $n > N_\omega$ and therefore $\sum_{n=1}^{\infty} \xi_n(\omega) < \infty$ as well.

Now lets assume $\sum_{n=1}^{\infty} \xi_n < \infty$. Since $\xi_n \wedge 1 \leq \xi_n$ we know that $\sum_{n=1}^{\infty} \xi_n < \infty$, so without loss of generality we can assume $0 \leq \xi_n \leq 1$.

$$
\begin{aligned}
0 < \mathbf{E}\left[e^{-\sum_{n=1}^{\infty} \xi_n}\right] &= \mathbf{E}\left[\prod_{n=1}^{\infty} e^{-\xi_n}\right] = \prod_{n=1}^{\infty} \mathbf{E}\left[e^{-\xi_n}\right] \\
&\leq \prod_{n=1}^{\infty} (1 - a\mathbf{E}[\xi_n]) && \text{where } a = 1 - e^{-1} \text{ by Lemma 5.1} \\
&\leq \prod_{n=1}^{\infty} e^{-a\mathbf{E}[\xi_n]} && \text{since } 1 + x \leq e^x \text{ by Lemma 5.1} \\
&= e^{-a\sum_{n=1}^{\infty} \mathbf{E}[\xi_n]}
\end{aligned}
$$

which shows that $\sum_{n=1}^{\infty} \mathbf{E}[\xi_n] < \infty$. □

Exercise 19.20. Suppose ξ is a random variable, let $\mathcal{F}$ be a σ-algebra and let A be a measurable set. Show that $\mathbf{E}[\xi \mid \mathcal{F}, A] = \frac{\mathbf{E}[\xi; A \mid \mathcal{F}]}{\mathbf{P}\{A \mid \mathcal{F}\}}$ on A.

Proof. Note by Localization we know that $\mathbf{1}_A \mathbf{E}[\xi \mid \mathcal{F}, A] = \mathbf{E}[\xi; A \mid \mathcal{F}, A]$, therefore we may assume that $\xi = \mathbf{1}_A \xi$ and show $\mathbf{E}[\xi \mid \mathcal{F}, A] = \mathbf{1}_A \frac{\mathbf{E}[\xi \mid \mathcal{F}]}{\mathbf{P}\{A \mid \mathcal{F}\}}$ almost surely.

Pick $F \in \mathcal{F}$ and calculate

$$
\begin{aligned}
\mathbf{E}\left[\mathbf{1}_A \frac{\mathbf{E}[\xi \mid \mathcal{F}]}{\mathbf{P}\{A \mid \mathcal{F}\}}; A \cap F\right] &= \mathbf{E}\left[\mathbf{E}\left[\frac{\xi; F}{\mathbf{P}\{A \mid \mathcal{F}\}} \mid \mathcal{F}\right]; A\right] && \text{by pushout} \\
&= \mathbf{E}\left[\mathbf{E}\left[\frac{\xi; F}{\mathbf{P}\{A \mid \mathcal{F}\}} \mid \mathcal{F}\right] \mathbf{P}\{A \mid \mathcal{F}\}\right] \\
&= \mathbf{E}[\mathbf{E}[\xi; F \mid \mathcal{F}]] && \text{by pushout} \\
&= \mathbf{E}[\xi; F] = \mathbf{E}[\xi; A \cap F] && \text{by tower property}
\end{aligned}
$$

and trivially

$$
\mathbf{E}\left[\mathbf{1}_A \frac{\mathbf{E}[\xi \mid \mathcal{F}]}{\mathbf{P}\{A \mid \mathcal{F}\}}; A^c \cap F\right] = 0 = \mathbf{E}[\xi; A^c \cap F]
$$

Since sets of the form $A \cap F$, $A^c \cap F$ and F for $F \in \mathcal{F}$ form a π-system that generate $\sigma(A, \mathcal{F})$ we have shown the result. □

Exercise 19.21. Let $A_1, A_2, \cdots$ be a disjoint partition of Ω and let $\mathcal{F} = \sigma(A_1, A_2, \dots)$. Show that for every integrable random variable ξ we have $\mathbf{E}[\xi \mid \mathcal{F}] = \sum_{\mathbf{P}\{A_n\} \neq 0} \frac{\mathbf{E}[\xi; A_n]}{\mathbf{P}\{A_n\}} \mathbf{1}_{A_n}$ almost surely.

Proof. First note that it is trivial that $\sum_{\mathbf{P}\{A_n\}\neq 0} \frac{\mathbf{E}[\xi;A_n]}{\mathbf{P}\{A_n\}}\mathbf{1}_{A_n}$ is $\mathcal{F}$-measurable. Because the A_n are a disjoint partition, they are a π-system and the it will suffice to show the averaging property for the sets A_n. Pick an A_m such that $\mathbf{P}\{A_m\} \neq 0$, they by disjointness of the A_n we get

$$\mathbf{E}\left[\sum_{\mathbf{P}\{A_n\}\neq 0} \frac{\mathbf{E}\left[\xi; A_n\right]}{\mathbf{P}\{A_n\}}\mathbf{1}_{A_n}; A_m\right] = \mathbf{E}\left[\frac{\mathbf{E}\left[\xi; A_m\right]}{\mathbf{P}\{A_m\}}\mathbf{1}_{A_m}\right] = \mathbf{E}\left[\xi; A_m\right]$$

For any A_m with $\mathbf{P}\{A_m\} = 0$ and again applying the disjointness of the A_n we get disjointness of the A_n that

$$0 = \mathbf{E}\left[\sum_{\mathbf{P}\{A_n\}\neq 0} \frac{\mathbf{E}\left[\xi; A_n\right]}{\mathbf{P}\{A_n\}}\mathbf{1}_{A_n}; A_m\right] = \mathbf{E}\left[\xi; A_m\right]$$

□

Exercise 19.22. Suppose ξ is a random element in S such that $\mathbf{P}\{\xi \in \cdot \mid \mathcal{F}\}$ has a regular version ν. Let $f : S \to T$ be measurable. Show that $\mathbf{P}\{f(\xi) \in \cdot \mid \mathcal{F}\}$ has a regular version given by $\nu \circ f^{-1}(\omega, A) = \nu(\omega, f^{-1}(A))$.

Proof. Our hypothesis is that for every A, $\mathbf{P}\{\xi \in A \mid \mathcal{F}\}(\omega) = \mu(\omega, A)$. We calculate

$$\begin{aligned}\mathbf{P}\{f(\xi) \in A \mid \mathcal{F}\}(\omega) &= \mathbf{E}\left[\mathbf{1}_{f^{-1}(A)}(\xi) \mid \mathcal{F}\right] \\ &= \int \mathbf{1}_{f^{-1}(A)}(s)\, d\mu(\omega, s) \qquad \text{by Theorem 11.26} \\ &= \mu(\omega, f^{-1}(A))\end{aligned}$$

and we are done. □

Exercise 19.23. Let ξ be a random element in S. Show that ξ is $\mathcal{F}$-measurable if and only if δ_ξ is a regular version of $\mathbf{P}\{\xi \in \cdot \mid \mathcal{F}\}$.

TODO: Refine this statement to include almost sureness...

Proof. $\mathcal{F}$-measurability of ξ is equivalent to $\mathcal{F}$-measurability of $\mathbf{1}_A(\xi)$ for all A which is equivalent to $\mathbf{P}\{\xi \in A \mid \mathcal{F}\} = \mathbf{1}_A(\xi)$ almost surely for all A. Evaluating the last equality at ω we see that

$$\begin{aligned}\mathbf{P}\{\xi \in A \mid \mathcal{F}\}(\omega) &= \begin{cases} 1 & \text{if } \xi(\omega) \in A \\ 0 & \text{if } \xi(\omega) \notin A \end{cases} \\ &= \delta_{\xi(\omega)}(A)\end{aligned}$$

The fact that δ_ξ is a probability kernel is simple. It is trivial that for fixed ω, $\delta_\xi(\omega)$ is a probability measure. If we fix A then $\delta_\xi(\omega)(A)$ is clearly seen to be measurable since it is just the characteristic function of the measurable set A. □

Exercise 19.24. Let ξ be an integrable random variable for which $\mathbf{E}\left[\xi \mid \mathcal{F}\right] \overset{d}{=} \xi$. Show that in fact $\mathbf{E}\left[\xi \mid \mathcal{F}\right] = \xi$ a.s.

Proof. Here is a simple and conceptual proof in the case that $\mathbf{E}\left[\xi \mid \mathcal{F}\right]$ (and therefore ξ) take finitely many values/are simple functions. Let $y_1 < \cdots < y_n$ be the values of ξ such that $\mathbf{P}\{\xi = y_i\} \neq 0$. Consider $A_1 = \{\mathbf{E}\left[\xi \mid \mathcal{F}\right] = y_1\}$. By definition of conditional expectation $\mathbf{E}\left[\xi; A_1\right] = \mathbf{E}\left[\mathbf{E}\left[\xi \mid \mathcal{F}\right]; A_1\right] = y_1\mathbf{P}\{A_1\}$. Because y_1 is the

minimum value of ξ it follows that we must have $\xi = y_1$ identically on A_1. Since $\xi \stackrel{d}{=} \mathbf{E}[\xi \mid \mathcal{F}]$, we know that $\mathbf{P}\{\xi = y_1\} = \mathbf{P}\{A_1\}$ and therefore $\xi \geq y_2$ almost surely off of A_1. Now induct.

If we want to apply standard machinery to go from the simple function case. Then we could approximate ξ by an increasing family of simple functions of the form $f_n(\xi)$ but then we know that $f_n(\xi) \stackrel{d}{=} f_n(\mathbf{E}[\xi \mid \mathcal{F}])$ but not necessarily that $f_n(\xi) \stackrel{d}{=} \mathbf{E}[f_n(\xi) \mid \mathcal{F}]$ which is what we would need in order to use the simple function case. All roads seem to lead to a need to show that $\mathbf{E}[f(\xi) \mid \mathcal{F}]$ and $f(\mathbf{E}[\xi \mid \mathcal{F}])$ are equal in some sense (either a.s. or in distribution).

The idea is to use Jensen's inequality. First note that we can find a strictly convex function f such that $0 \leq f(x) \leq |x|$. Therefore we know that $\mathbf{E}[f(\xi)] < \infty$.

Moreover, by Theorem 11.25 be have a regular version ν for $\mathbf{P}\{\xi \in A \mid \mathcal{F}\}$. By Theorem 11.26 we know that $\mathbf{E}[f(\xi) \mid \mathcal{F}] = \int s \, d\mu(s)$.

Because $\xi \stackrel{d}{=} \mathbf{E}[\xi \mid \mathcal{F}]$ we also know that $f(\xi) \stackrel{d}{=} f(\mathbf{E}[\xi \mid \mathcal{F}])$ which shows us that ...

TODO: I am aiming to show that $\mu \circ f^{-1}$ is a regular version for $\mathbf{P}\{f(\mathbf{E}[\xi \mid \mathcal{F}]) \in \cdot \mid \mathcal{F}\}$. If we could get that then we could calculate

$$\begin{aligned} f(\mathbf{E}[\xi \mid \mathcal{F}]) &= \mathbf{E}[f(\mathbf{E}[\xi \mid \mathcal{F}]) \mid \mathcal{F}] \\ &= \int s d\mu \circ f^{-1}(s) && \text{by Theorem 11.26} \\ &= \int f(s) \, d\mu(s) && \text{by Expectation Rule} \\ &= \mathbf{E}[f(\xi) \mid \mathcal{F}] && \text{by Theorem 11.26} \end{aligned}$$

Now apply the strictly convex case of Jensen's Inequality to conclude the result.

If we assume finite second moments then there should be a proof of this by showing that the conditional variance is 0. TODO: Define conditional variance and show the result. □

Exercise 19.25. Prove or disprove the following statement. Suppose $\xi \stackrel{d}{=} \eta$, show that for every A, $\mathbf{P}\{\xi \in A \mid \mathcal{F}\} = \mathbf{P}\{\eta \in A \mid \mathcal{F}\}$ a.s.

Proof. This is false. Let $\Omega = \{0, 1\}$ with uniform distribution and power set σ-algebra. Let $\xi(x) = x$ and let $\eta(x) = 1 - x$. Note that $\xi \stackrel{d}{=} \eta$ (both have a uniform distribution on $\{0, 1\}$). Now take $\mathcal{F} = \mathcal{A}$ so that $\mathbf{P}\{\xi \in A \mid \mathcal{F}\} = \mathbf{1}_{\xi \in A}$ and $\mathbf{P}\{\eta \in A \mid \mathcal{F}\} = \mathbf{1}_{\eta \in A}$ and take $A = \{0\}$ or $A = \{1\}$. □

Exercise 19.26. Find $\xi, \eta, \mathcal{F}$ such that $\xi \stackrel{d}{=} \eta$ but $\mathbf{E}[\xi \mid \mathcal{F}] \neq \mathbf{E}[\eta \mid \mathcal{F}]$ a.s.

Proof. Pick sets A, B, C such that $\mathbf{P}\{A\} = \mathbf{P}\{B\}$ but $\mathbf{P}\{A \cap C\} \neq \mathbf{P}\{B \cap C\}$. Even more trivially, take $\mathcal{F} = \mathcal{A}$ so that $\mathbf{E}[\xi \mid \mathcal{F}] = \xi$ and similarly with η. Now the statement is equivalent to show two random elements that not almost surely equal but have the same distribution. □

Exercise 19.27. Suppose $\xi, \tilde{\xi}$ are integrable random variables and $\eta, \tilde{\eta}$ are random elements in $(T, \mathcal{T})$ such that $(\xi, \eta) \stackrel{d}{=} (\tilde{\xi}, \tilde{\eta})$. Show that $\mathbf{E}[\xi \mid \eta] \stackrel{d}{=} \mathbf{E}\left[\tilde{\xi} \mid \tilde{\eta}\right]$.

Proof. First, note the intuition behind the statement. As a result of $(\xi, \eta) \stackrel{d}{=} (\tilde{\xi}, \tilde{\eta})$ we can also conclude that $\xi \stackrel{d}{=} \tilde{\xi}$ and $\eta \stackrel{d}{=} \tilde{\eta}$. However, we also expect that the

conditional distributions on T are equal (thinking heuristically of a formula like $\mathbf{P}\{A \mid B\} = \mathbf{P}\{A \cap B\}/\mathbf{P}\{B\}$). The first order of business is to formulate this intuition precisely and prove it.

By Theorem 11.25 there are probability kernels μ and $\tilde{\mu}$ such that $\mathbf{P}\{\xi \in A \mid \eta\} = \mu(\eta, A)$ and $\mathbf{P}\{\tilde{\xi} \in A \mid \tilde{\eta}\} = \tilde{\mu}(\tilde{\eta}, A)$ for all Borel sets A. Our first claim is that $\mu = \tilde{\mu}$ almost surely with respect to $\mathcal{L}\eta$.

Pick a Borel set A and let $B = \{t \in T \mid \mu(t, A) > \tilde{\mu}(t, A)\}$.

$$
\begin{aligned}
0 &= \mathbf{P}\{\xi \in A; \eta \in B\} - \mathbf{P}\{\tilde{\xi} \in A; \tilde{\eta} \in B\} && \text{by hypothesis} \\
&= \mathbf{E}\left[\int \mathbf{1}_{A\times B}(s,\eta)\, d\mu(\eta, s) - \int \mathbf{1}_{A\times B}(s,\tilde{\eta})\, d\tilde{\mu}(\eta, s)\right] && \text{by Theorem 11.26} \\
&= \mathbf{E}\left[\mathbf{1}_B(\eta)\mu(\eta, A) - \mathbf{1}_B(\tilde{\eta})\tilde{\mu}(\tilde{\eta}, A)\right] \\
&= \int \mathbf{1}_B(t)\mu(t, A) - \mathbf{1}_B(t)\tilde{\mu}(t, A)\, d\mathcal{L}(\eta)(t) && \text{by Lemma 3.52 and } \mathcal{L}(\eta) = \mathcal{L}(\tilde{\eta}).
\end{aligned}
$$

which by choice of B shows that $\mu(t, A) = \tilde{\mu}(t, A)$ almost surely $\mathcal{L}(\eta)$. We can show this almost surely for all $A = (-\infty, r]$ with $r \in \mathbb{Q}$ by taking the union of a countable number of null sets. This shows that $\mu = \tilde{\mu}$ a.s.

Having shown equality of the conditional distributions it follows from Theorem 11.26 that if we define $f(t) = \int s\, d\mu(t, s)$ then we have $\mathbf{E}[\xi \mid \eta] = f(\eta)$ and $\mathbf{E}\left[\tilde{\xi} \mid \tilde{\eta}\right] = f(\tilde{\eta})$. Since $\eta \overset{d}{=} \tilde{\eta}$ it follows that $f(\eta) \overset{d}{=} f(\tilde{\eta})$ and the result is proven. □

Exercise 19.28. Suppose ξ is a random element in a Borel space $(S, \mathcal{S})$, let $\mathcal{F}$ be a σ-algebra and let $\eta = \mathbf{P}\{\xi \in \cdot \mid \mathcal{F}\}$, show $\xi \perp\!\!\!\perp_\eta \mathcal{F}$.

Proof. First it is worth clarifying the question. Since we have assume S is Borel then by Theorem 11.25 be may assume that η is an $\mathcal{F}$-measurable random measure on S. We are asked to show conditional independence of ξ and $\mathcal{F}$ relative to this random measure.

By Lemma 11.15 it will suffice to show for every $A \in \mathcal{S}$,

$$\mathbf{E}[\xi \in A \mid \eta] = \mathbf{E}[\xi \in A \mid \eta, \mathcal{F}] = \mathbf{E}[\xi \in A \mid \mathcal{F}]$$

where the last equality follows from the $\mathcal{F}$-measurability of η. However this is easily verified since the σ-algebra on the space of probability measures $\mathcal{P}(S)$ is the smallest σ-algebra that makes evaluation maps $ev_B(\mu) = \mu(B)$ measurable (here $B \in \mathcal{S}$). Thus we have by definition of η, $\mathbf{E}[\xi \in A \mid \mathcal{F}] = ev_A(\eta)$ which shows that $\mathbf{E}[\xi \in A \mid \mathcal{F}]$ is in fact η-measurable. □

Exercise 19.29. Suppose $\xi \perp\!\!\!\perp_\eta \zeta$ and $\gamma \perp\!\!\!\perp (\xi, \eta, \zeta)$, show that $\xi \perp\!\!\!\perp_{\eta,\gamma} \zeta$ and $\xi \perp\!\!\!\perp_\eta (\zeta, \gamma)$.

Proof. By Lemma 11.16, $\xi \perp\!\!\!\perp_\eta (\zeta, \gamma)$ is equivalent to $\xi \perp\!\!\!\perp_\eta \zeta$ and $\xi \perp\!\!\!\perp_{\eta,\zeta} \gamma$. The fact that $\xi \perp\!\!\!\perp_\eta \zeta$ is a hypothesis whereas $\xi \perp\!\!\!\perp_{\eta,\zeta} \gamma$ follows from another application of Lemma 11.16 to show that $\gamma \perp\!\!\!\perp (\xi, \eta, \zeta)$ is equivalent to $\gamma \perp\!\!\!\perp \zeta$ and $\gamma \perp\!\!\!\perp_\zeta \eta$ and $\gamma \perp\!\!\!\perp_{\zeta,\eta} \xi$

Now by Lemma 11.16 we know $\xi \perp\!\!\!\perp_\eta (\gamma, \zeta)$ is equivalent to $\xi \perp\!\!\!\perp_\eta \gamma$ and $\xi \perp\!\!\!\perp_{\eta,\gamma} \zeta$ hence implies $\xi \perp\!\!\!\perp_{\eta,\gamma} \zeta$. □

Exercise 19.30. Suppose we are given random elements such that $(\xi, \eta, \zeta) \overset{d}{=} (\tilde{\xi}, \tilde{\eta}, \tilde{\zeta})$, then $\xi \perp\!\!\!\perp_\eta \zeta$ if and only if $\tilde{\xi} \perp\!\!\!\perp_{\tilde{\eta}} \tilde{\zeta}$.

Proof. First we

□

Exercise 19.31. Suppose τ and σ are discrete optional times with respect the filtration $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots$. Then $\sigma \wedge \tau$ and σ and $\sigma \vee \tau$ are optional times. In addition,

$$\mathcal{F}_{\tau\wedge\sigma} \subset \mathcal{F}_\sigma \subset \mathcal{F}_{\tau\vee\sigma}$$

Proof. First we show that $\tau\wedge\sigma$ and $\tau\vee\sigma$ are actually optional times. This is simple by noting

$$\{\tau \wedge \sigma \leq n\} = \{\tau \leq n\} \cup \{\sigma \leq n\} \in \mathcal{F}_n$$

and

$$\{\tau \vee \sigma \leq n\} = \{\tau \leq n\} \cap \{\sigma \leq n\} \in \mathcal{F}_n$$

If we are given $A \in \mathcal{F}_\sigma$ the by definition for all n, $A \cap \{\sigma \leq n\} \in \mathcal{F}_n$. Therefore since by definition of optional time we also have $\{\tau \leq n\} \in \mathcal{F}_n$ we have

$$A \cap \{\tau \vee \sigma \leq n\} = (A \cap \{\sigma \leq n\}) \cap \{\tau \leq n\} \in \mathcal{F}_n$$

which shows $A \in \mathcal{F}_{\sigma\vee\tau}$.

Now if we assume that $A \in \mathcal{F}_{\sigma\wedge\tau}$, then for all n we have

$$A \cap \{\tau \wedge \sigma \leq n\} = A \cap \{\tau \leq n\} \cup A \cap \{\sigma \leq n\} \in \mathcal{F}_n$$

Since we have $\{\sigma \leq n\}, \{\tau \leq n\} \in \mathcal{F}_n$, then we know that $\{\tau \leq n\} \setminus \{\sigma \leq n\} \in \mathcal{F}_n$ and so

$$(A \cap \{\tau \leq n\}) \cup (A \cap \{\sigma \leq n\}) \cup (\{\tau \leq n\} \setminus \{\sigma \leq n\})^c = A \cap \{\sigma \leq n\} \in \mathcal{F}_n$$

which shows $A \in \mathcal{F}_\sigma$. □

Exercise 19.32. Suppose τ is a discrete optional time with respect the filtration $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots$, then τ is $\mathcal{F}_\tau$-measurable.

Proof. For any n, m, we have

$$\{\tau = m\} \cap \{\tau \leq n\} = \begin{cases} \emptyset & \text{if } m > n \\ \{\tau = m\} & \text{if } m \leq n \end{cases}$$

hence in all cases is in $\mathcal{F}_n$. □

Exercise 19.33. Suppose τ and σ are discrete optional times with respect the filtration $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots$. Then each of $\{\sigma < \tau\}$, $\{\sigma \leq \tau\}$ and $\{\sigma = \tau\}$ is in $\mathcal{F}_\sigma \cap \mathcal{F}_\tau$.

Proof. It suffice to prove two of the three since the third set can be constructed using finite unions or intersections of the other two. First we show that $\{\sigma < \tau\} \in \mathcal{F}_\tau$. Pick an n and we calculate

$$\begin{aligned} \{\sigma < \tau\} \cap \{\tau \leq n\} &= \cup_{m\leq n}\{\sigma < \tau\} \cap \{\tau = m\} \\ &= \cup_{m\leq n}\{\sigma < m\} \cap \{\tau = m\} \end{aligned}$$

Now each $\{\sigma < m\} \in \mathcal{F}_m \subset \mathcal{F}_n$ and each $\{\tau = m\} \in \mathcal{F}_m \subset \mathcal{F}_n$ by definition of optional time so the union is and we have shown $\{\sigma < \tau\} \in \mathcal{F}_\tau$. The same argument clearly shows that the other sets are in $\mathcal{F}_\tau$ as well. To see that all sets are in $\mathcal{F}_\sigma$, it suffices to note for example that

$$\{\sigma < \tau\}^c = \{\tau \leq \sigma\} \in \mathcal{F}_\sigma$$

by what we have already proven. Apply the closure of σ-algebras under complement to get the result. □

Exercise 19.34. Let σ and τ be optional times with respect to the filtration $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots$. Show that

$$\mathbf{E}\left[\mathbf{E}\left[\xi \mid \mathcal{F}_\sigma\right] \mid \mathcal{F}_\tau\right] = \mathbf{E}\left[\mathbf{E}\left[\xi \mid \mathcal{F}_\tau\right] \mid \mathcal{F}_\sigma\right] = \mathbf{E}\left[\xi \mid \mathcal{F}_{\sigma\wedge\tau}\right]$$

Proof. The first thing to do is show how to calculate conditional expectations with respect to σ-algebras of the form $\mathcal{F}_\sigma$ for an arbitrary optional time σ. Given an integrable random variable ξ we let $M_n^\xi = \mathbf{E}\left[\xi \mid \mathcal{F}_n\right]$ be the martingale generated by ξ. We claim

$$\mathbf{E}\left[\xi \mid \mathcal{F}_\sigma\right] = M_\sigma^\xi$$

To see this, pick an $A \in \mathcal{F}_\sigma$ and then note that for every n, use the fact that $A \cap \{\sigma = n\} \in \mathcal{F}_n$ and the telescoping rule for conditional expectation to see

$$\mathbf{E}\left[\mathbf{1}_A \mathbf{1}_{\{\sigma=n\}}\xi\right] = \mathbf{E}\left[\mathbf{1}_A \mathbf{1}_{\{\sigma=n\}}\mathbf{E}\left[\xi \mid \mathcal{F}_n\right]\right] = \mathbf{E}\left[\mathbf{1}_A \mathbf{E}\left[\mathbf{1}_{\{\sigma=n\}}\xi \mid \mathcal{F}_n\right]\right]$$

which is easy to extend by linearity

$$\begin{aligned}\mathbf{E}\left[\mathbf{1}_A\xi\right] &= \sum_{n=0}^{\infty} \mathbf{E}\left[\mathbf{1}_A \mathbf{1}_{\{\sigma=n\}}\xi\right] = \sum_{n=0}^{\infty} \mathbf{E}\left[\mathbf{1}_A \mathbf{E}\left[\mathbf{1}_{\{\sigma=n\}}\xi \mid \mathcal{F}_n\right]\right] = \mathbf{E}\left[\mathbf{1}_A \sum_{n=0}^{\infty} \mathbf{E}\left[\mathbf{1}_{\{\sigma=n\}}\xi \mid \mathcal{F}_n\right]\right] \\ &= \mathbf{E}\left[\mathbf{1}_A M_\sigma^\xi\right]\end{aligned}$$

Using this formula twice we have

$$\begin{aligned}\mathbf{E}\left[\mathbf{E}\left[\xi \mid \mathcal{F}_\tau\right] \mid \mathcal{F}_\sigma\right] &= \mathbf{E}\left[M_\tau^\xi \mid \mathcal{F}_\sigma\right] \\ &= \sum_{n=0}^{\infty} \mathbf{1}_{\{\sigma=n\}}\mathbf{E}\left[M_\tau^\xi \mid \mathcal{F}_n\right] \\ &= \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} \mathbf{1}_{\{\sigma=n\}}\mathbf{E}\left[\mathbf{E}\left[\mathbf{1}_{\{\tau=m\}}\xi \mid \mathcal{F}_m\right] \mid \mathcal{F}_n\right]\end{aligned}$$

Now consider each term $\mathbf{1}_{\{\sigma=n\}}\mathbf{E}\left[\mathbf{E}\left[\mathbf{1}_{\{\tau=m\}}\xi \mid \mathcal{F}_m\right] \mid \mathcal{F}_n\right]$; there are two cases. If $m \leq n$ then since $\mathcal{F}_m \subset \mathcal{F}_n$ we can write

$$\mathbf{1}_{\{\sigma=n\}}\mathbf{E}\left[\mathbf{E}\left[\mathbf{1}_{\{\tau=m\}}\xi \mid \mathcal{F}_m\right] \mid \mathcal{F}_n\right] = \mathbf{1}_{\{\sigma=n\}}\mathbf{E}\left[\mathbf{1}_{\{\tau=m\}}\xi \mid \mathcal{F}_m\right] = \mathbf{1}_{\{\sigma=n\}}\mathbf{1}_{\{\tau=m\}}\mathbf{E}\left[\xi \mid \mathcal{F}_m\right]$$

If $n \leq m$ then because $\mathcal{F}_n \subset \mathcal{F}_m$ and the telescoping rule,

$$\mathbf{1}_{\{\sigma=n\}}\mathbf{E}\left[\mathbf{E}\left[\mathbf{1}_{\{\tau=m\}}\xi \mid \mathcal{F}_m\right] \mid \mathcal{F}_n\right] = \mathbf{E}\left[\mathbf{E}\left[\mathbf{1}_{\{\sigma=n\}}\mathbf{1}_{\{\tau=m\}}\xi \mid \mathcal{F}_m\right] \mid \mathcal{F}_n\right] = \mathbf{E}\left[\mathbf{1}_{\{\sigma=n\}}\mathbf{1}_{\{\tau=m\}}\xi \mid \mathcal{F}_n\right]$$

These two forms are a bit different and are not equivalent because we cannot ascertain the $\mathcal{F}_{m\wedge n}$-measurability of $\mathbf{1}_{\{\sigma=m\}}\mathbf{1}_{\{\tau=m\}}$. However, we do know that

$\{\sigma > m\} = \{\sigma \leq m\}^c$ is $\mathcal{F}_m$-measurable and $\{\tau > n\} = \{\tau \leq n\}^c$ is $\mathcal{F}_n$-measurable. So if we sum using the case $n \leq m$, we get,

$$\begin{aligned}\sum_{m>n} \mathbf{1}_{\{\sigma=n\}} \mathbf{E}\left[\mathbf{E}\left[\mathbf{1}_{\{\tau=m\}}\xi \mid \mathcal{F}_m\right] \mid \mathcal{F}_n\right] &= \sum_{m>n} \mathbf{E}\left[\mathbf{1}_{\{\sigma=n\}}\mathbf{1}_{\{\tau=m\}}\xi \mid \mathcal{F}_n\right] \\ &= \mathbf{E}\left[\mathbf{1}_{\{\sigma=n\}}\mathbf{1}_{\{\tau>n\}}\xi \mid \mathcal{F}_n\right] \\ &= \mathbf{1}_{\{\sigma=n\}}\mathbf{1}_{\{\tau>n\}}\mathbf{E}\left[\xi \mid \mathcal{F}_n\right] \\ &= \sum_{m>n} \mathbf{1}_{\{\sigma=n\}}\mathbf{1}_{\{\tau=m\}}\mathbf{E}\left[\xi \mid \mathcal{F}_n\right]\end{aligned}$$

So this shows us how to get everything into a common form if we break up the sum properly,

$$\begin{aligned}\mathbf{E}\left[\mathbf{E}\left[\xi \mid \mathcal{F}_\tau\right] \mid \mathcal{F}_\sigma\right] &= \sum_{n=0}^{\infty}\sum_{m=n+1}^{\infty} \mathbf{1}_{\{\sigma=n\}}\mathbf{E}\left[\mathbf{E}\left[\mathbf{1}_{\{\tau=m\}}\xi \mid \mathcal{F}_m\right] \mid \mathcal{F}_n\right] + \\ &\quad \sum_{m=0}^{\infty}\sum_{n=m}^{\infty} \mathbf{1}_{\{\sigma=n\}}\mathbf{E}\left[\mathbf{E}\left[\mathbf{1}_{\{\tau=m\}}\xi \mid \mathcal{F}_m\right] \mid \mathcal{F}_n\right] \\ &= \sum_{n=0}^{\infty}\sum_{m=n+1}^{\infty} \mathbf{1}_{\{\sigma=n\}}\mathbf{1}_{\{\tau=m\}}\mathbf{E}\left[\xi \mid \mathcal{F}_n\right] + \\ &\quad \sum_{m=0}^{\infty}\sum_{n=m}^{\infty} \mathbf{1}_{\{\sigma=n\}}\mathbf{1}_{\{\tau=m\}}\mathbf{E}\left[\xi \mid \mathcal{F}_m\right] \\ &= \sum_{m=0}^{\infty}\sum_{n=0}^{\infty} \mathbf{1}_{\{\sigma=n\}}\mathbf{1}_{\{\tau=m\}}\mathbf{E}\left[\xi \mid \mathcal{F}_{m\wedge n}\right] \\ &= M^\xi_{\sigma\wedge\tau} = \mathbf{E}\left[\xi \mid \mathcal{F}_{\sigma\wedge\tau}\right]\end{aligned}$$

□

Exercise 19.35. Show that a random variable ξ has subexponential tails if and only if there exists $C > 0$ such that $\mathbf{E}\left[|\xi|^k\right] \leq Ck^C$ for all integers $k > 0$.

Proof. TODO: Mimic the proof of Lemma 14.7. □

Exercise 19.36. Suppose we have σ-algebras $\mathcal{F}$, $\mathcal{G}_1$, $\mathcal{G}_2$, $\mathcal{H}$ with $\mathcal{G}_1 \subset \mathcal{G}_2$. If $\mathcal{F} \perp\!\!\!\perp_{\mathcal{H}} \mathcal{G}_1$ is it true that $\mathcal{F} \perp\!\!\!\perp_{\mathcal{H}} \mathcal{G}_2$? Prove or give a counterexample.

Proof. Here is a counterexample in which $\mathcal{G}_1$ is the trivial σ-algebra. Perform two independent Bernoulli trials with rate $1/2$. Thus we have sample space $\Omega = \{HH, HT, TT, TH\}$ with the uniform distribution. Let $A = \{HH, HT\}$ (and let $\mathcal{F} = \{\emptyset, \Omega, A, A^c\}$) and let $B = \{HT, TT\}$ (and let $\mathcal{H} = \{\emptyset, \Omega, B, B^c\}$). Note that A and B are independent. Now let $C = \{HH, TT\}$ (and let $\mathcal{G}_2 = \{\emptyset, \Omega, C, C^c\}$ and note that A and B are not conditionally independent given C because $\mathbf{P}\{A \cap B \mid C\} = 0$ whereas $\mathbf{P}\{A \mid C\} = 1/2$ and $\mathbf{P}\{B \mid C\} = 1/2$ □

Exercise 19.37. Suppose $\mathcal{F}$ is independent of $\mathcal{G}$ and $\mathcal{H}$, is it true that $\mathcal{F}$ is independent of $\sigma(\mathcal{G}, \mathcal{H})$? Prove or give a counterexample.

Proof. Note that $\mathcal{F}$ is independent of $\sigma(\mathcal{G}, \mathcal{H})$ if and only if $\mathcal{F} \perp\!\!\!\perp \mathcal{G}$ and $\mathcal{F} \perp\!\!\!\perp_{\mathcal{G}} \mathcal{H}$. Because of this equivalence the previous exercise is a counterexample here as well. Using the notation of the previous exercise, let $\mathcal{F} = \sigma(A)$ and let $\mathcal{G} = \sigma(C)$ and

note that A and C are independent by direct calculation (this is also intuitively clear). We also saw in the previous exercise that A and B are independent and that A is not conditionally independent of B given C; hence A is not indepndent of $\sigma(B, C)$.

Note that we can also show this directly without using the Lemma. A little work shows that $\sigma(B, C) = 2^{\Omega}$; it suffices to note that $B \cap C = \{TT\}$, $B^c \cap C^c = \{TH\}$, $B \cap C^c = \{HT\}$ and $B^c \cap C = \{HH\}$. Given this fact it is easy to see that A is not independent of $\sigma(B, C)$ by noting that, because $P(A) = 1/2$, it is not independent of itself.

Note also that the the key to the failure here is the fact that A, B and C are not jointly independent (they are pairwise independent), otherwise we could appeal to Lemma 7.13. To see the lack of joint independence consider $\mathbf{P}\{A \cap B \cap C\} = 0$. □

20. Techniques

This section is a place to collect some of the recurring proof techniques that one should be familiar with.

20.1. **Standard Machinery.** The standard measure theory arguments that proceed by showing a result for indicator functions, simple random variables and the positive random variables. TODO: There are a ton of examples of this such as Lemma 3.52 and Lemma 3.54.

20.1.1. *Monotone Class Arguments.* Part of the standard machinery that has independent utility is the monotone class argument. This allows one to demonstrate that a property holds for an entire σ-algebra of sets by showing that property holds for a simpler subclass of sets. Good examples are Lemma 3.65 and Lemma 7.12.

20.2. **Almost Sure Convergence.** When one needs to show almost sure convergence of a sequence of random variables the Borel Cantelli Theorem is a workhorse. Good examples of this are Lemma 7.29 and Lemma 8.9.

Another technique to use that is related is to show that the sum of the random variables is integrable. Then you can conclude that the sum of random variables is almost surely finite and therefore the terms of the sequence converge to zero a.s. Good examples of this are Lemma 8.22 and Lemma 8.9.

20.3. **Bounding Expectations.** A common task that one faces is to provide bounds for an expected value (or more generally a moment). For example, one may need to know that a random variable has a finite expectation for use with the Dominated Convergence Theorem.

20.3.1. *Using Tail Bound.* A problem I have encountered is trying to use a tail bound to prove that an expectation is finite. The problem that I sometime have is that I write:

$$\mathbf{E}\left[f(\xi)\right] = \mathbf{E}\left[\mathbf{1}_{\xi \leq \lambda} \cdot f(\xi)\right] + \mathbf{E}\left[\mathbf{1}_{\xi > \lambda} \cdot f(\xi)\right]$$

Often knowing $\xi \leq \lambda$ we can show that the first expectation is bounded (this is often easy). The problem is usually that one might be given a tail bound that controls $\mathbf{P}\{\xi > \lambda\}$ but there is no control over the behavior of $f(\xi)$ that allows one to provide a bound for the second expectation. Are there general approaches for dealing with this? Possible answer here is that one might need to take a different

approach and use Lemma 6.8. A good example of how to do this is with Lemma 14.7.

TODO: Passing from L^p convergence to almost sure convergence.