Probability Theory

David Blair

E-mail address: dblair@akamai.com

Contents

CHAPTER 1

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 1.1. 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 1.2. *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 1.3. 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 1.4. 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 1.5. *$f : (S, d) \to (S', d')$ 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 1.6. 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 1.7. 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 1.8. *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 1.9. *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 continuity 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 1.10. 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 1.11. *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 1.12. 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 1.13. A metric space (S, d) is said to be *complete* if every Cauchy sequence is convergent.

DEFINITION 1.14. 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 1.15. *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 1.16. *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 1.17. *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). $\square$

The following is a useful variant on the diagonal argument.

PROPOSITION 1.18. *Let (S, d) be a metric space and suppose we are given elements x, x_n and $x_{m,n}$ for $m, n \in \mathbb{N}$ satisfying,* $\lim_{m\to\infty} x_{m,n} = x_n$ *for all $n \in \mathbb{N}$ and* $\lim_{n\to\infty} x_n = x$*, then there exists a sequence l_m with* $\lim_{m\to\infty} l_m = \infty$ *such that* $\lim_{m\to\infty} x_{m,l_m} = x$.

PROOF. Let $k_1 = 1$ and for each $n \geq 2$ let k_n be chosen so that $k_n > k_{n-1}$ and $d(x_{m,n}, x_n) < \frac{1}{n}$ for all $m \geq k_n$. Define

$$l_m = \begin{cases} 1 & \text{for } k_1 \leq m < k_2 \\ 2 & \text{for } k_2 \leq m < k_3 \\ \dots \end{cases}$$

so that we have $\lim_{m\to\infty} l_m = \infty$ by construction and also $d(x_{m,l_m}, x_m) < 1/l_m$ for $m \geq 2$. To see that $\lim_{m\to\infty} x_{m,l_m} = x$, pick $\epsilon > 0$ and then pick $M \geq 2$ such that $l_m > 2/\epsilon$ and $d(x_m, x) < \epsilon/2$ for all $m \geq M$. Then by the definition of l_m and the triangle inequality for all $m \geq M$,

$$d(x_{m,l_m}, x) \leq d(x_{m,l_m}, x_m) + d(x_m, x) < 1/l_m + \epsilon/2 < \epsilon$$

$\square$

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 1.19. 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 1.20 (Taylor's Theorem). *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. $\square$

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 1.21. *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. If $f^{(n+1)}(x)$ is constant on the interval $[a, b]$ then by explicit integration we have the result for any $a < c < b$, so let us assume that $f^{(n+1)}(x)$ is not constant on $[a, b]$. 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 $M = \max_{x\in[a,b]} f^{(n+1)}(x)$. From this fact and the fact that $(b-x)^n$ is strictly positive on $[a, b)$ we have bounds

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

hence

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

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 $f^{(n+1)}(c) = \frac{(n+1)!}{(b-a)^{(n+1)}} \int_a^b \frac{(b-x)^n}{n!} f^{(n+1)}(x)\,dx$. Because the inequalities are strict and because $(b-x)^n$ is positive, it follows that in fact $c \in (a, b)$. $\square$

In addition to the integral form and the Lagrange form of the remainder it can also be useful to have an estimate on the remainder in hand.

Corollary 1.22. *Let $f : \mathbb{R} \to \mathbb{R}$ be a function which is m-times continuously differentiable. Then for all $1 \le n \le m$,*

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

where the remainder term satisfies

$$|r_n(b)| \leq \frac{\sup_{a\leq x\leq b}\left|f^{(n)}(x)-f^{(n)}(a)\right|}{n!}|b-a|^n$$

in particular we have $\lim_{b\to a}\frac{r_n(b)}{(b-a)^n}=0$.

Proof. By Taylor's Theorem we have

$$f(b)=\sum_{k=0}^{n-1}\frac{(b-a)^k}{k!}f^{(k)}(a)+R_{n-1}(b)$$

with

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

so that

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

and therefore

$$|r_n(b)| \leq \int_a^b \frac{|b-x|^{n-1}}{(n-1)!}\left|f^{(n)}(x)-f^{(n)}(a)\right|\,dx \leq \sup_{a\leq x\leq b}\left|f^{(n)}(x)-f^{(n)}(a)\right|\frac{|b-a|^n}{n!}$$

The last statement follows from the continuity of $f^{(n)}(x)$. □

Lemma 1.23. *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 1.24 (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 $\lambda'' \mid_{Y'} = \lambda'$.

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 $\left\|\tilde{\lambda}\right\| = \|\lambda\|$ (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).

1. Compactness

Definition 1.25. Let (S, d) be a metric space, then we say $K \subset S$ is *sequentially compact* if and only if for every sequence $x_1, x_2, \dots \in K$ there exists a convergent subsequence x_{n_j} such that $\lim_{j \to \infty} x_{n_j} \in K$.

Definition 1.26. 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 $U_1, \dots, U_n$ such that $\bigcup_{j=1}^n U_j \supset S$.

Definition 1.27. 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 1.28. 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 1.29. *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 $x_1, \dots, x_m$ 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 $x_1, x_2, \dots$ 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 $U_1, \dots, U_m$ then $x_n \notin \cup_{k=1}^m U_k$ for any $n > \max(N_1, \dots, N_m)$. 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 $A \cap B(y_1,1) \cap \dots \cap B(y_n, \frac{1}{n})$ 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 $x_1, x_2, \dots$ 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 $x_{n_1}, x_{n_2}, \dots$ 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 $j = 1, \dots, n-1$ 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 $U_1, \dots, U_n$ is a finite subcover of S and we are done. $\square$

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 1.30. *Every closed subset of a compact set is compact.*

Proof. Let B be a compact set and let $A \subset B$ be closed. Let U_α be an open cover of A, the we may append A^c to get an open cover of B. By compactness of B we may extract a finite subcover $U_{\alpha_1}, \dots, U_{\alpha_n}, A^c$ (there is no loss in generality in assuming that A^c is in the finite subcover). Clearly, $U_{\alpha_1}, \dots, U_{\alpha_n}$ is a finite subcover of A. $\square$

Theorem 1.31. *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 $f^{-1}(U_1), \dots, f^{-1}(U_n)$. It is easy to see that $U_1, \dots, U_n$ 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 $y \in U_i$. $\square$

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

Theorem 1.32. *[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 1.29 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 $\{(x_1/N, \dots, x_n/N) \mid -MN \leq x_j \leq MN\}$ and note that

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

□

Before we begin the proof we need a Lemma.

LEMMA 1.33. *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 1.30 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 1.34. 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 1.35. *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 1.29 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

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

which is the desired contradiction. □

LEMMA 1.36. *Let $K_1 \supset K_2 \supset \cdots$ be a nested collection of non-empty compact sets, then $\cap_{n=1}^{\infty} K_n$ is nonempty.*

PROOF. Pick $x_n \in K_n$ and note that by compactness there is a convergent subsequence. Let x be the limit of that convergent subsequence. By nestedness and closedness of each K_n we conclude that $x \in K_n$ for every n. $\square$

THEOREM 1.37. *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 1.31, we know that $f(S)$ is a closed bounded set. $\square$

A related notion is that of uniform convergence of functions.

DEFINITION 1.38. 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.

LEMMA 1.39. *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

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

$\square$

PROPOSITION 1.40. *Let (S, d) be a metric space and (T, d') a complete metric space. Suppose that $A \subset S$ and that $f : A \to T$ is a uniformly continuous function, then f has a unique continuous extension $\overline{f} : \overline{A} \to T$ to the closure of $A \subset S$.*

PROOF. Let $x \in \overline{A}$, pick a sequence x_n in A such that $\lim_{n\to\infty} x_n = x$ and observe that by uniform continuity of $f(x)$, for any $\epsilon > 0$ there exists a $\delta > 0$ such that $d(z, w) < \delta$ implies $d'(f(z), f(w)) < \epsilon$. If we pick $N > 0$ such that $d(x_n, x) < \delta/2$ for $n \geq N$ then $d(x_n, x_m) < \delta$ for all $n, m \geq N$ and thus $d'(f(x_n), f(x_m)) < \epsilon$ for all $n, m \geq N$. This shows that the sequence $f(x_n)$ is Cauchy and by completeness of T we can take the limit; we define $f(x) = \lim_{n\to\infty} f(x_n)$. We claim that this definition is independent of the sequence chosen. Indeed, let y_n be another sequence from A such that $\lim_{n\to\infty} y_n = x$. Pick an $\epsilon > 0$ and by uniform continuity of $f(x)$ let δ be chosen such that $d'(f(z), f(w)) < \epsilon/2$ whenever $d(z, w) < \delta$. There exists $N_1 > 0$ such that $d(y_n, x_n) < \delta$ for every $n > N_1$ and there exists $N_2 > 0$ such that $d'(f(x_n), f(x)) < \epsilon/2$ for all $n \geq N_2$. Then we have for all $n \geq N_1 \vee N_2$ by the triangle inequality $d'(f(y_n), f(x)) < \epsilon$. Note that this also shows that the extension $f(x)$ to $\overline{A}$ is continuous at $x \in \overline{A}$; since is was continuous at all points of A we know the extension is continuous. $\square$

2. Stone Weierstrass Theorem

LEMMA 1.41. *Let L be a lattice of continuous functions on a compact Hausdorff space X and suppose that the pointwise infimum $g(x) = \inf_{f\in L} f(x)$ is continuous. Then for every $\epsilon > 0$ there exists $f \in L$ such that* $0 \leq \sup\{x \in X \mid f(x) - g(x)\} < \epsilon$.

PROOF. For every $x \in X$ we can find an $f_x \in L$ such that $f_x(x) - g(x) < \epsilon/3$. By continuity of f_x and g we can find an open neighborhood U_x of x such that $|f_x(x) - f_x(y)| < \epsilon/3$ and $|g(x) - g(y)| < \epsilon/3$. By the triangle inequality it follows that $f_x(y) - g(y) < \epsilon$ for all $y \in U_x$. The U_x are an open cover of X so by compactness we may take a finite subcover $U_{x_1}, \dots, U_{x_n}$. Let $f = f_{x_1} \wedge \cdots \wedge f_{x_n}$ then for every $x \in X$ we have $x \in U_{x_j}$ for some x_j and

$$f(x) - g(x) \leq f_{x_j}(x) - g(x) < \epsilon$$

□

LEMMA 1.42. *Let L be a lattice of continuous functions on a compact Hausdorff space X such that*

(i) *L separates points (i.e. for every $x \neq y \in X$ there exists $f \in L$ such that $f(x) \neq f(y)$)*
(ii) *If $f \in L$ then for every $c \in \mathbb{R}$ we have $cf \in L$ and $f + c \in L$.*

Then for every continuous function g on X and $\epsilon > 0$ there exists $f \in L$ such $0 \leq \sup\{x \in X \mid f(x) - g(x)\} < \epsilon$.

PROOF. The first thing is to observe that for the lattice L we have complete control over the values of the function that separates points.

Claim 1: Suppose $x \neq y \in X$ and $a \neq b \in \mathbb{R}$ then there exists $f \in L$ such that $f(x) = a$ and $f(y) = b$.

To see the claim because L separates points we have an $h \in L$ such that $h(x) \neq h(y)$. Now it suffices to define

$$f(z) = \frac{a - b}{h(x) - h(y)} h(z) + \frac{bh(x) - ah(y)}{h(x) - h(y)}$$

and note that by (ii) we have $f \in L$.

Claim 2: For any closed set $F \subset X$, $y \notin F$ and $a \leq b \in \mathbb{R}$ we can find $f \in L$ such that $f \geq a$, $f(y) = a$ and $f(x) > b$ for all $x \in F$.

Pick an $x \in F$ then by Claim 1, we can find f_x such that $f_x(x) = b + 1$ and $f_x(y) = a$. By continuity of f_x we have an open neighborhood U_x of x such that $f(y) > b$ for all $y \in U_x$. Clearly the U_x form an open cover of F. Since F is closed and X is compact Hausdorff we know that F is also compact hence we can extract a finite open cover $U_{x_1}, \dots, U_{x_n}$ of F. Define

$$f = (f_{x_1} \vee \cdots \vee f_{x_n}) \wedge a$$

and observe that $f \in L$ since L is a lattice and by (ii) L contains the constant functions.

Now we can prove the Lemma proper. With g selected, let $L_g = \{f \in L \mid g \leq f\}$. Clearly L_g is a lattice so the result follows from Lemma 1.41 if we can show $g = \inf_{f \in L_g} f$. Pick an $\delta > 0$ and a $y \in X$, we try to find $f \in L_g$ such that $f(y) - g(y) \leq \delta$. First we find such and $f \in L$ and then show that in fact $f \in L_g$. Let $F = \{x \in X \mid g(x) + \delta\}$ which is closed by continuity of g. By compactness of X and continuity of g we know that g has a maximum value M. Using Claim 2 we know that we can find $f \in L$ such that $g(y)+\delta \geq f(x)$ for all $x \in X$, $g(y)+\delta = f(y)$ and $f(x) > M$ for all $x \in F$. To see that $f \in L_g$, note that by definition of F and construction of f for all $x \in X \setminus F$ we have $g(x) < g(y)+\delta \leq f(x)$ and for all $x \in F$ we have $g(x) \leq M < f(x)$. □

The Stone Weierstrass Theorem concerns the approximation properties of subalgebras of $C(X)$ but we have been describing the approximation properties of lattices of continuous functions. The connection will rely on the fact that we can uniformly approximate the absolute value function by a polynomial on a compact interval. We record that fact as the following

Lemma 1.43. *For every $\epsilon > 0$ there exists a polynomial $p(x)$ such that*

$$\sup\{x \in [-1,1] \mid |p(x) - |x||\} < \epsilon$$

Proof. TODO: □

Theorem 1.44 (Stone Weierstrass Theorem). *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. Let $\overline{A}$ be the uniform closure of A (that is to say the set of f such that for every $\epsilon > 0$ there exists $g \in A$ such that $\sup\{x \in X \mid |g(x) - f(x)|\} < \epsilon$. By Lemma 1.39 any such limit is continuous hence $\overline{A} \subset C(X)$. (TODO: The referenced result is stated for a metric space domain however the proof clearly works for a domain that is a general topological space).

TODO: Finish □

Corollary 1.45 (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 the 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. □

Corollary 1.46 (Weierstrass Approximation Theorem). *For every continuous function $f : [0,T] \to \mathbb{R}$ there exists a sequence of polynomials $p_n(x)$ such that $\lim_{n\to\infty} \sup_{0\leq x\leq T} |f(x) - p_n(x)| = 0$.*

3. Baire Category

The following theorem is surprisingly important; particularly in laying down the foundational results of functional analysis.

THEOREM 1.47. *Let U_n be a countable collection of open dense sets in a complete metric space (S, d), then $\cap_{n=1}^{\infty} U_n$ is dense.*

PROOF. To see that $\cap_{n=1}^{\infty} U_n$ is dense it suffices to show that intersects every open set in S. So let V be an open set be given.

By density of U_1 we know that $U_1 \cap V$ is nonempty and open. Pick an arbitrary point $x_1 \in U_1 \cap V$ and select an open ball $B(x_1; r_1)$ such that $\overline{B}(x_1, r_1) \subset U_1 \cap V$ and $r_1 \leq 1$. By density of U_2 we know that $U_2 \cap B(x_1, r_1) \neq \emptyset$ and therefore we can select an $x_2 \in U_2 \cap B(x_1, r_1)$. Since $U_2 \cap B(x_1, r_1)$ is open we may pick an $r_2 > 0$ such that $B(x_2, r_2) \subset U_2 \cap B(x_1, r_1)$ and moreover by assuming that $0 < r_2 < (r_1 - d(x_1, x_2))/2 \wedge 1/2$ we get $\overline{B}(x_2, r_2) \subset B(x_1, r_1)$ as well (since if $d(y, x_2) \leq r_2$ then by triangle inequality $d(x_1, y) \leq d(x_1, x_2) + d(x_2, y) < (r_1 + d(x_1, x_2))/2 < r_1$).

Suppose we have constructed $x_1, \dots, x_n$ and $r_1, \dots, r_n$ with the property that $B_j(x_j, r_j) \subset U_j \cap B(x_{j-1}, r_{j-1}) \subset U_1 \cap \dots \cap U_j$, $\overline{B}(x_j, x_j) \subset B(x_{j-1}, r_{j-1})$ and $r_j < 1/j$ for all $j = 2, \dots, n$. The arguing just as above by density of U_{n+1} we know that $B(x_n, r_n) \cap U_{n+1}$ is nonempty and open and therefore we may pick $x_{n+1} \in B(x_n, r_n) \cap U_{n+1}$ and $0 < r_{n+1} \leq 1/(n+1)$ small enough such that $\overline{B}(x_{n+1}, r_{n+1}) \subset B(x_n, r_n) \cap U_{n+1}$.

CLAIM 1.47.1. $x_1, x_2, \dots$ is a Cauchy sequence and $\lim_{n\to\infty} x_n \in \cap_{n=1}^{\infty} U_n$.

For a given $N \in \mathbb{N}$ and $n, m \geq N$, we know that $x_n, x_m \in B(x_N, r_N)$ and thus $d(x_m, x_n) < 2r_N \leq 2/N$; thus $x_1, x_2, \dots$ is Cauchy. By completeness of S we may take the limit $x = \lim_{n\to\infty} x_n$. Let $n \in \mathbb{N}$. By construction, $x_m \in B(x_n, r_n)$ for all $m \geq n$ and therefore $x \in \overline{B}(x_n, r_n)$. Also by construction $\overline{B}(x_{n+1}, r_{n+1}) \subset B(x_n, r_n) \subset U_n$ so we see that $x \in \cap_{n=1}^{\infty} U_n$. $\square$

CHAPTER 2

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 measurable, a means of measuring those sets and a theory of integrating suitable functions using that measure.

1. Measurable Spaces

Definition 2.1. A non-empty collection $\mathcal{A}$ of subsets of a set Ω is called a σ-algebra if given $A, A_1, A_2, \dots \in \mathcal{A}$ 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 2.2. *Let σ-algebra $\mathcal{A}$ in Ω, and $A_1, A_2, \dots \in \mathcal{A}$,*

(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 2.3. *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 $A, A_1, A_2, \dots$ 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 2.4. 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 2.5. Given a metric space S, the Borel σ-algebra $\mathcal{B}(S)$ is the σ-algebra generated by the open sets on S.

LEMMA 2.6. *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

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

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 2.7. *For $A, B, B_1, B_2, \dots \subset T$, 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 $B \setminus A = B \cap A^c$. $\square$

LEMMA 2.8. *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 $A, A_1, A_2, \dots \in \mathcal{S}'$. Since $\mathcal{S}' = f^{-1}\mathcal{T}$, there exist $B, B_1, B_2, \dots \in \mathcal{T}$ such that $A = f^{-1}(B)$ and $A_i = f^{-1}(B_i)$ for $i = 1, 2, \dots$. 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 $B, B_1, B_2, \dots \in \mathcal{T}'$ 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 2.9. *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 2.10. *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 $A_1, A_2, \dots$ 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. $\square$

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 2.11. 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 2.12. *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 $\mathcal{T}' = \left\{B \subset T; f^{-1}B \in S\right\}$. By our assumption, we have $\mathcal{C} \subset \mathcal{T}'$. Furthermore we know from Lemma 2.8 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. $\square$

Lemma 2.13. *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 $(g \circ f)^{-1}(B) = f^{-1}(g^{-1}(B))$ and the measurability of f and g. $\square$

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 2.14. *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. ☐

From the measurability of limits of real valued functions we can also generalize to measurability of limits in arbitrary metric spaces.

LEMMA 2.15. *Let (S, d) be a metric space and let $f_1, f_2, \dots$ be measurable functions $(\Omega, \mathcal{A})$ to $(S, \mathcal{B}(S)$, then $\lim_{n\to\infty} f_n$ is measurable if it exists.*

PROOF. Let $g : S \to \mathbb{R}$ be an arbitrary continuous function. Then g is Borel measurable and therefore $g \circ f_n$ are Borel measurable real valued functions. Moreover by continuity of g we know that $\lim_{n\to\infty} g \circ f_n = g \circ f$. Therefore by Lemma 2.14 we can conclude that $g \circ f$ is Borel measurable for all continuous $g : S \to \mathbb{R}$.

Now let $U \subset S$ be an open set and define $g_n(s) = nd(s, U^c) \wedge 1$ so that g_n are continuous functions such that $g_n \uparrow \mathbf{1}_U$. We know that $g_n \circ f$ are Borel measurable hence if follows that $\mathbf{1}_U \circ f$ is Borel measurable by another application of Lemma 2.14 which shows that $\{f \in U\}$ is measurable. Measurability of f follows from the fact that open sets generate the Borel σ-algebra and application of Lemma 2.12. ☐

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 2.16. 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 2.17. *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 $c_1, \dots, c_n$ 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 $c_1, \dots, c_n$ (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). $\square$

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 2.18. *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 2.17. $\square$

As an application of approximation by simple functions,

LEMMA 2.19. *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 2.14 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. $\square$

DEFINITION 2.20. Given two measurable functions f, g on the same measurable space $(\Omega, \mathcal{A})$, we say that f is g-measurable if $\sigma(f) \subset \sigma(g)$.

TODO: Where is the right place to introduce this concept? While the basic results of measure theory can be formulated in terms of general measurable spaces certain more advanced results require topological assumptions that prevent the wildness of set theory from taking over. For the results of this nature in which we are interested what is required is that the measure space look sufficiently like the Borel algebra on the $\mathbb{R}$. Somewhat surprisingly such a constraint isn't too severe (as we will show later) and for the purposes of these notes (and following the lead of Kallenberg) we will settle on the following definitions to capture these restrictions.

Definition 2.21. Two measure spaces $(S,\mathcal{S})$ and $(T,\mathcal{T})$ are said to be *Borel isomorphic* if there exists a bijection $f : S \to T$ such that both f and f^{-1} are measurable.

Definition 2.22. A measurable space $(S,\mathcal{S})$ is said to be a *Borel space* if it is Borel isomorphic to a Borel subset of $[0,1]$.

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 and gives us a chance to show how results may carry over from $\mathbb{R}$ to general Borel spaces.

Lemma 2.23. *Let $(S,\mathcal{S})$ be a Borel space and let $f : (\Omega,\mathcal{A}) \to S$ and $g : (\Omega,\mathcal{A}) \to (T,\mathcal{T})$ be measurable. Then f is g-measurable if and only if there exists measurable $h : T \to S$ such that $f = h \circ g$.*

Proof. For the if direction, assume $f = h \circ g$. Then for $B \in \mathcal{B}([0,1])$, 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, we first assume that $(S,\mathcal{S}) = ([0,1],\mathcal{B}([0,1]))$. Assume 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 $h = h_+ - h_-$.

Now let us assume that S is a Borel subset of $[0,1]$ and note that every measurable subset of S is over the form $A \cap S$ for a Borel subset $A \subset [0,1]$ and thus f is also g-measurable when considered as a function from Ω to $[0,1]$. By what we have just proven applied to f, we get $\tilde{h} : T \to [0,1]$ such that $\tilde{h} \circ g = f$. Because of the latter identity, we know that $\tilde{h}(g(\Omega)) \subset S$ however it is not necessarily the case that $tildeh(T) \subset S$. Since S is a Borel subset of $[0,1]$, we know that $\tilde{h}^{-1}(S)$ is $\mathcal{T}$-measurable and therefore we can pick an arbitrary point $s_0 \in S$ and define

$$h(t) = \begin{cases} \tilde{h}(t) & \text{if } t \in \tilde{h}^{-1}(S) \\ s_0 & \text{otherwise} \end{cases}$$

and note that we now have $h : T \to S$ and $f = h \circ g$.

It remains to extend the argument to general Borel spaces S. Assume that $j : S \to A \subset [0,1]$ is a Borel isomorphism to a Borel subset A. We can define

$\tilde{h} : T \to A$ such that $j \circ f = h \circ g$ by the above argument. Now let $h = j^{-1} \circ \tilde{h}$ so we have $h : T \to S$ and $f = h \circ g$. □

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

(i) $\Omega \in \mathcal{C}$.
(ii) for all $A, B \in \mathcal{C}$ if $A \subset B$, then $B \setminus A \in \mathcal{C}$.
(iii) for all $A_n \in \mathcal{C}$ if $A_1 \subset A_2 \subset \cdots$ and $A_n \uparrow A$, then $A \in \mathcal{C}$.

DEFINITION 2.25. 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 2.26. *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 $A_1, A_2, \dots \in \mathcal{C}$. 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 2.27 (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 2.26 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 $C \cap (A \setminus B) = (C \cap A) \setminus (C \cap B) \in \lambda(\mathcal{C})$. 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 2.27 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. Measures and Integration

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

Definition 2.28. 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 $A_1, A_2, \dots \in \mathcal{A}$ 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 2.29. 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 2.30. *Given a measure space $(\Omega, \mathcal{A}, \mu)$, and sets $A_1, A_2, \dots \in \mathcal{A}$.*

(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. $\square$

Lemma 2.31. *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 $A_1, A_2, \dots \in \mathcal{A}$.*

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 2.30 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 2.32. 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 2.33. *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 2.34. *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 $\int f \geq \int g$. $\square$

COROLLARY 2.35. *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 2.36. 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 2.37. *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. $\square$

Corollary 2.38. *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 2.39. *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 2.34.

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 2.40 (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 2.41 (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 2.42 (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 2.43. *[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 2.44. *[Tonelli's Theorem for Integrals and Sums]Given $f_1, f_2, \dots$ 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 2.45. *[Fatou's Lemma]Given $f_1, f_2, \dots$ 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

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

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 $n \to \infty$. $\square$

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

DEFINITION 2.46. 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 2.47. *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

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

$\square$

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

LEMMA 2.48. *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 2.49. 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 2.50. *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 2.51. *[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 2.52. *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 2.51. □

LEMMA 2.53. *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

$$
\begin{aligned}
\mu \circ f^{-1}\left(\bigcup_{i=1}^{\infty} A_i\right) &= \mu\left(\bigcup_{i=1}^{\infty} f^{-1}(A_i)\right) && \text{by Lemma 2.7} \\
&= \sum_{i=1}^{\infty} \mu(f^{-1}(A_i)) && \text{by countable additivity of measure} \\
&= \sum_{i=1}^{\infty} \mu \circ f^{-1}(A_i) && \text{by definition of push forward}
\end{aligned}
$$

□

DEFINITION 2.54. 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 2.55 (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 2.56. 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 2.57. *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. □

3.1. 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 2.55 and Lemma 2.57 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.

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 2.58. 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 countable 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)!

PROPOSITION 2.59. *Let $S_1, S_2, \dots$ be topological spaces then $\mathcal{B}(S_1) \otimes \mathcal{B}(S_2) \otimes \cdots \subset \mathcal{B}(S_1 \times S_2 \times \cdots)$. If every S_n is second countable then $\mathcal{B}(S_1) \otimes \mathcal{B}(S_2) \otimes \cdots = \mathcal{B}(S_1 \times S_2 \times \cdots)$.*

PROOF. $\mathcal{B}(S_1) \otimes \mathcal{B}(S_2) \otimes \cdots$ is the σ-algebra generated by cylinder sets $A_n \times \prod_{m \neq n} S_m$ for $A_n \in \mathcal{B}(S_n)$ so it suffices to show that such cylinder sets are in $\mathcal{B}(S_1 \times S_2 \times \cdots)$. This is clearly true for the case of A_n open since in this case we have a cylinder set for the product topology. On the other hand the set of all $A_n \subset S_n$ for which $A_n \times \prod_{m \neq n} S_m \in \mathcal{B}(S_1 \times S_2 \times \cdots)$ is easily seen to be a λ-system so we may apply the π-λ Theorem 2.27.

On the other hand, assume that each S_n is second countable. It follows that $S_1 \times S_2 \times \cdots$ is second countable and therefore every open set is a countable union

TODO: Finish... I think there is a result that $\mathcal{B}(S \times S) = \mathcal{B}(S) \otimes \mathcal{B}(S)$ implies that S is second countable (check Van der Vaart and Wellner). $\square$

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 2.60. *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 2.12. $\square$

TODO: Show that the collection $(f_{t_1}, \dots, f_{t_n}) \in A$ is a π-system (which is clearly generating by the previous Lemma). Use this fact in Lemma 9.6 and Theorem 13.11.

5. Null Sets and Completions of Measures

LEMMA 2.61. *Let $f \geq 0$ be a measurable function then $\int f \, d\mu = 0$ if and only if $f = 0$ μ-almost everywhere.*

PROOF. First suppose that f is simple with canonical representation $f = c_1 \mathbf{1}_{A_1} + \cdots + c_n \mathbf{1}_{A_n}$ where $c_i > 0$. Then $\int f \, d\mu = c_1 \mu(A_1) + \cdots + c_n \mu(A_n)$ and it follows the positivity of the c_i that $\int f \, d\mu = 0$ if and only if $\mu(A_1) = \cdots = \mu(A_n)$.

Now for a general non-negative measurable f we can find simple $0 \leq f_n \uparrow f$ such that $\int f \, d\mu = \lim_{n \to \infty} \int f_n \, d\mu$. If $\int f \, d\mu = 0$ then by monotonicity of integral and the result for simple functions we know that $f_n \neq 0$ almost everywhere. Taking the countable union of sets of measure zero we know that $f_n \neq 0$ for all n on a set of measure zero and therefore taking limits we conclude $f \neq 0$ on a set of measure zero. Conversely if $f \neq 0$ on a set of measure zero then since f_n is an increasing sequence if follows that each $f_n \neq 0$ on a set of measure zero and applying the result for simple functions $\int f_n \, d\mu = 0$ for all n. Taking the limits of the integrals we see that $\int f \, d\mu = 0$. ☐

LEMMA 2.62. *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, let $\mathcal{F}$ be a sub σ-algebra of $\mathcal{A}$ and let $\mathcal{F}^\mu$ be the μ-completion of $\mathcal{F}$. The for every $A \in \mathcal{F}^\mu$ there exist $A_-, A_+ \in \mathcal{F}$ such that $A_- \subset A \subset A_+$ and $\mu(A_-) = \mu(A_+)$.*

PROOF. TODO ☐

6. 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 2.63. 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 $A_1, A_2, \dots \subset \Omega$, then $\mu\left(\bigcup_{i=1}^\infty A_i\right) \leq \sum_{i=1}^\infty \mu(A_i)$.

DEFINITION 2.64. 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 2.65. 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 2.66. *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 $A_1, A_2, \dots$ 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. $\square$

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 2.67. *[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 2.68. *[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 $A_1, A_2, \dots \subset \mathbb{R}$. Pick $\epsilon > 0$ and then for each A_n we take a countable cover by intervals $I_{n1}, I_{n2}, \dots$ 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 1.32), 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 2.69. A subset $A \subset \mathbb{R}$ is *Lebesgue measurable* if A is λ-measurable with respect to the Lebesgue outer measure.

Lemma 2.70. *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 $I_1, I_2, \dots$ 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 2.71 (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 2.30)

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

Application of the π-λ Theorem (Theorem 2.27) 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 2.30) 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 2.67).

Proof. The existence of Lebesgue measure clearly follows from Lemma 2.66 applied to the outer measure constructed in Lemma 2.68. The fact that the σ-algebra of the restriction contains the Borel sets follows from Lemma 2.70.

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 2.71 to get uniqueness. □

TODO: Show that the σ-algebra of λ-measurable sets is the completion of the Borel σ-algebra.

Definition 2.72. 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 $\mu(\Omega_i) < \infty$.

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

Definition 2.73. 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 2.74. 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 $A_1, A_2, \dots \in \mathcal{A}_0$ 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 2.75. *A pre-measure is finitely additive and montonic. That is to say given any disjoint $A_1, \dots, A_n \in \mathcal{A}_0$ 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.

The construction of an outer measure from a set function requires almost no assumptions.

LEMMA 2.76. *Let $\mathcal{C}$ be collection of subsets of Ω such that $\emptyset \in \mathcal{C}$ and let $\mu_0 : \mathcal{C} \to \overline{\mathbb{R}}_+$ satisfy $\mu_0(\emptyset) = 0$ then the set function $\mu^* : 2^\Omega \to \overline{\mathbb{R}}_+$ defined by*

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

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 $B_n \in \mathcal{C}$, 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 $A_1, A_2, \dots$ 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 $B_{1n}, B_{2n}, \dots$ 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 2.77. *Let μ_0 be a set function on a Boolean algebra $(\Omega, \mathcal{A}_0)$ such that μ_0 is finitely additive on disjoint sets. If μ^* is the outer measure constructed in Lemma 2.76 and $A \in \mathcal{A}_0$ then A is μ^*-measurable.*

PROOF. Since $\emptyset \cap \emptyset = \emptyset$ thus by finite additivity $\mu_0(\emptyset) = 2\mu_0(\emptyset)$ and therefore $\mu_0(\emptyset) = 0$. Therefore we can indeed construct an outer measure from μ_0.

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 $B_1, B_2, \dots$ 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 2.78. *Given a pre-measure μ_0 on a Boolean algebra $(\Omega, \mathcal{A}_0)$ and the outer measure μ^* constructed in Lemma 2.76, 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 trivially true so we may assume $\mu^*(A) < \infty$. Let $\epsilon > 0$ be given and pick $A_1, A_2, \dots \in \mathcal{A}_0$ 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 2.79. 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 $C_1, \dots, C_n \in \mathcal{D}$ such that $A \setminus B = \cup_{j=1}^n C_j$

Example 2.80. The set of intervals $(a, b]$ with $a \leq b$ is a semiring. To be excruciatingly explicit we have the formulae

$$(a, b] \cap (c, d] = (a \vee c, (b \wedge d) \vee a \vee c]$$

and

$$(a, b] \setminus (c, d] = (a \wedge c, (a \vee c) \wedge b \wedge d] \cup a \vee c \vee (b \wedge d), c \vee d]$$

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

Definition 2.81. 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 2.82. *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 $\mathcal{D}$. ∎

Example 2.83. The set of disjoint unions of intervals $(a, b]$ with $a \leq b$ is a ring. This follows from the general result Lemma 2.82 but later on we shall have some use for the explicit formula

$$\begin{aligned}&(a, b] \cup (c, d] \\ &= (a \wedge c, (a \vee c) \wedge b \wedge d] \cup (a \vee c, (b \wedge d) \vee a \vee c] \cup a \vee c \vee (b \wedge d), c \vee d]\end{aligned}$$

which decomposes a union of half open intervals into a disjoint union of half open intervals.

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

Lemma 2.84. *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 2.8 proven for σ-algebras.

Lemma 2.85. *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 2.8 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 2.86. *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 $A_1, \dots, A_n \in \mathcal{D}$. 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. $\square$

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

Lemma 2.87 (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 $S_1, S_2, \dots$ 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 2.57 that for any measurable g, $\int g\, d\mu_n = \int g \mathbf{1}_{S_n}\, d\mu$. Putting these observations together,

$$\begin{aligned}
\int f(s,t)\, d\mu(s) &= \int f(s,t) \sum_{n=1}^{\infty} \mathbf{1}_{S_n}(s)\, d\mu(s) && \text{since } S_n \text{ is a partition of } S \\
&= \sum_{n=1}^{\infty} \int f(s,t) \mathbf{1}_{S_n}(s)\, d\mu(s) && \text{by Corollary 2.44} \\
&= \sum_{n=1}^{\infty} \int f(s,t)\, d\mu_n(s)
\end{aligned}$$

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. $\square$

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

THEOREM 2.88 (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 2.71 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 2.87. 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 $A_1, A_2, \dots \in \mathcal{S} \otimes \mathcal{T}$. 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 2.44) 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

$$\int \tilde{f} d(\mu\otimes\nu) = \int f d(\mu\otimes\nu)$$

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

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

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 2.89. 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 2.90 (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

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

Now fix $\mathcal{I}_2, \ldots, \mathcal{I}_n$ and consider $\nu(A) = \frac{1}{|\mathcal{I}_2| \cdots |\mathcal{I}_n|} \mu \left(A \times \mathcal{I}_2 \times \cdots \times \mathcal{I}_n \right)$ 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 $B_1 \in \mathcal{B}(\mathbb{R})$,

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

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 2.91. *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 $\lambda^n_P(A) = \lambda^a(PA)$ is translation invariant. To see this, assume we are given $h \in \mathbb{R}^n$ and note that

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

Therefore we know that $\lambda^n_P = c\lambda^n$ 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 2.92. *[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! ☐

6.3. Further Properties of Outer Measures. The results in this section will not be used until much later in the text. There will be little harm in the reader skipping these results and coming back to them when they are referenced.

One of the fundamental properties of countably additive measures is continuity (Lemma 2.30). It turns out that by adding a hypothesis we get a continuity property for outer measures as well.

DEFINITION 2.93. An outer measure μ^* is said to be *regular* if for each set A there exists a μ^*-measurable set B such that $A \subset B$ and $\mu^*(A) = \mu^*(B)$. Any such B is called a *measurable cover* of A (TODO: Compare with Van der Vaart and Wellner).

Examples of regular outer measures aren't too hard to come up with. Indeed, any outer measure that is constructed from a premeasure is regular.

PROPOSITION 2.94. *If an outer measure μ^* on a set Ω is induced from a finitely additive non-negative set function μ_0 on a Boolean algebra $\mathcal{A} \subset 2^\Omega$ then μ^* is regular. If in addition μ_0 is countably additive and there exists $\Omega_1, \Omega_2, \ldots \in \mathcal{A}$ with $\Omega = \cup_{n=1}^\infty \Omega_n$ and $\mu_0(\Omega_n) < \infty$ for all $n \in \mathbb{N}$ then a set A is μ^*-measurable if and only if there exists $B \in \mathcal{A}_{\delta\sigma}$ and a μ^*-null set N such that $B \cap N = \emptyset$ and $A = B \cup N$.*

PROOF. First assume only that μ_0 is finitely additive on disjoint sets. We establish a sequence of claims from which the result follows.

CLAIM 2.94.1. $\mathcal{A}_\sigma$ is closed under finite intersections

Let $A_1, A_2, \ldots \in \mathcal{A}$ and $B_1, B_2, \ldots \in \mathcal{A}$, let $A = \cup_{n=1}^\infty A_n$ and $B = \cup_{m=1}^\infty B_m$ and compute

$$\begin{aligned} A \cap B &= \cup_{n=1}^\infty \left(A_n \cap (\cup_{m=1}^\infty B_m)\right) \\ &= \cup_{n=1}^\infty \cup_{m=1}^\infty A_n \cap B_m \end{aligned}$$

Since $\mathcal{A}$ is a Boolean algebra it follows that $A_n \cap B_m \in \mathcal{A}$ for every m, n and therefore $A \cap B \in \mathcal{A}_\sigma$. The claim follows by a simple induction.

CLAIM 2.94.2. For any set $A \subset \Omega$ we have $\mu^*(A) = \inf\{\mu^*(B) \mid B \in \mathcal{A}_\sigma, A \subset B\}$

If $\mu^*(A) = \infty$ then for any $A \subset B$ we have $\mu^*(B)$ so the result follows. Suppose that $\mu^*(A) < \infty$ and let $\epsilon > 0$ be given. By definition of μ^* there exist $B_1, B_2, \ldots \in \mathcal{A}$ such that $\mu^*(A) \leq \sum_{n=1}^\infty \mu_0(B_n) < \mu^*(A) + \epsilon$. Let $B = \cup_{n=1}^\infty B_n$ and note that by subadditivity and monotonicity of μ^* and the fact that $\mu^* \leq \mu_0$ on $\mathcal{A}$ we have

$$\mu^*(A) \leq \mu^*(B) \leq \sum_{n=1}^\infty \mu^*(B_n) = \sum_{n=1}^\infty \mu_0(B_n) < \mu^*(A) + \epsilon$$

Since ϵ was arbitrary the claim follows.

CLAIM 2.94.3. For any set $A \subset \Omega$ there is a $B \in \mathcal{A}_{\sigma\delta}$ such that $A \subset B$ and $\mu^*(A) = \mu^*(B)$.

From the previous claim we may pick $C_1, C_2, \ldots \in \mathcal{A}_\sigma$ such that $A \subset C_n$ for all $n \in \mathbb{N}$ and $\lim_{n\to\infty} \mu^*(C_n) = \mu^*(A)$. Define $B_n = \cap_{j=1}^n C_j$. Since $\mathcal{A}_\sigma$ is closed under finite intersections $B_n \in \mathcal{A}_\sigma$ for every $n \in \mathbb{N}$. Also $A \subset B_n \subset C_n$ and therefore $\lim_{n\to\infty} \mu^*(B_n) = \mu^*(A)$. Let $B = \cap_{n=1}^\infty B_n \in \mathcal{A}_{\sigma\delta}$. Since the B_n are decreasing we have $A \subset B \subset B_n$ for every $n \in \mathbb{N}$ and thus by monotonicity and subadditivity of μ^* we get

$$\mu^*(A) \leq \mu^*(B) \leq \lim_{n\to\infty} \mu^*(B_n) = \mu^*(A)$$

and $\mu^*(A) = \mu^*(B)$ follows.

By Lemma 2.77 we know that every $A \in \mathcal{A}$ is μ^*-measurable and therefore every $A \in \mathcal{A}_{\sigma\delta}$ is μ^*-measurable. The regularity of μ^* follows from the previous claim.

Now assume that μ_0 is countably additive and let $\Omega_1, \Omega_2, \ldots \in \mathcal{A}$ be chosen so that $\Omega = \cup_{n=1}^\infty \Omega_n$ and $\mu_0(\Omega_n) < \infty$. By Lemma 2.78 we have $\mu_0 = \mu^*$ on $\mathcal{A}$. In particular $\mu^*(\Omega_n) = \mu_0(\Omega_n) < \infty$ for all $n \in \mathbb{N}$. Note that by passing to $\Omega_n \setminus (\Omega_1 \cup \cdots \cup \Omega_{n-1})$ we may assume that the Ω_n are disjoint. Let A be a μ^*-measurable set. For each $n \in \mathbb{N}$ we apply the last claim to the measurable set $A^c \cap \Omega_n$ and get a $B_n' \in \mathcal{A}_{\sigma\delta}$ such that $A^c \cap \Omega_n \cup B_n'$ and $\mu^*(B_n') = \mu^*(A^c \cap \Omega_n)$. Let $B_n = B_n' \cap \Omega_n$ and note that $B_n \in \mathcal{A}_{\sigma\delta}$, $A^c \cap \Omega_n \subset B_n \subset \Omega_n$ and $\mu^*(B_n) = \mu^*(A^c \cap \Omega_n)$. Let $N_n = B_n \setminus (A^c \cap \Omega_n)$ so that $\mu^*(N_n) = 0$. Then

$$\begin{aligned} \Omega_n \cap A &= \Omega_n \setminus A^c = \Omega_n \setminus (\Omega_n \cap A^c) \\ &= \Omega_n \setminus (B_n \setminus N_n) = \Omega_n \cap (B_n \cap N_n^c)^c = (\Omega_n \cap B_n^c) \cup (\Omega_n \cap N_n) \\ &= \Omega_n \setminus B_n \cup N_n \end{aligned}$$

Since $B_n \in \mathcal{A}_{\sigma\delta}$ it follows that $\Omega_n \setminus B_n \in \mathcal{A}_{\delta\sigma}$. Now write

$$A = \cup_{n=1}^{\infty} \Omega_n \cap A = \cup_{n=1}^{\infty} \Omega_n \setminus B_n \cup N_n$$

and use the fact that a countable union of sets in $\mathcal{A}_{\delta\sigma}$ is in $\mathcal{A}_{\delta\sigma}$ and a countable union of null sets is null. $\square$

PROPOSITION 2.95. *Let μ^* be an outer measure on a set Ω and let μ^+ be the outer measure induced by μ^* and the σ-algebra of μ^*-measurable sets. Then $\mu^* = \mu^+$ if and only if μ^* is regular.*

PROOF. First note that for an arbitrary outer measure μ^* we have $\mu^*(A) \leq \mu^+(A)$, for if $A_1, A_2, \dots$ is a countable cover of A then by countable subadditivity of μ^* we have $\mu^*(A) \leq \sum_{i=1}^{\infty} \mu^*(A_j)$. Now take the infimum of the right hand side over all countable subcovers with μ^*-measurable A_j.

Suppose μ^* is regular. If A is a set and B is its measurable cover we have $\mu^+(A) \leq \mu^*(B) = \mu^*(A)$ as well and it follows that $\mu^* = \mu^+$.

On the other hand suppose that $\mu^* = \mu^+$ and let A be a subset of Ω. If $\mu^*(A) = \infty$ then Ω is measurable cover so we may assume that $\mu * (A) < \infty$. For each $n \in \mathbb{N}$ we may pick μ^*-measurable $C_1, C_2,$ such that $A \subset \cup_{j=1}^{\infty} C_j$ and $\sum_{j=1}^{\infty} \mu^*(C_j) < \mu^*(A) + 1/n$. Since the μ^*-measurable sets are a σ-algebra the set $B_n = \cup_{j=1}^{\infty} C_j$ is μ^*-measurable and by subadditivity of μ^*

$$\mu^*(B_n) \leq \sum_{j=1}^{\infty} \mu^*(C_j) < \mu^*(A) + 1/n$$

Now let $B = \cap_{n=1}^{\infty} B_n$ which is μ^*-measurable and by continuity of measure $\mu^*(B) \leq \lim_{n\to\infty} \mu^*(A) + 1/n = \mu^*(A)$. On the other hand $A \subset B$ and therefore by monotonicity we get $\mu^*(A) = \mu^*(B)$. $\square$

TODO: Don't know if this result will be useful; maybe make an exercise (I got it from Biskup's notes)

PROPOSITION 2.96. *Let μ^* be a finite regular outer measure on Ω then A is μ^*-measurable if and only if*

$$\mu^*(\Omega) = \mu^*(A) + \mu^*(A^c)$$

PROOF. If A is μ^*-measurable then the result follows from the definition of μ^*-measurability applied to the set Ω so only the other direction requires proof.

Let E be a subset of Ω and let B be a measurable cover of E. By measurability of B we have

$$\begin{aligned}
\mu^*(A) &= \mu^*(A \cap B) + \mu^*(A \cap B^c) \\
\mu^*(A^c) &= \mu^*(A^c \cap B) + \mu^*(A^c \cap B^c) \\
\mu^*(\Omega) &= \mu^*(B) + \mu^*(B^c)
\end{aligned}$$

From these three facts we compute using subadditivity and monotonicity of μ^* to get

$$\begin{aligned}\mu^*(E) = \mu^*(B) &= \mu^*(\Omega) - \mu^*(B^c) = \mu^*(A) + \mu^*(A^c) - \mu^*(B^c) \\ &= \mu^*(A \cap B) + \mu^*(A \cap B^c) + \mu^*(A^c \cap B) + \mu^*(A^c \cap B^c) - \mu^*(B^c) \\ &\geq \mu^*(A \cap B) + \mu^*(A^c \cap B) \\ &\geq \mu^*(A \cap B) + \mu^*(A^c \setminus E)\end{aligned}$$

Since the opposite inequality follows by subadditivity, measurability of A follows. □

TODO: Show that an outer measure induced by a premeasure is regular (or more generally an outer measure constructed as in Proposition 2.76?)

PROPOSITION 2.97. *Let μ^* be a regular outer measure on a set Ω and let $A, A_1, A_2, \dots$ be subsets of Ω then if $A_i \uparrow A$ then $\mu^* A_i \uparrow \mu^* A$.*

PROOF. Pick μ^*-measurable sets C_k such that $A_k \subset C_k$ and $\mu^* A_k = \mu^* C_k$ and define $B_k = \cap_{j=k}^\infty C_j$. Clearly, B_k is μ^*-measurable and $A_k \subset B_k$. Also since $A_k \subset B_k$ we have $\mu^* A_k = \mu^* C_k \leq \mu^* B_k$ and since $B_k \subset C_k$ we have $\mu^* B_k \leq \mu^* C_k$; it follows that $\mu^* A_k = \mu^* B_k$ as well.Since $B_1 \subset B_2 \subset \cdots$ and $\cup_{k=1}^\infty A_k \subset \cup_{k=1}^\infty B_k$ we can use continuity of measure Lemma 2.30 and monotonicity of μ^* to see

$$\lim_{k \to \infty} \mu^* A_k = \lim_{k \to \infty} \mu^* B_k = \mu^* \cup_{k=1}^\infty B_k \geq \mu^* \cup_{k=1}^\infty A_k$$

On the other hand since $A_k \subset \cup_{j=1}^\infty A_j$ we have

$$\lim_{k \to \infty} \mu^* A_k \leq \mu^* \cup_{k=1}^\infty A_k$$

□

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.

DEFINITION 2.98. Let (S, d) be a metric space with an outer measure μ^* then we say that μ^* is a *metric outer measure* if and only if $\mu^*(A \cup B) = \mu^*(A) + \mu^*(B)$ for all A, B such that $d(A, B) > 0$.

TODO: Compare with Biskup's notes

LEMMA 2.99 (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 μ^* is metric.*

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. □

EXAMPLE 2.100. Let $\mathcal{C}$ be the set of closed squares in $\mathbb{R}^2$ and let μ_0 be the length of a side of each square. Now construct the outer measure μ^* as in Proposition 2.76. Let $0 < \epsilon < 1/4$ be chosen and consider the closed intervals $I_1 = [0, 1/2 - \epsilon]$ and $I_2 = [1/2 + \epsilon, 1]$. Define

$$A = I_1 \times I_1 \cup I_1 \times I_2 \cup I_2 \times I_1 \cup I_2 \times I_2$$

Consider covering A by countable numbers of squares. We claim that if A is not covered by a single square then it is covered by at least 3 squares and the sum of the sides of those squares exceeds 1. From this the minimal covering of A by squares is the single square $[0,1] \times [0,1]$ and $\mu^*(A) = 1$. (TODO: Show this more carefully).

On the other hand we claim $\mu^*(I_i \times I_j) = 1/2 - \epsilon$ for $i, j \in \{1, 2\}$ and from this we see that

$$\mu^*(A) = 1 < 4(1/2 - \epsilon) = \sum_{i,j} \mu^*(I_i \times I_j)$$

and it follows that μ^* is not metric.

7. 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 2.101. A *bounded signed measure* on a measurable space $(\Omega, \mathcal{A})$ is a bounded function $\nu : \mathcal{A} \to \mathbb{R}$ such that $\nu(\emptyset) = 0$ and for every disjoint $A_1, A_2, \dots \in \mathcal{A}$ such that $\sum_{n=1}^\infty |\nu(A_n)| < \infty$, we have $\nu(\bigcup_{n=1}^\infty A_n) = \sum_{n=1}^\infty \nu(A_n)$

Note that a bounded signed measure is finitely additive (just take infinitely many copies of the empty set use countable additivity). It is important to note that a bounded signed measure is continuous in the same way that an ordinary measure is.

Proposition 2.102. *Let ν be a bounded signed measure on the measurable space $(\Omega, \mathcal{A})$. If $A, A_1, A_2, \ldots \in \mathcal{A}$, $\sum_{n=1}^\infty |\nu(A_n \setminus A_{n-1})| < \infty$ and $A_n \uparrow A$ then $\nu(A) = \lim_{n\to\infty} \nu(A_n)$. If $A, A_1, A_2, \ldots \in \mathcal{A}$, $\sum_{n=1}^\infty |\nu(A_n \setminus A_{n+1})| < \infty$ and $A_n \downarrow A$ then $\nu(A) = \lim_{n\to\infty} \nu(A_n)$.*

Proof. Continuity follows from the same proof as Lemma 2.30. Defining $B_1 = A_1$ and $B_n = A_n \setminus A_{n-1}$ for $n > 1$ we see that B_n are disjoint, $A_n = \cup_{j=1}^n B_j$

and $A = \cup_{j=1}^\infty B_j$. By assumption, $\sum_{j=1}^\infty |\nu(B_j)| < \infty$ and therefore we may apply countable and finite additivity to see

$$\nu(A) = \sum_{j=1}^\infty \nu(B_j) = \lim_{n\to\infty} \sum_{j=1}^n \nu(B_j) = \lim_{n\to\infty} \nu(A_n)$$

To see continuity under decreasing sequences of sets

$$\begin{aligned}\nu(A_1) - \nu(\cap_{j=1}^\infty A_j) &= \nu(A_1 \setminus \cap_{j=1}^\infty A_j) = \nu(\cup_{j=1}^\infty A_j \setminus A_{j+1}) \\ &= \sum_{j=1}^\infty \nu(A_j \setminus A_{j+1}) = \lim_{n\to\infty} \sum_{j=1}^n \nu(A_j \setminus A_{j+1})\end{aligned}$$

□

Equally important to note is that monotonicity and subadditivity fail for signed measures.

EXAMPLE 2.103. Let $\Omega = \{1, 2, 3\}$ and define $\nu(1) = \nu(2) = 1$ and $\nu(3) = -1$ then $0 = \nu(\{1, 3\}) < \nu(\{1\}) = 1$ and $1 = \nu(\{1, 3\} \cup \{2, 3\}) > \nu(\{1, 3\}) + \nu(\{2, 3\}) = 0$.

DEFINITION 2.104. 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 2.105. Lebesgue measure and any Dirac measure on $\mathbb{R}$ are mutually singular.

EXAMPLE 2.106. 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 2.107. 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 2.108. 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 2.109. *[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 $\nu A_+ = c$. By continuity of measure Proposition 2.102 we expect to be able to show this by taking a limit over sets with $\nu(A)$ getting arbitrarily close to c. Reflecting for a moment one realizes there are two reasons that a set A may have measure less than c. The first reason is that the set may be missing some positive mass that can be added and the second reason is that the set may have some negative mass that can be subtracted; of course these two reasons are not mutually exclusive. In the first case taking a union of sets improves the approximation in the second case taking an intersection improves the approximation therefore we expect our limiting process to involve both unions and intersections.

A trick is that signed measures are not subadditive hence taking a union does not always increase measure. The first thing we need is a simple bound on the damage that taking a union can do to approximations to the supremum.

CLAIM 2.109.1. Suppose we are given $A, A' \in \mathcal{A}$ such that $\nu A \geq c - \epsilon$ and $\nu A' \geq c - \epsilon'$ then $\nu(A \cup A') \geq c - \epsilon - \epsilon'$. If $B \subset A \setminus A'$ then $-\epsilon \leq \nu(B) \leq \epsilon + \epsilon'$.

To see the first part of the claim,

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

For the second part of the claim, the lower bound actually holds for any subset of A; if $B \subset A$ then $\nu(B) + \nu(A \setminus B) = \nu(A) \geq c - \epsilon$ by $\nu(A \setminus B) \leq c$ hence $\nu(B) \geq -\epsilon$ follows. For the upper bound, if $B \subset A \setminus A'$ then by finite additivity and the two lower bounds already established,

$$\begin{aligned}
\nu(B) &= \nu(A \cup A') - \nu(A') - \nu(A \setminus A' \setminus B) \\
&\leq c - (c - \epsilon') - (-\epsilon) = \epsilon + \epsilon'
\end{aligned}$$

Now approximate the supremum by taking $A_1, A_2, \cdots \in \mathcal{A}$ such that $\nu A_n \geq c - 2^{-n}$. By the first part of the above claim and a simple induction we have $\nu(\cup_{j=n+1}^m A_j) \geq c - \sum_{j=n+1}^m 2^{-j}$ for all $n < m$. Note that

$$\cup_{j=n+1}^{m+1} A_j \setminus \cup_{j=n+1}^m A_j = A_{m+1} \setminus \cup_{j=n+1}^m A_j \subset A_{m+1} \setminus A_m$$

and therefore by the second part of the prior claim, $\left|\nu(\cup_{j=n+1}^{m+1} A_j \setminus \cup_{j=n+1}^m A_j)\right| \leq 2^{-m-1} + 2^{-m} < 2^{-m+1}$ so that we may apply continuity of measure (Proposition 2.102) to conclude

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

Let $A_+ = \bigcap_{n=1}^\infty \bigcup_{i=n+1}^\infty A_i$ and by the same argument as above we have

$$\bigcup_{i=n}^{\infty} A_i \setminus \bigcup_{i=n+1}^{\infty} A_i \subset A_n \setminus A_{n+1}$$

and therefore

$$\sum_{n=1}^{\infty} \left| \nu \bigcup_{i=n}^{\infty} A_i \setminus \bigcup_{i=n+1}^{\infty} A_i \right| \leq \sum_{n=1}^{\infty} 2^{-n+1} < \infty$$

and we may apply Proposition 2.102 to conclude

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

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

$$\begin{aligned}
\nu_+ B &= \nu(A_+ \cap B) \\
\nu_- B &= -\nu(A_- \cap B)
\end{aligned}$$

CLAIM 2.109.2. $\nu_\pm$ are both measures.

We prove this for ν_+, this will imply that ν_- is also a measure by considering $-\nu$. Since $\nu(A_+) = \sup_{A\in\mathcal{A}} \nu(A)$ it follows that $\nu_+(B) \geq 0$ for all B; if not then $\nu(A_+ \setminus B) = \nu(A_+) - \nu(A_+ \cap B) > \nu(A_+)$ which is a contradiction. Let $B_1, B_2, \ldots \in \mathcal{A}$ be disjoint then $\sum_{j=1}^n \nu(A_+ \cap B_j) \leq \nu(A_+)$ and is an increasing sequence thus $\sum_{j=1}^\infty |\nu(A_+ \cap B_j)| = \sum_{j=1}^\infty \nu(A_+ \cap B_j)$ exists and is finite. There since ν is a bounded signed measure

$$\sum_{j=1}^\infty \nu_+(B_j) = \sum_{j=1}^\infty \nu(A_+ \cap B_j) = \nu(A_+ \cap \cup_{j=1}^\infty B_j) = \nu_+(\cup_{j=1}^\infty B_j)$$

and the claim follows.

Since $\nu_+(A_+^c) = \nu(\emptyset) = 0$ and $\nu_-(A_+) = -\nu(\emptyset) = 0$ it follows that ν_+ and ν_- are mutually singular. Furthermore by finite additivity we know that $\nu(B) = \nu(B \cap A_+) + \nu(B \cap A_-) = \nu_+(B) - \nu_-(B)$.

If $\nu = \mu_+ - \mu_-$ with $\mu_\pm$ bounded mutually singular measures then pick B such that $\mu_+(B^c) = 0$ and $\mu_-(B) = 0$ then we have

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

from which it follows that $\mu_-(A_+ \cap B^c) = 0$ since otherwise

$$\mu_+(A_+ \cap B) = \nu(A_+ \cap B) > \nu(A_+) = \sup_{A\in\mathcal{A}} \nu(A)$$

therefore $\mu_-(A_+) = \mu_-(A_+ \cap B) + \mu_-(A_+ \cap B^c) = 0$. By a similar argument we conclude that $\mu_+(A_-) = 0$ thus we may assume that $B = A_+$ and it follows that $\mu_\pm = \nu \mid_{A_\pm} = \nu_\pm$. □

THEOREM 2.110 (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 2.111. 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 2.112 (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

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

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

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

It follows that we also have the contrapositive assertion $c < G(y)$ 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 follows by Lemma 2.71 as local finiteness obviously implies σ-finiteness on $\mathbb{R}$. $\square$

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 2.113 (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 $F' = f$.*

Before we give the proof of the Fundamental Theorem we need a couple of lemmas.

LEMMA 2.114. *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 $I_1, \dots, I_n$ such that $\sum_{i=1}^n |I_i| \geq \frac{\lambda G}{4}$.*

PROOF. We begin by finding a compact set to focus attention on.

CLAIM 2.114.1. There is a compact set K such that $K \subset G$ and $\lambda(K) \geq \frac{3}{4}\lambda(G)$.

We first note that G is a countable union of compact sets (in fact a countable union of closed bounded intervals). For an open interval this is easy to see by construction and since G is open it is a countable union of disjoint open intervals (Lemma 1.16 the claim follows. Thus we may write $G = \cup_{n=1}^\infty [a_n, b_n]$ and if we define $K_j = \cup_{n=1}^j [a_n, b_n]$ each K_j is compact and by continuity of measure $\lim_{j\to\infty} \lambda(K_j) = \lambda(G)$. For j sufficiently large we have $\lambda(K_j) \geq \frac{3}{4}\lambda(G)$.

Since K is compact there is a finite set of intervals $J_1, \dots, J_k$ with $J_i \in \mathcal{I}$ such that $K \subset J_1 \cup \dots \cup J_k$. Define I_1 to be the the largest interval among the J_i and inductively define I_j so that

$$\lambda(I_j) = \max\{\lambda(J_i) \mid J_i \cap I_l = \emptyset \text{ for } l = 1, \dots, j-1\}$$

where we stop the iteration if there is no J_i that is disjoint for each of the I_l for $l = 1, \dots, j-1$.

Now observe that for every J_i there is an I_k such that $J_i \cap I_k \neq \emptyset$ and $\lambda(I_k) \geq \lambda(J_i)$. To see this note that by construction there must be a smallest index k such that $J_i \cap I_k \neq \emptyset$. If $k = 1$ then again by construction it follows that $\lambda(I_1) = \lambda(I_k) \geq \lambda(J_i)$; if $k > 1$ then $J_i \cap I_l = \emptyset$ for $l = 1, \dots, k-1$ and by construction we know $\lambda(I_k) \geq \lambda(J_i)$. Define $\hat{I}_k$ to be the open interval with the same center as I_k but length three times as large; it follows from $J_i \cap I_k \neq \emptyset$ and $\lambda(I_k) \geq \lambda(J_i)$ that $J_i \subset \hat{I}_k$ and therefore we get

$$K \subset J_1 \cup \dots \cup J_k \subset \hat{I}_1 \cup \dots \cup \hat{I}_n$$

By subadditivity we get

$$\frac{3}{4}\lambda G \leq \lambda K \leq \sum_{j=1}^n \lambda(\hat{I}_j) = 3\sum_{j=1}^n \lambda(I_j)$$

□

Now we prove the fundamental theorem of calculus in the special case of measures that are mutually singular to Lebesgue measure.

LEMMA 2.115. *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

$$A_\epsilon = \{x \in A \mid \limsup_{h \to 0} \frac{F(x+h) - F(x-h)}{h} > \epsilon\}$$

so that for $x \in A_\epsilon$ there exist arbitrarily small $h > 0$ such that

$$\mu(x-h, x+h] = F(x+h) - F(x-h) > \epsilon h = \frac{1}{2}\epsilon\lambda(x-h, x+h]$$

Note that A_ϵ is measurable since

$$\limsup_{h \to 0} \frac{F(x+h) - F(x-h)}{h} = \limsup_{n \to \infty} n\left(F(x+1/n) - F(x-1/n)\right)$$

is measurable (Lemma 2.14).

By openness of G_δ and by the above remarks, for any $x \in A_\epsilon$ we can pick $h > 0$ small enough so that $I_x = (x-h, x+h] \subset G_\delta$ and $2\mu(I_x)/\epsilon > \lambda(I_x)$. Since $A_\epsilon \subset \bigcup_{x \in A_\epsilon} I_x$, by the previous Lemma 2.114 we pick a finite disjoint set $I_{x_1}, \dots, I_{x_n}$ and note that

$$\lambda A_\epsilon \leq \lambda \bigcup_{x \in A_\epsilon} I_x \leq 4 \sum_{k=1}^n |I_{x_k}| \leq 4 \sum_{k=1}^n \frac{2\mu I_{x_k}}{\epsilon} = \frac{8}{\epsilon}\mu \bigcup_{k=1}^n I_{x_k} \leq \frac{8\delta}{\epsilon}$$

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

Proof. From Lemma 2.115 if follows that $\frac{d}{dx} F_s = 0$ a.s. so it suffices to assume that $F(x) = \int_0^x f \, d\lambda$ for a non-negative locally integrable f. The trick is to reduce this case to Lemma 2.115 as well. Let $q \in \mathbb{Q}$ be chosen arbitrarily, let $F_q(x) = \int_0^x (f-q)_+ \, d\lambda$ and consider the locally finite measure $\mu_q(0, x] = F_q(x)$. Since $(f-q)_+ = 0$ on $\{f \leq q\}$ we know

$$\mu_q\{f \leq q\} = \int \mathbf{1}_{f \leq q} f \, d\lambda = 0$$

Apply Lemma 2.115 to conclude that $F_q'(x) = 0$ a.e. on $\{f \leq q\}$. On the other hand, $f \leq q + (f-q)_+$ and therefore

$$\begin{aligned}\limsup_{h \to 0} \frac{F(x+h) - F(x)}{h} &= \limsup_{h \to 0} h^{-1} \int_x^{x+h} f \, d\lambda \\ &\leq q + \limsup_{h \to 0} h^{-1} \int_x^{x+h} (f-q)_+ \, d\lambda = q \text{ a.e. on } \{f \leq q\}\end{aligned}$$

Via a union bound we conclude

$$\begin{aligned}\lambda\{f < \limsup_{h\to 0} \frac{F(x+h)-F(x)}{h}\} &= \lambda \bigcup_{q\in\mathbb{Q}} \{f \le q < \limsup_{h\to 0} \frac{F(x+h)-F(x)}{h}\} \\ &\le \sum_{q\in\mathbb{Q}} \lambda\{f \le q < \limsup_{h\to 0} \frac{F(x+h)-F(x)}{h}\} = 0\end{aligned}$$

which shows $\limsup_{h\to 0} \frac{F(x+h)-F(x)}{h} \le f$ a.e.

Arguing analogously we see that $\frac{d}{dx}\int_0^x (q-f)_+ \, d\lambda = 0$ a.e. on $\{f \ge q\}$ and from $f \ge q - (q-f)_+$ we get

$$\liminf_{h\to 0} \frac{F(x+h)-F(x)}{h} = q \text{ a.e. on } \{f \ge q\}$$

Again a union bound yields $\liminf_{h\to 0} \frac{F(x+h)-F(x)}{h} \ge f$ a.e. and it follows that

$$\liminf_{h\to 0} \frac{F(x+h)-F(x)}{h} = \limsup_{h\to 0} \frac{F(x+h)-F(x)}{h} = \lim_{h\to 0} \frac{F(x+h)-F(x)}{h} = f \text{ a.e.}$$

□

7.1. Functions of Bounded Variation. Recall that we have define $x_+ = x \vee 0$ and $x_- = |x| - x_+ = -(x \wedge 0)$. Given a real valued function F on $[a, b]$ we consider a partition $a = x_0 < x_1 < \cdots < x_n = b$ and define

$$\begin{aligned}p &= \sum_{j=1}^n (F(x_j) - F(x_{j-1}))_+ \\ n &= \sum_{j=1}^n (F(x_j) - F(x_{j-1}))_- \\ v &= \sum_{j=1}^n |F(x_j) - F(x_{j-1})|\end{aligned}$$

and note that $p + n = v$ and $p - n = F(b) - F(a)$. We define the *positive*, *negative* and *total variation* of F on $[a, b]$ to be the supremum of the above over all partitions of $[a, b]$:

$$\begin{aligned}P_a^b(F) &= \sup_{\substack{n\ge 1 \\ a=x_0<x_1<\cdots x_n=b}} \sum_{j=1}^n (F(x_j) - F(x_{j-1}))_+ \\ N_a^b(F) &= \sup_{\substack{n\ge 1 \\ a=x_0<x_1<\cdots x_n=b}} \sum_{j=1}^n (F(x_j) - F(x_{j-1}))_- \\ TV_a^b(F) &= \sup_{\substack{n\ge 1 \\ a=x_0<x_1<\cdots x_n=b}} \sum_{j=1}^n |F(x_j) - F(x_{j-1})|\end{aligned}$$

LEMMA 2.116. *For any function F defined on $[a, b]$ we have*

$$P_a^b(F) \vee N_a^b(F) \le TV_a^b(F)$$

and

$$TV_a^b(F) \leq P_a^b(F) + N_a^b(F)$$

PROOF. For any partition $a = x_0 < x_1 < \cdots < x_n = b$ we noted above $p + n \leq v$ and therefore $p \leq v$ which implies by taking the supremum on the right $p \leq TV_a^b(F)$ and then by taking the supremum on the left $P_a^b(F) \leq TV_a^b(F)$. The argument to show $N_a^b(F) \leq TV_a^b(F)$ is identical. Similarly from $v = p + n$, we can take two different suprema on the right to see that $v \leq P_a^b(F) + N_a^b(F)$ and then taking the supremum on the left we get $TV_a^b(F) \leq P_a^b(F) \leq TV_a^b(F)$. □

DEFINITION 2.117. We say that function F defined on $[a, b]$ has *bounded variation* on $[a, b]$ if $TV_a^b(F) < \infty$.

LEMMA 2.118. *If F has bounded variation on $[a, b]$ then*

$$TV_a^b(F) = P_a^b(F) + N_a^b(F)$$

and

$$F(b) - F(a) = P_a^b(F) - N_a^b(F)$$

PROOF. From 2.116, we know that F being of bounded variation implies that both the positive and negative variation are finite. Now with a fixed $a = x_0 < x_1 < \cdots < x_n = b$ we had $p = n + F(b) - F(a)$, so taking supremum on the right we get $p \leq N_a^b(F) + F(b) - F(a)$ and the taking suspremum on the left we get $P_a^b(F) \leq N_a^b(F) + F(b) - F(a)$. As noted the negative variation is finite and therefore we conclude $P_a^b(F) - N_a^b(F) \leq F(b) - F(a)$. Similarly we get from applying the same steps to $n = p + F(a) - F(b)$ that $N_a^b(F) \leq P_a^b(F) + F(a) - F(b)$ which gives us $F(b) - F(a) \leq P_a^b(F) - N_a^b(F)$ and therefore we conclude that $F(b) - F(a) = P_a^b(F) - N_a^b(F)$.

Now arguing from $p + n = v$ and taking the supremum on the right we have using $F(b) - F(a) = p - n$,

$$TV_a^b(F) \geq p + n = 2p + F(a) - F(b) = 2p + N_a^b(F) - P_a^b(F)$$

which upon taking another supremum gives

$$TV_a^b(F) \geq 2P_a^b(F) + N_a^b(F) - P_a^b(F) = P_a^b(F) + N_a^b(F)$$

Note a more hands on way of proving the this result is to note that we have a triangle inequality $(x + y)_+ leq x_+ + y_+$ and therefore if we are given a partition $a = x_0 < x_1 < \cdots < x_n = b$ and refine the partition by adding a new point then to create a new partition $a = \tilde{x}_0 < \tilde{x}_1 < \cdots < \tilde{x}_n = b$ then we have

$$\sum_{j=1}^{n} (F(x_j) - F(x_{j-1}))_+ \leq \sum_{j=1}^{n+1} (F(\tilde{x}_j) - F(\tilde{x}_{j-1}))_+$$

and similarly with the negative variation. Now let $\epsilon > 0$ be chosen and find partitions $a = x_0 < x_1 < \cdots < x_n = b$ such that

$$P_a^b(F) - \epsilon/2 < \sum_{j=1}^{n} (F(x_j) - F(x_{j-1}))_+ \leq P_a^b(F)$$

and $a = y_0 < y_1 < \cdots < y_m = b$ such that

$$N_a^b(F) - \epsilon/2 < \sum_{j=1}^{m} (F(y_j) - F(y_{j-1}))_+ \leq N_a^b(F)$$

By the above argument, both inequalities continue to hold if we take the common refinement of both partitions so we may in fact assume that $n = m$ and $x_j = y_j$ for $j = 0, \ldots, n$. Therefore by adding we get

$$\begin{aligned} P_a^b(F) + N_a^b(F) - \epsilon &< \sum_{j=1}^{n} (F(x_j) - F(x_{j-1}))_+ + (F(x_j) - F(x_{j-1}))_- \\ &= \sum_{j=1}^{n} |F(x_j) - F(x_{j-1})| \leq TV_a^b(F) \end{aligned}$$

and the result follows by taking the limit as ϵ goes to 0. □

THEOREM 2.119. *A function on $[a, b]$ is of bounded variation if an only if it is the difference of two non-decreasing functions.*

PROOF. First we show that a function of bounded variation is a difference of monotone functions. Consider a point $a \leq x \leq b$ and note that since every partition of $[a, x]$ can be extended to a partiton of $[a, b]$ we have $P_a^x(F) \leq P_a^b(F) \leq TV_a^b(F) < \infty$ and similarly with $N_a^x(F)$. The same argument for any pair $a \leq x \leq y \leq b$ shows that $P_a^x(F) \leq P_a^y(F)$ and similarly $N_a^x(F) \leq N_a^y(F)$. Therefore $P_a^x(F)$ and $N_a^x(F)$ are both non-decreasing functions and applying Lemma 2.118 on the interval $[a, x]$ we get $F(x) = P_a^x(F) - N_a^x(F) - F(a)$. Since $N_a^x(F) - F(a)$ is also a non-decreasing function we are done with this direction.

Now if $F(x) = G(x) - H(x)$ with both G and H monotone then for any partition $a = x_0 < x_1 < \cdots < x_n = b$ we have

$$\begin{aligned} \sum_{j=1}^{n} |F(x_j) - F(x_{j-1})| &= \sum_{j=1}^{n} |G(x_j) - G(x_{j-1}) - H(x_j) + H(x_{j-1})| \\ &\leq \sum_{j=1}^{n} (G(x_j) - G(x_{j-1})) + \sum_{j=1}^{n} (H(x_{j-1}) + H(x_j)) = G(b) - G(a) + H(a) - H(b) \end{aligned}$$

□

LEMMA 2.120. *Let f be a function of bounded variation on $[a, b]$, then for every $a < x < b$, $TV_a^x(f) + TV_x^b(f) = TV_a^b(f)$.*

PROOF. Pick partitions $a = x_0 < \cdots < x_n = x$ of $[a, x]$ and $x = y_0 < \cdots < y_m = b$ of $[x, b]$ and note that $a = x_0 < \cdots < x_n = y_0 < y_1 < \cdots < y_m = b$ is a partition of $[a, b]$. Therefore

$$\sum_{j=1}^{n} |f(x_j) - f(x_{j-1})| + \sum_{j=1}^{m} |f(y_j) - f(y_{j-1})| \leq TV_a^b(f)$$

which upon taking suprema over partitions of $[a, x]$ and $[x, b]$ shows $TV_a^x(f) + TV_x^b(f) \leq TV_a^b(f)$.

On the other hand, let $a = x_0 < \cdots < x_n = b$ be a partition of $[a, b]$. First assume that there exists an $0 < m < n$ such that $x_m = x$. It then follows that

$a = x_0 < \cdots < x_m = x$ is a partition of $[a, x]$ and $x = x_m < \cdots < x_n = b$ is a partition of $[x, b]$ and therefore

$$\sum_{j=1}^n |f(x_j) - f(x_{j-1})| = \sum_{j=1}^m |f(x_j) - f(x_{j-1})| + \sum_{j=m+1}^n |f(x_j) - f(x_{j-1})| \leq TV_a^x(f) + TV_x^b(f)$$

On the other hand, if x is not a member of the partition then we may add it and by the triangle inequality that can only increase the variation of the partition so the inequality still holds. Thus we may take the supremum over all partitions of $[a, b]$ and we get $TV_a^b(f) \leq TV_a^x(f) + TV_x^b(f)$ and the result is proven. ☐

LEMMA 2.121. *Let f be a left continuous function with bounded variation on $[a, b]$, then $TV_a^x(f)$ is a left continuous function of x. Similarly if f is right continuous (resp. continuous) then $TV_a^x(f)$ is a right continuous (resp. continuous).*

PROOF. We first suppose that f is left continuous and show that $TV_a^x(f)$ is left continuous at x. Pick $\epsilon > 0$ and select a partion $a = x_0 < x_1 < \cdots < x_n = x$ such that $\sum_{j=1}^n |f(x_j) - f(x_{j-1})| > TV_a^x(f) - \epsilon/2$. By left continuity of f at x we can pick a $\delta > 0$ such that $|f(x) - f(y)| < \epsilon/2$ for all $x - \delta < y < x$. Without loss of generality we may also assume that $\delta < x - x_{n-1}$. For any such y we define a new partition by adding the point y to the existing partition $x_0, \dots, x_n$; precisely define

$$\tilde{x}_j = \begin{cases} x_j & \text{for } j = 0, \dots, n-1 \\ y & \text{for } j = n \\ x & \text{for } j = n+1 \end{cases}$$

and note that by the triangle inequality,

$$TV_a^x(f) - \epsilon/2 < \sum_{j=1}^n |f(x_j) - f(x_{j-1})| \leq \sum_{j=1}^{n+1} |f(\tilde{x}_j) - f(\tilde{x}_{j-1})| \leq TV_a^x(f)$$

If restrict our attention to the partition $a = \tilde{x}_0 < \cdots < \tilde{x}_n = y$, by monotonicity of total variation and the choice of y we have

$$\begin{aligned} TV_a^x(f) \geq TV_a^y(f) &\geq \sum_{j=1}^n |f(\tilde{x}_j) - f(\tilde{x}_{j-1})| \\ &> TV_a^x(f) - \epsilon/2 - |f(x) - f(y)| > TV_a^x(f) - \epsilon \end{aligned}$$

which shows left continuity of $TV_a^x(f)$.

One could prove the case of right continuous f by an analogous argument that shows $TV_x^b(f)$ is a right continuous function of x and then observing $TV_a^x(f) = TV_a^b(f) - TV_x^b(f)$ Lemma 2.120 (do this as an exercise!). Here we take a slightly different approach and derive the case of right continuity from the case of left continuity. Given f a function on $[a, b]$, define the function $\tilde{f}(x) = f(b + a - x)$ on $[a, b]$. Note that f is right continuous if and only if $\tilde{f}$ is left continuous. Note also that the transformation $x \mapsto b + a - x$ is a bijection of $[a, y]$ and $[b + a - y, b]$ for every $a \leq y \leq b$ and therefore is a bijection of partitions of $[a, y]$ and $[b + a - y, b]$ for every such y. From this it follows that $TV_a^y(f) = TV_{b+a-y}^b(\tilde{f})$ for every $a \leq y \leq b$. In particular taking $y = b$, f is of bounded variation on $[a, b]$ if and only if $\tilde{f}$ is. Stitching all of these observations together, if f is right continuous, then $\tilde{f}$ is left continuous and therefore by the first part of the Lemma and Lemma 2.120 we know that $TV_y^b(\tilde{f}) = TV_a^b(\tilde{f}) - TV_a^y(\tilde{f})$ is a left continuous function of y. From this it follows that $TV_a^y(f) = TV_{b+a-y}^b(\tilde{f})$ is a right continuous function of y.

The case of f continuous follows immediately as a function is continuous if and only if it is both right continuous and left continuous. □

As an exercise, one should show that continuity of a function not only implies the continuity of the total variation but also the positive and negative variations (all we needed positivity and the triangle inequality of the absolute value; properties that are shared by the positive and negative part functions). (TODO: Can we instead derive the positive and negative variation cases from right continuity of total variation and f?) If we assume that we are given a right continuous function f of bounded variation, then by Lemma 2.121 we know that positive and negative variations are right continuous and therefore by Theorem 2.119 we see that f is a difference of monotone right continuous functions. By the construction of Lebesgue-Stieltjes measures this allows us to associate locally finite (signed) measures to f.

TODO: Define all of the measures involved and observe that $dF = dF_+ - dF_-$ is the Jordon decompostiion of the signed measure dF and that $dTV_a^s(F)$ is the absolute value of the signed measure dF.

Lemma 2.122. *Let F be a function of bounded variation of $[a,b]$ and let g be a measurable function then $\left|\int g\,dF\right| \leq \int |g|\,|dF|$.*

Proof. This is just a computation using the definitions and the triangle inequality

$$\begin{aligned}\left|\int g\,dF\right| &= \left|\int g\,dF_+ - \int g\,dF_-\right| \leq \left|\int g\,dF_+\right| + \left|\int g\,dF_-\right| \\ &\leq \int |g|\;dF_+ + \int |g|\;dF_- = \int |g|\;|dF|\end{aligned}$$

□

In addition to functions of bounded variation providing signed measures via the construction of Stieltjes measures integrals also provide a source of functions of bounded variation.

Definition 2.123. A function F is *absolutely continuous* on an interval $[a,b]$ if for every $\epsilon > 0$ there exists $\delta > 0$ such that for every $n > 0$ and every set of disjoint intervals $(a_j, b_j] \subset (a,b]$ for $j = 1, \ldots, n$ with $\sum_{j=1}^n (b_j - a_j) < \delta$ we have $\sum_{j=1}^n |F(b_j) - F(a_j)| < \epsilon$.

Lemma 2.124. *If F is absolutely continuous on $[a,b]$ then F is uniformly continuous on $[a,b]$ and has bounded variation on $[a,b]$.*

Proof. The fact that F is uniformly continuous is immediate by considering a single subinterval of $[a,b]$. Seeing that F has bounded variation is conceptually simple but notationally a little ugly. The idea is simply that any sufficiently fine partition of $[a,b]$ can be decomposed into a union of subpartitions of a subinterval of length less than any desired δ; this is enough to bound the total variation. To see the details, pick $N > 0$ so that $\sum_{j=1}^n (b_j - a_j) < (b-a)/N$ implies $\sum_{j=1}^n |F(b_j) - F(a_j)| < 1$. First assume that we have a partition $a = x_0 < x_1 <$

$\cdots < x_n = b$ such that for each $k = 0, \ldots, N$ there is an n_k with $x_{n_k} = (b-a)*k/N$. Then we have $\sum_{j=n_{k-1}+1}^{n_k}(x_j - x_{j-1}) < \delta$ for each k and therefore

$$\sum_{j=1}^{n} |F(b_j) - F(a_j)| = \sum_{k=1}^{(b-a)/N} \sum_{j=n_{k-1}+1}^{n_k} |F(b_j) - F(a_j)| < (b-a)/N < \infty$$

The assumption that $x_{n_k} = (b-a)*k/N$ can be arranged for by refining an arbitrary partition and noting that the total variation can only increase by doing so. □

To construct a general construction of absolutely continuous functions from Stieltjes measures we first prove the following fact about integrals on general measurable spaces.

Lemma 2.125. *Let $(S, \mathcal{S}, \mu)$ be a measure space and integrable function $f : S \to \mathbb{R}$, then for every $\epsilon > 0$ there exists a $\delta > 0$ such that for all $A \in \mathcal{S}$ such that $\mu(A) < \delta$ we have $\left|\int_A f \, d\mu\right| < \epsilon$.*

Proof. First assume that f is a positive integrable function. For each $n > 0$ define $f_n = f \wedge n$ and note that $f_n \uparrow f$; moreover $f_n \mathbf{1}_A \uparrow f \mathbf{1}_A$ for every $A \in \mathcal{S}$. By Monotone Convergence we know that $\int_A f_n \, d\mu \uparrow \int_A f \, d\mu$. Let $\epsilon > 0$ be given and choose $N > 0$ such that $\int f \, d\mu - \epsilon/2 < \int f_N \, d\mu \leq \int f \, d\mu$. Choose $\delta = \epsilon/2 * N$ and note that if $\mu(A) < \delta$ then

$$\int_A f \, d\mu = \int_A f_N \, d\mu + \int_A (f - f_N) \, d\mu \leq N\mu(A) + \int (f - f_N) \, d\mu < \epsilon$$

For general integrable f simply note that $\left|\int_A f \, d\mu\right| \leq \int_A |f| \, d\mu$ and apply the result just proved for positive integrable functions. □

Specializing to the case of locally finite signed measures on $\mathbb{R}$ we get

Corollary 2.126. *Let F be a right continuous function of bounded variation and let g be a measurable function that is integrable with respect to F then $\int_{-\infty}^{t} g \, dF$ has bounded variation. If F is also continuous then $\int_{-\infty}^{t} g \, dF$ is continuous.*

Proof. First assume that F is non-decreasing and right continuous. If g is integrable with respect to F then $\int_{-\infty}^{t} g_{\pm} \, dF$ is non-decreasing by monotonicity of integral and therefore $\int_{-\infty}^{t} g \, dF = \int_{-\infty}^{t} g_{+} \, dF - \int_{-\infty}^{t} g_{-} \, dF$ is a difference of non-decreasing functions and therefore is of bounded variation by Theorem 2.119. To extend to F of bounded variation, write

$$\int_{-\infty}^{t} g \, dF = \int_{-\infty}^{t} g_{+} \, dF_{+} + \int_{-\infty}^{t} g_{-} dF_{-} - \int_{-\infty}^{t} g_{-} \, dF_{+} - \int_{-\infty}^{t} g_{+} \, dF_{-}$$

and apply Theorem 2.119.

Now suppose that F is continuous and non-decreasing. For $\epsilon > 0$, pick $\delta > 0$ as in Lemma 2.125 and then as any union of intervals is measurable we get $\sum_{j=1}^{n}(F(b_j) - F(a_j)) < \delta$ implies $\left|\sum_{j=1}^{n} \int_{a_j}^{b_j} g \, dF\right| < \epsilon$. Let t be given and by continuity of F pick $\rho > 0$ such that $|s - t| < \rho$ implies $|F(s) - F(t)| < \delta$ and therefore

$$\left|\int_{-\infty}^{t} g \, dF - \int_{-\infty}^{s} g \, dF\right| = \left|\int_{s}^{t} g \, dF\right| < \epsilon$$

and continuity at t is proven. □

NOTE: It is not the case that every continuous function of bounded variation is absolutely continous. For that to be true we need to add the *Lusin N property* that says every the image of every Lebesgue null set is a null set. It turns out that absolute continuity is equivalent to continuity, bounded variation and the Lusin property.

TODO: Where to put this?

COROLLARY 2.127 (Integration By Parts). *Suppose f and g are absolutely continuous functions. Then*

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

8. 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 2.128. *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 2.129. *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 $n \geq \max(N_1, N_2)$

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

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

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

□

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

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 2.131. *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 2.129. By simple change of variables and Lemma 2.130 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

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

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 $f * \rho_n$. □

9. Daniell-Stone Integrals

We record some required facts about σ-rings that are completely analogous to corresponding facts about σ-algebras.

Lemma 2.132. *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. By virtue of the above Lemma we will refer to the σ-ring of Borel sets on $\mathbb{R}$ that do not contain 0 as $\mathcal{B}(\mathbb{R} \setminus \{0\})$. This is at the risk of potential confusion about whether we are considering this a σ-ring of subsets of $\mathbb{R}$ or a σ-algebra of subsets of $\mathbb{R} \setminus \{0\}$; pretty much we always have the former interpretation in mind. Our first order of business is to establish a simple generating set for Borel σ-ring on $\mathbb{R}$.

Lemma 2.133. *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 1.16). 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 2.6 to get the result. $\square$

One of the most often used facts from measure theory is the fact that measurable functions may be approximated by simple functions (Lemma 2.18). We need a small refinement of that Lemma that applies with σ-rings.

LEMMA 2.134. *For any function $f:\Omega\to\overline{\mathbb{R}}_+$ measurable with respect to a σ-ring $\mathcal{R}$, there exist a sequence of simple functions $f_1,f_2,\dots$ measurable with respect to $\mathcal{R}$ such that $0\le f_n\uparrow f$.*

PROOF. Recalling the proof of 2.18, define

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

and we know that f_n are simple functions $f_n\uparrow f$. The only thing to prove is that the f_n are $\mathcal{R}$-measurable; this follows because each preimage of f_n is either of the form $f^{-1}([k2^{-n},(k+1)2^{-n}))$, for $k=0,\dots,n2^n-1$ or $f^{-1}([n,\infty))$ and $f_n=0$ precisely on $f^{-1}([0,1/2^n))$. Therefore every preimage of a set in $\mathcal{B}(\mathbb{R}\setminus\{0\})$ is a union of sets $f^{-1}([k2^{-n},(k+1)2^{-n}))$, for $k=1,\dots,n2^n-1$ or $f^{-1}([n,\infty))$ and is therefore in $\mathcal{R}$ by the $\mathcal{R}$-measurability of f. $\square$

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

LEMMA 2.135. *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 $\{A\subset T \mid f^{-1}(A) \text{ is in the } \sigma\text{-ring generated by } f^{-1}(\mathcal{C})\}$ and is itself a σ-ring (Lemma 2.85). 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. $\square$

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 2.136. *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 $A_1, A_2, \dots \in \mathcal{R}$ 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

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

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 $i = 1, 2, \dots$ 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 $i = 1, 2, \dots$ we know that $A_i \cap C \in \mathcal{R}$ for $i = 1, 2, \dots$. 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. $\square$

EXAMPLE 2.137. 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 2.138. 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 2.139. *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 $f \wedge g = -(-f \vee -g)$. $\square$

DEFINITION 2.140. 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 $f_1, f_2, \dots \in \mathcal{L}$ 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 2.141. *Let $\mathcal{L}$ be a vector lattice of functions on a set S with a pre-integral I. For any $f, g \in \mathcal{L}$ such that $f \leq g$ we define*

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

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

For every $c > 0$, we let $M_c : S \times \mathbb{R} \to S \times \mathbb{R}$ be the mapping $M_c(s, t) = (s, ct)$. Then $M_c^{-1} : 2^{S \times \mathbb{R}} \to 2^{S \times \mathbb{R}}$ restricts to a bijection on the σ-algebra generated by $\mathcal{D}$ and furthermore for every set $A \in \sigma(\mathcal{D})$ and $c > 0$ we have $c\nu(M_c^{-1}A) = \nu(A)$.

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:).

Let $c > 0$ and consider the mapping $M_c(s, t) = (s, ct)$.

Claim 1: $M_c^{-1}(\sigma(\mathcal{D})) = \sigma(\mathcal{D})$.

Since M_c is a bijection it follows that $M_c^{-1} : 2^{S \times \mathbb{R}} \to 2^{S \times \mathbb{R}}$ is also a bijection. Furthermore if we consider a set of the form $[f, g)$ then

$$\begin{aligned} M_c^{-1}([f, g)) &= \{(s, t) \in S \times \mathbb{R} \mid f(s) \leq ct < g(s)\} \\ &= \{(s, t) \in S \times \mathbb{R} \mid (f/c)(s) \leq t < (g/c)(s)\} = [f/c, g/c) \end{aligned}$$

So if $[f, g) \in \mathcal{D}$ then it follows from the fact that $\mathcal{L}$ is a vector space that M_c^{-1} is bijection of $\mathcal{D}$ to itself. In particular, we know that $M_c^{-1}(\sigma(\mathcal{D}))$ is a σ-algebra containing $\mathcal{D}$ and therefore $M_c^{-1}(\sigma(\mathcal{D})) \supset \sigma(\mathcal{D})$. On the other hand, $(M_c)_*(\sigma(\mathcal{D})) = \{A \subset S \times \mathbb{R} \mid M_c^{-1}(A) \in \sigma(\mathcal{D})\}$ is also σ-algebra (Lemma 2.8) containing $\mathcal{D}$; hence $\sigma(\mathcal{D}) \subset (M_c)_*(\sigma(\mathcal{D}))$ which implies $M_c^{-1}(\sigma(\mathcal{D})) \subset \sigma(\mathcal{D})$.

Claim 2: For any $A \in \sigma(\mathcal{D})$ and $c > 0$ we have $c\nu(M_c^{-1}(A)) = \nu(A)$.

We start with considering $[f, g) \in \mathcal{D}$. We have already seen that $M_c^{-1}([f, g)) = [f/c, g/c)$ so we can just apply the definition to see the claim holds.

$$c\nu(M_c^{-1}([f, g))) = c\nu([f/c, g/c)) = cI(f/c - g/c) = I(f - g) = \nu([f, g))$$

To extend to the ring $\mathcal{R}$ generated by $\mathcal{D}$ we note that every element of the ring is a disjoint union of elements in $\mathcal{D}$. Furthermore M_c^{-1} preserves the Boolean algebra structure on $2^{S \times \mathbb{R}}$ (Lemma 2.7) therefore we have

$$\begin{aligned} c\nu(M_c^{-1}(\cup_{i=1}^n [f_i, g_i))) &= c\nu(\cup_{i=1}^n M_c^{-1}([f_i, g_i))) = c\sum_{i=1}^n \nu(M_c^{-1}([f_i, g_i))) \\ &= \sum_{i=1}^n \nu([f_i, g_i)) = \nu(\cup_{i=1}^n [f_i, g_i)) \end{aligned}$$

To extend to all of $\sigma(\mathcal{D})$ we use the fact that ν is defined as its associated outer measure $\nu(A) = \inf\{\nu(B) \mid B \supset A \text{ and } B \in \mathcal{R}\}$. Consider $A \in \sigma(\mathcal{D})$ and let $\epsilon > 0$. By definition we can find $B \in \mathcal{R}$ such that $B \supset A$ and $\nu(B) \leq \nu(A) + \epsilon$. Again applying Lemma 2.7 we see that $M_c^{-1}(B) \in \mathcal{R}$ and $M_c^{-1}(B) \supset M_c^{-1}(A)$ and therefore

$$c\nu(M_c^{-1}(A))c \leq M_c^{-1}(B) = \nu(B) \leq \nu(A) + \epsilon$$

Since $\epsilon > 0$ was arbitrary we conclude that $c\nu(M_c^{-1}(A)) \leq \nu(A)$. In the opposite direction for every $\epsilon > 0$ we can find $M_c^{-1}(B) \supset M_c^{-1}(A)$ such that $\nu(M_c^{-1}(B) \leq \nu(M_c^{-1}(A)) + \epsilon$. We know that $B \supset A$ and therefore

$$\nu(A) \leq \nu(B) = c\nu(M_c^{-1}(B) \leq c\nu(M_c^{-1}(A)) + \epsilon$$

so letting ϵ go to zero we conclude $\nu(A) \leq c\nu(M_c^{-1}(A))$ and we are done.

TODO: In the proof we use the fact that M_c^{-1} is a bijection on $\mathcal{R}$ which is a simple consequence of the fact M_c^{-1} is a bijection on $\mathcal{D}$ and Lemma 2.7; find the correct place to note this fact explicitly. □

THEOREM 2.142. *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 2.136. 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 2.135) 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 2.136). We define the measure by embedding $\mathcal{R}$ as sub-σ-ring in σ-algbra $\mathcal{A}$ constructed in Theorem 2.141. 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: For every $\mathcal{R}/\mathcal{B}(\mathbb{R} \setminus \{0\})$ simple function $f \geq 0$ of the form $f = \sum_{i=1}^n c_i \mathbf{1}_{A_i}$ we have $\int f \, d\mu = \nu([0, f))$.

To see this we know by Theorem 2.141 that for every $c > 0$ and $B \in \mathcal{A}$ we have $c\nu(M_c^{-1}(B) = \nu(B)$. We have shown that for every $A \in \mathcal{R}$, we have $A \times [0, c) \in \mathcal{A}$ and by definition $M_c^{-1}(A \times [0, c)) = A \times [0, 1)$; therefore $\nu(A \times [0, c)) = c\nu(A \times [0, 1)) = c\mu(A)$. It is also easy to see that $A \times [0, c) = [0, c\mathbf{1}_A)$, so we have for scalar multiples of characteristic functions $\int f \, d\mu = \nu([0, f))$. As for simple functions, each can be expressed as a sum $f = \sum_{i=1}^n c_i \mathbf{1}_{A_i}$ with $A_i \in \mathcal{R}$ and the A_i disjoint. Once again by definition we can see that $[0, f) = \cup_{i=1}^n [0, c\mathbf{1}_{A_i})$ where the disjointness of the A_i implies that the sets $[0, c\mathbf{1}_{A_i})$ are disjoint. Now by definition of the integral for a simple function and the additivity of the measure ν we get

$$\int f \, d\mu = \sum_{i=1}^n c_i \mu(A_i) = \sum_{i=1}^n \nu([0, c_i \mathbf{1}_{A_i})) = \nu([0, f))$$

Claim 4: For every $\mathcal{R}/\mathcal{B}(\mathbb{R} \setminus \{0\})$-measurable function $f \geq 0$ we have $\int f \, d\mu = \nu([0, f))$.

We take a sequence of positive simple functions $f_n \uparrow f$ which exists by Lemma 2.134. Since $[0, f_n) \uparrow [0, f)$ we can use the definition of integral with respect to μ, continuity of measure with respect to ν and Claim 3 to see

$$\int f \, d\mu = \lim_{n \to \infty} \int f_n \, d\mu = \lim_{n \to \infty} \nu([0, f_n)) = \nu([0, f))$$

By definition we have arranged for all $f \in \mathcal{L}$ to be $\mathcal{R}/\mathcal{B}(\mathbb{R} \setminus \{0\})$-measurable so by Claim 4 and the definition of ν, for $f \in \mathcal{L}$ with $f \geq 0$ we have $\int f \, d\mu = \nu([0, f)) = I(f)$. For arbitrary $f \in \mathcal{L}$ we write $f = f_+ - f_-$ with $f_+, f_- \in \mathcal{L}$ and $f_+, f_- \geq 0$ and use linearity of integral and pre-integral to conclude that $\int f \, d\mu = \int f_+ \, d\mu - \int f_- \, d\mu = I(f_+) - I(f_-) = I(f)$. □

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 2.142.

CHAPTER 3

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 3.1. 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 2.53 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 $\xi \overset{d}{=} \eta$ 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 3.2. 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 3.3. Given a random vector $\xi = (\xi_1, \dots, \xi_n)$ in $\mathbb{R}^n$ we define the *distribution function* to be

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

Lemma 3.4. *Let ξ and η be random vectors in $\mathbb{R}^n$ with distribution functions F and G, then $\xi \overset{d}{=} \eta$ if and only if $F = G$.*

Proof. This follows from Lemma 2.71 by noting that sets of the form $(-\infty, x_1] \times \cdots \times (-\infty, x_n]$ form a π-system that contains $\mathbb{R}^n$. $\square$

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

LEMMA 3.5. *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 2.112 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 3.6. 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 2.55 is the following

LEMMA 3.7 (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 2.55 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 3.8. *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 3.9 (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 $t \in \mathbb{R}$, $0 \leq \mathbf{E}\left[(t\xi + \eta)^2\right] = \mathbf{E}\left[\xi^2\right] t^2 + 2\mathbf{E}\left[\xi\eta\right] t + \mathbf{E}\left[\eta^2\right]$. The quadratic formula implies that $\sqrt{4\mathbf{E}\left[\xi^2\right]\mathbf{E}\left[\eta^2\right] - (2\mathbf{E}\left[\xi\eta\right])^2} \geq 0$ 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

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

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 3.10. 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 3.11 (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 $\left\|\hat{\xi}\right\|_p = \|\hat{\eta}\|_q = 1$. We now apply Young's Inequality to $\hat{\xi}$ and $\hat{\eta}$ to see

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

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 3.12. *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 $[1,\infty)$).

1. Convexity and Jensen's Inequality

DEFINITION 3.13. 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 3.14. *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 3.15. *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 3.14 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 3.14. 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 3.14

$$D^+\varphi(x) \leq \frac{\varphi(w) - \varphi(x)}{w - x} < \frac{\varphi(y) - \varphi(x)}{y - x}$$

The case of $D^-\varphi(y)$ follows analogously. □

COROLLARY 3.16. *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 3.15 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 3.17 (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. □

CHAPTER 4

Independence

Definition 4.1. 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 $t_1, \dots, t_n \in T$ 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 4.2. 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 4.3. 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 4.4. 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 4.5. *Suppose we are given a finite collection of random elements $\xi_1, \dots, \xi_n$ in measurable spaces $S_1, \dots, S_n$ with distributions $\mu_1, \dots, \mu_n$. The ξ_i are independent if and only if the distribution of $(\mu_1, \dots, \mu_n)$ 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

$$\begin{aligned}\mathbf{P}\{\xi_1^{-1}(B_1) \cap \cdots \cap \xi_n^{-1}(B_n)\} &= \mathbf{P}\{(\xi_1, \dots, \xi_n)^{-1}(B_1 \times \cdots \times B_n)\} \\ &= \mathbf{P}\{\xi_1^{-1}(B_1)\} \cdots \mathbf{P}\{\xi_n^{-1}(B_n)\}\end{aligned}$$

On the other hand, if we assume that the ξ_i are independent the above calculation shows that $(P \circ (\xi_1, \dots, \xi_n)^{-1}) = \mu_1 \otimes \cdots \otimes \mu_n$ 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 2.88. □

Having proven that the joint distribution of independent random elements is a product measure we can apply Fubini's Theorem to compute expectations of functions of independent random elements as iterated integrals. We make that statement precise in the following result. Note that there is an important generalization of this fact that eliminates the assumption of independence but requires the development of the notion of a conditional distribution (see Theorem 8.35). The result is much simpler than the notation required to state it.

Lemma 4.6. *Let ξ and η be independent random elements in measurable spaces $(S, \mathcal{S})$ and $(T, \mathcal{T})$ respectively. Let $f : S \times T \to \mathbb{R}$ be a measurable function and define $g(s) = \boldsymbol{E}[f(s, \eta)]$ and $h(s) = \boldsymbol{E}[|f(s, \eta)|]$. Suppose that either f is non-negative or $h(\xi)$ is integrable, then $\boldsymbol{E}[f(\xi, \eta)] = \boldsymbol{E}[g(\xi)] = \boldsymbol{E}[\boldsymbol{E}[f(s, \eta)] \mid_{s=\xi}]$.*

Proof. Let μ be the distribution of ξ and ν be the distribution of η; by Lemma 4.5 we know that the joint distribution of (ξ, η) is $\mu \otimes \nu$. Suppose that f is non-negative and use the Expectation Rule (Lemma 3.7) and Tonelli's Theorem 2.88 to calculate

$$\begin{aligned}
\mathbf{E}\left[f(\xi, \eta)\right] &= \int f(x, y)\, d(\mu \otimes \nu)(x, y) \\
&= \int \left[\int f(x, y)\, d\nu(y) \right] d\mu(x) = \int g(x)\, d\mu(x) = \mathbf{E}\left[g(\xi)\right]
\end{aligned}$$

If instead assuming $f \geq 0$, we assume that $h(\xi)$ is integrable. Applying the result just proven for the non-negative case to $|f|$ shows that in fact $\mathbf{E}\left[|f(\xi, \eta)|\right] < \infty$ so we may replay the same argument for f without the absolute value using Fubini's Theorem 2.88 in place of Tonelli's Theorem. □

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 4.7. 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 4.8. 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 4.9. *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 4.10. 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 4.11. *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 $A \in \mathcal{G}$,

$$\begin{aligned}
(\mu * \nu)(A) &= \int \mu(A-y)\,d\nu(y) \\
&= \int\int \mathbf{1}_{A-y}(x)f(x)g(y)\,d\lambda(x)d\lambda(y) \\
&= \int\int \mathbf{1}_{A}(x+y)f(x)g(y)\,d\lambda(x)d\lambda(y) \\
&= \int\int \mathbf{1}_{A}(x)f(x-y)g(y)\,d\lambda(x)d\lambda(y) \\
&= \int \mathbf{1}_{A}(x)\left[\int f(x-y)g(y)\,d\lambda(y)\right]d\lambda(x) \\
&= ((f*g)\cdot\lambda)(A)
\end{aligned}$$

$\square$

EXAMPLE 4.12. Let ξ and η be independent $N(0,1)$ random variables. Then $\xi+\eta$ is an $N(0,2)$ random variable. From Corollary 4.11, 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 4.13. *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 2.27) 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 $\sigma(\mathcal{T})$. $\square$

LEMMA 4.14. *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 4.13 shows the result. $\square$

In order to prove independence of a countable collection of σ-algebras it can be useful to reduce the task to showing a sequence of pairwise independent relationships as in the following Lemma.

LEMMA 4.15. *Let $\mathcal{A}_1, \mathcal{A}_2, \ldots$ be σ-algebras, then they are independent if and only if $\bigvee_{k=1}^{n} \mathcal{A}_k$ is independent of $\mathcal{A}_{n+1}$ for all $n \geq 1$.*

PROOF. The only if direction is an application of Lemma 4.14. The if direction will be shown by induction. To set notation, suppose that $A_{k_1} \in \mathcal{A}_{k_1}, \ldots, A_{k_m} \in \mathcal{A}_{k_m}$ are chosen and we must show $\mathbf{P}\{A_{k_1} \cap \cdots \cap A_{k_m}\} = \mathbf{P}\{A_{k_1}\} \cdots \mathbf{P}\{A_{k_m}\}$ where without log of generality we assume $1 \geq k_1 < \cdots < k_m$. If we let $n = k_1 \vee \cdots \vee k_m = k_m$ the induction variable is n. The case of $n = 1$ is trivial as there is nothing to prove, so suppose the result is true for $n-1$ and we are given $A_{k_1} \in \mathcal{A}_{k_1}, \ldots, A_{k_m} \in \mathcal{A}_{k_m}$ with $k_m = n$. Using the hypothesis, the fact that $k_{m-1} < n$ and induction hypothesis we know that

$$\begin{aligned}
\mathbf{P}\{A_{k_1} \cap \cdots \cap A_{k_m}\} &= \mathbf{P}\{A_{k_1} \cap \cdots \cap A_{k_{m-1}}\}\mathbf{P}\{A_{k_m}\} \\
&= \mathbf{P}\{A_{k_1}\} \cdots \mathbf{P}\{A_{k_m}\}
\end{aligned}$$

and the result is proven. $\square$

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 4.16. *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 4.17. *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 4.13. □

LEMMA 4.18. *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 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

$$\begin{aligned}
\mathbf{E}\left[f_1(\xi_1)\cdots f_n(\xi_n)\right] &= \sum_{k_1=1}^{m_1}\cdots\sum_{k_n=1}^{m_n} c_{k_1,1}\cdots c_{k_n,n}\mathbf{E}\left[\mathbf{1}_{A_{k_1,1}}(\xi_1)\cdots\mathbf{1}_{A_{k_n,n}}(\xi_n)\right] \\
&= \sum_{k_1=1}^{m_1}\cdots\sum_{k_n=1}^{m_n} c_{k_1,1}\cdots c_{k_n,n}\mathbf{E}\left[\mathbf{1}_{A_{k_1,1}}(\xi_1)\right]\cdots\mathbf{E}\left[\mathbf{1}_{A_{k_n,n}}(\xi_n)\right] \\
&= \sum_{k_1=1}^{m_1} c_{k_1,1}\mathbf{E}\left[\mathbf{1}_{A_{k_1,1}}(\xi_1)\right]\cdots\sum_{k_n=1}^{m_n} c_{k_n,n}\mathbf{E}\left[\mathbf{1}_{A_{k_n,n}}(\xi_n)\right] \\
&= \mathbf{E}\left[f_1(\xi_1)\right]\cdots\mathbf{E}\left[f_n(\xi_n)\right]
\end{aligned}$$

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 $f_{i,1}f_2\cdots f_n \uparrow f_1 f_2\cdots f_n$ we have

$$\begin{aligned}
\mathbf{E}\left[f_1(\xi_1)\cdots f_n(\xi_n)\right] &= \lim_{i\to\infty}\mathbf{E}\left[f_{i,1}(\xi_1)\cdots f_n(\xi_n)\right] && \text{by Monotone Convergence} \\
&= \lim_{i\to\infty}\mathbf{E}\left[f_{i,1}(\xi_1)\right]\cdots\mathbf{E}\left[f_n(\xi_n)\right] && \text{result for simple functions} \\
&= \mathbf{E}\left[f_1(\xi_1)\right]\cdots\mathbf{E}\left[f_n(\xi_n)\right] && \text{by Monotone Convergence}
\end{aligned}$$

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

$$\begin{aligned}
\mathbf{E}\left[f_1(\xi_1)f_2(\xi_2)\cdots f_n(\xi_n)\right] &= \mathbf{E}\left[f_1^+(\xi_1)f_2(\xi_2)\cdots f_n(\xi_n)\right] - \mathbf{E}\left[f_1^-(\xi_1)f_2(\xi_2)\cdots f_n(\xi_n)\right] \\
&= \mathbf{E}\left[f_1^+(\xi_1)\right]\mathbf{E}\left[f_2(\xi_2)\right]\cdots\mathbf{E}\left[f_n(\xi_n)\right] \\
&- \mathbf{E}\left[f_1^-(\xi_1)\right]\mathbf{E}\left[f_2(\xi_2)\right]\cdots\mathbf{E}\left[f_n(\xi_n)\right] \\
&= \mathbf{E}\left[f_1(\xi_1)\right]\mathbf{E}\left[f_2(\xi_2)\right]\cdots\mathbf{E}\left[f_n(\xi_n)\right]
\end{aligned}$$

Now perform induction on i to get the final result. □

Example 4.19. 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 4.20. *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 4.18, 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

$$\begin{aligned}\mathbf{E}[fg] &= \mathbf{E}[f_+g_+] - \mathbf{E}[f_+g_-] - \mathbf{E}[f_-g_-] + \mathbf{E}[f_-g_-] \\ &= \mathbf{E}[f_+]\,\mathbf{E}[g_+] - \mathbf{E}[f_+]\,\mathbf{E}[g_-] - \mathbf{E}[f_-]\,\mathbf{E}[g_-] + \mathbf{E}[f_-]\,\mathbf{E}[g_-] \\ &= (\mathbf{E}[f_+] - \mathbf{E}[f_-])(\mathbf{E}[g_+] - \mathbf{E}[g_-]) \\ &= \mathbf{E}[f]\,\mathbf{E}[g]\end{aligned}$$

□

EXAMPLE 4.21. 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 4.22. 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 4.23. *[Borel Cantelli Theorem]Let $(\Omega, \mathcal{A}, P)$ be a probability space and let $A_1, A_2, \dots \in \mathcal{A}$.*

(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 $1 + x \le e^x$ for $x \in \mathbb{R}$ from Lemma C.1 and assume that $\sum_{n=1}^{\infty} \mathbf{P}\{A_n\} = \infty$. By the calculation above we have

$$\begin{aligned}
\mathbf{P}\{\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k\} &= 1 - \lim_{n\to\infty} \lim_{m\to\infty} \left(\prod_{k=n}^{m} (1 - \mathbf{P}\{A_k\}) \right) \\
&\ge 1 - \lim_{n\to\infty} \lim_{m\to\infty} \left(\prod_{k=n}^{m} e^{-\mathbf{P}\{A_k\}} \right) \\
&= 1 - \lim_{n\to\infty} \lim_{m\to\infty} e^{-\sum_{k=n}^{m} \mathbf{P}\{A_k\}} \\
&= 1
\end{aligned}$$

But of course we know that $\mathbf{P}\{\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k\} \le 1$ so in fact we have shown that $\mathbf{P}\{A_n \text{ i.o.}\} = 1$. $\square$

EXAMPLE 4.24. 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 4.25. 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 4.26. 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 4.27 (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 $\mathbf{P}\{T\} = 0$ or $\mathbf{P}\{T\} = 1$.*

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 4.14 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 4.13 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 4.28. *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 4.29. *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 2.14 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 4.30. *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 4.31. *Let $\xi, \xi_1, \xi_2, \dots$ 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 3.12 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 4.32. *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 $(k_1, \dots, k_n) \in \{0,1\}^n$, 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 $(k_1, \dots, k_{n-1})$ we see

$$\mathbf{P}\{\xi_n(\vartheta(x)) = k_n\} = \sum_{(k_1, \dots, k_{n-1}) \in \{0,1\}^{n-1}} \mathbf{P}\{\cap_{j \leq n} \{\xi_j(\vartheta(x)) = k_j\}\} = \frac{1}{2}$$

which shows that each $\xi_n \circ \vartheta$ is a Bernoulli random variable with probability $\frac{1}{2}$.

In a similar vein, given $n_1, \dots, n_m$ and $k_{n_j} \in \{0,1\}$, let $n = \sup(n_1, \dots, n_m)$ for $j = 1, \dots, m$ and $A_n = \{(l_1, \dots, l_n) \mid l_{n_j} = k_{n_j} \text{ for } j = 1, \dots, m\}$ and we have

$$\begin{aligned} \mathbf{P}\{\cap_{j=1}^m \{\xi_{n_j}(\vartheta(x)) = k_{n_j}\}\} &= \sum_{(k_1, \dots, k_n) \in A_n} \mathbf{P}\{\cap_{j \leq n} \{\xi_j(\vartheta(x)) = k_j\}\} \\ &= 2^{n-m} \frac{1}{2^n} = \frac{1}{2^m} \end{aligned}$$

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 $\sum \xi_n \circ \vartheta \overset{d}{=} \sum \xi_n \circ \tilde{\vartheta}$. Thus we have shown that $\sum \xi_n \circ \vartheta$ is also $U(0,1)$. $\square$

LEMMA 4.33. *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 4.32. 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 4.32 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 4.14 we see that the $f_n \circ \vartheta$ are independent. $\square$

THEOREM 4.34. *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 4.33 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 2.112, we know that $\mathcal{L}(G_n \circ \vartheta_n) = \mu_n$ and by Lemma 4.16 we know that $G_n \circ \vartheta_n$ are still independent. $\square$

CHAPTER 5

Convergence of Random Variables

Definition 5.1. 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 $\xi_n \overset{a.s.}{\to} \xi$ 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 $\xi_n \overset{P}{\to} \xi$ 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 $\xi_n \overset{d}{\to} \xi$ 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 5.2. 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 $\xi_n \overset{L^p}{\to} \xi$ to denote convergence in L^p. We may also call convergence in L^p *convergence in* p^{th} *mean.*

Limits with respect to these forms of convergence are essentially unique.

Proposition 5.3. *Suppose that ξ_n is a sequence of random elements and suppose ξ and η are random elements such that $\xi_n \overset{a.s.}{\to} \xi$ and $\xi_n \overset{a.s.}{\to} \eta$ or $\xi_n \overset{P}{\to} \xi$ and $\xi_n \overset{P}{\to} \eta$. It follows that $\xi = \eta$ almost surely.*

Proof. For the case of almost sure convergence this follows from by taking the intersection of almost sure events to see that almost surely $\xi_n \to \xi$ and $\xi_n \to \eta$. By uniqueness of limits in S we have $\xi = \eta$ almost surely.

For the case of convergence in probability, let $\epsilon > 0$ be given and note that

$$\mathbf{P}\{d(\xi, \eta) > \epsilon\} \leq \mathbf{P}\{d(\xi, \xi_n) + d(\xi_n, \eta) > \epsilon\} \leq \mathbf{P}\{d(\xi, \xi_n) > \epsilon/2\} + \mathbf{P}\{d(\xi_n, \eta) > \epsilon/2\}$$

Now take the limit as $n \to \infty$ to conclude that $\mathbf{P}\{d(\xi, \eta) > \epsilon\} = 0$. By continuity of measure Lemma 2.30 we have $\mathbf{P}\{d(\xi, \eta) > 0\} = \lim_{n\to\infty} \mathbf{P}\{d(\xi, \eta) > 1/n\} = 0$. $\square$

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.

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 Theorem 4.23 a useful tool for proving almost sure convergence.

Lemma 5.4. *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 5.5. *Let $\xi, \xi_1, \xi_2, \ldots$ be random elements in the metric space (S, d). If $\xi_n \stackrel{a.s.}{\to} \xi$ then $\xi_n \stackrel{P}{\to} \xi$.*

PROOF. By Lemma 5.4 and continuity of measure, if $\xi_n \stackrel{a.s.}{\to} \xi$ 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. □

LEMMA 5.6. *Let $\xi, \xi_1, \xi_2, \ldots$ be random variables, if $\xi_n \stackrel{L^p}{\to} \xi$, then $\xi_n \stackrel{P}{\to} \xi$.*

PROOF. This is a simple application of Markov's Inequality (Lemma 10.1)

$$\mathbf{P}\{|\xi_n - \xi| > \epsilon\} = \mathbf{P}\{|\xi_n - \xi|^p > \epsilon^p\} \leq \frac{\mathbf{E}\left[|\xi_n - \xi|^p\right]}{\epsilon^p}$$

but the right hand side converges to 0 by assumption. □

TODO: Motivation for concept of convergence in probability. Almost sure convergence has a lot of intuitive appeal but it has a technical deficiency: it is not a topological notion. To see that this is true, consider $x, x_1, x_2, \ldots$ in a topological space, we first claim that if for every subsequence $N' \subset \mathbb{N}$ there is a further subsequence $N'' \subset N'$ such that $\lim_{\substack{n \to \infty \\ n \in N''}} x_n = x$ then it follows that $\lim_{n \to \infty} x_n = x$. This is easily seen by contradiction, for if $\lim_{n \to \infty} x_n \neq x$ then there is an open neighborhod $x \in U$ and a subsequence $N' \subset \mathbb{N}$ such that $x_n \notin U$ for all $n \in N'$. It is clear that x_n cannot converge to x along any subsequence of N'. We now give an example that shows that almost sure convergence does not have the above property.

EXAMPLE 5.7. [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 $\mathbf{1}_{I_n} \stackrel{P}{\to} 0$ if and only if $|I_n| \to 0$. For every $n > 0$ consider the events $A_{n,j} = [\frac{j-1}{2^n}, \frac{j}{2^n}]$ for $j = 1, \ldots, 2^n$. 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 $f_m \xrightarrow{P} 0$ 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 $\liminf_{m\to\infty} f_m(x) = 0$. On the other hand we claim that for every subsequence N' there is a further subsequence that does converge almost surely. To see this by the nesting structure of the intervals, note that there are an we can create a subsequence $N^{(1)}$ such that either $f_m \subset [0,1/2]$ for all $m \in N^{(1)}$ or $f_m \subset [1/2,1]$ for all $m \in N^{(2)}$; let $b_1 = 0$ in the former case and $b_1 = 1$ in the latter case. Then by induction we create nested subsequences $N^{(k)} \subset N^{(k-1)}$ such that there exists $b_k \in \{0,1\}$ such that if we define $j_k = 0.b_1 \cdots b_k$ we have $f_m \subset [j_k, j_k + 1/2^k]$ for all $m \in N^{(k)}$. Now take a diagonal subsequence. TODO: show that the diagonal subsequence converges to $\mathbf{1}_{0.b_1 b_2 \cdots}$.

Example 5.8 (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 5.7. 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 $f_m \xrightarrow{P} 0$ follows essentially unchanged; convergence in probability is insensitive the rescaling of the random variables. On the other hand, it is clear that $\mathbf{E}\left[f_m\right] = 1$ 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 5.9. *Let $\xi, \xi_1, \xi_2, \dots$ 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 10.1) 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 5.5: If $\xi_n \overset{a.s.}{\to} \xi$ then $d(\xi_n, \xi) \wedge 1 \overset{a.s.}{\to} 0$ and Dominated Convergence implies $\mathbf{E}[d(\xi_n, \xi) \wedge 1] \to 0$.

The relationship between almost sure convergence and convergence in probability can be made even tighter than Lemma 5.5. This suggests that perhaps convergence in probability is a topological notion; in fact we shal later see that there is a metric called the Ky-Fan metric which topologizes convergence in probability.

Lemma 5.10. *Suppose (S, d) is a metric space and let $\xi, \xi_1, \xi_2, \dots$ be random elements in S. Then $\xi_n \overset{P}{\to} \xi$ 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 $\xi_n \overset{P}{\to} \xi$. By Lemma 5.9, we know that $\lim_{n\to\infty} \mathbf{E}[d(\xi_n, \xi) \wedge 1] = 0$. Thus we can pick $n_k > 0$ such that $\mathbf{E}[d(\xi_{n_k}, \xi) \wedge 1] < \frac{1}{2^k}$. Therefore

$$\sum_{k=1}^{\infty} \mathbf{E}[d(\xi_{n_k}, \xi) \wedge 1] = \mathbf{E}\left[\sum_{k=1}^{\infty} d(\xi_{n_k}, \xi) \wedge 1\right] < \infty$$

where we have used Tonelli's Theorem 2.44. 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 $d(\xi_{n_k}, \xi) \wedge 1 \overset{a.s.}{\to} 0$ which in turn implies $d(\xi_{n_k}, \xi) \overset{a.s.}{\to} 0$

Here is an alternative proof of the first implication using Borel-Cantelli. Pick a sequence $n_1, n_2, \dots$ 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 $\xi_{n_k} \overset{a.s.}{\to} \xi$.

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 $\xi_n \overset{a.s.}{\to} \xi$ along N''. The claim is verified by using the fact (shown in the proof of Lemma 5.5) 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. $\square$

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 5.11. *Let $\xi, \xi_1, \xi_2, \dots$ be a random elements in a metrizable space S. The property $\xi_n \overset{P}{\to} \xi$ 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 5.12. *Let $\xi, \xi_1, \xi_2, \dots$ be a random elements in a metric space (S, d) such that $\xi_n \overset{P}{\to} \xi$. Let (T, d') be a metric space and let $f : S \to T$ be a continuous function, then $f(\xi_n) \overset{P}{\to} f(\xi)$.*

Proof. Pick a subsequence $N' \subset \mathbb{N}$ and note that by Lemma 5.10 we know there exists a subsequence $N'' \subset N'$ such that $\xi_n \overset{a.s.}{\to} \xi$ along N''. By the continuity of f, we know that $f(\xi_n) \overset{a.s.}{\to} f(\xi)$ along N'' hence another application of Lemma 5.10 shows that $f(\xi_n) \overset{P}{\to} f(\xi)$. □

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 5.13. *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 $(\xi_n, \eta_n) \overset{P}{\to} (\xi, \eta)$ if an only if $\xi_n \overset{P}{\to} \xi$ and $\eta_n \overset{P}{\to} \eta$.*

PROOF. Note that by Corollary 5.11 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 $(\xi_n, \eta_n) \overset{P}{\to} (\xi, \eta)$. 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 $\xi_n \overset{P}{\to} \xi$ and $\eta_n \overset{P}{\to} \eta$.

On the other hand if we assume that $\xi_n \overset{P}{\to} \xi$ and $\eta_n \overset{P}{\to} \eta$ 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 5.14. *Let $\xi, \xi_1, \xi_2, \dots$ and $\eta, \eta_1, \eta_2, \dots$ be sequences of random variables such that $\xi_n \overset{P}{\to} \xi$ and $\eta_n \overset{P}{\to} \eta$, then*

(i) $\xi_n + \eta_n \overset{P}{\to} \xi + \eta$
(ii) $\xi_n \eta_n \overset{P}{\to} \xi\eta$
(iii) $\xi_n / \eta_n \overset{P}{\to} \xi/\eta$ *if $\eta \neq 0$ a.e.*

PROOF. By Lemma 5.13 we know that $(\xi_n, \eta_n) \overset{P}{\to} (\xi, \eta)$ in $\mathbb{R}^2$. By continuity of algebraic operations and the Continuous Mapping Theorem the result holds. □

1. The Weak Law Of Large Numbers

THEOREM 5.15 (Weak Law of Large Numbers). *Let $\xi_1, \xi_2, \dots$ be independent and identically distributed random variables with*

$$\mu = \boldsymbol{E}[\xi_i] < \infty$$

Then

$$\frac{1}{n}\sum_{k=1}^{n} \xi_k \overset{P}{\to} \mu$$

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 10.1)

$$\mathbf{P}\{\left|\hat{S}_n\right| > \epsilon\} = \mathbf{P}\{\hat{S}_n^2 > \epsilon^2\} \leq \frac{\mathbf{Var}\left(\hat{S}_n\right)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

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 4.18, 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

$$\begin{aligned} \xi_{i,\leq N} &= f_{\leq N} \circ \xi_i \\ \xi_{i,>N} &= \xi_i - \xi_{i,\leq N} \end{aligned}$$

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

$$\mathbf{Var}\left(\xi_{i,\leq N}\right) = \mathbf{E}\left[\xi_{i,\leq N}^2\right] - \mathbf{E}\left[\xi_{i,\leq N}\right]^2 \leq \mathbf{E}\left[\xi_{i,\leq N}^2\right] \leq N\mathbf{E}\left[|\xi_i|\right] < \infty$$

which shows that $\xi_{i,\leq N}$ has finite variance. Let $\mu_N = \mathbf{E}\left[\xi_{i,\leq N}\right]$.

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

$$\begin{aligned} \lim_{N\to\infty} \mathbf{E}\left[\xi_{i,>N}\right] &= \lim_{N\to\infty} \mathbf{E}\left[|\xi_{i,>N}|\right] = 0 \\ \lim_{N\to\infty} \mathbf{E}\left[\xi_{i,\leq N}\right] &= \mathbf{E}\left[\xi_i\right] - \lim_{N\to\infty} \mathbf{E}\left[\xi_{i,>N}\right] = \mathbf{E}\left[\xi_i\right] \end{aligned}$$

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

$$\begin{aligned} \left|\mathbf{E}\left[\xi_{i,\leq N}\right] - \mathbf{E}\left[\xi_i\right]\right| &< \frac{\epsilon}{3} \\ \mathbf{E}\left[|\xi_{i,>N}|\right] &< \frac{\epsilon\delta}{3} \end{aligned}$$

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

$$\begin{aligned}\mathbf{P}\{\left|\frac{\sum_{i=1}^n \xi_i}{n} - \mu\right| \geq \epsilon\} &= \mathbf{P}\{\left|\frac{\sum_{i=1}^n \xi_{i,\leq N}}{n} - \mu_N + \mu_N - \mu + \frac{\sum_{i=1}^n \xi_{i,>N}}{n}\right| \geq \epsilon\} \\ &\leq \mathbf{P}\{\left|\frac{\sum_{i=1}^n \xi_{i,\leq N}}{n} - \mu_N\right| \geq \frac{\epsilon}{3}\} \\ &+ \mathbf{P}\{|\mu_N - \mu| \geq \frac{\epsilon}{3}\} + \mathbf{P}\{\left|\frac{\sum_{i=1}^n \xi_{i,>N}}{n}\right| \geq \frac{\epsilon}{3}\}\end{aligned}$$

Consider each of the three terms in turn. The first term we apply Chebyshev bounding

$$\mathbf{P}\{\left|\frac{\sum_{i=1}^n \xi_{i,\leq N}}{n} - \mu_N\right| \geq \frac{\epsilon}{3}\} \leq \frac{9\mathbf{Var}\left(\frac{\sum_{i=1}^n \xi_{i,\leq N}}{n}\right)}{\epsilon^2} \leq \frac{9N\mathbf{E}\left[|\xi_1|\right]}{n\epsilon^2} < \delta$$

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

$$\mathbf{P}\{\left|\frac{\sum_{i=1}^n \xi_{i,>N}}{n}\right| \geq \frac{\epsilon}{3}\} \leq \frac{3\mathbf{E}\left[\left|\frac{\sum_{i=1}^n \xi_{i,>N}}{n}\right|\right]}{\epsilon} \leq \frac{3\mathbf{E}\left[|\xi_{i,>N}|\right]}{\epsilon} < \delta$$

□

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 5.16. 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 $\mathbf{E}\left[\xi_n\right] = \mathbf{E}\left[|\xi_n|\right] = 1$ and therefore $\mathbf{E}\left[S_n\right] = 1$. 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.

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 5.17 (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*

$$\boldsymbol{P}\{\sup_n \left| \sum_{k=1}^n \xi_k - \boldsymbol{E}[\xi_k] \right| \geq \epsilon\} < \frac{1}{\epsilon^2} \sum_{k=1}^{\infty} \boldsymbol{Var}(\xi_k)$$

Proof. It is clear we may assume that $\mathbf{E}[\xi_n] = 0$ for all $n > 0$.

To clean up notation a bit we define $S_n = \sum_{k=1}^n \xi_k$. The proof of the result rests upon the following crucial fact.

Claim 5.17.1. For any $N > n > 0$, and any $A_n \in \sigma(\xi_1, \ldots, \xi_n)$ we have

$$\mathbf{E}[S_n; A_n] \leq \mathbf{E}[S_N; A_n]$$

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 4.14 we know $S_N - S_n$ is independent of $\sigma(\xi_1, \ldots, \xi_n)$. Therefore for any $A_n \in \sigma(\xi_1, \ldots, \xi_n)$ we have

$$\mathbf{E}\left[S_N^2 - S_n^2; A_n\right] \geq 2\mathbf{E}[(S_N - S_n)S_n; A_n] = 2\mathbf{E}[S_N - S_n]\,\mathbf{E}[S_n; A_n] = 0$$

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 $A_n = \{|S_k| < \epsilon$ for $1 \leq k < n$ 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 5.18 (Kolmogorov One-Series Criterion). *Let $\xi_1, \xi_2, \dots$ 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 5.17 to the sequence ξ_n for $n = N+1, N+2, \dots$, 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 5.17 to the sequence ξ_n for $n = N+1, N+2, \dots$, 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 5.10 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. $\square$

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 5.19. *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 5.20. *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 $n > \max(N_1, N_2)$,

$$\begin{aligned}\left| \frac{1}{b_n} \sum_{k=1}^{n} \beta_k s_k \right| &\leq \left| \frac{1}{b_n} \sum_{k=1}^{N_1} \beta_k s_k \right| + \left| \frac{1}{b_n} \sum_{k=N_1+1}^{n} \beta_k s_k \right| \\ &\leq \frac{b_{N_1} M}{b_n} + \frac{(b_n - b_{N_1})\epsilon}{2 b_n} \leq \epsilon\end{aligned}$$

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 5.21. *Assume that* $0 \leq b_1 \leq b_2 \leq \cdots$ *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 5.22 (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

$$\eta_n = \xi_{n,\leq n} = \xi_n \cdot \mathbf{1}_{[0,n]}(|\xi_n|)$$

Note

$$\begin{aligned}\sum_{n=1}^\infty \mathbf{P}\{\eta_n \neq \xi_n\} &= \sum_{n=1}^\infty \mathbf{P}\{|\xi_n| > n\}\\ &\leq \sum_{n=1}^\infty \int_{n-1}^n \mathbf{P}\{|\xi_n| \geq \lambda\}d\lambda \quad && \text{since } \mathbf{P}\{|\xi_n| \geq \lambda\} \text{ is decreasing}\\ &= \int_0^\infty \mathbf{P}\{|\xi| \geq \lambda\}d\lambda && \text{by i.i.d.}\\ &= \mathbf{E}[|\xi|] < \infty && \text{by Lemma 3.8}\end{aligned}$$

Now we apply the Borel Cantelli Theorem 4.23 to conclude that $\mathbf{P}\{\eta_n \neq \xi_n \text{ i.o.}\} = 0$. Stated conversely, $\mathbf{P}\{\text{there exists } N > 0 \text{ such that } \xi_n \leq n \text{ for all } n > N\} = 1$.

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 5.20

$$\frac{1}{n}\sum_{k=1}^{n} \mathbf{E}[\eta_n] = \lim_{n\to\infty} \mathbf{E}[\eta_n] = 0$$

Therefore if we can show that $\sum_{n=1}^{\infty} \frac{\mathbf{Var}(\eta_n)}{n^2} < \infty$, then by Corollary 5.21 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

$$\begin{aligned}
\sum_{n=1}^{\infty} \frac{\mathbf{Var}(\eta_n)}{n^2} &\leq \sum_{n=1}^{\infty} \frac{\mathbf{E}\left[\eta_n^2\right]}{n^2} \\
&= \sum_{n=1}^{\infty} \frac{\mathbf{E}\left[\xi_n^2; |\xi_n| \leq n\right]}{n^2} \\
&= \sum_{n=1}^{\infty} \sum_{k=1}^{n} \frac{\mathbf{E}\left[\xi^2; k-1 \leq |\xi| \leq k\right]}{n^2} \\
&= \sum_{k=1}^{\infty} \mathbf{E}\left[\xi^2; k-1 \leq |\xi| \leq k\right] \sum_{n=k}^{\infty} \frac{1}{n^2} \\
&\leq \sum_{k=1}^{\infty} \frac{C}{k} \mathbf{E}\left[\xi^2; k-1 \leq |\xi| \leq k\right] \\
&\leq C \sum_{k=1}^{\infty} \frac{k}{k} \mathbf{E}\left[|\xi|; k-1 \leq |\xi| \leq k\right] = C\mathbf{E}[|\xi|] < \infty
\end{aligned}$$

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 4.30, 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 4.16 the A_n are independent so the Borel Cantelli Theorem 4.23 implies $\sum_{n=1}^{\infty} \mathbf{P}\{A_n\} < \infty$. But now we can apply a tail bound

$$\begin{aligned}
\mathbf{E}\left[|\xi|\right] &= \int_0^\infty \mathbf{P}\{|\xi| \geq \lambda\}\, d\lambda && \text{by Lemma 3.8} \\
&\leq \sum_{n=0}^{\infty} \mathbf{P}\{|\xi| \geq n\} && \text{bounding by an upper Riemann sum} \\
&= 1 + \sum_{n=1}^{\infty} \mathbf{P}\{A_n\} < \infty && \text{by i.i.d.}
\end{aligned}$$

□

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 the Borel Cantelli Theorem 4.23 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 5.23 (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 2.44 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 $\eta_n \overset{a.s.}{\to} 0$. ☐

2.1. Empirical Distribution Functions 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 5.24. Given independent random variables $\xi_1, \xi_2, \dots$, 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 5.25. *Let $\xi_1, \xi_2, \dots$ be i.i.d. random variables with distribution function $F(x)$ and empirical distribution functions $\hat{F}_1(x), \hat{F}_2(x), \dots$. 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 4.16, 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 5.26 (Glivenko-Cantelli Theorem). *Let $\xi_1, \xi_2, \dots$ be i.i.d. random variables with distribution function $F(x)$ and empirical distribution functions $\hat{F}_1(x), \hat{F}_2(x), \dots$. Then,*

$$\lim_{n\to\infty} \sup_x \left| \hat{F}_n(x) - F(x) \right| = 0 \ \textit{a.s.}$$

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 2.112. 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 $k = 1, \dots, m-1$. 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 5.25 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 5.27. 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 $\mathrm{Med}(P) = \inf_x\{F(x) \geq \frac{1}{2}\}$. If ξ is a random variable then we will often write $\mathrm{Med}(\xi)$ for the median of the distribution of ξ.

Lemma 5.28. *Let $\xi_1, \xi_2, \dots$ be i.i.d. random variables and distribution function $F(x)$. Suppose that $F(x) > \frac{1}{2}$ for all $x > \mathrm{Med}\,\xi$. The sample median $\lim_{n\to\infty} \mathrm{Med}(P_n) = \mathrm{Med}(\xi)$ a.s.; one says that the sample median is a* strongly consistent *estimator of* $\mathrm{Med}(\xi)$.

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 $x > \mathrm{Med}(P)$. Given $\epsilon > 0$, pick $\delta > 0$ such that

$$\begin{aligned} F_P(\mathrm{Med}(P) - \epsilon) &< \mathrm{Med}(P) - \delta \\ F_P(\mathrm{Med}(P) + \epsilon) &> \mathrm{Med}(P) + \delta \end{aligned}$$

We claim that if Q satisfies $\sup_x |F_P(x) - F_Q(x)| \leq \delta$ then $|\mathrm{Med}(P) - \mathrm{Med}(Q)| \leq \epsilon$.

To see this first note that

$$F_P(\mathrm{Med}(Q)) \geq F_Q(\mathrm{Med}(Q)) - \delta \geq \frac{1}{2} - \epsilon$$

which implies that $\mathrm{Med}(Q) \geq \mathrm{Med}(P) - \epsilon$ by choice of δ and the increasing nature of $F_P(x)$. Secondly note that for any $x < \mathrm{Med}(Q)$ we have

$$F_P(x) \leq F_Q(x) + \delta < \frac{1}{2} + \epsilon$$

which implies $x < \mathrm{Med}(P) + \epsilon$ and therefore by arbitraryness of x, we have $\mathrm{Med}(Q) \leq \mathrm{Med}(P) + \epsilon$ 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. $\square$

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* which we'll prove later on in a couple of different ways (Theorem 12.32 and Theorem 15.33). 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.

3. Convergence In Distribution

As we have already remarked convergence in distribution is really a property of the laws of a sequence of random variables and therefore the limit of a sequence of random variables that converge in distribution can only be expected to be unique up to equality in distribution.

Lemma 5.29. *$\eta, \xi, \xi_1, \xi_2, \dots$ be a random elements in a metric space (S, d) such that $\xi_n \xrightarrow{d} \xi$ and $\xi_n \xrightarrow{d} \eta$, then $\eta \overset{d}{=} \xi$.*

Proof. Let F be a closed set in S and define $f_n(x) = nd(x, F) \wedge 1$. Then the f_n are bounded and continuous (look forward to Lemma 5.41 for a proof of a stronger result) and $f_n \downarrow \mathbf{1}_F$ thus by Monotone Convergence,

$$\mathbf{P}\{\xi \in F\} = \lim_{n\to\infty} \mathbf{E}\left[f_n(\xi)\right] = \lim_{n\to\infty} \lim_{m\to\infty} \mathbf{E}\left[f_n(\xi_m)\right] = \lim_{n\to\infty} \mathbf{E}\left[f_n(\eta)\right] = \mathbf{P}\{\eta \in F\}$$

Since the closed sets are a π-system that generate the Borel σ-algebra on S we have $\xi \overset{d}{=} \eta$ by montone classes (specifically Lemma 2.71). $\square$

This result will also follow from the fact that weak convergence of probability measures corresponds to convergence in a metric topology on the space of probability measures (proven later in this chapter).

Our next goal is to establish that convergence in distribution is implied by convergence in probability.

Lemma 5.30. *Let $\xi, \xi_1, \xi_2, \dots$ 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 $\mathbf{E}\left[f(\xi_n)\right]$. By Lemma 5.12 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. $\square$

EXAMPLE 5.31 (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 4.32 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 $i \neq j$,

$$\mathbf{E}\left[d(\xi_i, \xi_j) \wedge 1\right] = \mathbf{E}\left[d(\xi_i, \xi_j)\right] = \mathbf{P}\{\xi_i \neq \xi_j\} = \frac{1}{2}$$

and we conclude that ξ_i has no subsequence that is Cauchy in probability and hence ξ_i does not converge in probability.

EXAMPLE 5.32 (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 5.33. *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 5.9, 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 5.34. *Let $\xi, \xi_1, \xi_2, \dots$ 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 $j = 1, \dots, J$ 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 $N = \max(N_1, \dots, N_J)$ 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 5.35. *Let $\xi, \xi_1, \xi_2, \dots$ be sequence of random variables with distribution functions $F(x), F_1(x), F_2(x), \dots$. 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 $\lim_{n\to\infty} \mathbf{E}[f(\xi_n)] \to \mathbf{E}[f(\xi)]$. 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 5.36. 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 5.35 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 5.37. 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 5.38. 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 $d(f(x), f(y)) \leq C d(x, y)$ 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 5.39. 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 5.40. *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 5.41. *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 5.42. *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 5.43 (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 5.41 and Lemma 5.42 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 3.8, Fatou's Lemma (Theorem 2.45) and (iii) to see

$$\begin{aligned}\int f\,d\mu &= \int_0^\infty \mathbf{P}_\mu\{f > \lambda\}\,d\lambda \\ &\le \int_0^\infty \liminf_{n\to\infty} \mathbf{P}_{\mu_n}\{f > \lambda\}\,d\lambda \\ &\le \liminf_{n\to\infty} \int_0^\infty \mathbf{P}_{\mu_n}\{f > \lambda\}\,d\lambda \\ &= \liminf_{n\to\infty} \int f\,d\mu_n\end{aligned}$$

Now we play the same trick as in the proof of Dominated Convergence. Suppose f is bounded and continuous and suppose $|f| \le c$. By what we have just shown,

$$\int f\,d\mu = -c + \int (c+f)\,d\mu \le -c + \liminf_{n\to\infty} \int (c+f)\,d\mu_n = \liminf_{n\to\infty} \int f\,d\mu_n$$

$$-\int f\,d\mu = -c + \int (c-f)\,d\mu \le -c + \liminf_{n\to\infty} \int (c-f)\,d\mu_n = -\limsup_{n\to\infty} \int f\,d\mu_n$$

Therefore

$$\limsup_{n\to\infty} \int f\,d\mu_n \le \int f\,d\mu \le \liminf_{n\to\infty} \int f\,d\mu_n$$

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,

$$\liminf_{n\to\infty} \mu_n(A) \ge \liminf_{n\to\infty} \mu_n(\text{int}(A)) \ge \mu(\text{int}(A)) = \mu(A)$$

On the other hand,

$$\limsup_{n\to\infty} \mu_n(A) \le \limsup_{n\to\infty} \mu_n(\overline{A}) \le \mu(\overline{A}) = \mu(A)$$

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 $F_\epsilon = \{x \mid d(x,F) \le \epsilon\}$. 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 $\mu(\partial F_\epsilon) \ne 0$ for at most a countable number of ϵ. For every $n \ge 1$, there can only be a finite number F_ϵ with $\mu(\partial F_\epsilon) \ge \frac{1}{n}$ 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 1.17) 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

$$\limsup_{n\to\infty} \mu_n(F) \le \lim_{n\to\infty} \mu_n(F_{\epsilon_m}) = \mu(F_{\epsilon_m})$$

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. $\square$

DEFINITION 5.44. 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 5.45 (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

$$B^\epsilon_\delta = \{x \in S \mid \exists\, y \in S \text{ with } d(x,y) < \delta \text{ and } d'(g(x),g(y)) \geq \epsilon\}$$

Note that for $\delta' < \delta$ and fixed ϵ we have $B^\epsilon_{\delta'} \subset B^\epsilon_\delta$. Continuity of g implies that $\bigcap_{m=1}^\infty B^\epsilon_{\frac{1}{m}} = \emptyset$; and therefore by continuity of measure (Lemma 2.30) we know that $\lim_{m\to\infty} \mathbf{P}\{\xi \in B^\epsilon_{\frac{1}{m}}\} = 0$.

Now fix $\epsilon, \gamma > 0$ and note that for all $n, m > 0$, we have the bound

$$\mathbf{P}\{d'(g(\xi_n), g(\xi)) \geq \epsilon\} \leq \mathbf{P}\{d(\xi_n, \xi) \geq \frac{1}{m}\} + \mathbf{P}\{\xi \in B^\epsilon_{\frac{1}{m}}\}$$

By the previous observation, we can find an $m > 0$ such that $\mathbf{P}\{\xi \in B^\epsilon_{\frac{1}{m}}\} < \frac{\gamma}{2}$. 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 $g(\xi_n(\omega)) \to g(\xi(\omega))$. $\square$

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 5.46 (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 5.43) 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 $|f(x) - f(y)| \leq K d(x, y)$. 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 5.47 (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 5.45) to derive the result. To see $(\xi_n, \eta_n) \stackrel{d}{\to} (\xi, c)$, first note that $d((\xi_n, \eta_n), (\xi_n, c)) = d(\eta_n, c) \stackrel{P}{\to} 0$ by assumption. Therefore by the previous lemma, it suffices to show that $(\xi_n, c) \stackrel{d}{\to} (\xi, c)$. 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 $\lim_{n\to\infty} \mathbf{E}[f(\xi_n, c)] = \mathbf{E}[f(\xi, c)]$. □

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 5.48. A collection of random variables ξ_t for $t \in T$ is *uniformly integrable* if and only if $\lim_{M\to\infty} \sup_{t\in T} \mathbf{E}[|\xi_t| ; |\xi_t| > M] = 0$.

A very basic example of a uniformly integrable family is provided by i.i.d. sequences.

EXAMPLE 5.49. 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 2.55 which shows that the expectation is independent of n since $d\xi_n$ is independent of n.

The next example foreshadows the intimate relationship that uniform integrability has with limit theorems in the theory of integration.

EXAMPLE 5.50. Suppose η is an integrable random variable and ξ_t are random variables with $|\xi_t| \leq \eta$, then ξ_t are uniformly integrable. To see this let $\epsilon > 0$ be given and by Monotone Convergence choose $M > 0$ such that $\mathbf{E}[\eta; \eta > M] < \epsilon$. Then we have

$$\mathbf{E}[|\xi_t|; |\xi_t| > M] \leq \mathbf{E}[|\xi_t|; |\eta| > M] \leq \mathbf{E}[|\eta|; |\eta| > M] < \epsilon$$

so ξ_t is uniformly integrable.

The next example is often enough useful that we call it out in a Lemma.

LEMMA 5.51. *Let ξ_t be a collection of random variables such that for some $C > 0$ and $p > 1$ we have $\sup_{t\in T} \|\xi_t\|_p \leq C$. Then ξ_t is uniformly integrable.*

PROOF. This is a simple computation

$$\begin{aligned}
\lim_{M\to\infty} \sup_{t\in T} \mathbf{E}[|\xi_n|; |\xi_n| > M] &\leq \lim_{M\to\infty} \sup_{t\in T} \mathbf{E}\left[\frac{|\xi_t|^{p-1}}{M^{p-1}} |\xi_t|; |\xi_t| > M\right] \\
&\leq \lim_{M\to\infty} M^{1-p} \sup_{t\in T} \mathbf{E}[|\xi_t|^p] \\
&\leq C^p \lim_{M\to\infty} M^{1-p} = 0
\end{aligned}$$

□

LEMMA 5.52. *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 $t \in T$,

$$\begin{aligned}
\mathbf{E}[|\xi_t|] &= \mathbf{E}[|\xi_t|; |\xi_t| \leq M] + \mathbf{E}[|\xi_t|; |\xi_t| > M] \\
&\leq M + 1
\end{aligned}$$

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

$$\begin{aligned}
\mathbf{E}[|\xi_t|; A] &= \mathbf{E}[|\xi_t|; A \wedge |\xi_t| \leq M] + \mathbf{E}[|\xi_t|; A \wedge |\xi_t| > M] \\
&\leq M\delta + \mathbf{E}[|\xi_t|; |\xi_t| > M] \leq \epsilon
\end{aligned}$$

Now assume (i) and (ii). Pick $\epsilon > 0$ and $\delta > 0$ as in (ii) and let $M > 0$ be such that $\mathbf{E}[|\xi_t|] \leq M$ for all $t \in T$. Pick $N > \frac{M}{\delta}$ and note that

$$\mathbf{P}\{|\xi_t| > N\} \leq \frac{\mathbf{E}[|\xi_t|]}{N} \leq \frac{M}{N} < \delta$$

so by (ii), $\mathbf{E}[|\xi_t|; |\xi_t| > N] < \epsilon$ and uniform integrability is proven. □

Here are a few simple results that illustrates how the conditions for uniform integrability in the previous Lemma can often be more convenient than the definition.

LEMMA 5.53. *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 5.52 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 5.52 is shown.

To see condition (ii) of Lemma 5.52, 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

$$\begin{aligned}
\mathbf{E}\left[|a\xi_t + b\eta_t|^p ; A\right] &= \|a\xi_t \mathbf{1}_A + b\eta_t \mathbf{1}_A\|_p^p \\
&\leq \left(a \|\xi_t \mathbf{1}_A\|_p + b \|\eta_t \mathbf{1}_A\|_p\right)^p \\
&\leq \left(a \frac{\epsilon^{1/p}}{2a} + b \frac{\epsilon^{1/p}}{2b}\right)^p = \epsilon
\end{aligned}$$

□

LEMMA 5.54. *Suppose ξ_t for $t \in T$ is a uniformly integrable family of random variables and η is a bounded random variable, then $\eta\xi_t$ is a uniformly integrable family.*

PROOF. This is essentially trivial when using Lemma 5.52. We know that $\sup_t \mathbf{E}\left[|\xi_t|\right] \leq \|\eta\|_\infty \sup_t \mathbf{E}\left[|\xi_t|\right] < \infty$ and

$$\lim_{\mathbf{P}\{A\} \to 0} \sup_t \mathbf{E}\left[|\eta\xi_t| ; A\right] \leq \|\eta\|_\infty \lim_{\mathbf{P}\{A\} \to 0} \sup_t \mathbf{E}\left[|\xi_t| ; A\right] = 0$$

so uniformly integrability follows. □

Here is an example that shows that the condition (i) of Lemma 5.52 is not sufficient to guarantee uniform integrability (that is to say, this is an example of an L^1 bounded family of random variables that is not uniformly integrable).

EXAMPLE 5.55. 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 5.16. 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 $\sup_n \mathbf{E}\left[|\xi_n| ; |\xi_n| > M\right] = 1$.

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 5.56. *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:

$$\begin{aligned}
\mathbf{E}\left[\xi\right] &= \lim_{R \to \infty} \mathbf{E}\left[f_R(\xi)\right] && \text{by Monotone Convergence Theorem} \\
&= \lim_{R \to \infty} \lim_{n \to \infty} \mathbf{E}\left[f_R(\xi_n)\right] && \text{because } \xi_n \xrightarrow{d} \xi \\
&\leq \liminf_n \mathbf{E}\left[\xi_n\right] && \text{because } f_R(x) \leq x \text{ for all } x \geq 0
\end{aligned}$$

An alternative derivation is:

$$\begin{aligned}
\mathbf{E}\left[\xi\right] &= \int \mathbf{P}\{\xi > \lambda\}\, d\lambda && \text{by Lemma 3.8} \\
&\leq \int \liminf_n \mathbf{P}\{\xi_n > \lambda\}\, d\lambda && \text{by Portmanteau Lemma 5.43} \\
&\leq \liminf_n \int \mathbf{P}\{\xi_n > \lambda\}\, d\lambda && \text{by Fatou's Lemma (Theorem 2.45)} \\
&= \liminf_n \mathbf{E}\left[\xi_n\right] && \text{by Lemma 3.8}
\end{aligned}$$

Now assume that ξ_n is uniformly integrable. Then by what we have just proven and Lemma 5.52 we have

$$\mathbf{E}\left[\xi\right] \leq \liminf_n \mathbf{E}\left[\xi_n\right] \leq \sup_n \mathbf{E}\left[\xi_n\right] < \infty$$

So now we use the triangle inequality to write

$$|\mathbf{E}\left[\xi_n\right] - \mathbf{E}\left[\xi\right]| \leq |\mathbf{E}\left[\xi_n\right] - \mathbf{E}\left[f_R(\xi_n)\right]| + |\mathbf{E}\left[f_R(\xi_n)\right] - \mathbf{E}\left[f_R(\xi)\right]| + |\mathbf{E}\left[f_R(\xi)\right] - \mathbf{E}\left[\xi\right]|$$

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:

$$\begin{aligned}
\lim_{R \to \infty} \limsup_n |\mathbf{E}\left[\xi_n\right] - \mathbf{E}\left[f_R(\xi_n)\right]| &= \lim_{R \to \infty} \limsup_n \left(\mathbf{E}\left[\xi_n; \xi_n > R\right] - R\mathbf{P}\{\xi_n > R\}\right) \\
&\leq \lim_{R \to \infty} \limsup_n \mathbf{E}\left[\xi_n; \xi_n > R\right] - \lim_{R \to \infty} \liminf_n R\mathbf{P}\{\xi_n > R\} \\
&= 0
\end{aligned}$$

where in the last line we have used uniform integrability of ξ_n as well as the following

$$\lim_{R \to \infty} R \liminf_n \mathbf{P}\{\xi_n > R\} \leq \lim_{R \to \infty} R \sup_n \mathbf{P}\{\xi_n > R\} \leq \lim_{R \to \infty} \sup_n \mathbf{E}\left[\xi_n; \xi_n > R\right] = 0$$

The second term we have $\limsup_n |\mathbf{E}\left[f_R(\xi_n)\right] - \mathbf{E}\left[f_R(\xi)\right]| = 0$ because f_R is bounded continuous and $\xi_n \xrightarrow{d} \xi$. The third term we have $\lim_{R \to \infty} |\mathbf{E}\left[f_R(\xi)\right] - \mathbf{E}\left[\xi\right]| = 0$ by Monotone Convergence.

Putting the bounds on the three terms of the right hand side together we have $\limsup_{n \to \infty} |\mathbf{E}\left[\xi_n\right] - \mathbf{E}\left[\xi\right]| = 0$ which by positivity shows $\lim_{n \to \infty} |\mathbf{E}\left[\xi_n\right] - \mathbf{E}\left[\xi\right]| = 0$.

TODO: Here is an alternative proof of the same fact by approximating $x\mathbf{1}_{x\leq R}$ 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 $x\mathbf{1}_{x\geq R}$ 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

$$g_R(x) = x \wedge (R-x)_+ = \begin{cases} 0 & \text{if } x < 0 \text{ or } x > R \\ x & \text{if } 0 \leq x \leq \frac{R}{2} \\ R - x & \text{if } \frac{R}{2} < x \leq R \end{cases}$$

and note that

$$x - g_R(x) = \begin{cases} 0 & \text{if } x < \frac{R}{2} \\ 2x - R & \text{if } \frac{R}{2} \leq x \leq R \\ x & \text{if } R < x \end{cases}$$

so $x\mathbf{1}_{x\geq R} \leq x - g_R(x) \leq x$, and $\lim_{R\to\infty} x - g_R(x) = 0$. Putting these facts together we see

$$\begin{aligned}
&\lim_{R\to\infty} \limsup_{n\to\infty} \mathbf{E}[\xi_n; \xi_n \geq R] \\
&\leq \lim_{R\to\infty} \limsup_{n\to\infty} \mathbf{E}[\xi_n - g_R(\xi_n)] \\
&= \lim_{R\to\infty} \left(\lim_{n\to\infty} \mathbf{E}[\xi_n] - \lim_{n\to\infty} \mathbf{E}[g_R(\xi_n)] \right) \\
&= \lim_{R\to\infty} \mathbf{E}[\xi] - \mathbf{E}[g_R(\xi)] && \text{by assumption and } \xi_n \xrightarrow{d} \xi \\
&= \lim_{R\to\infty} \mathbf{E}[\xi - g_R(\xi)] = 0 && \text{by Dominated Convergence}
\end{aligned}$$

$\square$

Converge in mean and convergence of means become equivalent in the presence of almost sure convergence.

LEMMA 5.57. *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,

$$\lim_{n\to\infty} \|\xi_n\|_p \leq \lim_{n\to\infty} \|\xi_n - \xi\|_p + \|\xi\|_p = \|\xi\|_p$$

and

$$\|\xi\|_p \leq \lim_{n\to\infty} \|\xi_n - \xi\|_p + \lim_{n\to\infty} \|\xi_n\|_p = \lim_{n\to\infty} \|\xi_n\|_p$$

therefore $\lim_{n\to\infty} \|\xi_n\|_p = \|\xi\|_p$.

To see (ii), if $\xi_n \overset{a.s.}{\to} \xi$ and $\|\xi_n\|_p \to \|\xi\|_p$ then we know that $|\xi_n - \xi|^p \overset{a.s.}{\to} 0$ and we have the bound

$$|\xi_n - \xi|^p \leq (|\xi_n| + |\xi|)^p \leq 2^p \max(|\xi_n|^p, |\xi|^p) \leq 2^p (|\xi_n|^p + |\xi|^p)$$

and our assumption tells us that $\lim_{n\to\infty} 2^p \mathbf{E}[|\xi_n|^p + |\xi|^p] = 2^{p+1} \|\xi\|_p^p < \infty$. Therefore we can apply Dominated Convergence (Theorem 2.51) 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 5.58. *Let $\xi, \xi_1, \xi_2, \dots$ be random variables in L^p for $p > 0$ and suppose $\xi_n \overset{P}{\to} \xi$. Then the following are equivalent:*

(i) $\xi_n \overset{L^p}{\to} \xi$
(ii) $\|\xi_n\|_p \to \|\xi\|_p$
(iii) *The sequence of random variables $|\xi_1|^p, |\xi_2|^p, \dots$ is uniformly integrable.*

Proof. To see (i) implies (ii), this is the first part of Lemma 5.57.

Note that since $\xi_n \overset{P}{\to} \xi$ implies $\xi_n \overset{d}{\to} \xi$ we know that (ii) and (iii) are equivalent by Lemma 5.56.

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 $\xi_n \overset{P}{\to} \xi$ by Lemma 5.10 there is a further subsequence $N'' \subset N'$ such that $\xi_n \overset{a.s.}{\to} \xi$ along N''. However, Lemma 5.57 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 5.53 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

$$\begin{aligned}
\lim_{n\to\infty} \mathbf{E}[|\xi_n - \xi|^p] &= \lim_{R\to\infty} \limsup_{n\to\infty} (\mathbf{E}[|\xi_n - \xi|^p ; |\xi_n - \xi|^p \leq \epsilon] + \mathbf{E}[|\xi_n - \xi|^p ; \epsilon < |\xi_n - \xi|^p < R] + \mathbf{E}[|\xi_n - \xi|^p ; |\xi_n \\
&\leq \epsilon + \lim_{R\to\infty} \lim_{n\to\infty} R\mathbf{P}\{\epsilon < |\xi_n - \xi|^p\} + \lim_{R\to\infty} \sup_n \mathbf{E}[|\xi_n - \xi|^p ; |\xi_n - \xi|^p \geq R] \\
&= \epsilon
\end{aligned}$$

and since $\epsilon > 0$ was arbitrary, we have $\lim_{n\to\infty} \mathbf{E}[|\xi_n - \xi|^p] = 0$. □

TODO: Show how the proof of the Weak Law of Large Numbers applies to uniformly integrable sequences not just i.i.d.

Lemma 5.59. *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 3.8 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 Tonelli'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 2.113) 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). $\square$

5. Topology of Weak Convergence

We have defined convergence of a sequence of probability measures but have skirted describing the topology underlying this notion of convergence. An intuitively appealing approach would be to define a metric on the space of probability measures such that two measures are close if their values on some chosen collection of sets are close. A moments reflection on the Portmanteau Theorem 5.43 tells us that such a condition is likely to be too strong. For example, if we pick a closed set F then we know that it is possible for $\mu_n \xrightarrow{w} \mu$ but to have $\mu(F)$ strictly larger than all of the $\mu_n(F)$; even more precisely by considering the standard delta mass example it is possible for $\mu(F)$ to be equal to one but for all $\mu_n(F)$ to be zero.

As it turns out the intuitive idea can be rescued with a small modification. Again thinking about the delta mass example, we can see that while $\mu_n(F)$ remains zero for all n the mass of μ_n get arbitrary close to F so that we can potentially measure how close μ and μ_n are by looking at how much we have to *thicken* the set F to capture the mass of μ_n. Generalizing we may want to say that μ and ν are close if for every set A in some collection we don't have to thicken A very much for the $\mu(A)$ and $\nu(A)$ to be close; in fact the amount of thickening required may be a quantitative measure of closeness. We now proceed to make this idea precise.

Definition 5.60. Given a metric space (S, d) a subset $A \subset S$ and $\epsilon > 0$ define

$$A^\epsilon = \{x \in S \mid \inf_{y\in A} d(x, y) < \epsilon\}$$

Lemma 5.61. *For any set $A \subset S$ we have*

(i) *A^ϵ is an open set.*
(ii) $A^\epsilon = (\overline{A})^\epsilon$.
(iii) $(A^\epsilon)^\delta \subset A^{\epsilon+\delta}$

PROOF. To see (i) pick an $x \in A^\epsilon$ and pick $y \in A$ such that $d(x, y) < \epsilon$. Then by the triangle inequality for all $z \in S$ such that $d(x, z) < (\epsilon - d(x, y))/2$ we have

$$d(y, z) \leq d(x, y) + d(x, z) < (\epsilon + d(x, y))/2 < \epsilon$$

showing $z \in A^\epsilon$.

To see (ii), it is clear from the definition that $A^\epsilon \subset (\overline{A})^\epsilon$ since $A \subset \overline{A}$. To see the opposite inclusion suppose $x \in (\overline{A})^\epsilon$ and pick $y \in \overline{A}$ such that $d(x, y) < \epsilon$ then by density of A in $\overline{A}$ pick $z \in A$ such that $d(y, z) < (\epsilon - d(x, y))/2$. The triangle inequality as before shows $d(z, x) < \epsilon$ and therefore $x \in A^\epsilon$.

To see (iii) suppose $z \in (A^\epsilon)^\delta$ and pick $y \in A^\epsilon$ such that $d(z, y) < \delta$. Now pick $x \in A$ such that $d(x, y) < \epsilon$ and use the triangle inequality to conclude

$$d(x, z) \leq d(x, y) + d(y, z) < \epsilon + \delta$$

□

LEMMA 5.62 (Levy-Prohorov Metric). *Let (S, d) be a metric space and let $\mathcal{P}(S)$ denote the set of probability measures. Define*

$$\rho(\mu, \nu) = \inf\{\epsilon > 0 \mid \mu(F) \leq \nu(F^\epsilon) + \epsilon \text{ for all closed subsets } F \subset S\}$$

Then in fact

$$\rho(\mu, \nu) = \inf\{\epsilon > 0 \mid \mu(A) \leq \nu(A^\epsilon) + \epsilon \text{ for all Borel subsets } A \subset S\}$$

and furthermore ρ is a metric on $\mathcal{P}(S)$.

PROOF. Claim 1: For every $\alpha, \beta > 0$ if $\mu(F) \leq \nu(F^\alpha) + \beta$ for all closed subsets $F \subset S$ then $\nu(F) \leq \mu(F^\alpha) + \beta$ for all closed subsets $F \subset S$.

The proof of the claim relies on the observation that $F \subset (((F^\alpha)^c)^\alpha)^c$. To see the observation note that if $x \in F$ and $x \in ((F^\alpha)^c)^\alpha$ then we can find $y \notin F^\alpha$ such that $d(x, y) < \alpha$ which contradicts the definition of F^α. The claim follows by using inclusion in the observation in addition to the fact that F^α is open (hence $(F^\alpha)^c$ is closed) so

$$\nu(F) \leq \nu((((F^\alpha)^c)^\alpha)^c) = 1 - \nu(((F^\alpha)^c)^\alpha) \leq 1 - \mu((F^\alpha)^c) + \beta = \mu(F^\alpha) + \beta$$

With Claim 1 in hand symmetry of ρ now follows as the sets $\{\epsilon > 0 \mid \mu(F) < \nu(F^\epsilon) + \epsilon\}$ and $\{\epsilon > 0 \mid \nu(F) < \mu(F^\epsilon) + \epsilon\}$ are equal a fortiori the infimum are equal.

Clearly by continuity of measure (Lemma 2.30) we have $\mu(A) = \lim_{\epsilon \to 0} \mu(A^\epsilon)$ and therefore $\rho(\mu, \mu) = 0$. Conversely if $\rho(\mu, \nu) = 0$ then we pick a closed set F and for every $\epsilon > 0$ we have $\mu(F) < \nu(F^\epsilon) + \epsilon$. Again using continuity of measure we can conclude that $\mu(F) \leq \nu(F)$. By symmetry of ρ that we have already proven we can conclude $\nu(F) \leq \mu(F)$ and there $\mu(F) = \nu(F)$ for all closed set $F \subset S$. Since closed sets are a π-system that generate the Borel subsets of S we conclude that $\mu = \nu$ by a monotone class argument (Lemma 2.71).

To show the triangle inequality let μ, ν, ζ be probability measures and suppose $\epsilon > 0$ is such that $\mu(F) < \nu(F^\epsilon) + \epsilon$ and $\delta > 0$ is such that $\nu(F) < \zeta(F^\delta) + \delta$ for all closed sets $F \subset S$. Now choose a particular $F \subset S$ be closed then

$$\begin{aligned} \mu(F) &\leq \nu(F^\epsilon) + \epsilon \leq \nu(\overline{F^\epsilon}) + \epsilon \leq \zeta((\overline{F^\epsilon})^\delta) + \epsilon + \delta \\ &= \zeta((F^\epsilon)^\delta) + \epsilon + \delta \leq \zeta(F^{\epsilon+\delta}) + \epsilon + \delta \end{aligned}$$

Thus $\rho(\mu, \zeta) \leq \rho(\mu, \nu) + \rho(\nu, \zeta)$. TODO: This last conclusion is more or less obvious but there are some minor details that could be filled in here. □

TODO: Ky Fan metric that metrizes convergence in probability.

CHAPTER 6

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 6.1 (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 $\xi^2 \mathbf{1}_{|\xi|\geq\epsilon\sqrt{n}\sigma} \stackrel{a.s.}{\to} 0$. 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 4.34). 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 3.7) $\int \varphi \, d\gamma = \mathbf{E}\left[\varphi(\hat{T}_n)\right]$ 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 1.20) 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 4.18 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 1.21) 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

$$\begin{aligned}
\sum_{m=1}^n \left|\mathbf{E}\left[R_m(\hat{\xi}_m); \left|\hat{\xi}_m\right| > \epsilon\right]\right| &\leq \|\varphi''\|_\infty \sum_{m=1}^n \left|\mathbf{E}\left[\left|\hat{\xi}_m\right|^2; \left|\hat{\xi}_m\right| > \epsilon\right]\right| \\
&= \|\varphi''\|_\infty \frac{1}{\Sigma_n^2}\sum_{m=1}^n \left|\mathbf{E}\left[|\xi_m|^2; |\xi_m| > \epsilon\Sigma_n\right]\right| \\
&= \|\varphi''\|_\infty\, g_\epsilon(n)
\end{aligned}$$

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:

$$\begin{aligned}
\sum_{m=1}^n \left|\mathbf{E}\left[R_m(\hat{\eta}_m)\right]\right| &\leq \frac{\|\varphi'''\|_\infty}{6} \sum_{m=1}^n \left|\mathbf{E}\left[|\hat{\eta}_m|^3\right]\right| \\
&= \frac{\|\varphi'''\|_\infty}{6} \sum_{m=1}^n \frac{\sigma_m^3}{\Sigma_n^3} \left|\mathbf{E}\left[|\eta_m|^3\right]\right| \\
&= \frac{r_n \|\varphi'''\|_\infty}{6} \sum_{m=1}^n \frac{\sigma_m^2}{\Sigma_n^2} \left|\mathbf{E}\left[|\eta_m|^3\right]\right| \\
&= \frac{r_n \|\varphi'''\|_\infty}{6} \frac{2\sqrt{2}}{\sqrt{\pi}} < \frac{r_n \|\varphi'''\|_\infty}{2}
\end{aligned}$$

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 $1 \leq m \leq n$,

$$\begin{aligned}
\frac{\sigma_m^2}{\Sigma_n^2} &= \frac{1}{\Sigma_n^2}\left(\mathbf{E}\left[\xi_m^2; |\xi_m| < \epsilon\Sigma_n\right] + \mathbf{E}\left[\xi_m^2; |\xi_m| \geq \epsilon\Sigma_n\right]\right) \\
&\leq \frac{1}{\Sigma_n^2}\left(\epsilon^2\Sigma_n^2 + \Sigma_n^2 g_n(\epsilon)\right) = \epsilon^2 + g_n(\epsilon)
\end{aligned}$$

hence $r_n^2 = \max_{1 \leq m \leq n} \frac{\sigma_m^2}{\Sigma_n^2} \leq \epsilon^2 + g_n(\epsilon)$. ☐

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 6.2 (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 $\mathbf{Var}\left(\xi_n\right) = (-n)^2 \cdot \frac{1}{2n^2} + 0 \cdot (1 - \frac{1}{2n^2}) + n^2 \cdot \frac{1}{2n^2} = 1$. $\sum_{n=1}^\infty \mathbf{P}\{\xi_n \neq 0\} = \sum_{n=1}^\infty \frac{1}{n^2} < \infty$ 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 6.1, $\Sigma_n = \sqrt{n}$, 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 6.3. *Let $\xi, \xi_1, \xi_2, \dots$ be random vectors in $\mathbb{R}^N$, then $\xi_n \overset{d}{\to} \xi$ 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 $\xi_n \overset{d}{\to} \xi$ 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 2.129, 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,

$$\begin{aligned}
|\mathbf{E}[f_k(\xi) - f(\xi)]| &\leq \mathbf{E}[|f_k(\xi) - f(\xi)|; |\xi| \leq R] + \mathbf{E}[|f_k(\xi) - f(\xi)|; |\xi| > R] \\
&\leq \epsilon \mathbf{P}\{|\xi| \leq R\} + 2\|\xi\|_\infty \mathbf{P}\{|\xi| > R\} \\
&\leq \epsilon + \frac{\epsilon}{2} < 2\epsilon
\end{aligned}$$

and via the same calculation, for $n > N_1$

$$|\mathbf{E}[f_k(\xi_n) - f(\xi_n)]| \leq \epsilon + 2\|\xi_n\|_\infty \mathbf{P}\{|\xi_n| > R\} \leq 2\epsilon$$

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 $|\mathbf{E}[f_k(\xi_n) - f_k(\xi)]| < \epsilon$. Putting these three estimates together, we have for $n > \max(N_1, N_2)$,

$$|\mathbf{E}[f(\xi_n) - f(\xi)]| \leq 5\epsilon$$

$\square$

We are not going to prove the following but we should talk about it:

THEOREM 6.4. *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.

CHAPTER 7

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 7.1. 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 3.17, $\left|\int f \, d\mu\right| \leq \int |f| \, d\mu$. ☐

Definition 7.2. 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 7.3. *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

$$\begin{aligned}
\left|e^{i\langle u,x\rangle} - e^{i\langle v,x\rangle}\right|^2 &= \left|e^{i\langle u-v,x\rangle} - 1\right|^2 \\
&= (\cos(\langle u-v,x\rangle) - 1)^2 + \sin^2(\langle u-v,x\rangle) \\
&= 2(1 - \cos(\langle u-v,x\rangle) \\
&\leq \langle u-v,x\rangle^2 && \text{by Lemma C.1} \\
&\leq \|u-v\|_2^2 \, \|x\|_2^2 && \text{by Cauchy Schwartz}
\end{aligned}$$

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 7.4. 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 7.5. *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

$$\begin{aligned}
\left|\frac{1}{t}e^{i\langle u,x\rangle}\left(e^{itx_j}-1\right)\right|^2 &= \left|\frac{e^{itx_j}-1}{t}\right|^2 \\
&= \frac{\cos^2(tx_j)-2\cos(tx_j)+1+\sin^2(tx_j)}{t^2} \\
&= 2\left(\frac{1-\cos(tx_j)}{t^2}\right) \\
&\leq x_j^2 && \text{by Lemma C.1}
\end{aligned}$$

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 7.6. *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 2.88) 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 Tonelli's Theorem,

$$\begin{aligned} \int\int\int \left| e^{-\frac{\epsilon^2 u^2}{2}} f(x) e^{iux} e^{-iuy} \right| d\mu(y)\, dx\, du &= \int\int\int e^{-\frac{\epsilon^2 u^2}{2}} \left| f(x) \right| d\mu(y)\, dx\, du \\ &= \int \left| f(x) \right| dx \int e^{-\frac{\epsilon^2 u^2}{2}}\, du < \infty \end{aligned}$$

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 7.10, 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 2.131 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,

$$\lim_{\epsilon \to 0} \left| \int_{-\infty}^{\infty} f - f * \rho_\epsilon\, d\mu \right| \leq \lim_{\epsilon \to 0} \int_{-\infty}^{\infty} \left| f - f * \rho_\epsilon \right| d\mu \quad \leq \lim_{\epsilon \to 0} \sup_x \left| f - f * \rho_\epsilon \right| = 0$$

and by integrability of $\hat{f}(u)$, the fact that $|\hat{\mu}| \leq 1$ (Lemma 7.3) 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 7.7. *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 2.131 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 7.8. *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 4.18 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 $\varphi_\xi(u_1,\dots,u_n) = \prod_{j=1}^n \varphi_{\xi_j}(u_j)$, 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 7.7 we know that ξ and η have the same distribution. Thus the ξ_j are also independent by Lemma 4.5 and the equality of the distributions of each ξ_j and η_j. ☐

LEMMA 7.9. *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

$$\begin{aligned}\varphi_{\xi+\eta}(u) &= \mathbf{E}\left[e^{i\langle u,\xi+\eta\rangle}\right] = \mathbf{E}\left[e^{i\langle u,\xi\rangle}e^{i\langle u,\eta\rangle}\right] \\ &= \mathbf{E}\left[e^{i\langle u,\xi\rangle}\right]\mathbf{E}\left[e^{i\langle u,\eta\rangle}\right] = \varphi_\xi(u)\varphi_\eta(u) \qquad \text{by Lemma 4.18}\end{aligned}$$

☐

EXAMPLE 7.10. 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 7.5 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 7.3 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 7.11. *Let ξ be a random vector in $\mathbb{R}^d$ then for $a \in \mathbb{R}^r$ and A an $r \times d$ 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

$$\langle u, Av\rangle = u^*Av = (u^*Av)^* = v^*A^*u = \langle A^*u, v\rangle$$

$\square$

EXAMPLE 7.12. 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 7.11 and Example 7.10

$$\varphi_\xi(u) = e^{iu\mu}\varphi_\eta(\sigma u) = e^{iu\mu - \frac{1}{2}u^2\sigma^2}$$

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 7.13 (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 6.3 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 2.131 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,

$$\begin{aligned}\left|\int f\, d\mu_n - \int f\, d\mu\right| &\leq \left|\int (f - \rho_\epsilon * f)\, d\mu_n\right| + \left|\int \rho_\epsilon * f\, d\mu_n - \int \rho_\epsilon * f\, d\mu\right| + \left|\int (\rho_\epsilon * f - f)\, d\mu\right| \\ &\leq \delta + \frac{1}{2\pi}\left|\int \hat{f}(t)e^{-\frac{1}{2}\epsilon^2t^2}(\hat{\mu}_n(t) - \hat{\mu}(t))\, dt\right| + \delta\end{aligned}$$

where we have used the Plancherel Theorem (Theorem 7.6) 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 $\hat{f}(t) \leq \|f\|_\infty$ and together with Lemma 7.3 we see that

$$\left|\hat{f}(t)e^{-\frac{1}{2}\epsilon^2t^2}(\hat{\mu}_n(t) - \hat{\mu}(t))\right| \leq 2\|f\|_\infty e^{-\frac{1}{2}\epsilon^2t^2}$$

where the upper bound is an integrable function of t. Therefore by Dominated Convergence we see that $\limsup_{n\to\infty} \left|\int f\, d\mu_n - \int f\, d\mu\right| \leq 2\delta$. Since $\delta > 0$ was arbitrary, we have $\int f\, d\mu_n = \int f\, d\mu$. $\square$

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 7.14 (Central Limit Theorem). *Let $\xi, \xi_1, \xi_2, \dots$ be i.i.d. random variables with $\mu = \boldsymbol{E}[\xi]$ and $\sigma = \boldsymbol{Var}(\xi_n) < \infty$, then*

$$\sqrt{n}(\frac{1}{n}\sum_{i=1}^n \xi_i - \mu) \xrightarrow{d} N(0, \sigma^2)$$

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 7.13 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

$$\lim_{n \to \infty} \mathbf{E}\left[e^{itS_n/\sqrt{n}} \right] = \lim_{n \to \infty} \left(1 - \frac{t^2}{2n}(1 + \epsilon_n) \right)^n = \lim_{n \to \infty} e^{n \log(1 - \frac{t^2}{2n}(1+\epsilon_n))} = e^{-t^2/2}$$

□

It is also useful to call out a useful corollary of the continuity theorem that allows one to characterize convergence in distribution of random vectors by considering one dimensional projections.

COROLLARY 7.15 (Cramer Wold Device). *Let $\xi, \xi_1, \xi_2, \ldots$ be random vectors in $\mathbb{R}^d$. Then $\langle c, \xi_n \rangle \xrightarrow{d} \langle c, \xi \rangle$ for all $c \in \mathbb{R}^d$ if and only if $\xi_n \xrightarrow{d} \xi$.*

PROOF. Simply note that for all random vectors ξ, $c \in \mathbb{R}^d$ and $x \in \mathbb{R}$

$$\varphi_{\langle c, \xi \rangle}(x) = \mathbf{E}\left[e^{ix\langle c, \xi \rangle} \right] = \varphi_\xi(xc)$$

Therefore if $\langle c, \xi_n \rangle \xrightarrow{d} \langle c, \xi \rangle$ for all $c \in \mathbb{R}^d$ then by the Glivenko-Levy Continuity Theorem 7.13 we know that

$$\lim_{n \to \infty} \varphi_{\xi_n}(c) = \lim_{n \to \infty} \varphi_{\langle c, \xi_n \rangle}(1) = \varphi_{\langle c, \xi \rangle}(1) = \varphi_\xi(c)$$

and applying the Theorem again we conclude that $\xi_n \xrightarrow{d} \xi$. In a completely analogous way, if we assume that $\xi_n \xrightarrow{d} \xi$ then for all $c \in \mathbb{R}^d$ and $x \in \mathbb{R}$, then

$$\lim_{n \to \infty} \varphi_{\langle c, \xi_n \rangle}(x) = \lim_{n \to \infty} \varphi_{\xi_n}(xc) = \varphi_\xi(xc) = \varphi_{\langle c, \xi \rangle}(x)$$

from which we conclude that $\langle c, \xi_n \rangle \xrightarrow{d} \langle c, \xi \rangle$. □

THEOREM 7.16 (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 $\pm\infty$ due to mass escaping to infinity. If we assume continuity at 0, then we can prove tightness which keeps the mass from escaping and shows that the limits are $0, 1$ as required of a distribution function. Note that the pointwise limit of a sequence of characteristic functions is the characteristic function of a measure (though not necessarily a probability measure); this fact is often know as the Helly Selection Theorem. It can be restated in terms of a topology on the space of locally finite measures called the vague topology and the Helly Selection Theorem can be restated as saying that the space of probability measures is relatively sequentially compact in the vague topology on the locally finite measures on $\mathbb{R}^n$.

1. Gaussian Random Vectors and the Multidimensional Central Limit Theorem

There is a version of the Central Limit Theorem for random vectors in $\mathbb{R}^d$ 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 7.17. A random vector ξ in $\mathbb{R}^d$ is said to be a *Gaussian random vector* is for every $a \in \mathbb{R}^d$, the random variable $\langle a, \xi \rangle$ 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 7.18. *A random vector ξ in $\mathbb{R}^d$ is Gaussian if and only if there is a $\mu \in \mathbb{R}^d$ and a symmetric positive semi-definite matrix $Q \in \mathbb{R}^{d \times d}$ 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 $a \in \mathbb{R}^d$ 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 7.11 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 7.12 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 $\mu = (\mu_1, \dots, \mu_d) = \mathbf{E}[\xi]$ and let $Q = \mathbf{Cov}(\xi)$. Pick $a \in \mathbb{R}^d$ and note that

$$\begin{aligned}
\mathbf{E}[\langle a, \xi\rangle] &= \langle a, \mu\rangle \\
\mathbf{Var}(\langle a, \xi\rangle) &= \mathbf{E}\left[(\langle a, \xi\rangle - \mathbf{E}[\langle a, \xi\rangle])^2\right] \\
&= \mathbf{E}\left[(\langle a, \xi - \mu\rangle)^2\right] \\
&= \mathbf{E}\left[a^*(\xi - \mu)(\xi - \mu)^* a\right] \\
&= a^* \mathbf{E}\left[(\xi - \mu)(\xi - \mu)^*\right] a = \langle a, Qa\rangle
\end{aligned}$$

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 7.12, 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 7.11 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 7.7), 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 7.19. Let $\xi_1, \ldots, \xi_d$ be independent random variables with ξ_i being normal $N(\mu_i, \sigma_i^2)$. Then $\xi = (\xi_1, \ldots, \xi_d)$ is a Gaussian random vector. In fact, if we let $\mu = (\mu_1, \ldots, \mu_d)$ and

$$Q = \mathrm{Diag}(\sigma_1^2, \ldots, \sigma_d^2)$$

then $\xi = N(\mu, Q)$.

EXAMPLE 7.20. Let $(\xi_1, \ldots, \xi_d)$ be $\xi = N(\mu, Q)$. Let A be a $r \times d$ matrix then $A\xi$ is $N(A\mu, AQA^T)$.

To see this note that if $a \in \mathbb{R}^d$ then $\langle a, A\xi \rangle = \langle a^T A, \xi \rangle$ so we see that $A\xi$ is in fact Gaussian. The fact that $\mathbf{E}[A\xi] = A\mathbf{E}[\xi]$ is immediate from linearity of expectation. To calculate the covariance we

$$\mathbf{Cov}(A\xi) = \mathbf{E}\left[A(\xi - \mu)(\xi - \mu)^T A^T\right] = A\mathbf{E}\left[(\xi - \mu)(\xi - \mu)^T\right] A^T = AQA^T$$

Gaussian random vectors are easy to handle in many cases because independence is so easily verified; this is a case in which uncorrelated implies independent.

PROPOSITION 7.21. *Let ξ be a Gaussian random vector in $\mathbb{R}^d$ then for any subset $I \subset \{1, \ldots, d\}$ let $\xi_I = (\xi_{i_1}, \ldots, \xi_{i_r})$ be random vector in $\mathbb{R}^r$ where r is the cardinality of I. Given any two disjoint subsets I, J we have ξ_I and ξ_J are independent if and only if $\boldsymbol{Cov}(\xi_I, \xi_J) = 0$.*

PROOF. Since independence always implies uncorrelated it suffices to assume that ξ_I and ξ_J are uncorrelated and then show independence.

Concatenate ξ_I and ξ_J into a single random vector $\eta = (\xi_I, \xi_J)$. Clearly η is a Gaussian random vector since it is a projection of a Gaussian random vector; furthermore since ξ_I and ξ_J are uncorrelated it has a block diagonal covariance matrix

$$\mathbf{Cov}(\eta) = \begin{bmatrix} \mathbf{Cov}(\xi_I) & 0 \\ 0 & \mathbf{Cov}(\xi_J) \end{bmatrix}$$

By Theorem 7.18 we can calculate the characteristic function.

$$\varphi_\eta(u) = e^{i\langle u, \mu \rangle - \frac{1}{2}\langle u, \mathbf{Cov}(\eta) u \rangle} = e^{i\langle u_I, \mu_I \rangle - \frac{1}{2}\langle u_I, \mathbf{Cov}(\xi_I) u_I \rangle} e^{i\langle u_J, \mu_J \rangle - \frac{1}{2}\langle u_J, \mathbf{Cov}(\xi_J) u_J \rangle} = \varphi_{\xi_I}(u)\varphi_{\xi_J}(u)$$

which by Theorem 7.8 shows that ξ_I and ξ_J are independent. □

The characterization of Gaussian random vectors using characteristic functions allows us to see that almost sure limits of Gaussian random vectors are Gaussian random vectors. We will need this result when we construct Brownian motion later on.

LEMMA 7.22. *Let $\xi_1, \xi_2, \ldots$ be a sequence of random vectors in $\mathbb{R}^d$ 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 7.18 and the Glivenko-Levy Continuity Theorem (Theorem 7.13) 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^{ix}. Thus, using Lemma 7.18 again shows that ξ is $N(\mu, C)$. □

TODO: Gaussian Random Variables in $\mathbb{R}^n$ and the multidimensional CLT. Pretty sure this can be derived from the Cramer Wold device. 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.

One last thing we need in the sequel are estimates on the tails of normal random variables. These results are not required yet nor do they add anything significant to the conceptual picture so the reader can safely skip over them and return to them when they are referenced.

LEMMA 7.23. *Given an $N(0,1)$ random variable ξ we have for all $\lambda > 0$,*

$$\frac{\lambda}{\sqrt{2\pi}(1+\lambda^2)} e^{-\lambda^2/2} \leq \boldsymbol{P}\{\xi \geq \lambda\} \leq \frac{1}{\sqrt{2\pi}\lambda} e^{-\lambda^2/2}$$

PROOF. We start by showing the upper bound

$$\mathbf{P}\{\xi \geq \lambda\} = \frac{1}{\sqrt{2\pi}} \int_\lambda^\infty e^{\frac{-x^2}{2}} \, dx \leq \frac{1}{\sqrt{2\pi}} \int_\lambda^\infty \frac{x}{\lambda} e^{\frac{-x^2}{2}} \, dx = \frac{1}{\sqrt{2\pi}\lambda} e^{\frac{-\lambda^2}{2}}$$

Interestingly, the lower bound follows from the upper bound. Define

$$f(\lambda) = \lambda e^{-\lambda^2/2} - (1+\lambda^2) \int_\lambda^\infty e^{-x^2/2} \, dx$$

and notice that $f(0) = -\int_0^\infty e^{-x^2/2} \, dx = -\frac{\sqrt{2\pi}}{2} < 0$. Furthermore if we use the upper bound just proven

$$\lim_{\lambda\to\infty} f(\lambda) = \lim_{\lambda\to\infty} \lambda^2 \int_\lambda^\infty e^{-x^2/2} \, dx \leq \lim_{\lambda\to\infty} \lambda e^{-\lambda^2/2} = 0$$

and therefore $\lim_{\lambda\to\infty} f(\lambda) = 0$. In addition we have for $\lambda \geq 0$,

$$\begin{aligned} \frac{d}{d\lambda} f(\lambda) &= e^{-\lambda^2/2} - \lambda^2 e^{-\lambda^2/2} + (1+\lambda^2) e^{-\lambda^2/2} - 2\lambda \int_\lambda^\infty e^{-x^2/2} \, dx \\ &= 2\lambda \left(\frac{1}{\lambda} e^{-\lambda^2/2} - \int_\lambda^\infty e^{-x^2/2} \, dx \right) \geq 0 \end{aligned}$$

where the last inequality follows from the upper bound just proven. This shows that $f(\lambda) \geq 0$ for all $\lambda \geq 0$ and we are done. ☐

2. Laplace Transforms

It turns out to be useful to specialize characteristic functions for the case in which we have a measure that is supported on the positive orthant $\mathbb{R}_+^d$.

DEFINITION 7.24. Let μ be a probability measure on $\mathbb{R}_+^d$. Its *Laplace Transform* is denoted $\tilde{\mu}$ and is the function on $\mathbb{R}_+^d$ defined by

$$\tilde{\mu}(u) = \int e^{-\langle u,x \rangle} d\mu(x)$$

Next we observe that the behavior of the Laplace transform near zero corresponds to the behavior of the measure near infinity.

LEMMA 7.25. *Let μ be a probability measure on $\mathbb{R}_+^d$ and let $\mathbb{1} = (1, \dots, 1) \in \mathbb{R}_+^d$. Then for each $r > 0$ we have*

$$\mu\{|x| \geq r\} \leq 2(1 - \tilde{\mu}(\mathbb{1}/r))$$

PROOF. In order to see how simple the estimate is, first assume that $d = 1$. Observe that because e^{-ux} is a decreasing function of x for $u > 0$ we have $e^{-ux} \leq e^{-1} < 1/2$ for all $x \geq 1/u$ and $e^{-ux} \leq 1$ for all $x \geq 0$. Therefore for a fixed $r > 0$,

$$\begin{aligned}\tilde{\mu}(r) &= \int e^{-rx}\, d\mu(x) = \int \mathbf{1}_{[0,1/r)}(x) e^{-rx}\, d\mu(x) + \int \mathbf{1}_{[1/r,\infty)}(x) e^{-rx}\, d\mu(x) \\ &\leq \mu[0,1/r) + \frac{1}{2}\mu[1/r,\infty) = 1 - \frac{1}{2}\mu[1/r,\infty)\end{aligned}$$

To extend to the case of general d, we need a little bit more information. Note that minimum value of $\langle \mathbb{1}, x\rangle = \sum_{j=1}^d x_j$ on $\mathbb{R}_+^d \cap \{|x| \geq u\}$ is u (it occurs at the points $(0,\dots,0,u,0,\dots,0)$). TODO: Show this... Therefore we know that for all fixed $r \in \mathbb{R}_+$ we have $e^{-\langle r\mathbb{1},x\rangle} \leq e^{-1} < 1/2$ for all $x \in \mathbb{R}_+^d$ with $|x| \geq 1/r$. Now we can playback the same argument as the case $d = 1$:

$$\begin{aligned}\tilde{\mu}(r\cdot\mathbb{1}) &= \int e^{-r\langle \mathbb{1},x\rangle}\, d\mu(x) \\ &= \int \mathbf{1}_{|x|<1/r}(x) e^{-r\langle \mathbb{1},x\rangle}\, d\mu(x) + \int \mathbf{1}_{|x|\geq 1/r}(x) e^{-r\langle \mathbb{1},x\rangle}\, d\mu(x) \\ &\leq \mu\{|x| < 1/r\} + \frac{1}{2}\mu\{|x| \geq 1/r\} = 1 - \frac{1}{2}\mu\{|x| \geq 1/r\}\end{aligned}$$

□

LEMMA 7.26. *Let $\{\mu_\alpha\}$ be a family of probability measures on $\mathbb{R}_+^d$, then the family $\{\mu_\alpha\}$ is tight if and only if the family $\{\tilde{\mu}_\alpha\}$ is equicontinuous at 0. If this is true then $\{\tilde{\mu}_\alpha\}$ is uniformly equicontinuous on all of $\mathbb{R}_+^d$.*

PROOF. First we assume that the family $\{\mu_\alpha\}$ is equicontinuous and show tightness. To do this, note that if $\epsilon > 0$ is given, then by equicontinuity we can find $\delta > 0$ such that $1 - \tilde{\mu}_\alpha(u\mathbb{1}) < \epsilon/2$ for all $0 \leq u < \delta$ and all α. By Lemma 7.25 we get for every $r > 1/\delta$

$$\mu_\alpha\{|x| \geq r\} \leq 2(1 - \tilde{\mu_\alpha}(\mathbb{1}/r)) < \epsilon$$

and therefore tightness is proven.

Now assume that the family $\{\tilde{\mu}_\alpha\}$ is tight. For each α let ξ_α be a random vector with distribution μ_α. Using the elementary bound $|e^{-x} - e^{-y}| \leq |x - y| \wedge 1$ for $0 \leq x, y < \infty$ and Cauchy-Schwartz (Lemma 3.9) we see that for any $0 < \epsilon < 2$,

$$\begin{aligned}|\tilde{\mu}_\alpha(u) - \tilde{\mu}_\alpha(v)| &= \left|\mathbf{E}\left[e^{-\langle u,\xi_\alpha\rangle} - e^{-\langle v,\xi_\alpha\rangle}\right]\right| \\ &\leq \mathbf{E}\left[\left|e^{-\langle u,\xi_\alpha\rangle} - e^{-\langle v,\xi_\alpha\rangle}\right|\right] \\ &\leq \mathbf{E}\left[|\langle u - v, \xi_\alpha\rangle| \wedge 1\right] \\ &= \mathbf{E}\left[|\langle u - v, \xi_\alpha\rangle| \wedge 1; \langle u - v, \xi_\alpha\rangle < \epsilon/2\right] + \mathbf{E}\left[|\langle u - v, \xi_\alpha\rangle| \wedge 1; \langle u - v, \xi_\alpha\rangle \geq \epsilon/2\right] \\ &\leq \epsilon/2 + \mathbf{P}\{\langle u - v, \xi_\alpha\rangle \geq \epsilon/2\} \\ &\leq \epsilon/2 + \mathbf{P}\{|\xi_\alpha| \geq \frac{\epsilon}{2\,|u - v|}\}\end{aligned}$$

Thus by tightness for all $u, v \in \mathbb{R}_+^d$ with $|u - v|$ sufficiently small we have $|\tilde{\mu}_\alpha(u) - \tilde{\mu}_\alpha(v)| < \epsilon$ uniformly in α. Thus we see that the family $\{\tilde{\mu}_\alpha\}$ is uniformly equicontinuous on all of $\mathbb{R}_+^d$ and in particular at 0. □

The following result is analogous to the Glivenko-Levy Continuity Theorem 7.13 for characteristic functions. As with that result, here we point out that the assumption that μ_n converges to a probability measure is critical and we will return to the question of how to remove the assumption (using the notion of tightness) later on.

THEOREM 7.27 (Glivenko-Levy Continuity Theorem). *If $\mu, \mu_1, \mu_2, \dots$ are probability measures on $(\mathbb{R}_+^d, \mathcal{B}(\mathbb{R}_+^d))$, then μ_n converge weakly to μ if and only if $\tilde{\mu}_n(u)$ converge to $\tilde{\mu}(u)$ pointwise. Moreover if this is true then the convergence is uniform on bounded sets.*

PROOF. Since μ_n converge to μ weakly and $e^{-\langle u,x \rangle}$ is bounded and continuous we know that $\tilde{\mu}_n(u) \to \tilde{\mu}(u)$ pointwise. In fact, by Lemma 11.8 we know that the family μ_n is tight and therefore by Lemma 7.26 it is uniformly equicontinuous on $\mathbb{R}_+^d$. this convergence is uniform on every bounded set.

Now we assume that $\tilde{\mu}_n(u)$converges to $\tilde{\mu}(u)$ for every $u \in \mathbb{R}_+^d$. We now want to approximate general bounded continuous functions by functions $e^{-\langle u,x \rangle}$ in order derive weak convergence. To do this, we will consider $[0,\infty]^d$ which is a compact Hausdorff space and therefore amenable to application of the Stone Weierstrass Theorem 1.44. To use an approximation of functions on $[0,\infty]^d$ derive an effective approximation on $\mathbb{R}_+^d$ is going to require that we are able to control behavior of the measures μ_n at infinity and therefore we first show that μ_n is a tight family. Suppose $\epsilon > 0$ is given and use the continuity of $\tilde{\mu}(u)$ and the fact that $\tilde{\mu}(0) = 1$ to find $r_0 > 0$ such that $1 - \tilde{\mu}(\mathbb{1}/r_0) < \epsilon/2$. By pointwise convergence $\tilde{\mu}_n(\mathbb{1}/r_0) \to \tilde{\mu}(\mathbb{1}/r_0)$ we can find an $N > 0$ such that $1 - \tilde{\mu_n}(\mathbb{1}/r_0) < \epsilon$ for all $n \geq N$ and therefore by Lemma 7.25, $\mu_n(|x| \geq r) \leq 1 - \mu_n(\mathbb{1}/r_0) < \epsilon$. For each $n = 1, \dots, N-1$ by continuity of measure we can find $r_n > 0$ such that $\mu_n\{|x| \geq r_n\} < \epsilon$. Therefore taking the maximum $r = r_0 \vee r_1 \vee \cdots \vee r_{N-1}$ we get $\mu_n\{|x| \geq r\} < \epsilon$ for all n and we have shown μ_n is tight.

Suppose that $\epsilon > 0$ is given and pick $r > 0$ such that $\mu_n(\mathbb{R}_+^d \setminus B(0,r)) < \epsilon$ and $\mu(\mathbb{R}_+^d \setminus B(0,r)) < \epsilon$.

Having shown that μ_n is tight we return to the task of creating an approximation. Since $e^{-\langle u,x \rangle}$ has limits (either 0 or 1) as $x \to \infty$ we can extend each such function to a continuous function on $[0,\infty]^d$. Note also that the family $e^{-\langle k,x \rangle}$ for $k \in \mathbb{Z}_+^d$ contains the constant functions and separates points therefore we can apply the Stone Weierstrass Theorem to conclude that any continuous function on $[0,\infty]^d$ can be uniformly approximated by a linear combination of $e^{-\langle k,x \rangle}$.

Given a bounded continuous function $f : \mathbb{R}_+^d \to \mathbb{R}$ such that $|f(x)| \leq M$ we apply a continuous cutoff $1 - d(x, B(0,r)) \vee 0$ to create a function $\hat{f} : \mathbb{R}_+^d \to \mathbb{R}$ such that $\hat{f}(x) \leq M$ for all $x \in \mathbb{R}_+^d$, $f(x) = \hat{f}(x)$ for $x \in B(0,r)$ and $\hat{f}(x) = 0$ for $|x| > r + 1$. Note that for every n we have

$$\int \left| f - \hat{f} \right| \, d\mu_n = \int_{|x|<r} \left| f - \hat{f} \right| \, d\mu_n + \int_{|x|\geq r} \left| f - \hat{f} \right| \, d\mu_n < 2M\epsilon$$

and similarly for μ.

The function $\hat{f}$ can be extended by zero to define a continuous function on $[0,\infty]^d$ and therefore we can find some finite linear combination $g = \sum_k c_k e^{-\langle k,x \rangle}$

such that $\left|\hat{f}(x) - g(x)\right| < \epsilon$ for all $x \in [0,\infty]^d$ so *a fortiori* for all $x \in \mathbb{R}_+^d$. Therefore

$$\begin{aligned}
&\left|\int f\, d\mu_n - \int f\, d\mu\right| \\
&\leq \int \left|f - \hat{f}\right|\, d\mu_n + \int \left|\hat{f} - g\right|\, d\mu_n + \left|\int g\, d\mu_n - \int g\, d\mu\right| + \int \left|\hat{f} - g\right|\, d\mu + \int \left|f - \hat{f}\right|\, d\mu \\
&\leq 2M\epsilon + \epsilon + \left|\sum_k c_k(\mu_n(k) - \mu(k))\right| + \epsilon + 2M\epsilon
\end{aligned}$$

Now take the limit as $n \to \infty$ and use the fact that $\mu_n \to \mu$ pointwise (recall that the above sum over k is finite) and then let $\epsilon \to 0$. □

The Cramer-Wold device for Laplace transforms is a simple corollary.

Corollary 7.28 (Cramer Wold Device). *Let $\xi, \xi_1, \xi_2, \ldots$ be random vectors in $\mathbb{R}_+^d$. If $\langle c, \xi_n\rangle \xrightarrow{d} \langle c, \xi\rangle$ for all $c \in \mathbb{R}_+^d$ then it follows that $\xi_n \xrightarrow{d} \xi$.*

Proof. Since $\langle c, \xi_n\rangle \xrightarrow{d} \langle c, \xi\rangle$ we know that that $\mathbf{E}\left[e^{-\langle c, \xi_n\rangle}\right] \to \mathbf{E}\left[e^{-\langle c, \xi\rangle}\right]$ for all $c \in \mathbb{R}_+^d$ by definition of weak convergence and therefore by Theorem 7.27 we conclude $\xi_n \xrightarrow{d} \xi$. □

CHAPTER 8

Conditioning

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 8.1. 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 8.2 (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 3.11) to each of the terms on the right hand side and use the fact that $\frac{1}{p} + \frac{1}{q} = 1$ is equivalent to $p = (p-1)q$ to see

$$\int f(f+g)^{p-1}\, d\mu \leq \left(\int f^p \, d\mu \right)^{\frac{1}{p}} \left(\int (f+g)^{(p-1)q} \, d\mu \right)^{\frac{1}{q}} = \|f\|_p \, \|f+g\|_p^{p/q}$$

Applying this argument to the term $\int g(f+g)^{p-1}\, d\mu$ as well we get

$$\|f+g\|_p^p \leq (\|f\|_p + \|g\|_p) \cdot \|f+g\|_p^{p/q}$$

and dividing through by $\|f+g\|_p^{p/q}$ and using $p - \frac{p}{q} = 1$ we get $\|f+g\|_p \leq \|f\|_p + \|g\|_p$. $\square$

LEMMA 8.3. *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 $\left\|f_m - f_{n_j}\right\|_p \leq \frac{1}{2^j}$ for all $m > n_j$. In this way we get a subsequence f_{n_j} such that $\left\|f_{n_{j+1}} - f_{n_j}\right\|_p \leq \frac{1}{2^j}$ for all $j \in \mathbb{N}$. Now by applying Monotone Convergence and the triangle inequality we have

$$\begin{aligned}
\left\| \sum_{j=1}^{\infty} \left| f_{n_{j+1}} - f_{n_j} \right| \right\|_p &= \lim_{N \to \infty} \left\| \sum_{j=1}^{N} \left| f_{n_{j+1}} - f_{n_j} \right| \right\|_p \\
&\leq \lim_{N \to \infty} \sum_{j=1}^{N} \left\| f_{n_{j+1}} - f_{n_j} \right\|_p \\
&\leq \lim_{N \to \infty} \sum_{j=1}^{N} \frac{1}{2^j} < \infty
\end{aligned}$$

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 2.45) 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 $f_n \xrightarrow{L^p} f$. $\square$

We know that measurable functions can be approximated by simple functions (Lemma 2.18) with pointwise convergence. It is useful to extend this approximation to L^p spaces.

LEMMA 8.4. *Simple functions are dense in $L^p(\Omega, \mathcal{A}, \mu)$.*

PROOF. Pick a positive function $f \in L^p(\Omega, \mathcal{A}, \mu)$ and sequence of simple functions such that $0 \leq f_n \uparrow f$. Then it is also true that $f_n^p \uparrow f^p$ and Dominated Convergence tells us that $\lim_{n\to\infty} \|f_n\|_p = \|f\|_p$. By Lemma 5.57 we conclude that $f_n \xrightarrow{L^p} f$.

To finish the proof, take an arbitrary f and write it as $f = f_+ - f_-$. Now take positive simple functions $g_n \uparrow f_+$ and $h_n \uparrow f_-$ and use the triangle inequality to see that

$$\lim_{n\to\infty} \|f - (g_n - h_n)\|_p \leq \lim_{n\to\infty} (\|f_+ - g_n\|_p + \|f_- - h_n\|_p) = 0$$

□

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 8.3).

The following approximation result will be used only occasionaly.

LEMMA 8.5. $\cup_n L^p(\Omega, \mathcal{F}_n, \mu)$ *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 2.30) tells us that $\lim_{n\to\infty} \|\mathbf{1}_{A_n}\|_p = \|\mathbf{1}_A\|_p$ so Lemma 5.57 implies $\mathbf{1}_{A_n} \xrightarrow{L^p} \mathbf{1}_A$. Since V is closed we know $\mathbf{1}_A \in V$. □

LEMMA 8.6. *Let S be a metric space and let μ be a finite Borel measure on S. Then the space of bounded continuous functions is dense in $L^p(S, \mathcal{B}(S), \mu)$.*

PROOF. Note that the finiteness of μ guarantees that any bounded measurable function is also in $L^p(S, \mathcal{B}(S), \mu)$ so the proof will focus on establishing boundedness of functions involved and not concern itself with verifying p-integrability. Suppose we have $U \subset S$ an open set. Let $f_n(x) = (nd(x, U^c)) \wedge 1$. We know that $f_n(x)$ is increasing, bounded and continuous with $\lim_{n\to\infty} f_n(x) = \mathbf{1}_U(x)$ and therefore $f_n^p(x) \uparrow \mathbf{1}_U$ as well. By Monotone Convergence we have $\lim_{n\to\infty} \|f_n\|_p^p = \mu(U) = \|\mathbf{1}_U\|_p^p$ hence $f_n \xrightarrow{L^p} \mathbf{1}_U$. Now we extend to general Borel sets A by a monotone class argument. We claim that

$$\mathcal{C} = \{A \in \mathcal{B}(S) \mid \text{ there exist bounded continuous } f_n \text{ such that } f_n \xrightarrow{L^p} \mathbf{1}_A\}$$

is a λ-system. Supposing $A \subset B$ with $A, B \in \mathcal{C}$ we get bounded continuous f_n such that $f_n \xrightarrow{L^p} \mathbf{1}_A$ and bounded continuous g_n such that $g_n \xrightarrow{L^p} \mathbf{1}_B$ by Lemma 5.58.

Then

$$\lim_{n\to\infty} \left\|\mathbf{1}_{B\setminus A} - (g_n - f_n)\right\|_p \leq \lim_{n\to\infty} \left\|\mathbf{1}_B - g_n\right\|_p + \lim_{n\to\infty} \left\|\mathbf{1}_A - f_n\right\|_p = 0$$

and therefore $B \setminus A \in \mathcal{C}$. If $A_1 \subset A_2 \subset \cdots$ with $A_n \in \mathcal{C}$ then

$$\lim_{n\to\infty} \left\|\mathbf{1}_{A_n}\right\|_p = \lim_{n\to\infty} \mu(A_n)^{1/p} = \mu(\cup_{n=1}^\infty A_n)^{1/p} = \left\|\mathbf{1}_{\cup_{n=1}^\infty A_n}\right\|_p$$

Now for each A_n there exists a sequence bounded continuous $f_{n,m}$ with $\lim_{m\to\infty} \left\|\mathbf{1}_{A_n} - f_{n,m}\right\|_p = 0$. Now we can find a subsequence f_{n,m_n} such that $\lim_{n\to\infty} \left\|\mathbf{1}_{\cup_{n=1}^\infty A_n} - f_{n,m_n}\right\|_p = 0$ which shows $\cup_{n=1}^\infty A_n \in \mathcal{C}$. Now as open sets are clearly a π-system the π-λ Theorem 2.27 shows that $\mathcal{B}(S) \subset \mathcal{C}$. Now for any simple function $f = \sum_{j=1}^m c_j \mathbf{1}_{A_j}$ we can find $f_{j,n}$ such that $\lim_{n\to\infty} \left\|\mathbf{1}_{A_j} - f_{j,n}\right\|_p = 0$ and by the triangle inequality

$$\lim_{n\to\infty} \left\| f - \sum_{j=1}^m c_j f_{j,n} \right\|_p \leq \lim_{n\to\infty} \sum_{j=1}^m |c_j| \left\|\mathbf{1}_{A_j} - f_{j,n}\right\|_p = 0$$

and the fact that each $\sum_{j=1}^m c_j f_{j,n}$ is bounded and continuous is clear.

The last step is to use the fact that simple functions are dense in $L^p(S, \mathcal{B}(S), \mu)$ (Lemma 8.4). □

TODO: Pretty sure the above result will have an extension to σ-finite case.

TODO: Develop inner product and projection for L^2 spaces.

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 2.23 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 2.110) 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

$$\nu(A) = \mathbf{E}\left[\xi \mathbf{1}_A\right] = \mathbf{E}\left[\frac{d\nu}{d\mu}\mathbf{1}_A\right]$$

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 8.7 (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

$$\mathbf{E}\left[\xi; A\right] =< \xi, \mathbf{1}_A >=< \mathbf{E}^{\mathcal{F}}\xi, \mathbf{1}_A >= \mathbf{E}\left[\mathbf{E}^{\mathcal{F}}\xi; A\right]$$

If we define $A = \{\mathbf{E}^{\mathcal{F}}\xi \geq 0\}$ the above implies

$$\begin{aligned}
\mathbf{E}\left[\left|\mathbf{E}^{\mathcal{F}}\xi\right|\right] &= \mathbf{E}\left[\mathbf{E}^{\mathcal{F}}\xi; A\right] - \mathbf{E}\left[\mathbf{E}^{\mathcal{F}}\xi; A^c\right] && \text{by linearity of expectation} \\
&= \mathbf{E}\left[\xi; A\right] - \mathbf{E}\left[\xi; A^c\right] && \text{by (i)} \\
&\leq \mathbf{E}\left[|\xi|; A\right] + \mathbf{E}\left[|\xi|; A^c\right] && \text{since } \xi \leq |\xi| \text{ and } -\xi \leq |\xi| \\
&= \mathbf{E}\left[|\xi|\right] && \text{by linearity of expecation}
\end{aligned}$$

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 2.50 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 2.50 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 5.6 this implies that the converges is in probability and by Lemma 5.10 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 1.15 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

$$\mathbf{E}\left[\xi\mathbf{E}^{\mathcal{F}}\eta; A\right] = \mathbf{E}[\xi\eta; A]$$

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 8.8. *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,

$$\mathbf{E}[\xi; \cup_n A_n] = \lim_{n\to\infty} \mathbf{E}[\xi; A_n] = \lim_{n\to\infty} \mathbf{E}[\eta; A_n] = \mathbf{E}[\eta; \cup_n A_n]$$

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 2.27) to get the result. □

Occasionally it can be useful to extend the defining property of conditional expectation beyond indicator functions.

LEMMA 8.9. *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 $\xi \cdot \eta_n$,

$$\begin{aligned}
\mathbf{E}[\xi \cdot \eta] &= \lim_{n\to\infty} \mathbf{E}[\xi \cdot \eta_n] && \text{by Dominated Convergence} \\
&= \lim_{n\to\infty} \mathbf{E}\left[\mathbf{E}^{\mathcal{F}}\xi \cdot \eta_n\right] \\
&= \mathbf{E}\left[\mathbf{E}^{\mathcal{F}}\xi \cdot \eta\right] && \text{by Dominated Convergence}
\end{aligned}$$

For general integrable η split into its positive and negative parts $\eta = \eta_+ - \eta_-$ and use linearity of expectation. □

It is important to extend our basic limit theorems of integration theory to conditional expectations. We have already proven the analogue of montone convergence. Here we address the other cases of importance. The proofs are essentially identical to the non-conditional cases.

LEMMA 8.10 (Fatou's Lemma for Conditional Expectation). *Let $\xi_1, \xi_2, \ldots$ be positive random variables then*

$$\boldsymbol{E}\left[\liminf_{n\to\infty} \xi_n \mid \mathcal{F}\right] \leq \liminf_{n\to\infty} \boldsymbol{E}[\xi_n \mid \mathcal{F}]$$

PROOF. The proof is essentially identical to the case for ordinary expectations (Theorem 2.45) since we have montone convergence and monotonicity of conditional expectation

$$\begin{aligned}\mathbf{E}\left[\liminf_{n\to\infty} \xi_n \mid \mathcal{F}\right] &= \lim_{n\to\infty} \mathbf{E}\left[\inf_{k\geq n} \xi_k \mid \mathcal{F}\right] \\ &\leq \lim_{n\to\infty} \inf_{k\geq n} \mathbf{E}\left[\xi_k \mid \mathcal{F}\right] \\ &= \liminf_{n\to\infty} \mathbf{E}\left[\xi_n \mid \mathcal{F}\right]\end{aligned}$$

where all of the equalities and inequalities are taken to be almost sure. □

LEMMA 8.11 (Dominated Convergence for Conditional Expectation). *Let $\xi, \xi_1, \xi_2, \ldots$ be random variables such that $\xi_n \stackrel{a.s.}{\to} \xi$ and η be a positive random variables such that $|\xi_n| \leq \eta$, $\boldsymbol{E}[\eta] < \infty$ then*

$$\boldsymbol{E}[\xi \mid \mathcal{F}] = \lim_{n\to\infty} \boldsymbol{E}[\xi_n \mid \mathcal{F}] \ \ a.s.$$

PROOF. Note that $\eta \pm \xi_n \geq 0$ so we may apply Fatou's Lemma 8.10 to both sequences.

$$\begin{aligned}\mathbf{E}\left[\eta \mid \mathcal{F}\right] \pm \mathbf{E}\left[\xi \mid \mathcal{F}\right] &= \mathbf{E}\left[\eta \pm \xi \mid \mathcal{F}\right] \\ &= \mathbf{E}\left[\lim_{n\to\infty} \eta \pm \xi_n \mid \mathcal{F}\right] \\ &\leq \liminf_{n\to\infty} \mathbf{E}\left[\eta \pm \xi_n \mid \mathcal{F}\right] \\ &= \mathbf{E}\left[\eta \mid \mathcal{F}\right] + \liminf_{n\to\infty} \mathbf{E}\left[\pm\xi_n \mid \mathcal{F}\right]\end{aligned}$$

where all of the comparisons are in an almost sure sense. Now by integrability of η and the chain rule of conditional expectation we know that $\mathbf{E}\left[\mathbf{E}\left[\eta \mid \mathcal{F}\right]\right] = \mathbf{E}\left[\eta\right] < \infty$ and therefore $\mathbf{E}\left[\eta \mid \mathcal{F}\right] < \infty$ a.s. Thus it is permissible to subtract $\mathbf{E}\left[\eta \mid \mathcal{F}\right]$ from both sides of the inequality above and deduce the pair of inequalities

$$\pm\mathbf{E}\left[\xi \mid \mathcal{F}\right] \leq \liminf_{n\to\infty} \mathbf{E}\left[\pm\xi_n \mid \mathcal{F}\right] \ \text{a.s.}$$

Now using this pair of inequalities

$$\limsup_{n\to\infty} \mathbf{E}\left[\xi_n \mid \mathcal{F}\right] = -\liminf_{n\to\infty} \mathbf{E}\left[-\xi_n \mid \mathcal{F}\right] \leq \mathbf{E}\left[\xi \mid \mathcal{F}\right] \leq \liminf_{n\to\infty} \mathbf{E}\left[\xi_n \mid \mathcal{F}\right] \ \text{a.s.}$$

which shows us that $\mathbf{E}\left[\xi \mid \mathcal{F}\right] = \lim_{n\to\infty} \mathbf{E}\left[\xi_n \mid \mathcal{F}\right]$ a.s. □

LEMMA 8.12. *Suppose that ξ_t for $t \in T$ is a uniformly integrable family of random variables and then $\boldsymbol{E}[\xi_t \mid \mathcal{F}]$ is uniformly integrable. Moreover if ξ is a random variable and ξ_n is a uniformly integrable family of random variables such that $\xi_n \overset{a.s.}{\to} \xi$ then $\boldsymbol{E}[\xi_n \mid \mathcal{F}] \overset{a.s.}{\to} \boldsymbol{E}[\xi \mid \mathcal{F}]$.*

PROOF. To see uniform integrability of $\mathbf{E}[\xi_t \mid \mathcal{F}]$ we use Lemma 5.52. Since conditional expectation is an L^1 contraction, the L^1 boundedness of $\mathbf{E}[\xi_t \mid \mathcal{F}]$ follows from the L^1 boundedness of ξ_t. Now if we let A be measurable and pick $R > 0$, then by using monotonicity and the tower property of conditional expectation

$$\begin{aligned}\mathbf{E}[|\mathbf{E}[\xi_t \mid \mathcal{F}]|; A] &\leq \mathbf{E}[\mathbf{E}[|\xi_t| \mid \mathcal{F}]; A] \\ &= \mathbf{E}[\mathbf{E}[|\xi_t|; |\xi_t| \leq R \mid \mathcal{F}]; A] + \mathbf{E}[\mathbf{E}[|\xi_t|; |\xi_t| > R \mid \mathcal{F}]; A] \\ &\leq R\mathbf{P}\{A\} + \mathbf{E}[|\xi_t|; |\xi_t| > R]\end{aligned}$$

and therefore taking $\sup_t$, $\lim_{\mathbf{P}\{A\}\to 0}$ and $\lim_{R\to\infty}$ and using the uniform integrability of ξ_t we get uniform integrability of $\mathbf{E}[\xi_t \mid \mathcal{F}]$.

If we assume that $\xi_1, \xi_2, \dots$ are uniformly integrable and $\xi_n \overset{a.s.}{\to} \xi$ then picking a measurable A and using the first part of this lemma and Lemma 5.54 we know that both families $\xi_1 \mathbf{1}_A, \xi_2 \mathbf{1}_A, \dots$ and $\mathbf{E}[\xi_1 \mid \mathcal{F}]\mathbf{1}_A, \mathbf{E}[\xi_2 \mid \mathcal{F}]\mathbf{1}_A, \dots$ are uniformly integrable. So know using using Lemma 5.58 to justify exchanging limits and expectations we get

$$\mathbf{E}[\xi; A] = \lim_{n\to\infty} \mathbf{E}[\xi_n; A] = \lim_{n\to\infty} \mathbf{E}[\mathbf{E}[\xi_n \mid \mathcal{F}]; A] = \mathbf{E}\left[\lim_{n\to\infty} \mathbf{E}[\xi_n \mid \mathcal{F}]; A\right]$$

Since $\lim_{n\to\infty} \mathbf{E}[\xi_n \mid \mathcal{F}]$ is $\mathcal{F}$-measurable (Lemma 2.14) we know that $\mathbf{E}[\xi \mid \mathcal{F}] = \lim_{n\to\infty} \mathbf{E}[\xi_n \mid \mathcal{F}]$ by the defining property of conditional 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 8.13. 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 8.14. *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*

$$\boldsymbol{E}[\xi \mid \mathcal{F}] = \boldsymbol{E}[\eta \mid \mathcal{G}] \quad \textit{a.s. on } A$$

PROOF. We first claim that if $B \subset A$ and $B \in \mathcal{F} \vee \mathcal{G}$ then in fact $B \in \mathcal{F} \cap \mathcal{G}$. To see the claim, $A \cap \mathcal{F} \vee \mathcal{G}$ is a σ-algebra of subsets of A generated by $A \cap \mathcal{F} = A \cap \mathcal{G} = A \cap \mathcal{F} \cap \mathcal{G}$ hence $A \cap \mathcal{F} \vee \mathcal{G} \subset A \cap \mathcal{F} \cap \mathcal{G}$. The opposite inclusion is trivial.

Consider the set $\{\mathbf{E}[\xi \mid \mathcal{F}] > \mathbf{E}[\eta \mid \mathcal{G}]\} \cap A$ and observe by the above claim that it is contained in $\mathcal{F} \cap \mathcal{G}$. Therefore by monotonicity of conditional expectation, the averaging property of conditional expectation and the fact that $\xi = \eta$ almost surely on A we have

$$\begin{aligned}
0 &\leq \mathbf{E}[(\mathbf{E}[\xi \mid \mathcal{F}] - \mathbf{E}[\eta \mid \mathcal{G}]); \{\mathbf{E}[\xi \mid \mathcal{F}] > \mathbf{E}[\eta \mid \mathcal{G}]\} \cap A] \\
&= \mathbf{E}[\mathbf{E}[\xi \mid \mathcal{F}]; \{\mathbf{E}[\xi \mid \mathcal{F}] > \mathbf{E}[\eta \mid \mathcal{G}]\} \cap A] - \mathbf{E}[\mathbf{E}[\eta \mid \mathcal{G}]; \{\mathbf{E}[\xi \mid \mathcal{F}] > \mathbf{E}[\eta \mid \mathcal{G}]\} \cap A] \\
&= \mathbf{E}[\xi; \{\mathbf{E}[\xi \mid \mathcal{F}] > \mathbf{E}[\eta \mid \mathcal{G}]\} \cap A] - \mathbf{E}[\eta; \{\mathbf{E}[\xi \mid \mathcal{F}] > \mathbf{E}[\eta \mid \mathcal{G}]\} \cap A] \\
&= 0
\end{aligned}$$

which shows $\mathbf{E}[\xi \mid \mathcal{F}] \leq \mathbf{E}[\eta \mid \mathcal{G}]$ almost surely on A. Switching the roles of $\mathcal{F}$ and $\mathcal{G}$ yields the opposite inequality and the result follows. □

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 8.15. *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 8.16. *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 2.23 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.

LEMMA 8.17. *Let $\mathcal{F}$ be a σ-algebra and let ξ be integrable, then $\boldsymbol{E}[\xi \mid \mathcal{F}] = \boldsymbol{E}\left[\xi \mid \overline{\mathcal{F}}\right]$ a.s.*

PROOF. Let $A \in \overline{\mathcal{F}}$. We know from Lemma ??? that there exist $A_\pm \in \mathcal{F}$ such that $A_- \subset A \subset A_+$ and $\mathbf{P}\{A_+ \setminus A_-\} = 0$. It is clear that for any $\xi \geq 0$ we have

$$\mathbf{E}[\xi; A_-] \leq \mathbf{E}[\xi; A] \leq \mathbf{E}[\xi; A_+] = \mathbf{E}[\xi; A_-] + \mathbf{E}[\xi; A_+ \setminus A_-] = \mathbf{E}[\xi; A_-]$$

and therefore $\mathbf{E}[\xi; A_-] = \mathbf{E}[\xi; A] = \mathbf{E}[\xi; A_+]$. By linearity this clearly extends to integrable ξ. Therefore we get

$$\mathbf{E}[\xi; A] = \mathbf{E}[\xi; A_-] = \mathbf{E}[\mathbf{E}[\xi \mid \mathcal{F}]; A_-] = \mathbf{E}[\mathbf{E}[\xi \mid \mathcal{F}]; A]$$

which gives the result. □

3. Conditional Independence

DEFINITION 8.18. 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 4.13

LEMMA 8.19. *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 4.13. □

LEMMA 8.20. *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 8.8 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 $H \in \mathcal{H}$. $\square$

LEMMA 8.21. *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 8.20 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 8.20, 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 8.19 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 8.20 and the pull out rule to get

$$\begin{aligned}
\mathbf{P}\{H \mid \mathcal{G}, \mathcal{F}_1, \dots, \mathcal{F}_n\} &= \mathbf{E}\left[\mathbf{P}\{H \mid \mathcal{G}, \mathcal{F}_1, \mathcal{F}_2, \dots\} \mid \mathcal{G}, \mathcal{F}_1, \dots, \mathcal{F}_n\right] \\
&= \mathbf{E}\left[\mathbf{P}\{H \mid \mathcal{G}\} \mid \mathcal{G}, \mathcal{F}_1, \dots, \mathcal{F}_n\right] \\
&= \mathbf{P}\{H \mid \mathcal{G}\}
\end{aligned}$$

so in particular, for all $n \geq 0$,

$$\mathbf{P}\{H \mid \mathcal{G}, \mathcal{F}_1, \dots, \mathcal{F}_n\} = \mathbf{P}\{H \mid \mathcal{G}, \mathcal{F}_1, \dots, \mathcal{F}_{n+1}\}$$

Another application of Lemma 8.20 shows that $\mathcal{H} \perp\!\!\!\perp_{(\mathcal{G}, \mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_n)} \mathcal{F}_{n+1}$ for all $n \geq 0$. $\square$

LEMMA 8.22. *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 8.20, 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 8.20 tells us that $\mathcal{F} \perp\!\!\!\perp_{\mathcal{A}, \mathcal{G}} \mathcal{H}$. $\square$

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.

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 8.23. 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 8.24. 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 8.25. 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 each equipped with the power set σ-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, one 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*. Note that because we are using power set σ-algebras the measurability conditions in the definition of a kernel are trivially satisfied.

Many mappings on the space of probability measures are measurable.

LEMMA 8.26. *Let $(S, \mathcal{S})$ and $(T, \mathcal{T})$ be measurable spaces and let $f : S \to T$ be a measurable function then the following mappings are measurable:*

(i) $\mu \mapsto \mu_A$ *for every* $A \in \mathcal{S}$.
(ii) $\mu \mapsto \int f \, d\mu$ *for every measurable function* $f : S \to \mathbb{R}$.
(iii) $(\mu, \nu) \mapsto \mu \otimes \nu$.
(iii) $\mu \mapsto \mu \circ f^{-1}$.

PROOF. To see (i) simply note that for every $B \in \mathcal{S}$ and $C \in \mathcal{B}(\mathbb{R})$, we have $\{\mu \mid \mu_A(B) \in C\} = \{\mu \mid \mu(A \cap B) \in C\}$ which is measurable since $A \cap B \in \mathcal{S}$.

For (ii) note that for $f = \mathbf{1}_A$ an indicator function we have $\int f \, d\mu = \mu(A)$ is a measurable function of μ be definition of the σ-algebra on $\mathcal{P}(S)$. By Lemma 2.19 we then see that $\int f \, d\mu$ is measurable for simple functions. For positive functions f we take an increasing sequence of simple functions $f_n \uparrow f$ so that $\int f \, d\mu = \lim_{n \to \infty} \int f_n \, d\mu$ which is measurable by Lemma 2.14. For general f we write $f = f_+ - f_-$ and use Lemma 2.19 again.

To see (iii) we first note that for $A \in \mathcal{S}$ and $B \in \mathcal{T}$ we have $(\mu \otimes \nu)(A \times B) = \mu(A)\nu(B)$ which is a measurable function of (μ, ν) by definition of the σ-algebras on $\mathcal{P}(S)$ and $\mathcal{P}(T)$, definition of the product σ-algebra and continuity (hence Borel measurability) of multiplication on $\mathbb{R}$. Now we extend to general $A \in \mathcal{S} \otimes \mathcal{T}$ by a monotone class argument. Let $\mathcal{C} = \{A \in \mathcal{S} \otimes \mathcal{T} \mid \mu \otimes \nu(A)$ is a measurable function of $(\mu, \nu)\}$. We claim that $\mathcal{C}$ is a λ-system. If $A, B \in \mathcal{C}$ such that $A \subset B$ then $(\mu \otimes \nu)(B \setminus A) = (\mu \otimes \nu)(B) - (\mu \otimes \nu)(A)$ which is measurable by Lemma 2.19. If $A_1 \subset A_2 \subset \cdots$ with $A_n \in \mathcal{C}$ for $n = 1, 2, \ldots$ then by continuity of measure (Lemma 2.30) we have $(\mu \otimes \nu)(A) = \lim_{n \to \infty} (\mu \otimes \nu)(A_n)$ which is measurable by Lemma 2.14. Since the sets of the form $A \times B$ are a π-system generating $\mathcal{S} \otimes \mathcal{T}$ we can apply the π-λ Theorem (Theorem 2.27) to conclude $\mathcal{S} \otimes \mathcal{T} \subset \mathcal{C}$ and the claim is verified. By the result of the claim we now know that for every $C \in \mathcal{B}(\mathbb{R})$ and every $A \in \mathcal{S} \otimes \mathcal{T}$ we have

$$\{(\mu, \nu) \in \mathcal{P}(S) \times \mathcal{P}(T) \mid (\mu \otimes \nu)(A) \in C\} = \otimes^{-1}\{\mu \in \mathcal{P}(S \times T) \mid \mu(A) \in C\}$$

is a measurable subset of $\mathcal{P}(S) \times \mathcal{P}(T)$. Since sets of the form $\{\mu \in \mathcal{P}(S \times T) \mid \mu(A) \in C\}$ generate the σ-algebra on $\mathcal{P}(S \times T)$ we have that $\otimes$ is measurable (Lemma 2.12).

To see (iv), we know that $\mu \circ f^{-1}$ is indeed a probability measure (Lemma 2.53). To see the measurability of the pushforward, suppose $A \in \mathcal{T}$ and $B \in \mathcal{B}([0, 1])$ and note that

$$\{\mu \in \mathcal{P}(S) \mid \mu \circ f^{-1}(A) \in B\} = \{\mu \in \mathcal{P}(S) \mid \mu(f^{-1}(A)) \in B\}$$

which is measurable since $f^{-1}(A) \in \mathcal{S}$. Now the general result follows from Lemma 2.12. □

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 8.27. *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 2.12.

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 2.13).

The implication (i) implies (iii) is immediate. If we assume (iii) then we derive (i) by a monotone class argument. By Theorem 2.27 it suffices to show that $\mathcal{C} = \{A \mid \mu(s, A) : S \to [0, 1]$ 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 2.30) applied pointwise in s, we see $\mu(s, \cup_n A_n) = \lim_n \mu(s, A_n)$ which shows measurability by Lemma 2.14. □

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 write $\int f(x) \, d\mu(x)$. In a world with kernels the latter notation is unfortunate as it becomes difficult to construe whether the x dependence indicated for the measure means an integration variable or whether it indicates 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 $\int f(x) \, \mu(s, dx)$ which should be interpreted as the integral of $f(x)$ against the measure $\mu(s)$ for some particular value of s. The reader may already be wondering whether an expression such as this is a measurable function of the parameter s; we will state and prove a slightly more general fact below.

There is a useful generalization of the product measure construction involving kernels. It is a type of "twisted" product construction.

Definition 8.28. 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

$$\mu \otimes \nu(s, A) = \iint \mathbf{1}_A(t, u) \, d\nu(s, t, du) \, d\mu(s, dt)$$

We also have the special restriction $\mu\nu : S \times \mathcal{U} \to [0, 1]$ defined by $\mu\nu(s, B) = \mu \otimes \nu(s, T \times B)$.

The fact that this construction defines a probability kernel is the content of the next Lemma.

Lemma 8.29. *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 $g : S \times T \to U$ 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 $\mathcal{C} = \{C \in \mathcal{S} \otimes \mathcal{T} \mid \int \mathbf{1}_C(s,t)\, d\mu(s,dt) \text{ is measurable}\}$ 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 2.14). Now an application of Theorem 2.27 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 2.14. 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. $\square$

Example 8.30. This continues Example 8.25. 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

$$\begin{aligned}
(\mu \otimes \nu)_{s,t,u} = (\mu \otimes \nu)(s, \{(t,u)\}) &= \iint \mathbf{1}_{\{(t,u)\}}(x,y)\, d\nu(s,x,dy)\, d\mu(s,dx) \\
&= \int \mathbf{1}_{\{(t)\}}(x) \nu(s,x,\{u)\, d\mu(s,dx) \\
&= \mu_{s,t}\nu_{s,t,u}
\end{aligned}$$

There is a particularly important special case of this special case. Consider the case of $\mu : S \to \mathcal{P}(T)$ and $\nu : T \to \mathcal{P}(T)$. We can apply the kernel product $\mu \otimes \mu : S \to T \times U$ to sets of the form $T \times \{u\}$ for $u \in U$ and we get

$$\begin{aligned}
(\mu\nu)(s, \{u\}) &= (\mu \otimes \nu)(s, T \times \{u\}) \\
&= \sum_{t \in T} (\mu \otimes \nu)(s, \{(t,u)\}) \\
&= \sum_{t \in T} \mu_{s,t}\nu_{s,t,u}
\end{aligned}$$

so the product $\mu\nu$ is simply the matrix product.

It shall also be useful to show that we can construct a parameterized family of random elements whose distributions are given by a specified kernel.

LEMMA 8.31. *Let* $(S, \mathcal{S})$ *and* $(T, \mathcal{T})$ *be measurable spaces with* T *a Borel space and let* $\mu : T \times \mathcal{S} \to [0,1]$ *be a probability kernel. There exists a measurable function* $G : S \times [0,1] \to T$ *such if* ϑ *is a* $U(0,1)$ *random variable then* $G(s,\vartheta)$ *has distribution* $\mu(s,\cdot)$ *for all* $s \in S$.

PROOF. First assume that $T = [0,1]$; we replay the argument of Lemma 2.112 pointwise in S. Let

$$G(s,t) = \sup\{u \in [0,1] \mid \mu(s,[0,u]) < t\} \text{ for } s \in S \text{ and } t \in [0,1]$$

We claim that $G(s,t)$ is $\mathcal{S} \otimes \mathcal{B}([0,1])$-measurable. First note that if we define

$$G^{\mathbb{Q}}(s,t) = \sup\{u \in [0,1] \cap \mathbb{Q} \mid \mu(s,[0,u]) < t\} \text{ for } s \in S \text{ and } t \in [0,1]$$

then in fact $G^{\mathbb{Q}} = G$. To see this, it is clear that $G^{\mathbb{Q}} \leq G$. For the other inequality, let $s \in S$ and $t \in [0,1]$ be given and pick an arbitrary $\epsilon > 0$; let $u \in [0,1]$ be such that $G(s,t) - \epsilon < \mu(s,[0,u])$. Now take a sequence of $q_n \in [0,1] \cap \mathbb{Q}$ such that $q_n \downarrow u$ and use continuity of measure to conclude that $\lim_{n\to\infty} \mu(s,[0,q_n]) = \mu(s,[0,u]) < t$ so there is a $q \in [0,1] \cap \mathbb{Q}$ such that $q \geq x$ and $\mu(s,[0,q]) < t$. This proves that $G^{\mathbb{Q}}(s,t) \geq G(s,t) - \epsilon$ and since $\epsilon > 0$ was arbitrary we have the desired equality. Now for any $y \in [0,1]$ we can write

$$\{(s,t) \mid G(s,t) \leq y\} = \cap_{\substack{q \leq y \\ q \in \mathbb{Q}}} \{(s,t) \mid \mu(s,[0,q]) \leq y\}$$

and each $\{(s,t) \mid \mu(s,[0,q]) \leq y\}$ is measurable for fixed q since $\mu(s,[0,q])$ is a measurable function of s (e.g. observe $\{(s,t) \mid \mu(s,[0,q]) \leq y\} = \{(s,t) \mid t - \mu(s,[0,q]) \geq 0\}$ and use the measurability of the function $g(s,t) = t - \mu(s,[0,q])$).

Now note that

$$\mathbf{P}\{G(s,\vartheta) \leq u\} = \mathbf{P}\{\vartheta \leq \mu(s,[0,u])\} = \mu(s,[0,u])$$

and therefore $G(s,\vartheta) \overset{d}{=} \mu(s,\cdot)$ by Lemma 3.4.

To extend to general Borel spaces T, first suppose that $T \in \mathcal{B}([0,1])$. Given a probability kernel $\mu : S \times \mathcal{T} \to [0,1]$ we define $\tilde{\mu} : S \times \mathcal{B}([0,1]) \to [0,1]$ by $\tilde{\mu}(s,A) = \mu(s, A \cap T)$. It is clear that $\tilde{\mu}(s,\cdot)$ is a probability measure for all $s \in S$ and furthermore since $A \cap T \in \mathcal{T}$ we know that $\tilde{\mu}(s,A)$ is $\mathcal{S}$-measurable for every $A \in \mathcal{B}([0,1])$ hence $\tilde{\mu}$ is a probability kernel (Lemma 8.26 and Lemma 8.27). Note that by construction for all $s \in S$ we have $\tilde{\mu}(s,T^c) = \mu(s, T \cap T^c) = 0$. Applying the result for $[0,1]$ we get a measurable $\tilde{G} : S \times [0,1] \to [0,1]$ such that $\mathbf{P}\{\tilde{G}(s,\vartheta) \in A\} = \mu(s,A)$. Pick an arbitrary point $t_0 \in T$ and define

$$G(s,t) = \mathbf{1}_{\tilde{G}^{-1}(T)}(s,t) G(s,t) + t_0 \mathbf{1}_{\tilde{G}^{-1}(T^c)}(s,t)$$

$G(s,t)$ is a measurable function $G : S \times [0,1] \to T$. Furthermore for all $sinS$ and $A \in \mathcal{T}$,

$$\mathbf{P}\{G(s,\vartheta) \in A\} = \begin{cases} \mathbf{P}\{\tilde{G}(s,\vartheta) \in A\} & \text{if } t_0 \notin A \\ \mathbf{P}\{\tilde{G}(s,\vartheta) \in A\} + \mathbf{P}\{\tilde{G}(s,t) \in T^c\} & \text{if } t_0 \in A \end{cases}$$
$$= \mathbf{P}\{\tilde{G}(s,\vartheta) \in A\} = \tilde{\mu}(s,A) = \mu(s, A \cap T) = \mu(s,A)$$

proving the result for Borel subsets of $[0,1]$.

Lastly suppose T is Borel isomorphic to a Borel subset of $[0,1]$ and let $\mu : S \times \mathcal{T} \to [0,1]$ be a probability kernel. If $A \in \mathcal{B}([0,1])$ and $g : T \to A$ is a Borel isomorphism then note that $\mu \circ g^{-1}(s, A) = \mu(s, g^{-1}(A))$ defines a probability kernel $\mu \circ g^{-1} : S \times A \cap \mathcal{B}([0,1]) \to [0,1]$. It is clear that if select $G : S \times [0,1] \to A$ as above then $G \circ g^{-1} : S \times [0,1] \to T$ is measurable and

$$\begin{aligned}\mathbf{P}\{g^{-1}(G(s,\vartheta)) \in B\} &= \mathbf{P}\{G(s,\vartheta) \in g(B)\} \\ &= \mu \circ g^{-1}(s, g(B)) = \mu(s, B)\end{aligned}$$

so $G \circ g^{-1}$ proves the result. □

Lemma 8.32. *Let $(S, \mathcal{S})$ and $(T, \mathcal{T})$ be measurable spaces and let $f : S \to T$ be a Borel isomorphism then $f^{-1} : \mathcal{T} \to \mathcal{S}$ is a bijection.*

Proof. Since a Borel isomorphism is a bijection we know that $f^{-1} : 2^T \to 2^S$ is a bijection (Lemma 2.9). By measurability of f we know that $f^{-1}(\mathcal{T}) \subset \mathcal{S}$. Moreover for any $A \in \mathcal{S}$ by measurability of we know that $f^{-1}\ \{t \in T \mid f(t) \in A\} \in \mathcal{T}$ and clearly $f^{-1}\{t \in T \mid f(t) \in A\} = \{s \in S \mid f(f^{-1}(s)) \in A\} = A$ since f is a bijection. □

Lemma 8.33. *Let $(S, \mathcal{S})$ and $(T, \mathcal{T})$ be measurable spaces and let $f : S \to T$ be a Borel isomorphism, then the map $f_* : \mathcal{P}(S) \to \mathcal{P}(T)$ given by $f_*(\mu)(A) = \mu(f^{-1}(A))$ is a Borel isomorphism with $(f_*)^{-1} = (f^{-1})_*$.*

Proof. We first show f_* is measurable. Let $F \in \mathcal{T}$ and let $G \in \mathcal{B}([0,1])$ and consider the measurable set $\{\mu \mid \mu(F) \subset G\} \subset \mathcal{P}(T)$. Since f is measurable we know that $f^{-1}(F) \in \mathcal{S}$ and therefore

$$f_*^{-1}\{\mu \mid \mu(F) \subset G\} = \{\mu \mid f_*\mu(F) \subset G\} = \{\mu \mid \mu(f^{-1}(F)) \subset G\}$$

is measurable in $\mathcal{P}(S)$. Since the σ-algebra on $\mathcal{P}(T)$ is generated by sets of the form $\{\mu \mid \mu(F) \subset G\}$, measurability of f_* follows from Lemma 2.12.

Since f is a Borel isomorphism, we know $(f^{-1})_* : \mathcal{P}(T) \to \mathcal{P}(S)$ is well defined and measurable and we can compute that for all $A \in \mathcal{S}$ and $\mu \in \mathcal{P}(S)$ we have

$$(f^{-1})_* f_* \mu(A) = f_*\mu(f(A)) = \mu(f^{-1}(f(A))) = \mu(A)$$

so that $(f^{-1})_* \circ f_* = id$. By symmetry we have $f_* \circ (f^{-1})_* = id$ and the result is shown. □

Theorem 8.34. *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.

Now we show how to handle the case of general Borel S. Let A be a Borel subset of $\mathbb{R}$ and let $j : S \to A$ be a Borel isomorphism. We apply the result just proven to $j \circ \xi : \Omega \to A$ and get the existence of a probability kernel $\tilde{\mu} : T \to \mathcal{P}(A)$ such that $\mathbf{P}\{j \circ \xi \in B \mid \eta\} = \tilde{\mu}(\eta, B)$ for all Borel subsets $B \subset A$. By Lemma 8.33 we know that $j_* : \mathcal{P}(S) \to \mathcal{P}(A)$ is a Borel isomorphism so we can define $\mu = j_*^{-1} \circ \tilde{\mu}$ which is a probability kernel by Lemma 8.27. Because j is a Borel isomorphism, we know that every measurable subset of S is of the form $j^{-1}B$ for some Borel $B \subset A$ (Lemma 8.32) and we have

$$\mathbf{P}\{\xi \in j^{-1}B \mid \eta\} = \mathbf{P}\{j \circ \xi \in B \mid \eta\} = \tilde{\mu}(\eta, B) = \mu(\eta, j^{-1}B)$$

$\square$

The following theorem is an absolutely essential tool for computing conditional expectations. Suppose we are given a random variable that is a real valued function applied to a pair of random elements $f(\xi, \eta)$. In the case that ξ and η are indendent we applied the Expectation Rule and Fubini's Theorem in Lemma 4.6 to calculate the expected value of $f(\xi, \eta)$ as an iterated integral. Naively we might expect that in the general case we can calculate the expectation of f conditioned on η by fixing the value of η and then taking an "appropriate" expectation'. The appropriate notion of expectation is given by integration against the distribution of ξ conditional on η.

TODO: Show how this works in the discrete case

THEOREM 8.35. *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 2.27). 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 $\mathbf{E}[\mathbf{1}_{\cup_n C_n}(\xi, \eta)] = \lim_{n \to \infty} \mathbf{E}[\mathbf{1}_{C_n}(\xi, \eta)]$. 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

$$\mathbf{E}\left[\int \left|f(s,\eta)\right|\, d\nu(s)\right] = \mathbf{E}\left[\left|f(\xi,\eta)\right|\right] < \infty$$

Which shows us that $\int \left|f(s,\eta)\right|\, d\nu(s) < \infty$ 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 $\tilde{f}(s,t,u) = u f(s,t)$ 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 8.36 (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 8.34 we know $\mathbf{P}\{\xi \in \cdot \mid \mathcal{F}\}$ has regular version μ. Now by Theorem 8.35 and the ordinary Jensen Inequality (Lemma 3.17) applied pointwise we know that

$$\varphi(\mathbf{E}\left[\xi \mid \mathcal{F}\right]) = \varphi\left(\int s\, \mu(ds)\right) \leq \int \varphi(s)\, \mu(ds) = \mathbf{E}\left[\varphi(\xi) \mid \mathcal{F}\right]$$

TODO: The strictly convex/equality case □

As another application of Theorem 8.35 we give a little result about the interaction between conditional indpendence and conditional expectations.

Corollary 8.37. *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 8.20 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 8.35

$$\mathbf{E}[f(\xi) \mid \mathcal{G}] = \int f(s)\, \mu(ds) = \mathbf{E}[f(\xi) \mid \mathcal{F}, \mathcal{G}] \text{ a.s.}$$

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).

For random vectors, the existence of regular versions allows us to bring the theory of characteristic functions to bear on problems.

Lemma 8.38. *Let ξ be a random vector in $\mathbb{R}^n$ and let $\mathcal{F}$ be a σ-algebra. Suppose that $\phi : \mathbb{R}^n \times \Omega \to \mathbb{C}$ is a function such that for each fixed $u \in \mathbb{R}^n$ we have*

$$\phi(u, \omega) = \boldsymbol{E}\left[e^{i\langle u, \xi\rangle} \mid \mathcal{F}\right] \text{ a.s.}$$

If for every $\omega \in \Omega$ there is a probability measure $\mu(\omega)$ on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$ such that $\phi(u, \omega) = \int e^{i\langle u, x\rangle}\, \mu(\omega, dx)$ then it follows that for every $A \in \mathcal{B}(\mathbb{R}^n)$ we have

$$\boldsymbol{P}\{\xi \in A \mid \mathcal{F}\}(\omega) = \mu(\omega, A) \text{ a.s.}$$

Proof. By Theorem 8.34 we know that we may chose a regular version ν for $\mathbf{P}\{\xi \in \cdot \mid \mathcal{F}\}$. By Theorem 8.35 we know that for every fixed $u \in \mathbb{R}^n$ we have

$$\mathbf{E}\left[e^{i\langle u, \xi\rangle} \mid \mathcal{F}\right] = \phi(u, \omega) = \int e^{i\langle u, x\rangle}\, \nu(\omega, dx)$$

almost surely and by taking a countable intersection of almost sure events we may assume that $\phi(u, \omega) = \int e^{i\langle u, x\rangle}\, \nu(\omega, dx)$ for all $u \in \mathbb{Q}^n$ almost surely. For each fixed ω, both sides of this equation are characteristic functions of a probability measure hence each side is uniformly continuous (Lemma 7.3) and therefore equality on $\mathbb{Q}^n$ can be upgraded to equality on $\mathbb{R}^n$. Now the characteristic function uniquely identifies the underlying probability measure Theorem 7.7 and therefore

$$\mu(\omega, \cdot) = \nu(\omega, \cdot) = \mathbf{P}\{\xi \in A \mid \mathcal{F}\}(\omega) \text{ a.s.}$$

☐

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 given the first random element is the kernel.

LEMMA 8.39. *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 $\xi : \hat{\Omega} \to S$ 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 8.35. Note that if we succeed in constructing ξ then that result tells how to compute expectations on $\hat{\Omega}$. Following that lead, let $(\Omega, \mathcal{A}, P)$ be the probability space underlying the random element η and define $(\hat{\Omega}, \hat{\mathcal{A}}) = (S \times \Omega, \mathcal{S} \otimes \mathcal{A})$. Define the probability measure

$$\hat{P}(A) = \mathbf{E}\left[\int \mathbf{1}_A(s, \omega)\, d\mu(\eta(\omega), s)\right]$$

Note that $(\hat{\Omega}, \hat{\mathcal{A}}, \hat{P})$ is an extension of $(\Omega, \mathcal{A}, P)$ since for $A \in \mathcal{A}$,

$$\hat{P}(S \times A) = \mathbf{E}\left[\int \mathbf{1}_S(s)\mathbf{1}_A(\omega)\, d\mu(\eta(\omega), s)\right] = \mathbf{E}\left[\mathbf{1}_A(\omega)\right] = P(A)$$

Now define $\xi(s, \omega) = s$ and note that for $A \in \mathcal{S}$ and $B \in \mathcal{A}$,

$$\hat{P}(\xi \in A; B) = \mathbf{E}\left[\int \mathbf{1}_A(s)\mathbf{1}_B(\omega)\, d\mu(\eta(\omega), s)\right] = \mathbf{E}\left[\mu(\eta, A); B\right]$$

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 8.20. □

The next result is closely related to the previous lemma and provides an answer to a very natural question. Suppose that one is given measure μ on a product space $S \times T$. It is trivial that one can construct random elements ξ and η in S and T respectively such that the law of (ξ, η) is μ (just take the probability space to be $(S \times T, \mathcal{S} \times \mathcal{T}, \mu)$ and then use the identity map). Now suppose that one is given a random element η in T such that the law of η is the marginal of μ (i.e. $\mathbf{P}\{\eta \in B\} = \mu(S \times B)$); one may ask whether one can find a random element ξ in S such that the law of (ξ, η) is μ. If μ is a product measure then this follows whenever S is Borel by creating a ξ independent of η such that $\mathbf{P}\{\xi \in A\} = \mu(A \times T)$. In fact due to the existence of conditional distributions a similar proof works for general μ. The result is expressed in terms of random elements in $S \times T$ rather a probability measure μ and the proof shows that the construction can be done with a single uniform randomization variable in addition to η.

LEMMA 8.40. *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 η be a random element in T both defined on a probability space $(\Omega, \mathcal{A})$. Let $\tilde{\eta}$ be a random element in T defined on a probability space $(\tilde{\Omega}, \tilde{\mathcal{A}})$ and assume that $\eta \overset{d}{=} \tilde{\eta}$. Then there exists a measurable* *function* $f : T \times [0,1] \to S$ *such that if* ϑ *is a* $U(0,1)$ *random variable defined on* $(\tilde{\Omega}, \tilde{\mathcal{A}})$ *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. By Theorem 8.34 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 8.31 we can find measurable $f : T \times [0,1] \to S$ such that for every $t \in T$ 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

$$\begin{aligned}
&\mathbf{E}\left[g(\tilde{\xi}, \tilde{\eta})\right] \\
&= \mathbf{E}\left[g(f(\tilde{\eta}, \vartheta), \tilde{\eta})\right] \\
&= \mathbf{E}\left[\int_0^1 g(f(\tilde{\eta}, x), \tilde{\eta})\, dx\right] && \tilde{\eta} \perp\!\!\!\perp \vartheta, \text{ Lemma 4.6 and Lemma 3.7} \\
&= \mathbf{E}\left[\int_0^1 g(f(\eta, x), \eta)\, dx\right] && \text{since } \eta \overset{d}{=} \tilde{\eta} \\
&= \mathbf{E}\left[\int g(s, \eta)\, d\mu(\eta, s)\right] && \text{by Expectation Rule Lemma 3.7 and } \mathcal{L}(f(t, \vartheta)) = \mu(t) \\
&= \mathbf{E}\left[g(\xi, \eta)\right] && \text{by Theorem 8.35}
\end{aligned}$$

which shows in particular that $(\xi, \eta) \overset{d}{=} (\tilde{\xi}, \tilde{\eta})$. Note that in applying the fact that $\eta \overset{d}{=} \tilde{\eta}$ we are moving from taking expectations against $\tilde{\Omega}$ to taking expectations against Ω. □

LEMMA 8.41. *Let* S *and* T *be Borel spaces with* $f : S \to T$ *measurable and let* ξ *be a random element in* S *and* η *be a random element in* T *such that* $f(\xi) \overset{d}{=} \eta$. *Then there exists a random element in* S $\tilde{\xi}$ *such that* $\xi \overset{d}{=} \tilde{\xi}$ *and* $f(\tilde{\xi}) = \eta$ *a.s.*

PROOF. Since $f(\xi) \overset{d}{=} \eta$ and S is Borel by Lemma 8.40 we can find $\tilde{\xi} \overset{d}{=} \xi$ such that $(\xi, f(\xi)) \overset{d}{=} (\tilde{\xi}, \eta)$. Now applying the measurable function $f \times id : S \times T \to T \times T$ we conclude that $(f(\xi), f(\xi)) \overset{d}{=} (f(\tilde{\xi}), \eta)$. Because the diagonal $\Delta \subset T \times T$ is measurable (TODO: Do we really need Borel-ness for this?) we can conclude

$$\mathbf{P}\{f(\tilde{\xi}) = \eta\} = \mathbf{P}\{(f(\tilde{\xi}), \eta) \in \Delta\} = \mathbf{P}\{(f(\xi), f(\xi)) \in \Delta\} = 1$$

□

CHAPTER 9

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).

1. Stochastic Processes

We first begin with a very general notion of *stochastic process* which we rather quickly specialize.

DEFINITION 9.1. 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 $\tilde{X}$ on Ω to get an element $\hat{X} : T \to S^\Omega$. It is customary to write $\hat{X}(t)$ as X_t.

LEMMA 9.2. *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 X_t is a composition of measurable functions and it follows from Lemma 2.13 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 2.12 shows that X is measurable. □

It can be useful to know what operations on set functions are measurable with respect to the product topology on S^T. Here we record a simple fact that we will use.

LEMMA 9.3. *Let G be a measurable group and T be an index set, with group operations defined pointwise, $(G^T, \mathcal{G}^T)$ is a measurable group.*

PROOF. With the identity defined by the constant function $f(t) = e$, the fact that G^T is a group is immediate. To see measurability of the group operation, let $A \in \mathcal{G}$ and pick $t \in T$. Note that $(\pi_t, \pi_t) : \mathcal{G}^T \otimes \mathcal{G}^T / \mathcal{G} \otimes \mathcal{G}$ - measurable by definition of the product σ-algebra (both on G^T and generically) and we know the group operation is $\mathcal{G} \otimes \mathcal{G}/\mathcal{G}$-measurable therefore $\{(f, g) \mid (f \cdot g)(t) \in A\}$ is $\mathcal{G}^T \otimes \mathcal{G}^T$-measurable. The proof for the inverse operation follows similarly. □

DEFINITION 9.4. 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*.

Note that we do not require U to be a measurable subset of S^T (and in most case that we consider it will not be). According to this definition, a stochastic process is simply a random element in subset of a path space $(U, U \cap \mathcal{S}^T)$. As such it has a distribution $\mu \circ X^{-1}$ which is a measure on U; 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.

TODO: Build some intuition about the definition of a process. In particular, the reason for considering subsets $U \subset S^T$ is clear because S^T is just too big. It is very rare for one to be interested in arbitrary set functions; almost always one wants some kind of condition to be imposed such as continuity or at least some restriction on the discontinuities that can occur (e.g. allowing jump discontinuities but outlawing oscillatory discontinuities is common in stochastic processes). These subsets very often come with additional structure that either implies or constrains their measure theoretic structure (e.g. a metric topology that implies a Borel σ-algebra). A subtle point that shall come up is that one will want the implied measure theoretic structure be compatible with the measure theoretic structure of the general definition. TODO: Is there something deep about the use of the product σ-algebra in this context or is a technical convenience/least common denominator that allow one to prove general results? Well, arguably it is made in this way so that a stochastic process is just a family of random elements X_t indexed by T; where do we even state this fact?

LEMMA 9.5. *Let $(S, \mathcal{S})$ be a measurable space an let $U \subset S$ be a (not necessarily measurable) subset, then $U \cap \mathcal{S}$ is a σ-algebra on U. Furthermore if $\mathcal{C} \subset 2^S$ is a set of subsets of S that generates $\mathcal{S}$ then $\mathcal{D} = \{U \cap C \mid C \in \mathcal{C}\}$ generates $U \cap \mathcal{S}$. Lastly given a measurable space $(T, \mathcal{T})$ and an $\mathcal{S}/\mathcal{T}$-measurable function $f : S \to T$, the restriction $f|_U : U \to T$ is $U \cap \mathcal{S}/\mathcal{T}$-measurable.*

PROOF. The fact that $U \cap \mathcal{S}$ is a σ-algebra follows easily from the fact that $\mathcal{S}$ is a σ-algebra and the set theoretic identities $\cap_{i=1}(U \cap A_i) = U \cap \cap_{i=1} A_i$ and $U \setminus (U \cap A) = U \cap (U \cap A)^c = U \cap A^c$.

Given the generating set $\mathcal{C}$ for $\mathcal{S}$ and $\mathcal{D}$ defined as above it is immediate from the fact that $\mathcal{C} \subset \mathcal{S}$ that we have $\mathcal{D} \subset U \cap \mathcal{S}$. As we have just proven that $U \cap \mathcal{S}$ is a σ-algebra it follows that $\sigma(\mathcal{D}) \subset U \cap \mathcal{S}$.

On the other hand, let $\mathcal{E} = \{A \subset S \mid U \cap A \in \sigma(\mathcal{D})\}$. We claim that $\mathcal{E}$ is a σ-algebra. Indeed if $A, A_1, A_2, \ldots \in \mathcal{E}$ then we have $U \cap \cup_{i=1}^\infty A_i = \cup_{i=1}^\infty U \cap A_i \in \sigma(\mathcal{D})$ which implies $\cup_{i=1}^\infty A_i \in \mathcal{E}$ and $U \cap A^c = (U \cap U^c) \cup (U \cap A^c) = U \cap (U \cap A)^c \in \mathcal{D}$ which implies $A^c \in \mathcal{E}$. By the definition of $\mathcal{D}$, we know $\mathcal{C} \subset \mathcal{E}$ and therefore

$\mathcal{S} = \sigma(\mathcal{C}) \subset \sigma(\mathcal{E}) = \mathcal{E}$. Thus we have shown the reverse inclusion $U \cap \mathcal{S} \subset \sigma(\mathcal{D})$ and we have $U \cap \mathcal{S} = \sigma(\mathcal{D})$.

Lastly, $U \cap \mathcal{S}/\mathcal{T}$-measurability of the restriction of a $\mathcal{S}/mathcalT$-measurable f follows from the identity $(f|_U)^{-1}(A) = \{s \in S \mid f(s) \in A \text{ and } s \in U\} = U \cap f^{-1}(A)$ which shows $(f|_U)^{-1}(A) \in U \cap \mathcal{S}$ whenever $f^{-1}(A) \in \mathcal{S}$. $\square$

Lemma 9.6. *Let X be a stochastic process with values in $U \subset S^T$, then for every $t_1, \dots, t_n \in T$ then $(X_{t_1}, \dots, X_{t_n}) \in S^n$ is $\mathcal{S}^{\otimes n}$-measurable and the measures $\mu \circ (X_{t_1}, \dots, X_{t_n})^{-1}$ are called the* finite dimensional distributions *of X. Given $U \subset S^T$ then any two probability measures on $(U, U \cap \mathcal{S}^T)$ are equal if and only if their finite dimensional distributions are equal. In particular, if X and Y are two stochastic processes with values in $U \subset S^T$ then $X \overset{d}{=} Y$ if and only if their finite dimensional distributions are equal (written $X \overset{f.d.d.}{=} Y$). It is also that the case that $X \overset{f.d.d.}{=} Y$ if and only if $\mu \circ (X_{t_1}, \dots, X_{t_n})^{-1} = \mu \circ (Y_{t_1}, \dots, Y_{t_n})^{-1}$ for all $t_1, \dots, t_n \in T$ with the t_j distinct.*

Proof. Suppose that $t_1, \dots, t_n$ are given and define the n-dimensional projection $(\pi_{t_1}, \dots, \pi_{t_n}) : S^T \to S^n$. We claim that $(\pi_{t_1}, \dots, \pi_{t_n})$ is $\mathcal{S}^T/\mathcal{S}^{\otimes n}$ measurable. Indeed if we let $A_1 \times \cdots \times A_n \in \mathcal{S}^{\otimes n}$ then $(\pi_{t_1}, \dots, \pi_{t_n})^{-1}(A_1 \times \cdots \times A_n) = \pi_{t_1}^{-1}(A_1) \cap \cdots \cap \pi_{t_n}^{-1}(A_n)$, hence $(\pi_{t_1}, \dots, \pi_{t_n})^{-1}(A_1 \times \cdots \times A_n) \in \mathcal{S}^T$ by the measurability of each $\pi_{t_j}^{-1}(A_j)$ for $j = 1, \dots, n$. Since sets of the form $A_1 \times \cdots \times A_n$ generate $\mathcal{S}^{\otimes n}$ we see that $(\pi_{t_1}, \dots, \pi_{t_n})$ is measurable by application of Lemma 2.12.

The $\mathcal{S}^{\otimes n}$-measurability of $(X_{t_1}, \dots, X_{t_n})$ now follows directly from Lemma 2.13 and 9.5 since we can write $(X_{t_1}, \dots, X_{t_n}) = (\pi_{t_1}, \dots, \pi_{t_n}) \circ X$ as a composition of a $U \cap \mathcal{S}^T/\mathcal{S}^{\otimes n}$-measurable function $(\pi_{t_1}, \dots, \pi_{t_n})|_U$ and $U \cap \mathcal{S}^T$-measurable function X.

Suppose that μ and ν are probability measures on $(U, U \cap \mathcal{S}^T)$ whose finite dimensional projections are equal; that is to say for every $t_1, \dots, t_n \in T$ we have $\mu \circ (\pi_{t_1}, \dots, \pi_{t_n})^{-1} = \nu \circ (\pi_{t_1}, \dots, \pi_{t_n})^{-1}$. This fact that shows that μ and ν agree on all sets of the form $(\pi_{t_1}, \dots, \pi_{t_n})^{-1}(A)$ where $n > 0$, $t_1, \dots, t_n \in T$ and $A \subset \mathcal{S}^{\otimes n}$. Let the set of sets of this form be called $\mathcal{C}$. We claim $\mathcal{C}$ generates $U \cap \mathcal{S}^T$. Indeed it is the case that sets of the form $\pi_t^{-1}(A)$ for $t \in T$ and $A \subset \mathcal{S}$ generate $U \cap \mathcal{S}^T$. One can see this by observing that $\pi_t^{-1}(A) = U \cap \tilde{\pi}_t^{-1}(A)$ where $\tilde{\pi}_t : S^T \to S$ is the evaluation map extended to the entireity of S^T. By definition $\mathcal{S}^T$ is generated by the sets $\tilde{\pi}^{-1}(A)$ and therefore by Lemma 9.5 we conclude $U \cap \mathcal{S}^T$ is generated by sets of the form $U \cap \tilde{\pi}_t^{-1}(A) = \pi_t^{-1}(A)$.

Next we claim that $\mathcal{C}$ is a π-system. This follows immediately as we can write $(\pi_{t_1}, \dots, \pi_{t_n})^{-1}(A) \cap (\pi_{s_1}, \dots, \pi_{s_m})^{-1}(B) = (\pi_{t_1}, \dots, \pi_{t_n}, \pi_{s_1}, \dots, \pi_{s_m})^{-1}(A \times B)$. Now we may conclude $\mu = \nu$ by a monotone class argument (specifically Lemma 2.71).

The statement about stochastic processes follows by applying the fact just proven the laws of X and Y.

It is trivial that if $X \overset{d}{=} Y$ then the finite dimensional distributions with distinct t_j are equal. To see the converse note that the projection $(\pi_{t_1}, \dots, \pi_{t_n})$ for not necessarily distinct t_j may be written as a composition $i \circ (\pi_{s_1}, \dots, \pi_{s_m})$ with $s_1, \dots, s_m$ the set of distinct t_j and $i : S^m \to S_n$ that depends only on the t_j. Now the result follows from functoriality of the pushforward of a measure. $\square$

The previous result shows that the finite dimensional distributions uniquely characterize the distribution of a stochastic process. We now turn an associated existence problem. Namely given a family of distributions that are candiates to be the finite dimensional distributions of a stochastic process, is there in fact a stochastic process with these FDDs. In general this is not the case and the result requires topological assumptions. It is sufficient to assume that the spaces involved are Borel.

DEFINITION 9.7. Let $(S_1, \mathcal{S}_1), (S_2, \mathcal{S}_2), \dots$ be a sequence of measurable spaces, and for each $n \in \mathbb{N}$, let μ_n be a probability measure on on $S_1 \times \cdots \times S_n$. We say that the sequence of measures $\mu_1, \mu_2, \dots$ is *projective* if for every $n \in \mathbb{N}$ and every $A \in \mathcal{S}_1 \otimes \cdots \otimes \mathcal{S}_n$ we have $\mu_{n+1}(A \times S_{n+1}) = \mu_n(A)$.

THEOREM 9.8 (Daniell Theorem). *Let $(S_1, \mathcal{S}_1), (S_2, \mathcal{S}_2), \dots$ be a sequence of measurable spaces, with $S_2, S_3, \dots$ Borel and let $\mu_1, \mu_2, \dots$ be a projective sequence of measures then there exist random elements ξ_n in S_n for $n \in \mathbb{N}$ such that $\mathcal{L}(\xi_1, \dots, \xi_n) = \mu_n$ for all $n \in \mathbb{N}$. In particular, there exists a probability measure μ on $S_1 \times S_2 \times \cdots$ such that for every $n \in \mathbb{N}$ and $A \in \mathcal{S}_1 \otimes \cdots \otimes \mathcal{S}_n$ we have $\mu(A \times S_{n+1} \times \cdots) = \mu_n(A)$.*

PROOF. Trivially we can create ξ_1 with $\mathcal{L}(\xi_1) = \mu_1$ (just take $\Omega = S_1$ and define ξ_1 to be the identity). Now by extending Ω to $S_1 \times [0,1]$ we applying Lemma 4.33 we can find independent $U(0,1)$ random variables $\vartheta_2, \vartheta_3, \dots$ which are also independent of ξ_1. We construct the remaining $\xi_2, \xi_3, \dots$ by an induction argument using Lemma 8.40.

Suppose that we have constructed $\xi_1, \dots, \xi_n$ where for each $m > 1$ there exists a measurable function f_m such that $\xi_m = f_m(\xi_1, \vartheta_2, \dots, \vartheta_m)$. Let $\eta_1, \dots, \eta_{n+1}$ we arbitrary random elements such that $\mathcal{L}(\eta_1, \dots, \eta_{n+1}) = \mu_{n+1}$ (e.g. define $\tilde{\Omega} = S_1 \times \cdots \times S_{n+1}$ with probability measure μ_{n+1} and define $\eta_m(s_1, \dots, s_{n+1}) = s_m$). By the induction hypothesis and the projective property of the sequence μ_n we have for each $A \in \mathcal{S}_1 \otimes \cdots \otimes \mathcal{S}_n$

$$\begin{aligned}
\mathbf{P}\{(\eta_1, \dots, \eta_n) \in A\} &= \mathbf{P}\{(\eta_1, \dots, \eta_n, \eta_{n+1}) \in A \times S_{n+1}\} = \mu_{n+1}(A \times S_{n+1}) \\
&= \mu_n(A) = \mathbf{P}\{(\eta_1, \dots, \eta_n) \in A\}
\end{aligned}$$

and therefore $(\eta_1, \dots, \eta_n) \stackrel{d}{=} (\xi_1, \dots, \xi_n)$. Now we may apply Lemma 8.40 to conclude that there is a measurable function $g : S_1 \times \cdots \times S_n \times [0,1] \to S_{n+1}$ such that $\xi_{n+1} = g(\xi_1, \dots, \xi_n, \vartheta_{n+1})$ satisfies

$$\mathcal{L}(\xi_1, \dots, \xi_{n+1}) = \mathcal{L}(\eta_1, \dots, \eta_{n+1}) = \mu_{n+1}$$

Moreover we may define

$$f_{n+1}(x_1, \dots, x_{n+1}) = g(x_1, f_2(x_1, x_2), \dots, f_n(x_1, \dots, x_n), x_{n+1})$$

so that $\xi_{n+1} = f_{n+1}(\xi_1, \vartheta_2, \dots, \vartheta_{n+1})$.

For the last part of the theorem, define $\mu = \mathcal{L}(\xi_1, \xi_2, \dots)$. It then follows that for every $n \in \mathbb{N}$ and $A \in \mathcal{S}_1 \otimes \cdots \otimes \mathcal{S}_n$ we have

$$\begin{aligned}
\mu(A \times S_{n+1} \cdots) &= \mathbf{P}\{(\xi_1, \xi_2, \dots) \in A \times S_{n+1} \cdots\} \\
&= \mathbf{P}\{(\xi_1, \dots, \xi_n) \in A\} = \mu_n(A)
\end{aligned}$$

$\square$

We now generalize the Daniell Theorem to arbitrary index sets T. First we generalize the notion of a projective sequence of measures to a projective family on an arbitrary index set.

Definition 9.9. Let T be a set and suppose we are given a measurable space $(S_t, \mathcal{S}_t)$ for every $t \in T$ and for every finite subset $I \subset T$ we are given a probability measure μ_I on $\times_{t \in I} S_t$. For any subset $U \subset T$ define $(S_U, \mathcal{S}_U) = (\times_{t \in U} S_t, \otimes t \in U \mathcal{S}_t)$. We say that $\{\mu_I\}$ is a *projective family* if for every finite subset $J \subset T$ and $I \subset J$ we have $\mu_J(\cdot \times S_{J \setminus I}) = \mu_I(\cdot)$. If in the definition above we replace the set of finite subsets of T by the set of countable subsets of T the we say μ_I is a *countable projective family*.

Before we attack the theorem we give a description of the structure of the infinite product σ-algebra that will prove useful in the proof of the extension theorem.

Lemma 9.10. *Let T be a set and $(S_t, \mathcal{S}_t)$ be a family of measurable spaces then the σ-algebra $\otimes_{t \in T} \mathcal{S}_t$ is precisely the set of sets of the form $A \times S_{T \setminus U}$ where $U \subset T$ is a countable subset and $A \in \otimes_{t \in U} \mathcal{S}_t$.*

Proof. We claim that

$$\mathcal{C} = \{A \times S_{T \setminus U} \mid U \subset T \text{ is countable and } A \in \otimes_{t \in U} \mathcal{S}_t\}$$

is a σ-algebra. Obviously $\mathcal{C}$ is non-empty. To see that $\mathcal{C}$ is closed under set complement take $A \times S_{T \setminus U}$ with $U \subset T$ countable and $A \in \mathcal{S}_U$. Then $A^c \in \mathcal{S}_U$ and moreover $(A \times S_{T \setminus U})^c = A^c \times S_{T \setminus U} \in \mathcal{C}$. Given a sequence $C_1, C_2, \ldots \in \mathcal{C}$ with $C_j = A_j \times S_{T \setminus U_j}$, again by passing to the union $\cup_{j=1}^{\infty} U_j$ we may assume that the U_j are all the same countable subset of T and therefore $C_j = A_j \times S_{T \setminus U}$ with $A_j \in \mathcal{S}_U$. It follows that $\cup_{j=1}^{\infty} A_j \in \mathcal{S}_U$ and therefore $\cup_{j=1}^{\infty} C_j \in \mathcal{C}$. Closure under countable intersection follows by De Morgans Law. Since it is clear that each π_t is $\mathcal{C}$-measurable we see that $\otimes_{t \in T} \mathcal{S}_t \subset \mathcal{C}$.

To see the reverse conclusion fix a countable subset $U \subset T$ and we need to show that for every $A \in \mathcal{S}_U$ we have $A \times S_{T \setminus U} \in \otimes_{t \in T} \mathcal{S}_t$. We note that the set of all $A \subset S_U$ such that $A \times S_{T \setminus U} \in \mathcal{S}_T$ is a σ-algebra since it is precisely the pullback and pushforward of $\mathcal{S}_U$ under the projection $\pi_U : S_T \to S_U$ (Lemma 2.8). It clearly contains all sets of the form $B \times S_{U \setminus \{t\}}$ for $t \in U$ and $B \in \mathcal{S}_t$. Such sets generate the σ-algebra $\mathcal{S}_U$ and therefore we conclude that $A \times S_{T \setminus U} \in \mathcal{S}_T$. ☐

Theorem 9.11 (Daniell-Kolmogorov Theorem). *Let T be a set, $(S_t, \mathcal{S}_t)$ for $t \in T$ be a family of Borel sets and μ_I be a projective family of probability measures. There exists a random element ξ_t in S_t for all $t \in T$ such that for every $I \subset T$ we have $\mathcal{L}(\xi_I) = \mu_I$.*

Proof. Let $\overline{T}$ be the set of countable subsets of T. It is clear that the restriction of the projective family μ_I to any subset $U \subset T$ is a projective sequence and therefore we can apply Theorem 9.8 to construct a probability measure μ_U on S_U such that for every finite subset $J \subset U$ we have $\mu_U(\cdot \times S_{U \setminus J}) = \mu_J(\cdot)$.

Now assume that we have a *countable* subset $V \subset U$ and consider the probability measure $\mu_U(\cdot \times S_{U \setminus V}$ on S_V. From what we have just shown, for every finite subset $J \subset V$ and every $A \in \mathcal{S}_J$ we have

$$\mu_U(A \times S_{V \setminus J} \times S_{U \setminus V}) = \mu_U(A \times S_{U \setminus J}) = \mu_J(A)$$

and therefore $\mu_U(\cdot \times S_{U\setminus V})$ and μ_V have the same finite dimensional distributions and therefore by Lemma 9.6 we know that $\mu_U(\cdot \times S_{U\setminus V}) = \mu_V$. Thus we have extended the projective family μ_I to a countable projective family. Since by Lemma 9.10 we know that $\otimes_{t\in T}\mathcal{S}_t$ is the precisely the set of countable cylinder sets, we can define a measure as a set function on said sets. Pick $U \subset T$ and let $A \in \mathcal{S}_U$ then we define μ by $\mu(A \times S_{T\setminus U}) = \mu_U(A)$. We first claim that the μ is well defined. Suppose we have countable subset $U, V \subset T$, $A \in \mathcal{S}_U$ and $B \in \mathcal{S}_V$ such that $A \times S_{T\setminus U} = B \times S_{T\setminus V}$. We can write

$$A \times S_{T\setminus U} = A \times S_{V\setminus U} \times S_{T\setminus(U\cup V)}$$

and

$$B \times S_{T\setminus V} = B \times S_{U\setminus V} \times S_{T\setminus(U\cup V)}$$

from which it follows that $A \times S_{V\setminus U} = B \times S_{U\setminus V}$. Using this equality along with projectivity we get

$$\begin{aligned}\mu(A \times S_{T\setminus U}) &= \mu_U(A) = \mu_{U\cup V}(A \times S_{V\setminus U}) \\ &= \mu_{U\cup V}(B \times S_{U\setminus V}) = \mu_V(B) = \mu(B \times S_{T\setminus V})\end{aligned}$$

which shows that μ is well defined.

It is clear that $\mu(\emptyset) = \mu_I(\emptyset)$ for any $I \subset T$ and therefore $\mu(\emptyset) = 0$.

To see countable additivity of μ suppose we are given set $A_1 \times S_{T\setminus U_1}, A_2 \times S_{T\setminus U_2}, \dots$ where each $U_j \subset T$ is a countable subset and $A_j \in \mathcal{S}_{U_j}$ for $j \in \mathbb{N}$. If we define $U = \cup_{j=1}^\infty U_j$ and redefine A_j as $A_j \times S_{U\setminus U_j}$ then we may assume that the U_j are all the same. Therefore we now have by countable additivity of μ_U,

$$\begin{aligned}\mu(\cup_{j=1}^\infty(A_j \times S_{T\setminus U})) &= \mu((\cup_{j=1}^\infty A_j) \times S_{T\setminus U}) \\ &= \mu_U(\cup_{j=1}^\infty A_j) = \sum_{j=1}^\infty \mu_U(A_j) = \sum_{j=1}^\infty \mu(A_j \times S_{T\setminus U})\end{aligned}$$

Having defined μ on $(S_T, \mathcal{S}_T)$ we now define ξ_t to be the projection $\xi_t : S_T \to S_t$ for each $t \in T$. It is clear from the definition of μ that for every finite subset of $I \subset T$ and $A \in \mathcal{S}_I$ we have

$$\mathbf{P}\{\xi_I \in A\} = \mu(A \times S_{T\setminus I}) = \mu_I(A)$$

and therefore $\mathcal{L}(\xi_I) = \mu_I$. □

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 9.12. 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 observations of said process.

DEFINITION 9.13. 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 9.14. 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 9.15. 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 9.16. 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) $\mathbf{E}[M_t \mid \mathcal{F}_s] = M_s$ a.s. for all $s, t \in T$ with $s \leq t$.

If we replace the condition (iii) by the condition $M_s \leq \mathbf{E}[M_t \mid \mathcal{F}_s]$ a.s., then M is said to be a *submartingale* and if we replace it with $M_s \geq \mathbf{E}[M_t \mid \mathcal{F}_s]$ 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 9.17. *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* $M_t = \boldsymbol{E}[\xi \mid \mathcal{F}_t]$ *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

$$\mathbf{E}[M_t \mid \mathcal{F}_s] = \mathbf{E}[\mathbf{E}[\xi \mid \mathcal{F}_t] \mid \mathcal{F}_s] = \mathbf{E}[\xi \mid \mathcal{F}_s] = M_s$$

□

A martingale that can be expressed in the form given by the Lemma is referred to as a *closed* martingale. We call the reader's attention the the fact that the index set in the above Lemma is allowed to include ∞. For in that case, we might as well assume that ξ is $\mathcal{F}_\infty$-measurable (or equivalently assume that $\xi = M_\infty$ a.s.) This consideration points to an arguably less transparent defintion definition of a closed martingale as one for which $\sup T \in T$ (see Kallenberg for example; what we call a closed martingale he calls a *closable* martingale). Also thinking about the case in which $\infty \in T$ (or more generally when $\sup T \in T$) suggests a relationship between a closing ξ and a limit $\lim_{t\to\infty} M_t$. Such a relationship indeed exists and is explained in Martingale convergence theorems that follow.

The unbiased random walk provides one of the simplest examples of a martingale.

Example 9.18. 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 martingale out of the variance of a base martinagle.

Example 9.19. Suppose we have the setup of Example 9.18 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

$$\begin{aligned}
\mathbf{E}\left[M_n^2 - n\sigma^2 \mid \mathcal{F}_{n-1}\right] &= \mathbf{E}\left[M_{n-1}^2 + 2M_{n-1}\xi_n + \xi_n^2 - n\sigma^2 \mid \mathcal{F}_{n-1}\right] \\
&= M_{n-1}^2 + 2M_{n-1}\mathbf{E}\left[\xi_n \mid \mathcal{F}_{n-1}\right] + \mathbf{E}\left[\xi_n^2 \mid \mathcal{F}_{n-1}\right] - n\sigma^2 \\
&= M_{n-1}^2 + 2M_{n-1}\mathbf{E}\left[\xi_n\right] + \mathbf{E}\left[\xi_n^2\right] - n\sigma^2 \\
&= M_{n-1}^2 - (n-1)\sigma^2
\end{aligned}$$

The last example shows that martingales arise by "centering" an arbitrary integrable process.

Example 9.20. Let η_n be an arbitrary sequence of integrable random variables. Define $\mathcal{F}_n = \sigma(\eta_0, \dots, \eta_n)$ and

$$\xi_n = \sum_{j=1}^{n} \left(\eta_j - \mathbf{E}\left[\eta_j \mid \mathcal{F}_{j-1}\right]\right)$$

then ξ_n is an $\mathcal{F}$-martingale. It is clear that ξ_n is integrable and $\mathcal{F}$-adapted. The martingale property is an easy calculation

$$\begin{aligned}
\mathbf{E}\left[\xi_n \mid \mathcal{F}_{n-1}\right] &= \sum_{j=1}^{n} \left(\mathbf{E}\left[\eta_j \mid \mathcal{F}_{n-1}\right] - \mathbf{E}\left[\eta_j \mid \mathcal{F}_{j-1}\right]\right) \\
&= \sum_{j=1}^{n-1} \left(\eta_j - \mathbf{E}\left[\eta_j \mid \mathcal{F}_{j-1}\right]\right) = \xi_{n-1}
\end{aligned}$$

Returning to our gambling interpretation of martingales we discussed in Example 9.18, 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 9.21. 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 9.22. *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 $\tilde{M}_n = \sum_{j=0}^{n} A_j \left(M_j - M_{j-1}\right)$ *(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 $j \le n$. 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 9.23. *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 $\mathbf{E}[M_s] = \mathbf{E}[\mathbf{E}[M_t \mid \mathcal{F}_s]] = \mathbf{E}[M_t]$. □

Proposition 9.24. *Let X_t be an $\mathbb{R}^d$ valued $\mathcal{F}$-adapted and let $f : \mathbb{R}^d \to \mathbb{R}$ be convex such that $f(X_t)$ is integrable for every $t \in T$. If either X is a martingale or X is a submartingale and f is non-decreasing then $f(X_n)$ is an $\mathcal{F}$-submartingale.*

Proof. The fact that $f(X_n)$ is $\mathcal{F}$-adapted follows immediately from the $\mathcal{F}$-adaptedness of M and Lemma 2.13. Integrability of $f(M_n)$ is an explicit hypothesis. By Jensen's Inequality for conditional expectations (Theorem 8.36), if X_t is a martingale then for $s < t$,

$$f(X_s) = f(\mathbf{E}[X_t \mid \mathcal{F}_s]) \leq \mathbf{E}[f(X_t) \mid \mathcal{F}_s]$$

and if f is non-decreasing and X_t is a submartingale and $s < t$ then

$$f(X_s) \leq f(\mathbf{E}[X_t \mid \mathcal{F}_s]) \leq \mathbf{E}[f(X_t) \mid \mathcal{F}_s]$$

□

2. Optional Times

DEFINITION 9.25. 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 9.26. Given an $\mathcal{F}$-optional time τ, we define

$$\mathcal{F}_\tau = \{A \in \mathcal{A} \mid A \cap \{\tau \leq t\} \in \mathcal{F}_t \text{ for all } t \in T\}$$

Before going any further we have to validate that we have indeed defined a σ-algebra.

LEMMA 9.27. *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 $A_1, A_2, \dots \in \mathcal{F}_\tau$, 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 2.12. □

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 9.28. *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 $A \cap \{\tau \leq t\} = (A \cap \{\sigma \leq t\}) \cap \{\tau \leq t\} \in \mathcal{F}_t$. $\square$

LEMMA 9.29. *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

$$\{\tau = t\} = \{\tau \leq t\} \cap \left(\cup_{s<t}\{\tau \leq s\}\right)^c$$

which shows that $\{\tau = t\} \in \mathcal{F}_t$. $\square$

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 9.30. *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: We actually prove this later on... $\square$

The σ-algebra $\mathcal{F}_\tau$ maybe thought of as being constructed by patching together the individual σ-algebras $\mathcal{F}_t$ of the filtration; many arguments make use of this idea. A precise statement that allows localization of conditional expectations with respect to $\mathcal{F}_\tau$ is given here. The reader should translate the following lemma into the intuitively obvious prose assertion "given that $\tau = t$, an event A happens before τ if and only if A happens before t".

LEMMA 9.31. *Given a filtration $\mathcal{F}_t$ and an $\mathcal{F}$-optional time τ, for every $t \in T$, the σ-algebras $\mathcal{F}_t$ and $\mathcal{F}_\tau$ agree on the set $\{\tau = t\}$.*

PROOF. Suppose $A \in \mathcal{F}_\tau$ and $A \subset \{\tau = t\}$. Then by definition of $\mathcal{F}_\tau$ we know that $A = A \cap \{\tau \leq t\} \in \mathcal{F}_t$. On the other hand, if $A \in \mathcal{F}_t$ we know that for all $s \in T$,

$$A \cap \{\tau \leq s\} = A \cap \{\tau = t\} \cap \{\tau \leq s\} = \begin{cases} A & \text{if } s \geq t \\ \emptyset & \text{if } s < t \end{cases} \in \mathcal{F}_s$$

$\square$

Another useful fact is

Proposition 9.32. *Let σ and τ be $\mathcal{F}$-optional times then $\mathcal{F}_\sigma \cap \{\sigma \leq \tau\} \subset \mathcal{F}_{\sigma\wedge\tau} = \mathcal{F}_\sigma \cap \mathcal{F}_\tau$. In particular, $\mathcal{F}_\sigma = \mathcal{F}_\tau$ on $\{\sigma = \tau\}$.*

Proof. Let $A \in \mathcal{F}_\sigma$ and let $t \in T$.

$$A \cap \{\sigma \leq \tau\} \cap \{\tau \leq t\} = (A \cap \{\sigma \leq t\}) \cap \{\sigma \leq \tau\} \cap \{\sigma \wedge t \leq \tau \wedge t\} \in \mathcal{F}_t$$

where we have used the facts that σ and τ is $\mathcal{F}$-optional and both $\sigma \wedge t$ and $\tau \wedge t$ are $\mathcal{F}_t$-measurable (Lemma 9.27 and Lemma 9.28). The first containment follows from this fact applied to σ and $\sigma \wedge \tau$ (observing that $\{\sigma \leq \sigma \wedge \tau\} = \{\sigma \leq \tau\}$). Now applying this fact again,

$$\mathcal{F}_{\sigma\wedge\tau} = \mathcal{F}_{\sigma\wedge\tau} \cap \{\sigma \wedge \tau \leq \tau\} \subset \mathcal{F}_\tau$$

and thus $\mathcal{F}_{\sigma\wedge\tau} \subset \mathcal{F}_\sigma \cap \mathcal{F}_\tau$ by reversing the roles of σ and τ. On the other hand if $A \in \mathcal{F}_\sigma \cap \mathcal{F}_\tau$ and $t \in T$, then

$$A \cap \{\sigma \wedge \tau \leq t\} = (A \cap \{\sigma \leq t\}) \cup (A \cap \{\tau \leq t\}) \in \mathcal{F}_t$$

which shows $\mathcal{F}_{\sigma\wedge\tau} = \mathcal{F}_\sigma \cap \mathcal{F}_\tau$.

The last statement follows from observing $\{\sigma \geq \tau\} \in \mathcal{F}_\sigma \cap \mathcal{F}_\tau \subset \mathcal{F}_\sigma$ thus if $A \in \mathcal{F}_\sigma$, we also have $A \cap \{\sigma \geq \tau\} \in \mathcal{F}_\sigma$. Therefore if we write

$$A \cap \{\sigma = \tau\} = (A \cap \{\sigma \geq \tau\}) \cap \{\sigma \leq \tau\}$$

this shows $\mathcal{F}_\sigma \cap \{\sigma = \tau\} \subset \mathcal{F}_\sigma \cap \{\sigma \leq \tau\} \subset \mathcal{F}_\tau$ which in turn shows $\mathcal{F}_\sigma \cap \{\sigma = \tau\} \subset \mathcal{F}_\tau \cap \{\sigma = \tau\}$. Now the result follows by the same argument with the role of σ and τ reversed. ∎

Corollary 9.33. *Let τ be an $\mathcal{F}$-optional time then for every $t \in T$ we have $\{\tau \leq t\} \in \mathcal{F}_{\tau\wedge t}$.*

Proof. By definition, $\{\tau \leq t\} \in \mathcal{F}_t$ and by $\mathcal{F}_\tau$-measurability of τ (Lemma 9.27) we know that $\{\tau \leq t\} \in \mathcal{F}_\tau$. Now the result follows from $\mathcal{F}_{\tau\wedge t} = \mathcal{F}_\tau \cap \mathcal{F}_t$ (Proposition 9.32). ∎

3. 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 9.34. *Let $\mathcal{F}_n$ be a filtration and M_n be a sequence of $\mathcal{F}$-adapted integrable random variables. If $\boldsymbol{E}[M_n \mid \mathcal{F}_{n-1}] = M_{n-1}$ for $n > 0$ then M_n is an $\mathcal{F}$-martingale.*

Proof. We only have to show that $\mathbf{E}[M_n \mid \mathcal{F}_m] = M_m$ for all $m \leq n$. Because we know M_n is $\mathcal{F}_n$-measurable then we have $\mathbf{E}[M_n \mid \mathcal{F}_n] = M_n$. If $m < n-1$, then we proceed by induction assuming the result is true for $m+1$,

$$\begin{aligned}
\mathbf{E}[M_n \mid \mathcal{F}_m] &= \mathbf{E}[\mathbf{E}[M_n \mid \mathcal{F}_{m+1}] \mid \mathcal{F}_m] \\
&= \mathbf{E}[M_{m+1} \mid \mathcal{F}_m] && \text{by induction hypothesis} \\
&= M_m && \text{by hypothesis}
\end{aligned}$$

□

Furthermore in discrete time we have a simple version of a construction of a useful class of optional times.

DEFINITION 9.35. 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 9.36. *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 9.37. *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 9.38 (Optional Sampling Theorem). *Let σ and τ be bounded $\mathcal{F}$-optional times and let M_n be a martingale, then*

$$\boldsymbol{E}[M_\tau \mid \mathcal{F}_\sigma] \geq M_{\sigma \wedge \tau} \ a.s.$$

TODO: The assumption that σ is bounded can be removed (see Kallenberg's proof for a demonstration of that). How to fix up the proof below or amend them?

TODO: This result assumes that we have a martingale on $\mathbb{Z}$; the result holds with arbitrary countable index sets.

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 with σ deterministic 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 $\mathbf{E}[M_\tau \mid \mathcal{F}_k] = M_k$. We do this by induction on $m-k$. For $m-k=0$, the result is trivial since in this case $M_\tau = M_k$. 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

$$
\begin{aligned}
\mathbf{E}[M_\tau \mid \mathcal{F}_k] &= \mathbf{E}[M_{\tau \vee k+1} \mid \mathcal{F}_k] + \mathbf{E}[(M_k - M_{k+1})\mathbf{1}_{\tau=k} \mid \mathcal{F}_k] \\
&= \mathbf{E}[\mathbf{E}[M_{\tau \vee k+1} \mid \mathcal{F}_{k+1}] \mid \mathcal{F}_k] + \mathbf{1}_{\tau=k}\mathbf{E}[(M_k - M_{k+1}) \mid \mathcal{F}_k] \\
&= \mathbf{E}[M_{k+1} \mid \mathcal{F}_k] + 0 = M_k
\end{aligned}
$$

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 $0 \leq n \leq N$.

$$
\begin{aligned}
\mathbf{E}[M_\tau; A] &= \sum_{n=0}^N \sum_{m=0}^N \mathbf{E}[M_n \mathbf{1}_{\tau=n} \mathbf{1}_{\sigma=m} \mathbf{1}_A] \\
&= \sum_{n=0}^N \left(\sum_{m=n}^N \mathbf{E}[M_m \mathbf{1}_{\tau=m} \mathbf{1}_{\sigma=n} \mathbf{1}_A] + \sum_{m=n+1}^N \mathbf{E}[M_n \mathbf{1}_{\tau=n} \mathbf{1}_{\sigma=m} \mathbf{1}_A] \right) \\
&= \sum_{n=0}^N \mathbf{E}[(M_{\tau \vee n} - M_n \mathbf{1}_{\tau<n}) \mathbf{1}_{\sigma=n} \mathbf{1}_A] + \mathbf{E}[M_n \mathbf{1}_{\tau=n} \mathbf{1}_{\sigma \geq n+1} \mathbf{1}_A] \\
&= \sum_{n=0}^N \mathbf{E}[M_n \mathbf{1}_{\tau \geq n} \mathbf{1}_{\sigma=n} \mathbf{1}_A] + \mathbf{E}[M_n \mathbf{1}_{\tau=n} \mathbf{1}_{\sigma \geq n+1} \mathbf{1}_A] \\
&= \sum_{n=0}^N \mathbf{E}[M_n \mathbf{1}_{\tau \wedge \sigma=n} \mathbf{1}_A] \\
&= \mathbf{E}[M_{\tau \wedge \sigma}; A]
\end{aligned}
$$

and therefore by the defining property of conditional expectations we are done.

Here is another rather direct proof that seems quite transparent and is completely self contained. Suppose $\sigma, \tau \leq N$. Pick $A \in \mathcal{F}_\sigma$ and compute

$$
\begin{aligned}
\mathbf{E}\left[M_\tau; A\right] &= \sum_{n=0}^{N} \sum_{m=0}^{N} \mathbf{E}\left[M_n; A \cap \{\tau = n\} \cap \{\sigma = m\}\right] \\
&= \sum_{n=0}^{N-1} \sum_{m=n+1}^{N} \mathbf{E}\left[M_n; A \cap \{\tau = n\} \cap \{\sigma = m\}\right] + \sum_{n=0}^{N} \sum_{m=n}^{N} \mathbf{E}\left[M_m; A \cap \{\tau = m\} \cap \{\sigma = n\}\right] \\
&= \sum_{n=0}^{N-1} \mathbf{E}\left[M_n; A \cap \{\tau = n\} \cap \{\sigma > n\}\right] + \sum_{n=0}^{N} \sum_{m=n}^{N} \mathbf{E}\left[M_N; A \cap \{\tau = m\} \cap \{\sigma = n\}\right] \\
&= \sum_{n=0}^{N-1} \mathbf{E}\left[M_n; A \cap \{\tau = n\} \cap \{\sigma > n\}\right] + \sum_{n=0}^{N} \mathbf{E}\left[M_N; A \cap \{\tau \geq n\} \cap \{\sigma = n\}\right] \\
&= \sum_{n=0}^{N-1} \mathbf{E}\left[M_n; A \cap \{\tau = n\} \cap \{\sigma > n\}\right] + \sum_{n=0}^{N} \mathbf{E}\left[M_n; A \cap \{\tau \geq n\} \cap \{\sigma = n\}\right] \\
&= \sum_{n=0}^{N} \mathbf{E}\left[M_n; A \cap \{\tau \wedge \sigma = n\}\right] \\
&= \mathbf{E}\left[M_{\tau \wedge \sigma}; A\right]
\end{aligned}
$$

□

Corollary 9.39. *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 Sampling as $\tau \wedge n$ and $n-1$ are both bounded optional times and therefore

$$
\mathbf{E}\left[M_{\tau \wedge n} \mid \mathcal{F}_{n-1}\right] = M_{\tau \wedge n \wedge (n-1)} = M_{\tau \wedge (n-1)}
$$

Note that this can also be proven by a direct computation using the fact that $\{\tau \geq n\} = \{\tau \leq n-1\}^c \in \mathcal{F}_{n-1}$:

$$
\begin{aligned}
\mathbf{E}\left[M_{\tau \wedge n} \mid \mathcal{F}_{n-1}\right] &= \sum_{m=0}^{n-1} \mathbf{E}\left[M_m \mathbf{1}_{\tau = m} \mid \mathcal{F}_{n-1}\right] + \mathbf{E}\left[M_n \mathbf{1}_{\tau \geq n} \mid \mathcal{F}_{n-1}\right] \\
&= \sum_{m=0}^{n-1} M_m \mathbf{1}_{\tau = m} + M_{n-1} \mathbf{1}_{\tau \geq n} \\
&= \sum_{m=0}^{n-2} M_m \mathbf{1}_{\tau = m} + M_{n-1} \mathbf{1}_{\tau \geq n-1} = M_{\tau \wedge (n-1)}
\end{aligned}
$$

□

Lemma 9.40 (Doob Decomposition). *Let X_n be a submartingale, then there exists a martingale M_n and an almost surely increasing non-negative $\mathcal{F}$-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. Non-negativity of A_n follows from the facts that A_n is increasing and $A_0 = 0$. □

The Doob Decomposition is generally a useful tool to transfer results about martingales over to submartingales. As a first illustration we present the following optional sampling theorem to submartingales

Corollary 9.41. *Let X_n be a submartingale and let σ and τ be bounded optional times, then $\boldsymbol{E}[X_\tau \mid \mathcal{F}_\sigma] \geq X_{\sigma\wedge\tau}$ a.s.*

TODO: See comments about relaxing boundedness hypotheses.

Proof. We write $X_n = M_n + A_n + X_0$ with M_n a martingale and A_n positive increasing previsible. Applying optional sampling (Theorem 9.38) 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. □

There are other decomposition results that are of use. While the Doob Decomposition shows that a submartingale is bounded below by a martingale the following shows that an L^1-bounded submartingale is bounded above by a martingale as well.

Lemma 9.42 (Krickeberg Decomposition). *Let X_n be an L^1-bounded submartingale then there exists an L^1-bounded martingale M_n and a nonnegative L^1-bounded supermartingale A_n such that $X_n = M_n - A_n$.*

PROOF. Fix an $m \geq 0$ and for every $n \geq m$ define $M_{n,m} = \mathbf{E}[X_n \mid \mathcal{F}_m]$. Note that by the submartingale property

$$M_{n,m} = \mathbf{E}[X_n \mid \mathcal{F}_m] \leq \mathbf{E}[\mathbf{E}[X_{n+1} \mid \mathcal{F}_n] \mid \mathcal{F}_m] = \mathbf{E}[X_{n+1} \mid \mathcal{F}_m] = M_{n+1,m} \text{ a.s.}$$

and therefore we can define $M_m = \lim_{n\to\infty} M_{n,m}$. Furthermore we know that

$$\mathbf{E}[|M_{n,m}|] \leq \mathbf{E}[\mathbf{E}[|X_n| \mid \mathcal{F}_m]] = \mathbf{E}[|X_n|] \leq \sup_n \mathbf{E}[|X_n|] < \infty$$

and therefore we can apply the Monotone Convergence Theorem to conclude $\mathbf{E}[|M_m|] < \infty$ so M_m are integrable. Clearly by definition of conditional expectation, each $M_{n,m}$ is $\mathcal{F}_m$-measurable and therefore by Lemma 2.14 we know that M_m is $\mathcal{F}_m$-measurable showing M_m is $\mathcal{F}$-adapted. Lastly applying the monotone convergence property of conditional expectation and the tower rule for conditional expectation we get

$$\mathbf{E}[M_{m+1} \mid \mathcal{F}_m] = \lim_{n\to\infty} \mathbf{E}[\mathbf{E}[X_n \mid \mathcal{F}_{m+1}] \mid \mathcal{F}_m] = \lim_{n\to\infty} \mathbf{E}[X_n \mid \mathcal{F}_m] = M_m$$

which shows that M_m is indeed a non-negative martingale. L^1-boundedness of M_m follows from the argument that showed M_m was integrable.

Now define $A_n = M_n - X_n$ and note that

$$A_n = \lim_{m\to\infty} \mathbf{E}[X_m \mid \mathcal{F}_n] - X_n \geq 0 \text{ a.s.}$$

by the submartingale property of X_n. To see that A_n is an L^1-bounded supermartingale, note that integrability and L^1-boundedness of A_n follow by the triangle inequality and the corresponding properties of X_n and M_n, $\mathcal{F}$-adaptedness follows from the $\mathcal{F}$-adaptedness of X_n and M_n and the supermartingale property follows using the submartingale and martingale properties of X_n and M_n respectively

$$\mathbf{E}[A_{n+1} \mid \mathcal{F}_n] = \mathbf{E}[M_{n+1} \mid \mathcal{F}_n] - \mathbf{E}[X_{n+1} \mid \mathcal{F}_n] \leq M_n - X_n = A_n$$

□

LEMMA 9.43. *Let M_n be an L^1-bounded martingale then there exist non-negative martingales Y_n^+ and Y_n^- such that $M_n = Y_n^+ - Y_n^-$ a.s. and $\|Y_n^\pm\|_1 \leq \|M_n\|_1$.*

PROOF. This is a corollary of the proof of Lemma 9.42. If we apply that construction to each of the submartingales $M_n^\pm$ we get that $Y_n^\pm = \lim_{m\to\infty} \mathbf{E}[M_m^\pm \mid \mathcal{F}_m]$ defines a pair of nonnegative martingales. By linearity of conditional expectation and the martingale property of M_n we see that

$$Y_n^+ - Y_n^- = \lim_{m\to\infty} \mathbf{E}\left[M_m^+ - M_m^- \mid \mathcal{F}_n\right] = M_n^+ - M_n^- = M_n \text{ a.s.}$$

□

3.1. 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. Probably the most important such property is that boundedness of a martingale or submartingale implies convergence (analogous to the fact that a bounded increasing sequence in $\mathbb{R}$ must converge).

There are several fundamental inequalities that describe these intuitions 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 5.17) for a special class of dependent random variables. The reader may want to take a look at the proof of the Kolmogorov Maximal Inequality to see that the critical property that the proof relied on was the fact that the square of a random walk is a submartingale; beyond that independence was not utilized. Indeed the proof of the following maximal inequality is essentially the same as our proof of the Kolmogorov Maximal Inequality.

LEMMA 9.44 (Doob's Maximal Inequality). *Let X_t be a submartingale on a countable index set T, then for every $\lambda > 0$ and $t \in T$,*

$$\lambda \boldsymbol{P}\{\sup_{s \le t} X_s \ge \lambda\} \le \boldsymbol{E}\left[X_t; \sup_{s\le t} X_s \ge \lambda\right] \le \boldsymbol{E}\left[X_t^+\right]$$

where $X_t^+ = X_t \vee 0$.

PROOF. First we assume that T is a finite set. By reindexing we may as well assume that $T = \{0, \dots, n\}$ for some $n \ge 0$. Now pick an $n \in T$. The first thing to note is that for any submartingale X_n, $n \ge m$ and $A_m \in \mathcal{F}_m$,

$$\mathbf{E}\left[X_n; A_m\right] = \mathbf{E}\left[\mathbf{E}\left[X_n \mid \mathcal{F}_m\right]; A_m\right] \ge \mathbf{E}\left[X_m; A_m\right] \tag{5}$$

Now we can use (5) together with a decomposition of the event $\{\sup_{1\le k\le n} X_k \ge \lambda\}$ as the disjoint union of the events in which the first of the X_k equals or exceeds λ,

$$\begin{aligned}
\lambda \mathbf{P}\{\sup_{0\le k\le n} X_k \ge \lambda\} &= \lambda \mathbf{P}\{\cup_{k=0}^n \{X_0 < \lambda; \cdots; X_{k-1} < \lambda; X_k \ge \lambda\}\} \\
&= \lambda \sum_{k=0}^n \mathbf{P}\{X_0 < \lambda; \cdots; X_{k-1} < \lambda; X_k \ge \lambda\} \\
&\le \sum_{k=0}^n \mathbf{E}\left[X_k; X_0 < \lambda; \cdots; X_{k-1} < \lambda; X_k \ge \lambda\right] \\
&\le \sum_{k=0}^n \mathbf{E}\left[X_n; X_0 < \lambda; \cdots; X_{k-1} < \lambda; X_k \ge \lambda\right] \\
&= \mathbf{E}\left[X_n; \sup_{0\le k\le n} X_k \ge \lambda\right]
\end{aligned}$$

The above argument may be nicely rephrased in terms of optional times; it is a good exercise in the language and tools we have developed so we provide a second proof. Define

$$\tau = \min\{n \mid X_n \ge \lambda\}$$

where we assume the minimum of the empty set is positive infinity. Note that $\{\sup_{0\le k\le n} X_k \ge \lambda\} = \{\tau \le n\}$. and if we consider the stopped process $X_{\tau\wedge n}$ we have

$$X_{\tau\wedge n} \mathbf{1}_{\tau\le n} = \sum_{k=0}^n X_{\tau \wedge n} \mathbf{1}_{\tau = k} = \sum_{k=0}^n X_k \mathbf{1}_{\tau=k} \ge \lambda \sum_{k=0}^n \mathbf{1}_{\tau = k} = \lambda \mathbf{1}_{\tau \le n}$$

Use these two facts, the Optional Sampling Theorem 9.41 and the fact that $\{\tau \leq t\} \in \mathcal{F}_{\tau \wedge t}$ (Corollary 9.33) we get

$$\lambda \mathbf{P}\{\sup_{0 \leq k \leq n} X_k \geq \lambda\} = \lambda \mathbf{E}\left[\mathbf{1}_{\tau \leq n}\right] \leq \mathbf{E}\left[X_{\tau \wedge n} \mathbf{1}_{\tau \leq n}\right] \leq \mathbf{E}\left[\mathbf{E}\left[X_n \mid \mathcal{F}_{\tau \wedge n}\right] \mathbf{1}_{\tau \leq n}\right]$$
$$= \mathbf{E}\left[X_n \mathbf{1}_{\tau \leq n}\right] = \mathbf{E}\left[X_n; \sup_{0 \leq k \leq n} X_k \geq \lambda\right]$$

The second inequality is true because nonnegativity of X_n^+ implies

$$0 \leq X_n \mathbf{1}_{\sup_{0 \leq k \leq n} X_k \geq \lambda} \leq X_n^+$$

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

$$\{\sup_{\substack{s \leq t \\ s \in T'}} X_s \geq \lambda\} \subset \{\sup_{\substack{s \leq t \\ s \in T''}} X_s \geq \lambda\}$$

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 2.30) we have

$$\mathbf{P}\{\sup_{\substack{s \leq t \\ s \in T}} X_s \geq \lambda\} = \lim_{m \to \infty} \mathbf{P}\{\sup_{\substack{s \leq t \\ s \in T_m}} X_s \geq \lambda\}$$

and by the integrability of X_t and the bound $\left| X_t \mathbf{1}_{\sup_{\substack{s \leq t \\ s \in T}} X_s \geq \lambda} \right| \leq |X_t|$ we can apply Dominated Convergence to conclude

$$\mathbf{E}\left[X_t; \sup_{\substack{s \leq t \\ s \in T}} X_s \geq \lambda\right] = \lim_{m \to \infty} \mathbf{E}\left[X_t; \sup_{\substack{s \leq t \\ s \in T_m}} X_s \geq \lambda\right]$$

proving the result for countable T. ☐

A lesser known inequality is

Lemma 9.45 (Doob's Minimal Inequality). *Let X_t be a submartingale on a countable index set T, then for every every interval $[s, t] \subset T$ and $\lambda > 0$,*

$$\lambda \boldsymbol{P}\{\inf_{s \leq q \leq t} X_q \leq -\lambda\} \leq \boldsymbol{E}\left[X_t^+\right] - \boldsymbol{E}\left[X_s\right]$$

where $X_t^+ = X_t \vee 0$.

Proof. We start by assuming that T is finite and as in the proof of the Maximal inequality we assume that $T = \{0, \ldots, n\}$. Let $\tau = \min\{k \mid X_k \leq -\lambda\}$ be the hitting time for the interval $(-\infty, -\lambda]$ and note that it is an optional time. Furthermore we have by this definition $\{\min_{0 \leq k \leq n} X_k \leq -\lambda\} = \{\tau \leq n\}$. By Optional Sampling Theorem 9.38 we know that

$$\mathbf{E}\left[X_0\right] \leq \mathbf{E}\left[X_{\tau \wedge n}\right]$$

We write $X_{\tau \wedge n} = X_\tau \mathbf{1}_{\tau \leq n} + X_n \mathbf{1}_{\tau > n}$ and note that $X_n \mathbf{1}_{\tau > n} \leq X_n^+ \mathbf{1}_{\tau > n} \leq X_n^+$. Putting these facts together,

$$\begin{aligned}
\mathbf{E}\left[X_0\right] &\leq \mathbf{E}\left[X_{\tau \wedge n}\right] \\
&= \mathbf{E}\left[X_\tau ; \tau \leq n\right] + \mathbf{E}\left[X_n ; \tau > n\right] \\
&\leq -\lambda \mathbf{P}\{\tau \leq n\} + \mathbf{E}\left[X_n^+\right]
\end{aligned}$$

and the result is proven.

TODO: Extend from finite to countable as in the proof of the Maximal Inequality Lemma 9.44. □

Having proven a tail inequality it is often a good idea to see what it might say about expectations via Lemma 3.8. 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 9.46 (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 $t \in T$,*

$$\left\| \sup_{s \leq t} X_s \right\|_p \leq \frac{p}{p-1} \left\| X_t \right\|_p$$

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 $\mathbf{E}\left[(\sup_{0 \leq k \leq n} X_k)^p\right] < \infty$. With this assumption in place we can apply Lemma 3.8, the Maximal Inequality Lemma 9.44 and Tonelli's Theorem 2.88 to get

$$\begin{aligned}
\mathbf{E}\left[(\sup_{0 \leq k \leq n} X_k)^p\right] &= p \int_0^\infty \lambda^{p-1} \mathbf{P}\{\sup_{0 \leq k \leq n} X_k \geq \lambda\}\, d\lambda \\
&\leq p \int_0^\infty \lambda^{p-2} \mathbf{E}\left[X_n ; \sup_{0 \leq k \leq n} X_k \geq \lambda\right] d\lambda \\
&= p \mathbf{E}\left[X_n \int_0^\infty \lambda^{p-2} \mathbf{1}_{\sup_{0 \leq k \leq n} X_k \geq \lambda}\, d\lambda\right] \\
&= p \mathbf{E}\left[X_n \int_0^{\sup_{0 \leq k \leq n} X_k} \lambda^{p-2}\, d\lambda\right] \\
&= \frac{p}{p-1} \mathbf{E}\left[X_n (\sup_{0 \leq k \leq n} X_k)^{p-1}\right] \\
&\leq \frac{p}{p-1} \left\| X_n \right\|_p \mathbf{E}\left[(\sup_{0 \leq k \leq n} X_k)^p\right]^{\frac{p-1}{p}} \qquad \text{by Hölder's Inequality}
\end{aligned}$$

But now, we can divide both sides by $\mathbf{E}\left[(\sup_{0 \leq k \leq n} X_k)^p\right]^{\frac{p-1}{p}}$ to get the result.

It remains to remove the assumption that $\mathbf{E}\left[(\sup_{0 \leq k \leq n} X_k)^p\right] < \infty$. Obviously if $\left\| X_n \right\|_p = \infty$ then the result is trivially true so we may assume that $\left\| X_n \right\|_p < \infty$. Now we have for all $k \leq n$, by the submartingale property, Jensen's Inequality (Theorem 8.36) and the tower rule for conditional expectation

$$\mathbf{E}\left[X_k^p\right] \leq \mathbf{E}\left[\mathbf{E}\left[X_n \mid \mathcal{F}_k\right]^p\right] \leq \mathbf{E}\left[\mathbf{E}\left[X_n^p \mid \mathcal{F}_k\right]\right] = \mathbf{E}\left[X_n^p\right] < \infty$$

which shows that $\|X_k\|_p < \infty$ for all $0 \leq k \leq n$. But this implies that $\left\|\sup_{0\leq k\leq n} X_k\right\|_p < \infty$ (e.g. for any $\xi, \eta \in L^p$, write $\xi \vee \eta = \xi \mathbf{1}_{\xi > \eta} + \eta \mathbf{1}_{\xi \leq \eta}$ 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

$$\sup_{\substack{s\leq t\\ s\in T'}} X_s \leq \sup_{\substack{s\leq t\\ s\in T''}} X_s$$

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

$$\mathbf{E}\left[\sup_{\substack{s\leq t\\ s\in T}} X_s\right] = \lim_{n\to\infty} \mathbf{E}\left[\sup_{\substack{s\leq t\\ s\in T_n}} X_s\right] \leq \frac{p}{p-1} \|X_t\|_p$$

□

It is worthwhile emphasizing that the results above cover the case in which $\infty \in T$.

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 9.47. *Let $\mathcal{F}_n$ be a filtration, M_n be a $\mathcal{F}$- adapted process on $\mathbb{Z}_+$ and let $a < b$ be real numbers. Let $\tau_0 = 0$ and for each $j \geq 0$ define inductively*

$$\begin{aligned}\sigma_j &= \inf\{n \mid t \geq \tau_j \text{ and } M_n \leq a\}\\ \tau_{j+1} &= \inf\{n \mid t \geq \sigma_j \text{ and } M_n \geq b\}\end{aligned}$$

then each τ_j and σ_j is an $\mathcal{F}$-optional time (note that we treat the infimum of the empty set to be infinity). Furthermore if we define

$$\begin{aligned}&U_a^b(n) = \sup\{m \mid \tau_m \leq n\}\\ &= \sup\{m \mid \exists j_1 < k_1 < \cdots < j_m < k_m \leq n \text{ such that } X_{j_i} \leq a \text{ and } X_{k_i} \geq b \text{ for all } i = 1, \dots, m\}\end{aligned}$$

to be the number of upcrossings of X_m before n, then each $U_a^b(n)$ is $\mathcal{F}_n$-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

$$\{\sigma_n = m\} = \bigcup_{k<m} \left(\{\tau_n = k\} \cap \bigcap_{k<l<m} \{X_l > a\} \right) \cap \{X_m \leq a\}$$

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

$$\{\tau_{n+1} = m\} = \bigcup_{k<m} \left(\{\sigma_n = k\} \cap \bigcap_{k<l<m} \{X_l < b\} \right) \cap \{X_m \geq b\}$$

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 the $\mathcal{F}_n$-measurability of $U_a^b(n)$ we just express for $n \in \mathbb{Z}_+$

$$\{U_a^b(n) = m\} = \{\tau_m \leq n\} \cap \bigcap_{k>m} \{\tau_k > n\}$$

and

$$\{U_a^b(n) = \infty\} = \cap_{m=1}^\infty \{\tau_m \leq n\}$$

both of which are $\mathcal{F}_n$-measurable because we have just shown each τ_m is an optional time.

To see the last equality let the $\tilde{U}_a^b(n)$ be the right hand side of the equality to be shown. First note that if $\tau_m \leq n$ then taking $j_i = \sigma_{i-1}$ and $k_i = \tau_i$ we get an upcrossing sequence $j_1 < k_1 < \ldots < j_m < k_m \leq n$; therefore $U_a^b(n) \leq \tilde{U}_a^b(n)$. On the other hand, given such an upcrossing sequence we claim that this implies $\sigma_i \leq j_i$ and $\tau_i \leq k_i$ for $i = 1, \ldots, m$ so in particular, $\tau_m \leq n$. This follows from an induction argument that has two cases. First if $\tau_{i-1} \leq k_i$ and $X_{j_i} \leq a$ then we clearly see $\sigma i \leq j_i$. On the other hand if $\sigma_i \leq j_i$ then we clearly see that $\tau_i \leq k_i$. From $\tau_m \leq n$ it follows that $\tilde{U}_a^b(n) \leq U_a^b(n)$ so the desired equality is proven. □

The second definition of $U_a^b(n)$ provided in the previous result generalizes nicely to arbitrary time indexes; in particular for countable time indexes we get a workable definition and measurability.

Corollary 9.48. *Let $\mathcal{F}_t$ be a filtration, M_t be a $\mathcal{F}$- adapted process with a countable time index T and let $a < b$ be real numbers. If we define*

$$U_a^b(t)$$
$$= \sup\{m \mid \exists j_1 < k_1 < \cdots < j_m < k_m \leq t \text{ such that } X_{j_i} \leq a \text{ and } X_{k_i} \geq b \text{ for all } i = 1, \ldots, m\}$$

to be the number of upcrossings of X_s before t, then each $U_a^b(t)$ is $\mathcal{F}_t$-measurable.

Proof. The previous result shows the $\mathcal{F}_t$-measurability for finite time indexes T that contain t. Now write $T = \cup_{n=0}^\infty T_n$ where $t \in T_0 \subset T_1 \subset \cdots$ is a nested sequence of finite sets. It is easy to see that $U(t, a, b, T) = \lim_{n\to\infty} U(t, a, b, T_n)$ and therefore the result follows from Lemma 9.47 and Lemma 2.14. □

Lemma 9.49 (Doob's Upcrossing Inequality). *Let X_t be a submartingale with a countable time index T and let $U_a^b(t)$ be the number of upcrossings up to time $t \in T$. Then*

$$\boldsymbol{E}\left[U_a^b(t)\right] \leq \frac{\boldsymbol{E}\left[(X_t - a)_+\right]}{b - a}$$

Proof. As the first reduction note that we may assume that T is in fact finite. To see why let us temporarily change notation to make the dependence of $U_a^b(t)$ on the time index T explicit by writing $U(t, a, b, T)$. As noted in the proof of Corollary 9.48 if we consider a nested set of finite time indexes $t \in T_0 \subset T_1 \subset \cdots$ such that

$T = \cup_{n=0}^{\infty} T_n$ then in fact $\lim_{n\to\infty} U(t,a,b,T_n) = U(t,a,b,T)$. Now by Monotone Convergence we get $\lim_{n\to\infty} \mathbf{E}\left[U(t,a,b,T_n)\right] = \mathbf{E}\left[U(t,a,b,T)\right]$ and the result for T will follow from the result for finite time indexes.

The second 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 from Proposition 9.24 that $(X_t - a)_+$ is a positive submartingale. Furthermore $X_t \geq b$ if and only if $(X_t - a)_+ \geq b - a$ and $X_t \geq a$ if and only if $(X_t - a)_+ = 0$ and therefore the number of upcrossings of X_t between a and b is the same as the number of upcrossings of $(X_t - a)_+$ between 0 and $b - a$. Therefore the result is proven if we show that for every positive submartingale X_t and $b > 0$ we have

$$U_0^b(t) \leq \frac{\mathbf{E}\left[X_t\right]}{b}$$

To finish the proof, let n be the cardinality of T so that we know $\sigma_n = \tau_n = \infty$ and we can write the finite telescoping sum

$$X_t = X_{\tau_0 \wedge t} + \sum_{j=0}^{n} \left(X_{\sigma_j \wedge t} - X_{\tau_j \wedge t} \right) + \sum_{j=0}^{n} \left(X_{\tau_{j+1} \wedge t} - X_{\sigma_j \wedge t} \right)$$

Taking expectations we note that from the positivity of X_t we have $\mathbf{E}\left[X_{\tau_0 \wedge t}\right] \geq 0$ and because $\sigma_j \geq \tau_j$ and the optional sampling theorem for submartingles (Corollary 9.41) we have

$$\mathbf{E}\left[X_{\sigma_j \wedge t} - X_{\tau_j \wedge t}\right] = \mathbf{E}\left[\mathbf{E}\left[X_{\sigma_j \wedge t} - X_{\tau_j \wedge t} \mid \mathcal{F}_{\tau_j \wedge t}\right]\right] \geq \mathbf{E}\left[X_{\tau_j \wedge t} - X_{\tau_j \wedge t}\right] = 0$$

and we also have

$$\begin{aligned} &X_{\tau_{j+1} \wedge t} - X_{\sigma_j \wedge t} \geq b && \text{if } \tau_{j+1} \leq n \\ &X_{\tau_{j+1} \wedge t} - X_{\sigma_j \wedge t} = X_n \geq 0 && \text{if } \sigma_j \leq n < \tau_{j+1} \\ &X_{\tau_{j+1} \wedge t} - X_{\sigma_j \wedge t} = 0 && \text{if } n < \sigma_j \end{aligned}$$

so by considering only the terms in sum for which $\tau_{j+1} \leq n$ we get $\sum_{j=0}^{n} \left(X_{\tau_{j+1} \wedge t} - X_{\sigma_j \wedge t} \right) \geq bU_0^b(n)$. Putting this all together

$$\begin{aligned} \mathbf{E}\left[X_n\right] &= \mathbf{E}\left[X_{\tau_0 \wedge t}\right] + \sum_{j=0}^{n} \mathbf{E}\left[X_{\sigma_j \wedge t} - X_{\tau_j \wedge t}\right] + \sum_{j=0}^{n} \mathbf{E}\left[X_{\tau_{j+1} \wedge t} - X_{\sigma_j \wedge t}\right] \\ &\geq \sum_{j=0}^{n} \mathbf{E}\left[X_{\tau_{j+1} \wedge t} - X_{\sigma_j \wedge t}\right] \\ &\geq b\mathbf{E}\left[U_0^b(n)\right] \end{aligned}$$

and therefore the result is proved. $\square$

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 since one is expecting $X_{\sigma_j} \leq a < b \leq X_{\tau_j}$. 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 moreover $X_{\sigma_j \wedge n} = X_n \geq X_{\tau_j \wedge n}$ and not $X_{\sigma_j \wedge n} = X_{\sigma_j} \leq a$. 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 $(b-a)U_a^b(n)$).

The Upcrossing Lemma leads immediately to a proof that L^1-bounded submartingales converge almost surely. This result is usually stated for discrete submartingales X_n but with a little attention to details we get a stronger result that applies over countable time indexes (e.g. $\mathbb{Q}_+$) and paves the way for consideration of continuous time indexes such as $\mathbb{R}_+$.

Theorem 9.50 (Submartingale Convergence Theorem). *Let X_t be a $\mathcal{F}$-submartingale with a countable time index T such that $\sup_{t\in T}\|X_t\|_1 < \infty$ then there exists an $A \in \mathcal{F}_\infty$ with $\boldsymbol{P}\{A\} = 1$ such that for every increasing or decreasing sequence t_n in T there exists an integrable random variable X such that $X_{t_n} \to X$ on A (so in particular $X_{t_n} \overset{a.s.}{\to} X$).*

Proof. The first order of business here is leverage the Doob Upcrossing Inequality to show that X_t is not oscillatory almost surely 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 $\mathbf{E}\left[(X_t - a)_+\right] \leq \sup_{t\in T}\|X_t\|_1 + |a| < \infty$. Supposing $a, b \in \mathbb{R}$ with $a < b$ and $U_a^b(t)$ be the number of upcrossings of $[a,b]$ before t, we can see that $U_a^b(t)$ is positive and increasing in t and Lemma 9.49 and Monotone Convergence tell us that if we pick any sequence $t_1, t_2, \dots$ such that $\lim_{n\to\infty} t_n = \sup T$ then

$$\mathbf{E}\left[\lim_{n\to\infty} U_a^b(t_n)\right] = \lim_{n\to\infty}\mathbf{E}\left[U_a^b(t_n)\right] \leq \lim_{n\to\infty}\frac{\|X_{t_n}\|_1 + |a|}{b-a} \leq \frac{\sup_{t\in T}\|X_t\|_1 + |a|}{b-a} < \infty$$

If we let $U_a^b(\infty) = \lim_{n\to\infty} U_a^b(t_n) = \sup_{t\in T} U_a^b(t)$ be the number of upcrossing on T, then $U_a^b(\infty)$ is $\mathcal{F}_\infty$-measurable by Lemma 2.14, $U_a^b(\infty)$ is integrable and therefore almost surely finite.

Let $A = \cap_{\substack{a<b \\ a,b\in\mathbb{Q}}}\{U_a^b(\infty) < \infty\}$ which is a countable intersection of $\mathcal{F}_\infty$-measurable sets of probability one hence is a $\mathcal{F}_\infty$-measurable set of probability one. Let t_n be any increasing or decreasing sequence in T. For each $a, b \in \mathbb{Q}$ with $a < b$ define

$$\Lambda_a^b = \{\liminf_{n\to\infty} X_{t_n} < a < b < \limsup_{n\to\infty} X_{t_n}\}$$

and note that $\Lambda_a^b \subset \{U_a^b(\infty) = \infty\}$ (we can pick subsequences N and M such that X_{t_n} converges to $\liminf_{n\to\infty} X_{t_n}$ along N and $\limsup_{n\to\infty} X_{t_n}$ along M and in this way construct an infinite number of upcrossings of $[a,b]$; it is here that we require that the sequence t_n is increasing or decreasing). Thus

$$\begin{aligned}
\{\liminf_{n\to\infty} X_{t_n} < \limsup_{n\to\infty} X_{t_n}\} &= \bigcup_{\substack{a,b\in\mathbb{Q} \\ a<b}} \Lambda_a^b \\
&\subset \bigcup_{\substack{a,b\in\mathbb{Q} \\ a<b}} \{U_a^b(\infty) = \infty\} \\
&= A^c
\end{aligned}$$

and therefore $\lim_{n\to\infty} X_{t_n}$ exists on the set A (in particular almost surely since $\mathbf{P}\{A^c\} = 0$).

Let $X = \lim_{n\to\infty} X_{t_n}$ on A and for concreteness define it to be 0 on A^c. Our last task is to show that X is integrable (hence almost surely finite as well). This follows from Fatou's Lemma

$$\mathbf{E}[|X|] \leq \liminf_{n\to\infty} \mathbf{E}[|X_{t_n}|] \leq \sup_{t\in T} \|X_t\|_1 < \infty$$

and we are done. $\square$

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 9.17 for the definition of closed martingales).

Theorem 9.51 (Martingale Closure Theorem). *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 $X_n = \boldsymbol{E}[X \mid \mathcal{F}_n]$ almost surely.*

Proof. To see (i) implies (ii) we know from Lemma 5.52 that X_n uniformly integrable implies L^1 boundedness, hence we can apply Theorem 9.50 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 5.5) which together with uniform integrability implies $X_n \xrightarrow{L^1} X$ (Lemma 5.58).

To that (ii) implies (iii) suppose that $\epsilon > 0$ is given and let $N > 0$ be such that $\|X_n - X\|_1 = \mathbf{E}[|X_n - X|] < \epsilon$ 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

$$\begin{aligned}
|\mathbf{E}[X; A] - \mathbf{E}[X_m; A]| &= |\mathbf{E}[X; A] - \mathbf{E}[X_n; A]| \qquad \text{since } \mathbf{E}[X_n \mid \mathcal{F}_m] = X_m \\
&\leq \mathbf{E}[|X - X_n|; A] \\
&\leq \mathbf{E}[|X - X_n|] < \epsilon
\end{aligned}$$

and since ϵ is arbitrary, we conclude $\mathbf{E}[X; A] = \mathbf{E}[X_m; A]$ and therefore $\mathbf{E}[X \mid \mathcal{F}_m] = X_m$ a.s.

To see that (ii) implies (iii), we use Lemma 5.52. First note that by contraction property of conditional expectation, we have $\sup_n \mathbf{E}[|\mathbf{E}[X \mid \mathcal{F}_n]|] \leq \mathbf{E}[|X|]$ so the first condition of the lemma holds. To see the second condition, let $\epsilon > 0$ be fixed and pick $R > 0$ such that $\mathbf{E}[|X|; |X| > R] < \frac{\epsilon}{2}$ and pick A such that $\mathbf{P}\{A\} < \frac{\epsilon}{2R}$. Now, for every n,

$$\begin{aligned}
|\mathbf{E}[\mathbf{E}[X \mid \mathcal{F}_n]; A]| &\leq \mathbf{E}[\mathbf{E}[|X| \mid \mathcal{F}_n]; A] \\
&= \mathbf{E}[|X| \cdot \mathbf{E}[\mathbf{1}_A \mid \mathcal{F}_n]] \\
&= \mathbf{E}[|X| \cdot \mathbf{E}[\mathbf{1}_A \mid \mathcal{F}_n]; |X| \leq R] + \mathbf{E}[|X| \cdot \mathbf{E}[\mathbf{1}_A \mid \mathcal{F}_n]; |X| > R] \\
&\leq R\mathbf{E}[\mathbf{E}[\mathbf{1}_A \mid \mathcal{F}_n]] + \mathbf{E}[|X|; |X| > R] \\
&\leq \epsilon
\end{aligned}$$

and therefore we have condition (ii) of Lemma 5.52 satisfied and uniform integrability is shown. $\square$

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 9.52. *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.

Just for grins here is the proof that Kallenberg gives that is very similar up to a point to the proof in the previous result but instead of using the uniform integrability of ξ to make the elementary argument invokes some standard workhorse theorems. The resulting argument seems to me to be more difficult to understand. Maybe there is a problem with my argument but I don't see it. He says just as we do that for any $\mathcal{F}$,

$$|\mathbf{E}[\mathbf{E}[\xi \mid \mathcal{F}];A]| \leq \mathbf{E}[\mathbf{E}[|\xi| \mid \mathcal{F}];A] = \mathbf{E}[|\xi| \cdot \mathbf{E}[\mathbf{1}_A \mid \mathcal{F}]]$$

Now he observes that if the right hand side doesn't converge to zero uniformly in $\mathcal{F}$ as $\mathbf{P}\{A\} \to 0$ then there exists an $\epsilon > 0$, σ-algebras $\mathcal{F}_n$ and A_n with $\lim_{n\to\infty} \mathbf{P}\{A_n\} = 0$ such that

$$\mathbf{E}[|\xi| \cdot \mathbf{E}[\mathbf{1}_{A_n} \mid \mathcal{F}_n]] \geq \epsilon \text{ for all } n$$

so in particular no subsequence can converge to zero. Now we derive a contradiction. We know that $\mathbf{E}[\mathbf{E}[\mathbf{1}_{A_n} \mid \mathcal{F}_n]] = \mathbf{P}\{A_n\} \to 0$ and therefore $\mathbf{E}[\mathbf{1}_{A_n} \mid \mathcal{F}_n] \xrightarrow{P} 0$ (Lemma 5.6) and $\mathbf{E}[\mathbf{1}_{A_n} \mid \mathcal{F}_n] \xrightarrow{a.s.} 0$ along some subsequence N (Lemma 5.10). Now since ξ is integrable it is almost surely finite and therefore $|\xi|\,\mathbf{E}[\mathbf{1}_{A_n} \mid \mathcal{F}_n] \xrightarrow{a.s.} 0$ along the subsequence N and by Dominated Convergence we get $\mathbf{E}[|\xi| \cdot \mathbf{E}[\mathbf{1}_{A_n} \mid \mathcal{F}_n]] \to 0$ along N which is a contradiction. ☐

Convergence of martingales in L^p spaces with $p > 1$ is equivalent to boundedness. An even stronger condition holds, if a martingale converges in L^1 to a p-integrable limit then the convergence can be upgraded to L^p convergence.

THEOREM 9.53 (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. In fact, if $M_n \xrightarrow{L^1} M$ with $M \in L^p$ then $M_n \xrightarrow{L^p} M$ and 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 9.51 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

$$\left\| \sup_{0\leq k\leq n} |M_k| \right\|_p \leq \frac{p}{p-1} \left\|M_n\right\|_p < \frac{p}{p-1} \sup_n \left\|M_n\right\|_p < \infty$$

therefore by Monotone Convergence we have $\left\|\sup_{0\leq k\leq\infty} |M_k|\right\|_p = \lim_{n\to\infty} \left\|\sup_{0\leq k\leq n} |M_k|\right\|_p < \infty$. 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 assume that $M_n \stackrel{L^1}{\to} M$ with $M \in L^p$ and $p > 1$. Theorem 9.51 implies that $M_n = \mathbf{E}[M \mid \mathcal{F}_n]$ a.s. for every n. Now convexity of x^p for $p > 1$ and Jensen's Inequality (Theorem 8.36) imply

$$\mathbf{E}[|M_n|^p] = \mathbf{E}[|\mathbf{E}[M \mid \mathcal{F}_n]|^p] \leq \mathbf{E}[\mathbf{E}[|M|^p \mid \mathcal{F}_n]] = \|M\|_p^p < \infty$$

which shows that not only is M_n p-integrable but that the martingale M_n is L^p-bounded. The first part of the Theorem shows that $M_n \stackrel{L^p}{\to} M$. □

Martingale convergence also allows us to extend the optional sampling theorem to unbounded optional times.

Lemma 9.54. *Let M_n be a uniformly integrable martingale and let σ and τ be optional times, then M_τ is integrable and $\boldsymbol{E}[M_\tau \mid \mathcal{F}_\sigma] = M_{\sigma\wedge\tau}$.*

Proof. To see integrability of M_τ we use the Martingale Convergence Theorem 9.51 to conclude that there exists integrable M_∞ such that $M_n = \mathbf{E}[M_\infty \mid \mathcal{F}_n]$. By Lemma 8.14 and Lemma 9.31 for every n we can compute

$$M_\tau = M_n = \mathbf{E}[M_\infty \mid \mathcal{F}_n] = \mathbf{E}[M_\infty \mid \mathcal{F}_\tau] \text{ on } \{\tau = n\}$$

and therefore $M_\tau = \mathbf{E}[M_\infty \mid \mathcal{F}_\tau]$ proving integrability. Note that this was proven for arbitrary optional times so in particular $M_{\tau\wedge\sigma}$ is integrable as well.

To show the optional sampling equality we first observe by the result in the bounded case that for every n, $\mathbf{E}[M_{\tau\wedge n} \mid \mathcal{F}_\sigma] = M_{\sigma\wedge\tau\wedge n}$ and we just need to justify taking limits in the equality. Pick $A \in \mathcal{F}_\sigma$. We know that $M_n \stackrel{a.s.}{\to} M_\infty$ as well and therefore we have $M_{\tau\wedge n}\mathbf{1}_A \stackrel{a.s.}{\to} M_\tau \mathbf{1}_A$ and $M_{\tau\wedge\sigma\wedge n}\mathbf{1}_A \stackrel{a.s.}{\to} M_{\tau\wedge\sigma}\mathbf{1}_A$. To show

$$\mathbf{E}[M_\tau \mathbf{1}_A] = \lim_{n\to\infty} \mathbf{E}[M_{\tau\wedge n}\mathbf{1}_A] = \lim_{n\to\infty} \mathbf{E}[M_{\tau\wedge\sigma\wedge n}\mathbf{1}_A] = \mathbf{E}[M_{\tau\wedge\sigma}\mathbf{1}_A]$$

it will suffice to show that $M_{\tau\wedge n}$ is uniformly integrable for an arbitrary optional time τ. Suppose $\epsilon > 0$ is given. By the integrability of M_τ we can find $R_1 > 0$ such that $\mathbf{E}[|M_\tau| ; |M_\tau| > R_1] < \epsilon/2$ and by uniform integrability of M_n we can find $R_2 > 0$ such that $\sup_n \mathbf{E}[|M_n| ; |M_n| > R_2] < \epsilon/2$. Now let $R = R_1 \vee R_2$ and compute

$$\begin{aligned}
\sup_n \mathbf{E}[|M_{\tau\wedge n}| ; |M_{\tau\wedge n}| > R] &= \sup_n \mathbf{E}[|M_{\tau\wedge n}| ; |M_{\tau\wedge n}| > R \text{ and } \tau \leq n] + \\
&\quad \sup_n \mathbf{E}[|M_{\tau\wedge n}| ; |M_{\tau\wedge n}| > R \text{ and } \tau > n] \\
&\leq \mathbf{E}[|M_\tau| ; |M_\tau| > R] + \sup_n \mathbf{E}[|M_n| ; |M_n| > R] \\
&< \epsilon
\end{aligned}$$

□

Corollary 9.55. *Let M_n be a uniformly integrable martingale, then the set of random variables $\{M_\tau \mid \tau \text{ is an optional time}\}$ is uniformly integrable.*

Proof. By uniform integrability there is M_∞ such that $M_n \to M_\infty$ a.s. and in L^1. By the previous result we have $M_\tau = \mathbf{E}[M_\infty \mid \mathcal{F}_\tau]$ and therefore the result follows from Corollary 9.52. □

We now give a result that we'll use in the transition to continuous time.

THEOREM 9.56. *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. If $\cdots \subset \mathcal{F}_{-1} \subset \mathcal{F}_0$ is a filtration then as $n \to -\infty$, $\boldsymbol{E}[\xi \mid \mathcal{F}_n]$ converges to $\boldsymbol{E}[\xi \mid \bigcap_n \mathcal{F}_n]$ both almost surely and in L^1.*

PROOF. First we take the unbounded above case. We know from the tower property of conditional expectation and Corollary 9.52 that $\mathbf{E}[\xi \mid \mathcal{F}_n]$ 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 $\mathbf{E}[\xi \mid \bigvee_n \mathcal{F}_n] = M$ 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 9.51 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

$$\begin{aligned}
\mathbf{E}[M; A] &= \mathbf{E}[\mathbf{E}[M \mid \mathcal{F}_n]; A] \\
&= \mathbf{E}[\mathbf{E}[\xi \mid \mathcal{F}_n]; A] \\
&= \mathbf{E}\left[\mathbf{E}\left[\mathbf{E}\left[\xi \mid \bigvee_n \mathcal{F}_n\right] \mid \mathcal{F}_n\right]; A\right] \\
&= \mathbf{E}\left[\mathbf{E}\left[\xi \mid \bigvee_n \mathcal{F}_n\right]; A\right]
\end{aligned}$$

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 8.8) we conclude that $M = \mathbf{E}[\xi \mid \bigvee_n \mathcal{F}_n]$ almost surely.

Now we treat the unbounded below case. As before we know that $M_n = \mathbf{E}[\xi \mid \mathcal{F}_n]$ is a uniformly integrable martingale. By Theorem 9.51 we know that there is an integrable $M_{-\infty}$ such that $\lim_{n \to -\infty} M_n = M_{-\infty}$ a.s. and by uniform integrability and Lemma 5.58 we know that the convergence is also in L^1. We need to show that $M_{-\infty} = \mathbf{E}[\xi \mid \bigcap_n \mathcal{F}_n]$ a.s. The first step is to observe that since $\mathcal{F}_n$ is a filtration $\bigcap_n \mathcal{F}_n$ is the tail σ-algebra and therefore $M_{-\infty}$ is $\bigcap_n \mathcal{F}_n$-measurable. If we let $A \in \bigcap_n \mathcal{F}_n$ then for all $n \leq 0$ we have $\mathbf{E}[\xi; A] = \mathbf{E}[M_n; A]$. Since M_n is uniformly integrable it follows that $M_n \mathbf{1}_A$ is uniformly integrable as well and therefore can conclude that $\mathbf{E}[\xi; A] = \lim_{n \to -\infty} \mathbf{E}[M_n; A] = \mathbf{E}[M_{-\infty}; A]$. The result is proven. □

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. However it is easy to rescue the situation by adapting 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 9.17 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 9.36) 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 2.30) and Montone Convergence.

Here is the result from Stroock We know from Lemma 9.17 that $\mathbf{E}[\xi \mid \mathcal{F}_n]$ is an $\mathcal{F}$-martingale. By Doob's Maximal Inequality (Lemma 9.44), the $\mathcal{F}_n$-measurability of $\{\sup_{0\leq k\leq n} \mathbf{E}[\xi \mid \mathcal{F}_k] \geq \lambda\}$ and another application of the tower property we know that

$$\begin{aligned}\lambda \mathbf{P}\{\sup_{0\leq k\leq n} \mathbf{E}[\xi \mid \mathcal{F}_k] \geq \lambda\} &\leq \mathbf{E}\left[\mathbf{E}[\xi \mid \mathcal{F}_n]\,; \sup_{0\leq k\leq n} \mathbf{E}[\xi \mid \mathcal{F}_k] \geq \lambda\right] \\ &= \mathbf{E}\left[\xi; \sup_{0\leq k\leq n} \mathbf{E}[\xi \mid \mathcal{F}_k] \geq \lambda\right]\end{aligned}$$

By continuity of measure (Lemma 2.30) we know that

$$\mathbf{P}\{\sup_k \mathbf{E}[\xi \mid \mathcal{F}_k] \geq \lambda\} = \lim_{n\to\infty} \mathbf{P}\{\sup_{0\leq k\leq n} \mathbf{E}[\xi \mid \mathcal{F}_k] \geq \lambda\}$$

and by Dominated Convergence

$$\mathbf{E}\left[\xi; \sup_k \mathbf{E}[\xi \mid \mathcal{F}_k] \geq \lambda\right] = \lim_{n\to\infty} \mathbf{E}\left[\xi; \sup_{0\leq k\leq n} \mathbf{E}[\xi \mid \mathcal{F}_k] \geq \lambda\right]$$

so we have shown

$$\lambda \mathbf{P}\{\sup_k \mathbf{E}[\xi \mid \mathcal{F}_k] \geq \lambda\} \leq \mathbf{E}\left[\xi; \sup_k \mathbf{E}[\xi \mid \mathcal{F}_k] \geq \lambda\right]$$

End result from Stroock

To show almost sure convergence, we let $\mathcal{G}$ denote the set of all integrable ξ such that $\mathbf{E}[\xi \mid \mathcal{F}_n] \stackrel{a.s.}{\to} \xi$. 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 8.5) 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 $\xi_n \stackrel{L^1}{\to} \xi$. We show that $\mathbf{E}[\xi \mid \mathcal{F}_n] \stackrel{a.s.}{\to} \xi$ by using Lemma 5.4. Suppose $\epsilon > 0$ is given, then for every m, n

$$\begin{aligned}\mathbf{P}\{\sup_{k\geq m} |\mathbf{E}[\xi \mid \mathcal{F}_k] - \xi| > \epsilon\} &\leq \mathbf{P}\{\sup_{k\geq m} |\mathbf{E}[\xi - \xi_n \mid \mathcal{F}_k]| > \frac{\epsilon}{3}\} + \\ &\quad \mathbf{P}\{\sup_{k\geq m} |\mathbf{E}[\xi_n \mid \mathcal{F}_k] - \xi_n| > \frac{\epsilon}{3}\} + \mathbf{P}\{|\xi_n - \xi| > \frac{\epsilon}{3}\} \\ &\leq \frac{6}{\epsilon}\mathbf{E}[|\xi - \xi_n|] + \mathbf{P}\{\sup_{k\geq m} |\mathbf{E}[\xi_n \mid \mathcal{F}_k] - \xi_n| > \frac{\epsilon}{3}\}\end{aligned}$$

where the first term is bounded by our claim at the beginning of proof applied to $\xi_n - \xi$ and the third term is bounded by the Markov Inequality (Lemma 10.1).

Taking the limit as m goes to infinity and using our assumption that $\xi_n \in \mathcal{G}$ and the characterization of almost sure convergence from Lemma 5.4 we see that $\lim_{m\to\infty} \mathbf{P}\{\sup_{k\geq m} |\mathbf{E}[\xi_n \mid \mathcal{F}_k] - \xi_n| > \frac{\epsilon}{3}\} = 0$. Therefore

$$\lim_{m\to\infty} \mathbf{P}\{\sup_{k\geq m} |\mathbf{E}[\xi \mid \mathcal{F}_k] - \xi| > \epsilon\} \leq \frac{6}{\epsilon}\mathbf{E}[|\xi - \xi_n|]$$

and by taking the limit as n goes to infinity we get

$$\lim_{m\to\infty} \mathbf{P}\{\sup_{k\geq m} |\mathbf{E}[\xi \mid \mathcal{F}_k] - \xi| > \epsilon\} = 0$$

so $\mathbf{E}[\xi \mid \mathcal{F}_n] \stackrel{a.s.}{\to} \xi$ by another application of Lemma 5.4.

Since we know that the family $\mathbf{E}[\xi \mid \mathcal{F}_n]$ is uniformly integrable by Corollary 9.52, $\mathbf{E}[\xi \mid \mathcal{F}_n] \stackrel{L^1}{\to} \xi$ follows from the almost sure convergence and Lemma 5.58. □

3.2. Martingale Central Limit Theorem. Many of the important classical results of probability theory are statements about i.i.d. sequences of random variables (e.g. the Law of Large Numbers and the Central Limit Theorem). In our presentation of the Lindeberg Central Limit Theorem we actually showed that the result holds under a weaker condition than identical distribution. It is a natural question to understand whether one can also relax the condition of independence and still have Gaussian convergence. Indeed this is true and the introduction of martingales was partly motivated by a desire to have a well defined dependence structure that could facilitate such investigations. Indeed one of the first appearances of the martingale condition (which predated the use of the term martingale) was in Levy's proof of a martingale central limit theorem. Our next task is to extend the Lindeberg Central Limit Theorem to the case of *martingale differences.* The proof we give uses characteristic functions but we note that with stronger hypotheses we can extend the proof technique of Theorem 6.1 to martingale differences as Levy did.

Definition 9.57. Let $\mathcal{F}_n$ be a filtration and let X_n be a sequence of random variables, then we say that X_n is a *martingale difference sequence* if and only if $M_n = \sum_{j=0}^n X_j$ is an $\mathcal{F}$-martingale.

Proposition 9.58. *X_n is a martingale difference sequence if and only if X_n is $\mathcal{F}$-adapted, each X_n is integrable and $\boldsymbol{E}[X_n \mid \mathcal{F}_{n-1}] = 0$ a.s. for all $n \in \mathbb{N}$.*

PROOF. Clearly if X_n is a martingale difference then we let $M_n = \sum_{j=0}^n X_j$ and write $X_n = M_n - M_{n-1}$ to see that X_n is $\mathcal{F}$-adapted and integrable. Moreover the martingale property of M_n show

$$\mathbf{E}\left[X_n \mid \mathcal{F}_{n-1}\right] = \mathbf{E}\left[M_n - M_{n-1} \mid \mathcal{F}_{n-1}\right] = \mathbf{E}\left[M_n \mid \mathcal{F}_{n-1}\right] - M_{n-1} = 0$$

On the other hand if X_n is $\mathcal{F}$-adapted, integrable and $\mathbf{E}\left[X_n \mid \mathcal{F}_{n-1}\right] = 0$ then if we define $M_n = \sum_{j=1}^n X_j$ it is easily seen that M_n is $\mathcal{F}$ adapted and integrable and moreover

$$\mathbf{E}\left[M_n \mid \mathcal{F}_{n-1}\right] = \mathbf{E}\left[X_n \mid \mathcal{F}_{n-1}\right] + \mathbf{E}\left[M_{n-1} \mid \mathcal{F}_{n-1}\right] = M_{n-1}$$

which shows that M_n is an $\mathcal{F}$-martingale. $\square$

Martingale difference sequences posses orthogonality properties that make them behave to sequence of independent random variables. Martingale convergence theorems can be viewed as laws of large numbers from this point of view. By comparison here we call out a simple fact that amounts to the statement about how martingale difference sequences are uncorrelated.

PROPOSITION 9.59. *Let X_n be a martingale difference sequence with $\boldsymbol{E}\left[X_n^2\right] < \infty$ then it follows that $\boldsymbol{E}[X_n X_m] = 0$ for $m \neq n$.*

PROOF. Assume $m < n$ then it follows that from Cauchy Schwartz that $\mathbf{E}\left[|X_m|\,|X_n|\right] \leq \mathbf{E}\left[X_m^2\right]^{1/2} \mathbf{E}\left[X_n^2\right]^{1/2} < \infty$ so $X_m X_n$ is integrable. Thus we can compute using conditional expectations and the martingale difference property

$$\mathbf{E}\left[X_m X_n\right] = \mathbf{E}\left[X_m \mathbf{E}\left[X_n \mid \mathcal{F}_m\right]\right] = 0 \tag{6}$$

$\square$

A much deeper

THEOREM 9.60 (Martingale Central Limit Theorem). *Let $\mathcal{F}_n$ be a filtration and let X_n be a martingale difference sequence such that there is a $\sigma^2 > 0$ such that $n^{-1} \sum_{j=1}^n \boldsymbol{E}\left[X_j^2 \mid \mathcal{F}_{j-1}\right] \xrightarrow{P} \sigma^2$ and*

$$\frac{1}{n} \sum_{j=1}^n \boldsymbol{E}\left[X_j^2 \boldsymbol{1}_{|X_j| > \epsilon\sqrt{n}} \mid \mathcal{F}_{j-1}\right] \xrightarrow{P} 0$$

for every $\epsilon > 0$ then it follows that $\frac{1}{\sqrt{n}} \sum_{j=1}^n X_j \xrightarrow{d} N(0, \sigma^2)$.

PROOF. The first step of the proof is to set up a truncation so that we can reduce to a case in which we have some boundedness. For $k = 1, \dots, n$, let $A_k^n = \{n^{-1} \sum_{j=1}^k \mathbf{E}\left[X_j^2 \mid \mathcal{F}_{j-1}\right] \leq 2\sigma^2\}$ and to simplify notation in a few spots we define $A_m^n = \emptyset$ for any $m > n$. Note that A_k^n is $\mathcal{F}_{k-1}$ measurable, that $A_n^n \subset A_{n-1}^n \subset \cdots \subset A_1^n$ and

$$\lim_{n \to \infty} \mathbf{P}\{A_n^n\} \leq \lim_{n \to \infty} \mathbf{P}\{\left| n^{-1} \sum_{k=1}^n \mathbf{E}\left[X_k^2 \mid \mathcal{F}_{k-1}\right] - \sigma^2 \right| \leq \sigma^2\} = 1$$

Now define $X_{n,k} = \frac{1}{\sqrt{n}} X_k \mathbf{1}_{A^n_k}$ and observe that because A^n_k is $\mathcal{F}_{k-1}$-measurable $X_{n,k}$ is also a martingale difference sequence. In addition we can recast the hypotheses of the theorem in terms of the $X_{n,k}$:

$$\sum_{k=1}^n \mathbf{E}\left[X^2_{n,k} \mid \mathcal{F}_{k-1}\right] \leq 2\sigma^2$$
$$\sum_{k=1}^n \mathbf{E}\left[X^2_{n,k} \mid \mathcal{F}_{k-1}\right] \xrightarrow{P} \sigma^2$$
$$\sum_{k=1}^n \mathbf{E}\left[X^2_{n,k} \mathbf{1}_{|X_{n,k}|>\epsilon} \mid \mathcal{F}_{k-1}\right] \xrightarrow{P} 0$$

To see the first inequality we use $\mathcal{F}_{k-1}$-measurability of A^n_k, the pull out rule of conditional expectations and the definition of A^n_k to compute

$$\sum_{k=1}^n \mathbf{E}\left[X^2_{n,k} \mid \mathcal{F}_{k-1}\right](\omega) = \frac{1}{n}\sum_{k=1}^n \mathbf{E}\left[X^2_k \mathbf{1}_{A^n_k} \mid \mathcal{F}_{k-1}\right](\omega) = \frac{1}{n}\sum_{k=1}^n \mathbf{E}\left[X^2_k \mid \mathcal{F}_{k-1}\right](\omega)\mathbf{1}_{A^n_k}(\omega)$$
$$= \begin{cases} \frac{1}{n}\mathbf{1}_{A^n_j}(\omega)\sum_{k=1}^j \mathbf{E}\left[X^2_k \mid \mathcal{F}_{k-1}\right](\omega) & \text{for } \omega \in A^n_j \setminus A^n_{j+1} \text{ and } j = 1, \dots, n \\ 0 & \text{for } \omega \notin A^n_1 \end{cases}$$
$$\leq 2\sigma^2$$

To see the second claim use Exercise 13, the fact that $\sum_{k=1}^n \mathbf{E}\left[X^2_{n,k} \mid \mathcal{F}_{k-1}\right]$ equals $\frac{1}{n}\sum_{k=1}^n \mathbf{E}\left[X^2_k \mid \mathcal{F}_{k-1}\right]$ on A^n_n, the fact that $\mathbf{P}\{A^n_n\} \to 1$ and the hypothesis that $\frac{1}{n}\sum_{k=1}^n \mathbf{E}\left[X^2_k \mid \mathcal{F}_{k-1}\right] \xrightarrow{P} \sigma^2$.

The third fact follows from the fact that $X^2_{n,k} \leq \frac{1}{n} X^2_k$ and the Lindeberg-like condition (6).

Having defined the truncated triangular array $X_{n,k}$ and given some estimates for it, we claim that it suffices to show that $\sum_{k=1}^n X_{n,k} \xrightarrow{d} N(0,\sigma^2)$. Indeed for any $\epsilon > 0$ we have

$$\begin{aligned}\mathbf{P}\{\left|\sum_{k=1}^n X_{n,k} - \frac{1}{\sqrt{n}}\sum_{k=1}^n X_k\right| > \epsilon\} &\leq \mathbf{P}\{\sum_{k=1}^n X_{n,k} \neq \frac{1}{\sqrt{n}}\sum_{k=1}^n X_k\} \\ &\leq 1 - \mathbf{P}\{A^n_n\} \to 0\end{aligned}$$

thus the claim follows by Slutsky's Theorem 5.46.

Recall from Taylor's Theorem we have the $\left|e^{ix} - 1 - ix\right| \leq \frac{x^2}{2}$ and moreoer $\left|e^{ix} - 1 - ix + \frac{x^2}{2}\right| = x^2 R(x)$ with $\lim_{x\to 0} R(x) = 0$ and $|R(x)| \leq 1$ (see Theorem C.2 where these specific estimates are worked out). The first estimate is good for large x and the second is good for small x so we combine them into single estimate. Let $\epsilon > 0$ be given and pick $\delta > 0$ such that $|R(x)| < \epsilon$ for $|x| \leq \delta$ then we have the estimate

$$\left|e^{ix} - 1 - ix + \frac{x^2}{2}\right| \leq x^2 \mathbf{1}_{|x|\geq\delta} + \epsilon x^2$$

Define

$$R_{n,k}(u) = \mathbf{E}\left[e^{iuX_{n,k}} - 1 - iuX_{n,k} \mid \mathcal{F}_{k-1}\right]$$

and we have estimates

$$|R_{n,k}(u)| \leq \mathbf{E}\left[\left|e^{iuX_{n,k}} - 1 - iuX_{n,k}\right| \mid \mathcal{F}_{k-1}\right] \leq \frac{u^2}{2}\mathbf{E}\left[\left|X_{n,k}^2\right| \mid \mathcal{F}_{k-1}\right]$$

$$\sum_{k=1}^n |R_{n,k}(u)| \leq \frac{u^2}{2}\sum_{k=1}^n \mathbf{E}\left[\left|X_{n,k}^2\right| \mid \mathcal{F}_{k-1}\right] \leq u^2\sigma^2$$

Claim: For every $u \in \mathbb{R}$ we have $\max_{1\leq k\leq n} |R_{n,k}(u)| \stackrel{L^1}{\to} 0$.

We have the following estimate

$$\begin{aligned}
\max_{1\leq k\leq n} |R_{n,k}(u)| &\leq \frac{u^2}{2} \max_{1\leq k\leq n} \mathbf{E}\left[X_{n,k}^2 \mid \mathcal{F}_{k-1}\right] \\
&= \frac{u^2}{2} \max_{1\leq k\leq n} \{\mathbf{E}\left[X_{n,k}^2(\mathbf{1}_{|X_{n,k}|>\delta} + \mathbf{1}_{|X_{n,k}|>\delta}) \mid \mathcal{F}_{k-1}\right]\} \\
&= \frac{u^2}{2} \max_{1\leq k\leq n} \{\mathbf{E}\left[X_{n,k}^2\mathbf{1}_{|X_{n,k}|>\delta^2} \mid \mathcal{F}_{k-1}\right] + \delta\} \\
&= \frac{u^2}{2}\left(\sum_{k=1}^n \mathbf{E}\left[X_{n,k}^2\mathbf{1}_{|X_{n,k}|>\delta^2} \mid \mathcal{F}_{k-1}\right] + \delta\right) \stackrel{P}{\to} \frac{u^2\delta^2}{2}
\end{aligned}$$

Now we can let $\delta \to 0$ to conclude that $\max_{1\leq k\leq n} |R_{n,k}(u)| \stackrel{P}{\to} 0$. We also have $\max_{1\leq k\leq n} |R_{n,k}(u)| \leq \sum_{k=1}^n |R_{n,k}(u)| < u^2\sigma^2$. In particular $\max_{1\leq k\leq n} |R_{n,k}(u)|$ is uniformly integrable and thus by Lemma 5.58 $\max_{1\leq k\leq n} |R_{n,k}(u)| \stackrel{L^1}{\to} 0$.

Claim: $\lim_{n\to\infty} \mathbf{E}\left[\sum_{k=1}^n |R_{n,k}(u)|^2\right] = 0$.

We have

$$\begin{aligned}
\lim_{n\to\infty} \mathbf{E}\left[\sum_{k=1}^n |R_{n,k}(u)|^2\right] &\leq \lim_{n\to\infty} \mathbf{E}\left[\max_{1\leq k\leq n} |R_{n,k}(u)| \sum_{k=1}^n |R_{n,k}(u)|\right] \\
&\leq u^2\sigma^2 \lim_{n\to\infty} \mathbf{E}\left[\max_{1\leq k\leq n} |R_{n,k}(u)|\right] = 0
\end{aligned}$$

Claim: $\sum_{k=1}^n R_{n,k} \stackrel{P}{\to} -u^2\sigma^2/2$.

We start with the estimate

$$\begin{aligned}
\sum_{k=1}^n \left|R_{n,k} + \frac{u^2}{2}\mathbf{E}\left[X_{n,k}^2 \mid \mathcal{F}_{k-1}\right]\right| &= \sum_{k=1}^n \left|\mathbf{E}\left[e^{iuX_{n,k}} - 1 - iuX_{n,k} + \frac{u^2X_{n,k}^2}{2} \mid \mathcal{F}_{k-1}\right]\right| \\
&\leq \sum_{k=1}^n \mathbf{E}\left[u^2X_{n,k}^2\mathbf{1}_{|X_{n,k}|>\delta/|u|} + \epsilon u^2 X_{n,k}^2 \mid \mathcal{F}_{k-1}\right] \\
&\leq u^2 \sum_{k=1}^n \mathbf{E}\left[X_{n,k}^2\mathbf{1}_{|X_{n,k}|>\delta/|u|} \mid \mathcal{F}_{k-1}\right] + 2\epsilon u^2\sigma^2 \stackrel{P}{\to} 2\epsilon u^2\sigma^2
\end{aligned}$$

Now let $\epsilon \to 0$ to conclude that $\sum_{k=1}^n \left|R_{n,k} + \frac{u^2}{2}\mathbf{E}\left[X_{n,k}^2 \mid \mathcal{F}_{k-1}\right]\right| \stackrel{P}{\to} 0$. The fact that $\sum_{k=1}^n \mathbf{E}\left[X_{n,k}^2 \mid \mathcal{F}_{k-1}\right] \stackrel{P}{\to} \sigma^2$ means that the claim is proven.

Claim: $\prod_{k=1}^n (1 - R_{n,k}) \xrightarrow{L^1} e^{u^2\sigma^2/2}$ and in addition $\prod_{k=1}^m |1 - R_{n,k}| \leq e^{u^2\sigma^2}$ for all $n \in \mathbb{N}$ and $1 \leq m \leq n$.

By Taylor's Theorem (specifically Corollary 1.22) we may write $\log(1-x) = -x + xS(x)$ with $\lim_{x\to 0} S(x) = 0$. Let $\epsilon > 0$ be given and select $\delta > 0$ such that $|S(x)| < \epsilon$ for $|x| < \delta$. It follows that

$$\lim_{n\to\infty} \mathbf{P}\{\max_{1\leq k\leq n} |S(R_{n,k}(u))| \geq \epsilon\} \leq \lim_{n\to\infty} \mathbf{P}\{\max_{1\leq k\leq n} |R_{n,k}(u)| \geq \delta\} = 0$$

so $\max_{1\leq k\leq n} |S(R_{n,k}(u))| \xrightarrow{P} 0$. From this it follows that

$$\begin{aligned}\left|\sum_{k=1}^n R_{n,k}(u)S(R_{n,k}(u))\right| &\leq \max_{1\leq k\leq n} |S(R_{n,k}(u))| \sum_{k=1}^n |R_{n,k}(u)| \\ &\leq u^2\sigma^2 \max_{1\leq k\leq n} |S(R_{n,k}(u))| \xrightarrow{P} 0\end{aligned}$$

Therefore (TODO: What is the deal with what branch we need to be on? Don't we need to know this is non-zero as well?)

$$\prod_{k=1}^n (1 - R_{n,k}) = e^{\sum_{k=1}^n \log(1-R_{n,k})} = e^{-\sum_{k=1}^n R_{n,k}} e^{\sum_{k=1}^n R_{n,k}S(R_{n,k})} \xrightarrow{P} e^{u^2\sigma^2/2}$$

We also have

$$\prod_{k=1}^n |1 - R_{n,k}(u)| \leq \prod_{k=1}^n (1 + |R_{n,k}(u)|) \leq \prod_{k=1}^n e^{|R_{n,k}(u)|} = e^{\sum_{k=1}^n |R_{n,k}(u)|} \leq e^{u^2\sigma^2}$$

From this we also know that $\prod_{k=1}^n (1 - R_{n,k})$ is uniformly integrable and therefore we can upgrade the convergence in probability to L^1 convergence (Lemma 5.58).

Claim: $\lim_{n\to\infty} \left|\mathbf{E}\left[\prod_{k=1}^n e^{iuX_{n,k}}(1 - R_{n,k}(u)) - 1\right]\right| = 0$.

First note that because $R_{n,k}$ is $\mathcal{F}_{k-1}$-measurable and because $X_{n,k}$ is a martingale difference sequence we have for every $1 \leq k \leq n$,

$$\begin{aligned}\mathbf{E}\left[e^{iuX_{n,k}}(1 - R_{n,k}(u)) \mid \mathcal{F}_{k-1}\right] &= (1 - R_{n,k}(u))\mathbf{E}\left[e^{iuX_{n,k}} \mid \mathcal{F}_{k-1}\right] \\ &= (1 - R_{n,k}(u))\mathbf{E}\left[e^{iuX_{n,k}} - 1 - iuX_{n,k} + 1 \mid \mathcal{F}_{k-1}\right] \\ &= (1 - R_{n,k}(u))(1 + R_{n,k}(u)) = 1 - R_{n,k}^2(u)\end{aligned}$$

Applying this we can compute

$$\begin{aligned}&\mathbf{E}\left[\prod_{k=1}^n e^{iuX_{n,k}}(1 - R_{n,k}(u))\right] \\ &= \mathbf{E}\left[\mathbf{E}\left[\prod_{k=1}^n e^{iuX_{n,k}}(1 - R_{n,k}(u)) \mid \mathcal{F}_{n-1}\right]\right] \\ &= \mathbf{E}\left[\prod_{k=1}^{n-1} e^{iuX_{n,k}}(1 - R_{n,k}(u))\mathbf{E}\left[e^{iuX_{n,n}}(1 - R_{n,n}(u)) \mid \mathcal{F}_{n-1}\right]\right] \\ &= \mathbf{E}\left[\prod_{k=1}^{n-1} e^{iuX_{n,k}}(1 - R_{n,k}(u))\right] - \mathbf{E}\left[\prod_{k=1}^{n-1} e^{iuX_{n,k}}(1 - R_{n,k}(u))R_{n,n}^2(u)\right]\end{aligned}$$

This argument can be repeated $n-1$ more times to give us the identity

$$\begin{aligned}&\mathbf{E}\left[\prod_{k=1}^{n} e^{iuX_{n,k}}(1-R_{n,k}(u))\right] \\ &= 1 - \sum_{m=1}^{n} \mathbf{E}\left[\prod_{k=1}^{m-1} e^{iuX_{n,k}}(1-R_{n,k}(u))R^2_{n,m}(u)\right]\end{aligned}$$

from which we conclude

$$\begin{aligned}\left|\mathbf{E}\left[\prod_{k=1}^{n} e^{iuX_{n,k}}(1-R_{n,k}(u))\right]-1\right| &\leq \sum_{m=1}^{n} \mathbf{E}\left[\left|\prod_{k=1}^{m-1} e^{iuX_{n,k}}(1-R_{n,k}(u))R^2_{n,m}(u)\right|\right] \\ &\leq \sum_{m=1}^{n} \mathbf{E}\left[\prod_{k=1}^{m-1} \left|1-R_{n,k}(u)\right| \left|R_{n,m}(u)\right|^2\right] \\ &\leq e^{u^2\sigma^2} \sum_{m=1}^{n} \mathbf{E}\left[\left|R_{n,m}(u)\right|^2\right] \to 0\end{aligned}$$

Now we have everything to finish the proof of theorem. We get

$$\begin{aligned}\lim_{n\to\infty}\left|\mathbf{E}\left[e^{u^2\sigma^2/2}\prod_{k=1}^{n} e^{iuX_{n,k}}\right]-1\right| &\leq \lim_{n\to\infty}\left|\mathbf{E}\left[e^{u^2\sigma^2/2}\prod_{k=1}^{n} e^{iuX_{n,k}} - \prod_{k=1}^{n} e^{iuX_{n,k}}(1-R_{n,k}(u))\right]\right| \\ &+ \lim_{n\to\infty}\left|\mathbf{E}\left[\prod_{k=1}^{n} e^{iuX_{n,k}}(1-R_{n,k}(u))\right]-1\right| \\ &\leq \lim_{n\to\infty}\mathbf{E}\left[\prod_{k=1}^{n}\left|e^{u^2\sigma^2/2}-(1-R_{n,k}(u))\right|\right] = 0\end{aligned}$$

from which we conclude that $\lim_{n\to\infty}\mathbf{E}\left[e^{iu\sum_{k=1}^n X_{n,k}}\right] = e^{-u^2\sigma^2/2}$ which by the Glivenko-Levy Continuity Theorem 7.13 shows $\sum_{k=1}^n X_{n,k} \xrightarrow{d} N(0,\sigma^2)$. $\square$

Now we turn to the inequalities of Burkholder that show an intimitate relationship between a martingale and the sum of the squares of the associated martingale differences (the so-called quadratic variation). These ideas will come up again in the context of stochastic integration in the Burkholder-Davis-Gundy inequalities Theorem 14.32.

Definition 9.61. Let X_n be a process then the process

$$[X]_n = \sum_{i=1}^{n}(X_i - X_{i-1})^2$$

is called the *quadratic variation* of X.

It is trivial that if X is $\mathcal{F}$-adapted then so if $[X]$.

Theorem 9.62. *Let $1 < p < \infty$ then there exist constants C_1 and C_2 so that for any $\mathcal{F}$-martingale M_n we have*

$$C_1 \boldsymbol{E}\left[[M]_n^{p/2}\right] \leq \boldsymbol{E}\left[|M_n|^p\right] \leq C_2 \boldsymbol{E}\left[[M]_n^{p/2}\right]$$

The proof depends on a few lemmas.

LEMMA 9.63. *Let X_n have $X_0 = 0$ and be either an $\mathcal{F}$-martingale or a non-negative $\mathcal{F}$-submartingale. For $\lambda > 0$ and $n \in \mathbb{N}$ define the $\mathcal{F}$-optional time $\tau = \min\{m \mid X_m > \lambda\} \wedge (n+1)$ then*

$$\boldsymbol{E}\left[\sum_{i=1}^{\tau-1}(X_i - X_{i-1})^2\right] + \boldsymbol{E}\left[X_{\tau-1}^2\right] \leq 2\lambda \boldsymbol{E}\left[|X_n|\right]$$

PROOF. For arbitrary $m \in \mathbb{N}$

$$\begin{aligned}
X_{m-1}^2 + \sum_{i=1}^{m-1}(X_i - X_{i-1})^2 &= X_{m-1}^2 + \left(\sum_{i=1}^{m-1}(X_i - X_{i-1})\right)^2 - 2\sum_{1 \leq i < j \leq m-1}(X_i - X_{i-1})(X_j - X_{j-1}) \\
&= 2X_{m-1}^2 - 2\sum_{1 \leq i < j \leq m-1}(X_i - X_{i-1})(X_j - X_{j-1}) \\
&= 2X_{m-1}^2 - 2\sum_{j=2}^{m-1}\left(\sum_{i=1}^{j-1}(X_i - X_{i-1})\right)(X_j - X_{j-1}) \\
&= 2X_{m-1}^2 - 2\sum_{j=2}^{m-1} X_{j-1}(X_j - X_{j-1}) \\
&= 2X_{m-1}X_m - 2X_{m-1}(X_m - X_{m-1}) - 2\sum_{j=2}^{m-1} X_{j-1}(X_j - X_{j-1}) \\
&= 2X_{m-1}X_m - 2\sum_{j=2}^{m} X_{j-1}(X_j - X_{j-1})
\end{aligned}$$

Therefore if σ is an arbitrary bounded $\mathcal{F}$-optional time with $\sigma \leq M$,

$$X_{\sigma-1}^2 + \sum_{i=1}^{\sigma-1}(X_i - X_{i-1})^2 \leq 2X_{\sigma-1}X_\sigma - 2\sum_{j=2}^{\sigma} X_{j-1}(X_j - X_{j-1})$$

Now note that

$$\begin{aligned}
\mathbf{E}\left[\sum_{j=2}^{\sigma} X_{j-1}(X_j - X_{j-1})\right] &= \sum_{j=2}^{n+1}\mathbf{E}\left[X_{j-1}(X_j - X_{j-1}); j \leq \sigma\right] \\
&= \sum_{j=2}^{M}\mathbf{E}\left[X_{j-1}\mathbf{E}\left[X_j - X_{j-1} \mid \mathcal{F}_{j-1}\right]; j-1 > \sigma\right] \geq 0
\end{aligned}$$

where in fact the inequality is equality in the martingale case.

Now the result of the Lemma only depends on the values of $X_0, \dots, X_n$ so by redefining X to be X^n we may assume that $X_m = X_n$ for all $m \geq n$. Returning to τ note that under either hypothesis on X_n we know that $|X_n|$ is a submartingale and therefore

$$\begin{aligned}
\mathbf{E}\left[X_{\tau-1}^2\right] + \mathbf{E}\left[\sum_{i=1}^{\tau-1}(X_i - X_{i-1})^2\right] &\leq 2\mathbf{E}\left[X_{\tau-1}X_\tau\right] \\
&\leq 2\mathbf{E}\left[|X_{\tau-1}X_\tau|\right] \leq 2\lambda\mathbf{E}\left[|X_\tau|\right] \leq 2\lambda\mathbf{E}\left[\mathbf{E}\left[|X_n| \mid \mathcal{F}_\tau\right]\right] = 2\lambda\mathbf{E}\left[|X_n|\right]
\end{aligned}$$

$\square$

The second lemma gives the left hand Burkholder inequality for the case of nonnegative submartingales.

LEMMA 9.64. *Let X_n be a nonnegative $\mathcal{F}$-submartingale with $X_0 = 0$, let $n \in \mathbb{N}$, $\theta > 0$ be given and define*

$$Y = \theta \left(\sum_{i=1}^n (X_i - X_{i-1})^2 \right)^{1/2} \vee \max_{1 \leq i \leq n} X_i$$

Then for each $\lambda > 0$ we have

$$\lambda \boldsymbol{P}\{Y > \lambda \sqrt{1 + 2\theta^2}\} \leq 3\boldsymbol{E}[X_n; Y > \lambda]$$

and for every $1 < p < \infty$

$$\left\| \left(\sum_{i=1}^n (X_i - X_{i-1})^2 \right)^{1/2} \right\|_p \leq 9p^{1/2} q \left\| X_n \right\|_p$$

where $\frac{1}{p} + \frac{1}{q} = 1$.

PROOF. We begin with a simple union bound followed by an application of Doob's Maximal Inequality (Lemma 9.44)

$$\begin{aligned}
\mathbf{P}\{Y > \lambda\sqrt{1+2\theta^2}\} &\leq \mathbf{P}\{Y > \lambda\sqrt{1+2\theta^2}; \max_{1 \leq i \leq n} X_i < \lambda\} + \mathbf{P}\{Y > \lambda\sqrt{1+2\theta^2}; \max_{1 \leq i \leq n} X_i \geq \lambda\} \\
&\leq \mathbf{P}\{\theta \left(\sum_{i=1}^n (X_i - X_{i-1})^2 \right)^{1/2} > \lambda\sqrt{1+2\theta^2}; \max_{1 \leq i \leq n} X_i < \lambda\} + \mathbf{P}\{\max_{1 \leq i \leq n} X_i \geq \lambda\} \\
&\leq \mathbf{P}\{\theta \left(\sum_{i=1}^n (X_i - X_{i-1})^2 \right)^{1/2} > \lambda\sqrt{1+2\theta^2}; \max_{1 \leq i \leq n} X_i < \lambda\} + \lambda^{-1}\mathbf{E}\left[X_n; \max_{1 \leq j \leq n} X_j \geq \lambda\right] \\
&\leq \mathbf{P}\{\theta \left(\sum_{i=1}^n (X_i - X_{i-1})^2 \right)^{1/2} > \lambda\sqrt{1+2\theta^2}; \max_{1 \leq i \leq n} X_i < \lambda\} + \lambda^{-1}\mathbf{E}\left[X_n; Y \geq \lambda\right]
\end{aligned}$$

For $1 \leq m \leq n$ let

$$Z_m = X_m \mathbf{1}_{\theta\left(\sum_{i=1}^m (X_i - X_{i-1})^2\right)^{1/2} > \lambda}$$

and define $Z_0 = 0$ and $Z_m = Z_n$ for $m > n$. It is clear from the fact that X_n is a submartingale that Z_n is $\mathcal{F}$-adapted and each Z_n is integrable. We compute

$$\begin{aligned}
\mathbf{E}\left[Z_m \mid \mathcal{F}_{m-1}\right] &= \mathbf{E}\left[X_m \mathbf{1}_{\theta\left(\sum_{i=1}^m (X_i - X_{i-1})^2\right)^{1/2} > \lambda} \mid \mathcal{F}_{m-1}\right] \\
&\geq \mathbf{E}\left[X_m \mathbf{1}_{\theta\left(\sum_{i=1}^{m-1} (X_i - X_{i-1})^2\right)^{1/2} > \lambda} \mid \mathcal{F}_{m-1}\right] \\
&= \mathbf{E}\left[X_m \mid \mathcal{F}_{m-1}\right] \mathbf{1}_{\theta\left(\sum_{i=1}^{m-1} (X_i - X_{i-1})^2\right)^{1/2} > \lambda} \\
&= X_{m-1} \mathbf{1}_{\theta\left(\sum_{i=1}^{m-1} (X_i - X_{i-1})^2\right)^{1/2} > \lambda} = Z_{m-1}
\end{aligned}$$

and it follows that Z is a nonnegative $\mathcal{F}$-submartingale.

CLAIM 9.64.1. Let

$$A = \left\{ \theta \left(\sum_{i=1}^{n} (X_i - X_{i-1})^2 \right)^{1/2} > \lambda\sqrt{1+2\theta^2}; \max_{1 \le i \le n} X_i < \lambda \right\}$$
$$B = \left\{ \left(\sum_{i=1}^{n} (Z_i - Z_{i-1})^2 \right)^{1/2} > \lambda; \max_{1 \le i \le n} Z_i < \lambda \right\}$$

then $A \subset B$.

Since by definition $Z_i \le X_i$ for $1 \le i \le n$ we clearly have $A \subset \{\max_{1 \le i \le n} Z_i < \lambda\}$ Define the optional time

$$\tau = n \wedge \left\{ m \mid \theta \left(\sum_{i=1}^{m} (X_i - X_{i-1})^2 \right)^{1/2} > \lambda \right\}$$

Observe that by the nonnegativity of X_n and the fact that $\max_{1 \le i \le n} X_i < \lambda$ on A we have

$$|X_\tau - X_{\tau-1}| \le X_\tau \vee X_{\tau-1} < \lambda \text{ on } A$$

Also observe that by definition of Z and τ we also have $Z_m = X_m$ when $\tau + 1 \le m \le n$ (of course when $\tau = n$ the statement is vacuous). From these two observations and definition of A we have

$$\begin{aligned}
(1+2\theta^2)\lambda^2 &< \theta^2 \sum_{i=1}^{n} (X_i - X_{i-1})^2 \\
&\le \theta^2 \sum_{i=1}^{\tau-1} (X_i - X_{i-1})^2 + \theta^2 (X_\tau - X_{\tau-1})^2 + \theta^2 \sum_{i=\tau+1}^{n} (X_i - X_{i-1})^2 \\
&< \lambda^2 + \theta^2\lambda^2 + \theta^2 \sum_{i=1}^{n} (Z_i - Z_{i-1})^2
\end{aligned}$$

on A. Subtracting $(1+\theta^2)\lambda^2$ and dividing by θ^2 we see that $\sum_{i=1}^{n} (Z_i - Z_{i-1})^2 > \lambda^2$ on A and the claim is proven.

Let $\tau = (n+1) \wedge \min\{i\{|Z_i| > \lambda\}$ as in Lemma 9.63 and note that $\tau = n+1$ on $\max_{1 \leq i \leq n} Z_i < \lambda$. By the previous claim and Lemma 9.63 we get the bound

$$\begin{aligned}
\mathbf{P}\{\theta \left(\sum_{i=1}^n (X_i - X_{i-1})^2\right)^{1/2} > \lambda\sqrt{1+2\theta^2}; \max_{1 \leq i \leq n} X_i < \lambda\} &\leq \mathbf{P}\{\left(\sum_{i=1}^n (Z_i - Z_{i-1})^2\right)^{1/2} > \lambda; \max_{1 \leq i \leq n} Z_i < \lambda\} \\
&\leq \lambda^{-2}\mathbf{E}\left[\sum_{i=1}^n (Z_i - Z_{i-1})^2; \left(\sum_{i=1}^n (Z_i - Z_{i-1})^2\right)^{1/2} > \lambda; \right. \\
&\leq \lambda^{-2}\mathbf{E}\left[\sum_{i=1}^n (Z_i - Z_{i-1})^2; \max_{1 \leq i \leq n} Z_i < \lambda\right] \\
&\leq \lambda^{-2}\mathbf{E}\left[\sum_{i=1}^{\tau-1} (Z_i - Z_{i-1})^2; \max_{1 \leq i \leq n} Z_i < \lambda\right] \\
&\leq \lambda^{-2}\left(\mathbf{E}\left[\sum_{i=1}^{\tau-1} (Z_i - Z_{i-1})^2\right] + \mathbf{E}\left[Z_{\tau-1}^2\right]\right) \\
&\leq 2\lambda^{-1}\mathbf{E}\left[Z_n\right] \\
&= 2\lambda^{-1}\mathbf{E}\left[X_n; \theta\left(\sum_{i=1}^n (X_i - X_{i-1})^2\right)^{1/2} > \lambda\right] \\
&\leq 2\lambda^{-1}\mathbf{E}\left[X_n; Y > \lambda\right]
\end{aligned}$$

Therefore we have (TODO: We need to convert a $\geq$ to a $>$)

$$\mathbf{P}\{Y > \lambda\sqrt{1+2\theta^2}\} \leq 2\lambda^{-1}\mathbf{E}\left[X_n; Y > \lambda\right] + \lambda^{-1}\mathbf{E}\left[X_n; Y \geq \lambda\right] \leq 3\lambda^{-1}\mathbf{E}\left[X_n; Y > \lambda\right]$$

Now we turn this maximal inequality into the L^p inequality using Lemma 3.8, Tonelli's Theorem and the Hölder Inequality together with the fact that $q(p-1) = p$:

$$\begin{aligned}
\mathbf{E}\left[Y^p\right] &= p\int_0^\infty \lambda^{p-1}\mathbf{P}\{Y > \lambda\}\, d\lambda \\
&= p(1+2\theta^2)^{p/2}\int_0^\infty \lambda^{p-1}\mathbf{P}\{Y > \lambda\sqrt{1+2\theta^2}\}\, d\lambda \\
&\leq 3p(1+2\theta^2)^{p/2}\int_0^\infty \lambda^{p-2}\mathbf{E}\left[X_n; Y > \lambda\right]\, d\lambda \\
&\leq 3p(1+2\theta^2)^{p/2}\mathbf{E}\left[X_n\int_0^Y \lambda^{p-2}\, d\lambda\right] \\
&\leq 3p(p-1)^{-1}(1+2\theta^2)^{p/2}\mathbf{E}\left[X_n Y^{p-1}\right] \\
&\leq 3q(1+2\theta^2)^{p/2}\mathbf{E}\left[X_n^p\right]^{1/p}\mathbf{E}\left[Y^p\right]^{1/q}
\end{aligned}$$

so $\|Y\|_p \leq 3q(1+2\theta^2)^{p/2} \|X_n\|_p$. By definition of Y we have

$$\theta \left\| \left(\sum_{i=1}^n X_i^2 \right)^{1/2} \right\|_p \leq \|Y\|_p \leq 3q(1+2\theta^2)^{p/2} \|X_n\|_p$$

Now choose $\theta = p^{-1/2}$ and note that $(1+2\theta^2)^{p/2} = (1+2/p)^{p/2} \leq e < 3$ and the lemma follows. □

Now we give the proof of the Burkholder inequality

PROOF. Since M is a martingale we let $X_n = (M_n)_+$ and $Y_n = (M_n)_-$ and note that by Proposition 9.24 each of X and Y is a nonnegative $\mathcal{F}$-submartingale. By the triangle inequality in $\mathbb{R}^n$ be have

$$\begin{aligned}
[M]_n^{1/2} &= \left(\sum_{i=1}^n (M_i - M_{i-1})^2 \right)^{1/2} = \left(\sum_{i=1}^n ((X_i - X_{i-1}) - (Y_i - Y_{i-1}))^2 \right)^{1/2} \\
&\leq \left(\sum_{i=1}^n (X_i - X_{i-1})^2 \right)^{1/2} + \left(\sum_{i=1}^n (Y_i - Y_{i-1})^2 \right)^{1/2} = [X]_n^{1/2} + [Y]_n^{1/2}
\end{aligned}$$

and by the triangle inequality in L^p (Lemma 8.2) and Lemma 9.64 we get

$$\left\| [M]_n^{1/2} \right\|_p \leq \left\| [X]_n^{1/2} \right\|_p + \left\| [Y]_n^{1/2} \right\|_p \leq 9p^{1/2} q (\|X_n\|_p + \|Y_n\|_p) \leq 18 p^{1/2} q \|M_n\|_p$$

which gives the left hand inequality.

The right hand inequality is derived from the left hand inequality by defining the random variable

$$N_n = \operatorname{sgn} M_n \, |M_n|^{p-1} / \|M_n\|_p^{p-1}$$

and considering the closed martingale $N_m = \mathbf{E}[N_n \mid \mathcal{F}_m]$. We shall need the simple fact that

$$\|N_n\|_q = \left(\mathbf{E}\left[|M_n|^{q(p-1)} / \|M_n\|_p^{q(p-1)} \right] \right)^{1/q} = \left(\mathbf{E}\left[|M_n|^p \right] / \|M_n\|_p^p \right)^{1/q} = 1$$

We calculate using Proposition 9.58, Cauchy-Schwartz for sequences, the Hölder inequality and the left hand Burkholder inequality and above calculation of the

q-norm of N_n

$$\begin{aligned}
\|M_n\|_p &= \mathbf{E}\left[|M_n|^p\right] / \|M_n\|_p^{p-1} = \mathbf{E}\left[M_n \operatorname{sgn} M_n \, |M_n|^{p-1}\right] / \|M_n\|_p^{p-1} = \mathbf{E}\left[M_n N_n\right] \\
&= \mathbf{E}\left[\sum_{i=1}^n (M_i - M_{i-1}) \sum_{j=1}^n (N_j - N_{j-1})\right] \\
&= \mathbf{E}\left[\sum_{i=1}^n (M_i - M_{i-1})(N_i - N_{i-1})\right] + \mathbf{E}\left[\sum_{i=1}^n \sum_{j=1}^{i-1} \mathbf{E}\left[M_i - M_{i-1} \mid \mathcal{F}_j\right](N_j - N_{j-1})\right] + \\
&\qquad \mathbf{E}\left[\sum_{j=1}^n \sum_{i=1}^{j-1} (M_i - M_{i-1}) \mathbf{E}\left[N_j - N_{j-1} \mid \mathcal{F}_i\right]\right] \\
&= \mathbf{E}\left[\sum_{i=1}^n (M_i - M_{i-1})(N_i - N_{i-1})\right] \\
&\leq \mathbf{E}\left[\left(\sum_{i=1}^n (M_i - M_{i-1})^2\right)^{1/2} \left(\sum_{i=1}^n (N_i - N_{i-1})^2\right)^{1/2}\right] \\
&\leq \left\|[M]_n^{1/2}\right\|_p \left\|[N]_n^{1/2}\right\|_q \leq 18pq^{1/2} \left\|[M]_n^{1/2}\right\|_p \|N_n\|_q = 18pq^{1/2} \left\|[M]_n^{1/2}\right\|_p
\end{aligned}$$

□

TODO: Discuss i.i.d. implies martingale difference sequence implies white noise (uncorrelated). Philosophically when thinking about mean zero discrete time processes: i.i.d. signifies complete unpredictability, martingale difference sequence signifies unpredictability in the mean (but possible predictability of higher moments: GARCH gives examples of this), uncorrelated means linearly unpredictable (but possibly non-linearly predictable).

TODO: Given example of a non-i.i.d. martingale difference sequence (there are martingale that is not a random walk is an example, but it might be nice to come up with examples that show predictable higher moments) and uncorrelated non-martingale difference sequences (e.g. $y_n = e_n + e_{n-1}e_{n-2}$).

4. Applications of Discrete Time Martingales

The theme of this section is the utility of martingales. That is to say we are concerned with scenarios in which martingales aren't the primary objects of study rather martingales arise as a tool in the study of seemingly unrelated problems.

As a warm up we give a martingale proof of the Kolmogorov $0-1$ Law (Theorem 4.27).

Proof. Suppose that $\xi_1, \xi_2, \dots$ are independent random variables and let $\mathcal{A}_n = \sigma(\xi_m; m \geq n)$ and $\mathcal{T}_\infty = \cap_{n=1}^\infty \mathcal{A}_n$. Let $A \in \mathcal{T}_\infty$. Let $\mathcal{F}_n = \sigma(\xi_1, \dots, \xi_n)$ and observe that for every $n \in \mathbb{N}$ we have $A \in \sigma(\xi_{n+1}, \xi_{n+2}, \dots)$ and by Lemma 4.14 we know $\mathcal{F}_n \perp\!\!\!\perp \sigma(\xi_{n+1}, \xi_{n+2}, \dots)$ thus $\mathbf{P}\{A \mid \mathcal{F}_n\} = \mathbf{P}\{A\}$. On the other hand we know that $\mathbf{P}\{A \mid \mathcal{F}_n\}$ is a closed martingale and therefore by the Levy Upward Theorem 9.56 and hence $\mathbf{P}\{A \mid \mathcal{F}_n\} \stackrel{a.s.}{\to} \mathbf{P}\{A \mid \mathcal{F}_\infty\} = \mathbf{1}_A$ where we have used the fact that $A \in \mathcal{F}_\infty = \sigma(\xi_1, \xi_2, \dots)$. Putting everything together we see that $\mathbf{P}\{A\} = \mathbf{1}_A$ a.s. hence $\mathbf{P}\{A\} \in \{0, 1\}$. □

4.1. Exchangability and the Strong Law of Large Numbers. Here we see how almost sure convergence of bounded martingales implies the strong law of large numbers. In fact we shall prove an generalization of the strong law to the class of *exchangable* random variables. The notion of exchangability plays a role in the conceptual foundations of Bayesian statistics so it is worth developing some of the ideas related to it.

TODO: Urn processes, parametric models. Which of the following (or both?) is a proper example of a parametric model: Give yourself a bag of coins with different biases, pick a coin and then perform an infinite number of coin flips -or- do an infinite number of times: pick a coin with replacement and flip.

Definition 9.65. Let $T = \{t_1, t_2, \dots\} \subset \mathbb{N}$ be a finite or countably infinite set, let $(S, \mathcal{S})$ be a measurable space and suppose we are given random elements ξ_t in S for $t \in T$. Then ξ_t is said to be *exchangeable* if for every $N \in \mathbb{N}$, every permutation $\pi : T \to T$ such that $\pi_t = t$ for all $t \geq N$ we have

$$(\xi_{t_1}, \xi_{t_2}, \dots) \stackrel{d}{=} (\xi_{\pi t_1}, \xi_{\pi t_2}, \dots)$$

Note that exchangeability is a generalization of i.i.d.

Proposition 9.66. *Let ξ_t be an i.i.d. set of random elements, then ξ_t is exchangable.*

Proof. Since ξ_t is i.i.d. then $\mathcal{L}(\xi_t) = \mathcal{L}(\xi_{t_1})$ for all $t \in T$, thus for all permutations π of T, we have

$$\mathcal{L}(\xi_{\pi(t_1)}, \xi_{\pi(t_2)}, \dots) = \mathcal{L}(\xi_{\pi(t_1)}) \otimes \mathcal{L}(\xi_{\pi(t_2)}) = \mathcal{L}(\xi_{t_1})^{\mathrm{card}(T)}$$

which shows ξ_t is exchangeable. ∎

Theorem 9.67. *Let ξ_n be an exchangeable sequence of integrable random variables then there exists a σ-algebra $\mathcal{I}_\infty$ such that*

$$\frac{1}{n} \sum_{j=1}^{\infty} \xi_j \stackrel{a.s.}{\to} \boldsymbol{E}[\xi_1 \mid \mathcal{I}_\infty]$$

If the ξ_n are i.i.d. then $\boldsymbol{E}[\xi_1 \mid \mathcal{I}_\infty] = \boldsymbol{E}[\xi_1]$ almost surely.

Proof. The key idea of the proof is to realize that sample averages $\frac{1}{n} \sum_{j=1}^{\infty} \xi_j$ are in fact conditional expectations. For fixed $n \in \mathbb{N}$ consider the set $\mathcal{C}_n$ of Borel subsets $A \subset \mathbb{R}^\infty$ such that $\pi A = A$ for all permutations π of the naturals $\mathbb{N}$ such that $\pi(j) = j$ for $j > n$. In the following we will refer to such a permutation as an n-permutation.

Claim 9.67.1. $\mathcal{C}_n$ is a σ-algebra.

Observe that we can write $\mathcal{C}_n = \cap_\pi \mathcal{C}_\pi$ where $\mathcal{C}_\pi$ is the set of Borel sets invariant under a single n-permutation; it therefore suffices to show that $\mathcal{C}_\pi$ is a σ-algebra. It is elementary that every permutation π induces a bijection on $\mathbb{R}^\infty$: if $(x_{pi(1)}, x_{\pi(2)}, \dots) = (y_{pi(1)}, y_{\pi(2)}, \dots)$ the for any $n \in \mathbb{N}$ we know that$x_n = x_{\pi(\pi^{-1}(n)} = y_{\pi(\pi^{-1}(n)} = y_n$ and given $(x_1, x_2, \dots)$ define $y_n = x_{\pi^{-1}(n)}$ and observe $\pi(x_1, x_2, \dots) = (y_1, y_2, \dots)$. From the fact that π is a bijection and Lemma 2.7 we see $\pi A = A$ then also $\pi A^c = (\pi A)^c = A^c$. Suppose that $A_1, A_2, \dots \in \mathcal{C}_\pi$ then since π is a bijection and Lemma 2.7

$$\pi(A_1 \cap A_2 \cap \cdots) = \pi A_1 \cap \pi A_2 \cap \cdots = A_1 \cap A_2 \cap \cdots$$

Now let $\boldsymbol{\xi} = (\xi_1, \xi_2, \dots)$ and we define $\mathcal{I}_n = \boldsymbol{\xi}^{-1}\mathcal{C}_n$ and we see from Lemma 2.8 that $\mathcal{I}_n$ is a σ-algebra.

CLAIM 9.67.2. $\mathbf{E}[\xi_1 \mid \mathcal{I}_n] = \frac{1}{n}\sum_{j=1} \xi_j$ a.s.

First we validate that $\sum_{j=1}^n \xi_j$ is $\mathcal{I}_n$ measurable. To see this, let $\phi : \mathbb{R}^\infty \to \mathbb{R}$ be defined by $\phi(x_1, x_2, \dots) = x_1 + \cdots + x_n$. If$\pi$ is permutation of $\mathbb{N}$ such that $\pi(j) = j$ for $j > n$ then

$$\phi(x_1, x_2, \dots) = x_1 + \cdots + x_n = x_{\pi(1)} + \cdots + x_{\pi(n)} = \phi(x_{\pi(1)}, x_{\pi(2)}, \dots)$$

and we see that $\phi^{-1}A \in \mathcal{C}_n$ for all Borel sets $A \subset \mathbb{R}$. Writing $\sum_{j=1}^n \xi_j = \phi \circ \boldsymbol{\xi}$ we see that the sum is $\mathcal{I}_n$ measurable.

Next we show the defining property of conditional expectation, so let $A \in \mathcal{I}_n$ which we may write as $\boldsymbol{\xi}^{-1}B$ with $B \in \mathcal{C}_n$. Pick an arbitrary $j \in \mathbb{N}$ with $j \leq n$ and an arbitrary n-permutation π such that $\pi(j) = 1$ then we have by the expectation rule Lemma 3.7, the fact that $\pi B = B$ and exchangeability

$$\begin{aligned}\mathbf{E}[\xi_j; A] &= \int x_j \mathbf{1}_B(x)\, \boldsymbol{\xi}(dx) = \int x_1 \mathbf{1}_{\pi B}(\pi(x))\, \boldsymbol{\xi}(dx) \\ &= \int x_1 \mathbf{1}_B(x)\, (\pi \circ \boldsymbol{\xi})(dx) = \int x_1 \mathbf{1}_B(x)\, \boldsymbol{\xi}(dx) = \mathbf{E}[\xi_1; A]\end{aligned}$$

Therefore by linearity of conditional expectation, $\mathbf{E}\left[\frac{1}{n}\sum_{j=1}^n \xi_j; A\right] = \mathbf{E}[\xi_1; A]$ and the claim is verified.

Now defining $\mathcal{I}_\infty = \cap_{n=1}^\infty \mathcal{I}_n$, the almost sure convergence follows from the Levy Downward Theorem 9.56. In the i.i.d. case we know from the Kolmogorov $0-1$ Law (specifically Corollary 4.30) we know that the limit $\mathbf{E}[\xi_1 \mid \mathcal{A}_\infty]$ is almost surely constant in which case it must be equal to $\mathbf{E}[\xi_1]$. ☐

5. 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 9.68. 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 9.69. *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

$$\{\tau \leq t\} = \bigcap_{\substack{r \in \mathbb{Q} \\ r > t}} \{\tau < r\} \qquad \{\tau < t\} = \bigcup_{\substack{r \in \mathbb{Q} \\ r < t}} \{\tau \leq r\}$$

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. □

In the previous section we identified a useful class of optional times that we called hitting times. Hitting times can be defined in continuous time but there are more stringent requirements on when they are optional times.

LEMMA 9.70. *Let $\mathcal{F}$ be a filtration on $\mathbb{R}_+$, let X_t be an $\mathcal{F}$-adapted process with values in a measurable space $(S, \mathcal{S})$ where S is topological and $\mathcal{S}$ contains the Borel σ-algebra, $B \in \mathcal{S}$ and $\tau_B = \inf\{t > 0 \mid X_t \in B\}$. Then if S is a metric space, B*

is closed and X_t is continuous τ_B is $\mathcal{F}$-optional and if B is open and X_t is right continuous then τ_B is weakly $\mathcal{F}$-optional.

Proof. To see the first case, by the countability and density of the rationals in $\mathbb{R}_+$, continuity of X_t and closedness of B we know that $X_{\tau_B} \in B$ and therefore $\tau_B \leq t$ if and only if there is an $0 < s \leq t$ such that $X_s \in B$. This latter statement is true if and only if there is an integer $m > 0$ and points X_q with $q \in \mathbb{Q}$ and $1/m \leq q \leq t$ that are arbitrarily close to B. Translating this observation into set operations we get

$$\{\tau_B \leq t\} = \cup_{m=1}^{\infty} \cap_{n=1}^{\infty} \cup_{\substack{1/m \leq q \leq t \\ q \in \mathbb{Q}}} \{d(X_q, B) < \frac{1}{n}\} \in \mathcal{F}_t$$

because each $\{x \in S \mid d(x, B) < \frac{1}{n}\}$ is open and thus $\{d(X_q, B) < \frac{1}{n}\} \in \mathcal{F}_t$ because X is $\mathcal{F}$-adapted. To see the second case note that by similar considerations $\tau_B < t$ if and only if there exists a $q \in \mathbb{Q}$ such that $0 \leq q < t$ with $X_q \in B$ thus

$$\{\tau_B \leq t\} = \cup_{\substack{0 \leq q < t \\ q \in \mathbb{Q}}} \{X_q \in B\} \in \mathcal{F}_t$$

□

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 9.71. *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 9.72. *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 9.69, 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 9.69. □

We shall have a need for the following characterization of uniform integrability for martingales on $\mathbb{Z}_-$ (sometimes called a *backward submartingale*).

Lemma 9.73. *Let X_n be an $\mathcal{F}$-submartingale on $\mathbb{Z}_-$, then $\boldsymbol{E}[X_n]$ is bounded if and only if X_n is uniformly integrable.*

Proof. As a first simple observation, we know that since X_n is a submartingale then

$$\mathbf{E}[X_n] = \mathbf{E}[\mathbf{E}[X_n \mid \mathcal{F}_{n-1}]] \geq \mathbf{E}[X_{n-1}]$$

so boundedness of $\mathbf{E}[X_n]$ is equivalent to $\lim_{n \to -\infty} \mathbf{E}[X_n] = \inf_n \mathbf{E}[X_n] > -\infty$.

Assume that $\mathbf{E}[X_n]$ is L^1 bounded. We proceed by constructing the analogue of the Doob Decomposition for time index $\mathbb{Z}_-$ and then invoking results for martingales. Recall in the Doob Decomposition we write a submartingale X_n on $\mathbb{Z}_+$ as $M_n + A_n$ where M_n is a martingale and $A_n = \sum_{m=1}^n \mathbf{E}[X_m \mid \mathcal{F}_{m-1}] - X_{m-1}$. So to make this work for $\mathbb{Z}_-$ we have to handle the fact that the desired definitions now involve an infinite sum which must converge for things to make sense. To that end, define for $n \leq 0$,

$$\alpha_n = \mathbf{E}[X_n \mid \mathcal{F}_{n-1}] - X_{n-1} = \mathbf{E}[X_n - X_{n-1} \mid \mathcal{F}_{n-1}] \geq 0$$

so that α_n is a predictable process. Observe that by Monotone Convergence

$$\begin{aligned}
\mathbf{E}\left[\sum_{n \leq 0} \alpha_n\right] &= \lim_{m \to \infty} \sum_{-m \leq n \leq 0} \mathbf{E}[\alpha_n] \\
&= \lim_{m \to \infty} \sum_{-m \leq n \leq 0} \mathbf{E}[X_n] - \mathbf{E}[X_{n-1}] \\
&= \mathbf{E}[X_0] - \inf_n \mathbf{E}[X_n] < \infty
\end{aligned}$$

Therefore we know that $\sum_{n \leq 0} \alpha_n$ is almost surely finite. With that in hand we can define for each $n \leq 0$

$$A_n = \sum_{m \leq n} \alpha_m = \sum_{m \leq n} \mathbf{E}[X_m \mid \mathcal{F}_{m-1}] - X_{m-1}$$

so that A_n is integrable. Moreover since A_n is almost surely increasing we know that $\sup_n A_n \leq A_0$ and therefore the sequence A_n is uniformly integrable (e.g. see Example 5.50). Now we define

$$M_n = X_n - A_n$$

so that by integrability of A_n we have M_n is integrable and moreover

$$\begin{aligned}\mathbf{E}[M_n \mid \mathcal{F}_{n-1}] &= \mathbf{E}[X_n \mid \mathcal{F}_{n-1}] - A_n \\ &= \mathbf{E}[X_n \mid \mathcal{F}_{n-1}] - \mathbf{E}[X_n \mid \mathcal{F}_{n-1}] + X_{n-1} - A_{n-1} = M_{n-1}\end{aligned}$$

so that M_n is a martingale. Since M_n is closed we conclude from Theorem 9.51 that M_n is uniformly integrable. The uniform integrability of A_n and M_n together imply the uniform integrability of X_n (Lemma 5.53).

Now if we assume that X_n is uniformly integrable then it follows that X_n is L^1 bounded (Lemma 5.52) and therefore $\mathbf{E}[X_n]$ is bounded since $|\mathbf{E}[X_n]| \leq \mathbf{E}[|X_n|]$. ☐

The martingale results for discrete time tell us quite a bit about what can happen in continuous time as well. If we are given a submartingale on $\mathbb{R}_+$ then we can restrict it to $\mathbb{Q}_+$ and ask what we know about the restricted process; as we'll soon see we know quite a lot! The first issue which we examine gets to the heart of whether we can extrapolate from the discrete case to the continuous case. If there are no restrictions on the regularity/continuity of sample paths then there is very little that we can say about what happens on $\mathbb{R}_+ \setminus \mathbb{Q}_+$ based on what is happening on $\mathbb{Q}_+$. Thus our first task is to understand the ways in which we can modify a continuous time submartingale to get a different submartingale that has some continuity in sample paths. Note here that the use of the word modify is quite a bit subtle: we mean to use the word both in its colloquial sense of *how can we change continuous time submartingale to make it have regular sample paths* as well as the technical sense of *when are the changes that we make to a continuous time submartingale a modification of the stochastic process.* The specific type of regularity we aim for is that sample paths of the submartingale are right continuous and have left limits. It is traditional to refer to such paths as cadlag which is an acronym derived from the French phrase *continue à droits limite à gauche.* The reader may encounter the acronym rcll derived from the English but the general consensus is that cadlag is more euphonious and is therefore preferred. Moreover the French lends itself to the acronym caglad to describe paths that are continuous on the left and have right limits. Processes with caglad paths are less common than those with cadlag paths but will come up as integrands in the theory of stochastic integration.

One last subtle point is the difference between sample path properties being assumed to hold for every sample path versus only holding for almost every sample path. As a general principle one might expect that we gloss over the distinction. There is a subtle danger in doing so. In the event that a process has a particular sample path property almost surely and is adapted to a given filtration $\mathcal{F}$, there is an indistinguishable process that has the sample path property everywhere but the latter process may no longer be adapted to $\mathcal{F}$. Scenarios do come up in which the filtration must be respected and therefore we call out the case of a process posessing cadlag sample paths almost surely as a separate case from the one in which it possess cadlag sample paths surely.

Definition 9.74. A stochastic process X with time scale an subinterval of $\mathbb{R}$ is said to be *cadlag* if every path X_t has finite left limits and is continuous on the right. Such a process will be said to be *almost surely cadlag* if these path properties hold for almost every sample path.

The formal development of these ideas comes with a lot of technical baggage so before we jump into the details let's step back and think about what we can expect. We discuss the martingale case here even though almost all of what we say applies equally to submartingales. Let's suppose that we have an $\mathcal{F}$-martingale X on $\mathbb{R}_+$. If we restrict a X to $\mathbb{Q}_+$ then the Martingale Convergence Theorem (and at its core the upcrossing lemma) tells us that almost surely on any bounded interval the restricted martingale has limits along all montone sequences. Therefore at worst the restricted martingale on $\mathbb{Q}_+$ has jump discontinuities (almost surely!). This gives us hope that we can modify (in the colloquial sense) X to create a new process Y on $\mathbb{R}_+$ such that Y is cadlag: simply define Y_t for $t \in \mathbb{R}_+$ such that $Y_t = \lim_{\substack{q \downarrow t \\ q \in \mathbb{Q}}} X_q$. This gives us a process to be sure but it isn't even $\mathcal{F}$-adapted in general: by the definition of Y_t as limit of X_q for $q > t$ we know that Y_t is $\mathcal{F}_t^+$-measurable but it is not necessarily $\mathcal{F}_t$-measurable. This introduces one of the key ideas: if we have any hope of changing X to get an adapted cadlag Y we had either be prepared to pass to the filtration $\mathcal{F}^+$ or start with a right continuous one.

We've already glossed over an issue that brings up a second key idea. The construction described only works *almost surely*; we have to come up with a different plan on the null set where X_q doesn't have limits. The easiest thing to do is just to set $Y_t \equiv 0$ when this occurs. Since the event of X on $\mathbb{Q}_+$ being ill-behaved has probability zero whatever it is we decide to do won't prevent Y from being a version of X. The issue is that the event of X on $\mathbb{Q}_+$ being ill-behaved depends on all of X_t for all $t \geq 0$ hence is in $\mathcal{F}_\infty$; thus as we modify X_t to get Y_t, in accounting for the ill-behavedness of X_t we are changing each Y_t on an event in $\mathcal{F}_\infty$ further destroying adaptedness of Y. The good news is that we do know that the event in question is a null event and therefore we come to the second key idea: to get a cadlag Y from X we had either be prepared to add all of the null events of $\mathcal{F}_\infty$ to each $\mathcal{F}_t^+$ or assume that they are there to begin with. The filtration that is right continuous and has null sets added is referred to as the *partial augmentation* of $\mathcal{F}$ (it is distinguished from the full augmentation in that it does not assume that the filtration is complete).

These first two ideas are enough to get us an adapted process Y but more is true: Y is a martingale with respect the partially augmented filtration. This is not obvious and requires checking using discrete time results. The remaining issue and question is whether Y is indeed a version of X. The following example shows that this may not be true without further hypotheses on X.

EXAMPLE 9.75. Let $\Omega = \{-1, 1\}$ with the probability measure $\mathbf{P}\{1\} = \mathbf{P}\{-1\} = \frac{1}{2}$. Let $\mathcal{F}_t = \{\Omega, \emptyset\}$ for $0 \leq t \leq 1$ and $\mathcal{F}_t = \{\Omega, \emptyset, \{1\}, \{-1\}\}$ for $t > 1$ let

$$X_t(\omega) = \begin{cases} 0 & \text{for } 0 \leq t \leq 1 \\ \omega & \text{for } t > 1 \end{cases}$$

It is easy to see that X_t is an $\mathcal{F}$-martingale. Now define

$$Y_t(\omega) = \begin{cases} 0 & \text{for } 0 \leq t < 1 \\ \omega & \text{for } t \geq 1 \end{cases}$$

and note that Y_1 is not $\mathcal{F}_1$-measurable. However, it is easy to see that $\mathcal{F}_t^+ = \{\Omega, \emptyset\}$ for $0 \leq t < 1$, $\mathcal{F}_t^+ = \{\Omega, \emptyset, \{1\}, \{-1\}\}$ for $t \geq 1$ and Y_t is an $\mathcal{F}^+$-martingale. Note however that $\mathbf{P}\{X_1 = Y_1\} = 0 \neq 1$ and thus Y is not a version of X.

Note that X is not a $\mathcal{F}^+$-martingale (or $\mathcal{F}^+$-sub/supermartingale) as for $t > 1$ we have $\mathbf{E}\left[X_t \mid \mathcal{F}_1^+\right] = X_t$ and therefore $\mathbf{E}\left[X_t \mid \mathcal{F}_1^+\right] > X_1$ with probability $1/2$ (i.e. when $\omega = 1$) and $\mathbf{E}\left[X_t \mid \mathcal{F}_1^+\right] < X_1$ with probability $1/2$ (i.e. when $\omega = -1$).

Example 9.75 shows that there is a limit to what we can accomplish by taking a martingale with respect to an arbitrary filtration and trying find a version that is cadlag. Nonetheless, the method we've outlined to make a cadlag process Y from an arbitrary process X can be shown to result in a version if X is assumed to be a martingale with respect to the the right continuous filtration in the first place (plus some extra conditions if X is only assumed to be a submartingale). Thus the impediment to the existence of a cadlag version in Example 9.75 is in the final comment about X not being a martingale with respect to the right continuous filtration.

THEOREM 9.76. *Let X_t be a $\mathcal{F}$-submartingale on $\mathbb{R}_+$ and let Y_q denote the restriction to $\mathbb{Q}_+$.*

(i) *There exists a set $A \subset \mathcal{F}_\infty$ with $\boldsymbol{P}\{A\} = 1$ on which $\lim_{q \to t^+} Y_q$ and $\lim_{a \to t^-} Y_q$ exist for all $t \in \mathbb{R}_+$. If we define*
$$Z_t(\omega) = \begin{cases} \lim_{q \to t^+} Y_q(\omega) & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}$$
then Z is a cadlag $\overline{\mathcal{F}_+}$-submartingale.

(ii) *X has a right continuous version if and only if Z is a version of X.*

(iii) *If $\mathcal{F}$ is right continuous then Z is a version of X if and only if $t \mapsto \boldsymbol{E}[X_t]$ is right continuous. Moreover in this case, there is a version $\tilde{Z}$ with almost surely cadlag paths which is an $\mathcal{F}$-martingale.*

PROOF. Pick $t \geq 0$ and note that since Y_q^+ is a submartingale we have for all $q \in [0,t] \cap \mathbb{Q}$
$$\mathbf{E}\left[Y_q^+\right] \leq \mathbf{E}\left[\mathbf{E}\left[Y_t^+ \mid \mathcal{F}_q\right]\right] = \mathbf{E}\left[Y_t^+\right]$$
and by the same reasoning using the fact that Y_q is a submartingale we know that $\mathbf{E}[Y_0] \leq \mathbf{E}[Y_q]$. Together these imply
$$\mathbf{E}[|Y_q|] = \mathbf{E}\left[Y_q^+\right] + \mathbf{E}\left[Y_q^-\right] = 2\mathbf{E}\left[Y_q^+\right] - \mathbf{E}[Y_q] \leq 2\mathbf{E}\left[Y_t^+\right] - \mathbf{E}[Y_0]$$
which implies that Y_q restricted to $[0,t]$ is L^1-bounded. We can apply Theorem 9.51 to the restricted submartingale Y_q to construct $A_t \in \mathcal{F}_t$ with $\mathbf{P}\{A_t\} = 1$ such that for all increasing and decreasing sequences q_n in $\mathbb{Q} \cap [0,t]$ we have Y_{q_n} converges on A_t. So in particular $\lim_{q \to t^-} Y_q$ and $\lim_{q \to t^+} Y_q$ exist for every $t \in [0,t]$ on A_t. Taking the intersection of $A = \cap_{N=1}^\infty A_N$ we see that $\lim_{q \to t^-} Y_q$ and $\lim_{q \to t^+} Y_q$ exist for every $t \in \mathbb{R}_+$ on $A \in \mathcal{F}_\infty$. Note that the process Z_t is $\overline{\mathcal{F}_+}$-adapted (in fact is adapted with respect to the smaller filtration $\mathcal{G}_t = \sigma(A, \mathcal{F}_t^+)$). We claim that the process Z_t is cadlag. Clearly sample paths on A^c are continuous since they are constant. So we consider a sample path on A. If we fix $t \geq 0$ and $\epsilon > 0$ then by definition of Z_t on A we may find a $\delta > 0$ such that $|Y_q - Z_t| < \epsilon/2$ for all $t < q < t + \delta$ and $q \in \mathbb{Q}$. If we take an arbitrary $t < s < t + \delta$ then again applying the definition of Z_s we may pick a $s < q < t + \delta$ such that $|Y_q - Z_s| < \epsilon/2$,

therefore by the triangle inequality we have $|Z_t - Z_s| < \epsilon$ and right continuity is established. Similarly if we define $Y_t^- = \lim_{q \to t^-} Y_q$, we may find a δ such that $\left|Y_t^- - Y_q\right| < \epsilon/2$ for all $t - \delta < q < t$. For any $t - \delta < s < t$ by the definition of Z_s we may pick $s < q < t$ such that $|Y_q - Z_s| < \epsilon/2$ and again the triangle inequality implies $\left|Y_t^- - Z_s\right| < \epsilon$. This shows $\lim_{s \to t^-} Z_s = \lim_{q \to t^-} Y_q$ and in particular Z_t has left limits.

Now to see that Z is a submartingale, let $0 \leq s < t < \infty$ be arbitrary and pick a decreasing sequence $t_n \in \mathbb{Q}_+$ such that $t_n \downarrow t$ and a decreasing sequence $s_n \in \mathbb{Q}_+$ such that $s_n < t$ for all n and $s_n \downarrow s$. For each n and m we have $Y_{s_m} \leq \mathbf{E}\left[Y_{t_n} \mid \mathcal{F}_{s_m}\right]$ a.s. by the submartingale property of X. By the Levy Downward Theorem 9.56 we know that $\lim_{m \to \infty} \mathbf{E}\left[Y_{t_n} \mid \mathcal{F}_{s_m}\right] = \mathbf{E}\left[Y_{t_n} \mid \mathcal{F}_s^+\right]$ a.s. and by definition of Z we know $Z_s = \lim_{m \to \infty} Y_{s_m}$ a.s. therefore $Z_s \leq \mathbf{E}\left[Y_{t_n} \mid \mathcal{F}_s^+\right]$ a.s. Again, by the definition of Z we have $Y_{t_n} \stackrel{a.s.}{\to} Z_t$, furthermore as the sequence t_n is bounded we have already shown Y_{t_n} is L^1-bounded. This allows us to apply Lemma 9.73 to conclude that Y_{t_n} is uniformly continuous hence $Y_{t_n} \stackrel{L^1}{\to} Z_t$ by Lemma 5.58. Thus we have

$$Z_s \leq \lim_{n \to \infty} \mathbf{E}\left[Y_{t_n} \mid \mathcal{F}_s^+\right] = \mathbf{E}\left[Z_t \mid \mathcal{F}_s^+\right] = \mathbf{E}\left[Z_t \mid \overline{\mathcal{F}}_s^+\right]$$

where the last equality follows by Lemma 8.17.

It is worth noting that the submartingale property also holds with respect to the smaller filtration $\mathcal{G}_t$ alluded to above; the fact that the result is expressed in terms of the augmented filtration is something of a tradition and is due to the fact that the augmented filtration proves to be necessary in subsequent theory. The tradition is not without its shortcomings; when we get to the discussion of Girsanov theory it will be inconvenient to require the full completion of $\mathcal{F}^+$.

Suppose that X has a right continuous version W. Then by taking an intersection of almost sure events, we see that almost surely $Y_q = W_q$ for all $q \in \mathbb{Q}_+$ (Y continues to denote the restriction of X to $\mathbb{Q}$). If we fix a particular $t \geq 0$ and use the fact that W is a version of X, the right continuity of W_q and the definition of Z to see that almost surely we see

$$X_t = W_t = \lim_{q \to t^+} W_q = \lim_{q \to t^+} Y_q = Z_t$$

and therefore Z is a version of X.

We now assume that $\mathcal{F}$ is right continuous. Before proceeding to show that Z is a version of X if and only if $\mathbf{E}\left[X_t\right]$ is right continuous we need two small computations. We have already observed that for every sequence $t_n \downarrow t$ with $t \geq 0$ and $t_n \in \mathbb{Q}_+$ we have not only does $Y_{t_n} \stackrel{a.s.}{\to} Z_t$ but also $Y_{t_n} \stackrel{L^1}{\to} Z_t$. From this fact and the definition of Y we get

$$\lim_{t_n \to t} \mathbf{E}\left[X_{t_n}\right] = \lim_{t_n \to t} \mathbf{E}\left[Y_{t_n}\right] = \mathbf{E}\left[Z_t\right]$$

Moreover using the submartingale property of X, the definition of Y, the fact that $Y_{t_n} \stackrel{L^1}{\to} Z_t$, Lemma 8.17 and the $\overline{\mathcal{F}}$-adaptedness of Z we get

$$X_t \leq \lim_{t_n \to t} \mathbf{E}\left[X_{t_n} \mid \mathcal{F}_t\right] = \lim_{t_n \to t} \mathbf{E}\left[Y_{t_n} \mid \mathcal{F}_t\right] = \mathbf{E}\left[Z_t \mid \mathcal{F}_t\right] = \mathbf{E}\left[Z_t \mid \overline{\mathcal{F}}_t\right] = Z_t \text{ a.s.}$$

Now we suppose that $\mathbf{E}\left[X_t\right]$ is a right continuous function of t and we want to show that Z is a version of X. From the above two observations and the right continuity of $\mathbf{E}[X_t]$ we get

$$\mathbf{E}[|Z_t - X_t|] = \mathbf{E}[Z_t - X_t] = \mathbf{E}[Z_t] - \mathbf{E}[X_t] = 0$$

which shows $X_t = Z_t$ a.s. (i.e. Z is a version of X).

Now if we assume that Z is a version of X then playing the above argument backward, we conclude that

$$\mathbf{E}[X_t] = \mathbf{E}[Z_t - X_t] + \mathbf{E}[Z_t] = \mathbf{E}[Z_t] = \lim_{t_n \to t} \mathbf{E}[X_{t_n}]$$

which shows that $\mathbf{E}[X_t]$ is right continuous.

To see that last piece, for each $t \geq 0$ define $\tilde{A}_t = \cap_{s>t} A_s$ and $\tilde{Z}_t = \mathbf{1}_{\tilde{A}_t} \lim_{q \to t^+} X_q$. Note that $\tilde{Z}$ is $\mathcal{F}$ adapted (by right continuity). Furthermore $\tilde{Z}$ and Z are indistinguishable (specifically they agree on A). From the properties of Z it follows that $\tilde{Z}$ is a version of X, has cadlag paths almost surely and is an $\mathcal{F}$ martingale. □

Since the condition of right continuity of $\mathbf{E}[X_t]$ is trivially satisfied in the case of a martingale we know that given any martingale X on a right continuous filtration we may find a version of X that is a cadlag martingale on the completion of that filtration (actually the completion is quite a bit more than is required as seen from the proof). For the most part we don't worry too much about the fact that the filtration has to be enlarged and in much of the theory we define the problem away by assuming that our filtration is both right continuous and complete to begin with. However, there are cases in which passing to the completion can cause real issues (e.g. one loses the Borel space property by adding in additional sets). Note that need to enlarge the space to the completion arises from handling those places in which right limits do not exist. If one knows for some other reason (or is willing to assume) that these limits exist then one can dispense with the addition of null sets and it suffices to take the right continuous filtration. Alternatively if may be possible to make due with a version that is cadlag almost surely instead of everywhere and in that case we can make due with a right continuous filtration $\mathcal{F}$ by the last statement in the theorem.

LEMMA 9.77. *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 9.44) we get the first result. By Claim 1 and the countable index set L^p inequality we get the second result. □

Lemma 9.78 (Doob's L^p Inequality). *Let X_t be a non-negative submartingale on $\mathbb{R}_+$ with X_t and $\mathcal{F}$ right continuous, then for all $p > 1$ and $0 \leq t < \infty$,*

$$\left\| \sup_{0 \leq s \leq t} X_s \right\|_p \leq \frac{p}{p-1} \left\| X_t \right\|_p$$

Proof. TODO: □

Theorem 9.79 (L^1 Submartingale Convergence Theorem). *Let X_t be a cadlag $\mathcal{F}$-submartingale on $\mathbb{R}_+$ such that $\sup_{0 \leq t < \infty} \left\| X_t \right\|_1 < \infty$ then there exists an $X \in L^1$ such that $X_t \overset{a.s.}{\to} X$ a.s.*

Proof. Restricting X_t to $\mathbb{Q}_+$ and applying Theorem 9.50 we know that there exists X such that $\lim_{\substack{q \to \infty \\ q \in \mathbb{Q}_+}} X_q = X$ almost surely. By right continuity of X we also get that $\lim_{t \to \infty} X_t = X$ almost surely (let $\epsilon > 0$ be given, for almost every ω we pick N_ω such that $|X_q(\omega) - X(\omega)| \leq \epsilon$ for all $q > N_\omega$ then for any $t > N_\omega$ we have $|X_t(\omega) - X(\omega)| = \lim_{\substack{q \downarrow t \\ q \in \mathbb{Q}_+}} |X_t(\omega) - X(\omega)| \leq \epsilon$). □

NOTE: It is also true that a UI submartingale is L^1 convergent but it is not true that an L^1 convergent submartingale must be UI (while that is true in discrete time). Make this into a result somewhere and provide the counterexample (I am pretty sure Rogers and Williams has one).

Theorem 9.80 (Martingale Closure Theorem). *Let X_t be a cadlag martingale then the following are equivalent*

(i) *X_t is uniformly integrable*
(ii) *there exists an integrable X such that $X_t \overset{L^1}{\to} X$*
(iii) *there exists an integrable X such that $X_t = \boldsymbol{E}[X \mid \mathcal{F}_t]$ almost surely.*

Proof. Given the result Theorem 9.79, the proof is essentially identical to the discrete time case. To see (i) implies (ii) we know from Lemma 5.52 that X_t uniformly integrable implies L^1 boundedness, hence we can apply Theorem 9.79 to conclude the existence of an integrable X such that $X_t \overset{a.s.}{\to} X$. However almost sure convergence implies convergence in probability (Lemma 5.5) which together with uniform integrability implies $X_t \overset{L^1}{\to} X$ (Lemma 5.58).

To see that (ii) implies (iii) from $X_t \xrightarrow{L^1} X$ we get for any σ-algebra $\mathcal{G}$,

$$\lim_{t\to\infty} \|\mathbf{E}[X_t \mid \mathcal{G}] - \mathbf{E}[X \mid \mathcal{G}]\|_1 \leq \lim_{t\to\infty} \mathbf{E}[\mathbf{E}[|X_t - X| \mid \mathcal{G}]] = \lim_{t\to\infty} \mathbf{E}[|X_t - X|] = 0$$

and therefore for any fixed $s \geq 0$ and the martingale property $X_s = \mathbf{E}[X_t \mid \mathcal{F}_s]$ a.s. we have

$$\begin{aligned}\|X_s - \mathbf{E}[X \mid \mathcal{F}_s]\|_1 &\leq \lim_{t\to\infty} \|X_s - \mathbf{E}[X_t \mid \mathcal{F}_s]\|_1 + \lim_{t\to\infty} \|\mathbf{E}[X_t \mid \mathcal{F}_s] - \mathbf{E}[X \mid \mathcal{F}_s]\|_1 \\ &= \lim_{t\to\infty} \|\mathbf{E}[X_t \mid \mathcal{F}_s] - \mathbf{E}[X \mid \mathcal{F}_s]\|_1 = 0\end{aligned}$$

and we get that $X_s = \mathbf{E}[X \mid \mathcal{F}_s]$ a.s.

To see that (ii) implies (iii), we simply invoke Corollary 9.52. ☐

The L^p version of the Martingale Convergence Theorem follows just as in the discrete case.

THEOREM 9.81 (L^p Martingale Convergence). *Given a martingale M_t on $\mathbb{R}_+$, then for $p > 1$, there exists an $M \in L^p$ such that $M_t \xrightarrow{L^p} M$ if and only if M_t is L^p bounded. In fact, if $M_t \xrightarrow{L^1} M$ with $M \in L^p$ then $M_n \xrightarrow{L^p} M$ and M_n is L^p bounded.*

PROOF. TODO: Make this an exercise. ☐

THEOREM 9.82. *Let X_t be an $\mathcal{F}$-submartingale on $\mathbb{R}_+$ with X_t and $\mathcal{F}$ right continuous, let σ and τ be optional times with τ bounded, then X_τ is integrable and*

$$\boldsymbol{E}[X_\tau \mid \mathcal{F}_\sigma] \leq X_{\tau\wedge\sigma} \text{ a.s.}$$

PROOF. For each $n > 0$, restrict X to the dyadic rationals $X_{k/2^n}$ on the filtration $\mathcal{F}_{k/2^n}$. Is is immediate that this is a discrete submartingale.

Define the discrete approximations of optional times $\tau_n = \frac{1}{2^n}\lfloor 2^n\tau + 1\rfloor$ and $\sigma_n = \frac{1}{2^n}\lfloor 2^n\sigma + 1\rfloor$ so that τ_n and σ_n are optional times such that $\tau_n \downarrow \tau$ and $\sigma_n \downarrow \sigma$ (Lemma 9.71) and furthermore $\mathcal{F}_\sigma = \cap_n \mathcal{F}_{\sigma_n}$ (Lemma 9.72 and right continuity of $\mathcal{F}$). We can now apply the Optional Sampling Theorem Corollary 9.41 to conclude that for each $m, n > 0$,

$$\mathbf{E}[X_{\tau_n} \mid \mathcal{F}_{\sigma_m}] \geq X_{\tau_n\wedge\sigma_m} \text{ a.s.}$$

Now holding n fixed we note that since σ_m is decreasing in m we have $\mathcal{F}_{\sigma_1} \supset \mathcal{F}_{\sigma_2} \supset \cdots$ and therefore we can apply the downward Levy-Jessen Theorem 9.56 and right continuity of X_t to conclude

$$\mathbf{E}[X_{\tau_n} \mid \mathcal{F}_\sigma] = \lim_{m\to\infty} \mathbf{E}[X_{\tau_n} \mid \mathcal{F}_{\sigma_m}] \geq \lim_{m\to\infty} X_{\tau_n\wedge\sigma_m} = X_{\tau_n\wedge\sigma} \text{ a.s.}$$

Now we need to justify taking the limit $n \to \infty$. To do this, we claim that the sequence of random variable X_{τ_n} is a backward submartingale; that is to say if we consider $X_{\tau_{-n}}$ and the filtration $\mathcal{F}_{\tau_{-n}}$ for every $n < 0$ then $X_{\tau_{-n}}$ is an $\mathcal{F}_{\tau_{-n}}$-submartingale on $\mathbb{Z}_-$. The fact that $X_{\tau_{-n}}$ is a submartingale follows from the fact that τ_n is decreasing and Optional Sampling (Corollary 9.41) together with our awkward indexing

$$\mathbf{E}\left[X_{\tau_{-n}} \mid \mathcal{F}_{\tau_{-(n-1)}}\right] \geq X_{\tau_{-(n-1)}} \text{ a.s.}$$

Furthermore we can show that $\mathbf{E}[X_{\tau_n}]$ is bounded because τ is bounded. If we pick $T > 0$ such that $0 \leq \tau \leq T$ then we have an upper bound

$$\mathbf{E}[X_{\tau_n}] = \mathbf{E}[\mathbf{E}[X_T \mid \mathcal{F}_{\tau_n}]] = \mathbf{E}[X_T] < \infty$$

and a lower bound from

$$-\infty < \mathbf{E}[X_0] \le \mathbf{E}[\mathbf{E}[X_{\tau_n} \mid \mathcal{F}_0]] = \mathbf{E}[X_{\tau_n}]$$

By Lemma 9.73 we can now conclude that X_{τ_n} is uniformly integrable. So now we pick $A \in \mathcal{F}_\sigma$ and by right continuity of X_t we have

$$\lim_{n\to\infty} X_{\tau_n}\mathbf{1}_A = X_\tau \mathbf{1}_A \text{ a.s.}$$

and

$$\lim_{n\to\infty} X_{\tau_n\wedge\sigma}\mathbf{1}_A = X_{\tau\wedge\sigma}\mathbf{1}_A \text{ a.s.}$$

and therefore by uniform integrability and Lemma 5.58 we get

$$\mathbf{E}[X_\tau; A] = \lim_{n\to\infty} \mathbf{E}[X_{\tau_n}; A] \ge \lim_{n\to\infty} \mathbf{E}[X_{\tau_n\wedge\sigma}; A] = \mathbf{E}[X_{\tau\wedge\sigma}; A]$$

which shows $\mathbf{E}[X_\tau \mid \mathcal{F}_\sigma] \ge X_{\tau\wedge\sigma}$ a.s. by the defining property and monotonicity of conditional expectation. □

6. 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 possible to talk about the regularity of sample paths.

As we make additional assumptions about the structure of the time scale T, we can discuss measurability, continuity and even differentiability of sample paths. For the moment, there is a very mild restriction that we make that uses just the structure of a measure space on the time scale.

Definition 9.83. Let $(\Omega, \mathcal{A})$, $(S, \mathcal{S})$ and $(T, \mathcal{T})$ be measurable spaces. A process X on T with values in S is said to be *jointly measurable* or simply *measurable* if $X : \Omega \times T \to S$ is $\mathcal{A} \otimes \mathcal{T}/\mathcal{S}$ measurable.

When the time scale and state space of a process are topological spaces then we can discuss continuity of sample paths. Here we begin with most important case of metric spaces and show that it implies joint measurability.

Lemma 9.84. *Let S be a metric space and let T be a separable metric space both given the Borel σ-algebra. Suppose a process X on T with values in S has continuous sample paths, then X is jointly measurable.*

Proof. Let $\{t_n\}$ be a countable dense set of points in T. We use this dense set to provide a sequence of approximations to X. To this end, for each $n \ge 1$ and $k \ge 1$ define

$$\begin{aligned} B_{n,k} &= \{t \in T \mid d(t, t_k) < 1/n\} \\ V_{n,k} &= B_{n,k} \setminus \cup_{j=1}^{k-1} B_{n,j} \end{aligned}$$

Clearly the $V_{n,k}$ are Borel measurable since the $B_{n,k}$ are open. By construction they are disjoint and by density of the t_k they cover T. Define

$$X_t^n(\omega) = X_{t_k}(\omega) \text{ for } t \in V_{n,k}$$

Since for any $A \in \mathcal{B}(S)$ we have $\{(\omega, t) \mid X_t^n(\omega) \in A\} = \cup_{k=1}^\infty V_{n,k} \times \{X_{t_k} \in A\}$ we see that X^n is jointly measurable.

Now by density of X^n and the continuity of sample paths of X we have $\lim_{n\to\infty} X^n = X$ and joint measurability of X follows from Lemma 2.15. □

Definition 9.85. 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 9.86. 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 9.87. *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, in particular a progressive process is jointly 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 $A_1, A_2, \dots \in \mathcal{PM}$, 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 9.88. 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 $\{\tau \leq t\} = [0, t] \in \mathcal{B}([0, 1])$ is seen to be an optional time. Because $\mathcal{F}_t = \mathcal{B}([0, 1])$ we see that for every Borel measurable A, $A \cap \{\tau \leq t\} = A \cap [0, t] \in \mathcal{F}_t$ 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 2.87) a progressively measurable process is adapted.

LEMMA 9.89. *Let X be a process on $\mathbb{R}_+$ with values in a metric space $(S, \mathcal{B}(S))$ adapated to the filtration $\mathcal{F}$. Suppose X has left or right continuous sample paths, then X is $\mathcal{F}$-progressively measurable.*

PROOF. The proof is analogous to Lemma 9.84. We give the proof for right continuous sample paths with the case of left continuous sample paths being very similar.

Let $t \geq 0$ be given X^n be the process on $[0, t]$ be defined by $X^n_s(\omega) = X_{\frac{k+1}{2^n} \wedge t}(\omega)$ for $\frac{k}{2^n} < s \leq \frac{k+1}{2^n} \wedge t$. It is clear that for any $A \in \mathcal{B}(S)$ we have

$$\begin{aligned}
&\{(\omega, s) \mid 0 \leq s \leq t; X^n_s(\omega) \in A\} \\
&= \bigcup_{k=0}^{\lfloor 2^n t \rfloor} \{(\omega, s) \mid 0 \leq s \leq t; \frac{k}{2^n} < s \leq \frac{k+1}{2^n} \wedge t; X_{\frac{k+1}{2^n} \wedge t}(\omega) \in A\} \\
&= \bigcup_{k=0}^{\lfloor 2^n t \rfloor} \{X_{\frac{k+1}{2^n} \wedge t}(\omega) \in A\} \times (\frac{k}{2^n} < s \leq \frac{k+1}{2^n}] \in \mathcal{F}_t \otimes \mathcal{B}[0, t]
\end{aligned}$$

which shows that X^n is progressively measurable. By right continuity, we see that $\lim_{n \to \infty} X^n = X \mid_{\Omega \times [0,t]}$ and therefore by Lemma 2.15 we have X is progressively measurable. ☐

LEMMA 9.90. *Let X be an $\mathcal{F}$-progressively measurable process on $\mathbb{R}_+$ with values in a measurable space $(S, \mathcal{S})$ and let τ be an $\mathcal{F}$-optional time, then X_τ is $\mathcal{F}_\tau$-measurable. Moreover, the stopped process X^τ is $\mathcal{F}$-progressively measurable, in particular $X_{\tau \wedge t}$ is $\mathcal{F}_t$ measurable for all $t \geq 0$.*

PROOF. We first claim that if we can prove $X_{\tau \wedge t}$ is $\mathcal{F}_t$-measurable for all $t \geq 0$ then it follows that X_τ is $\mathcal{F}_\tau$-measurable. This follows from picking a measurable set $A \in \mathcal{S}$ and noting that

$$\{\tau \leq t\} \cap \{X_\tau \in A\} = \{\tau \leq t\} \cap \{X_{\tau \wedge t} \in A\}$$

which is $\mathcal{F}_t$ since τ is $\mathcal{F}$-optional and we have assumed $\{X_{\tau \wedge t} \in A\} \in \mathcal{F}_t$.

To see that X^τ is an $\mathcal{F}$-progressively measurable process, pick a $t \geq 0$ and consider the restriction of X^τ to $\Omega \times [0, t]$. Note that by replacing τ with $\tau \wedge t$, we can assume that $\tau \leq t$ which implies τ is $\mathcal{F}_t$-measurable (to see this note that for $s \leq t$, $\{\tau \leq s\} \in \mathcal{F}_s \subset \mathcal{F}_t$ and for $s > t$, $\{\tau \leq s\} = \Omega$). Now we can factor the restriction of $X_{\tau \wedge t}$ to $\Omega \times [0,t]$ as $X^\tau = X \mid_{\Omega \times [0,t]} \circ T^t$ where $T^t : \Omega \times [0,t] \to \Omega \times [0,t]$ is defined by $T^t(\omega, s) = (\omega, \tau(\omega) \wedge s)$. We claim that T^t is $\mathcal{F}_t \otimes \mathcal{B}([0,t])$-measurable. This follows from the $\mathcal{F}_t \otimes \mathcal{B}([0,t])$-measurability of $(\omega, s) \mapsto \tau(\omega) \wedge s$ which follows by noting that for every $0 \leq u \leq t$,

$$\{\tau \wedge s \leq u\} = \{\tau \leq u\} \times [0,t] \cup \Omega \times [0,u] \in \mathcal{F}_t \otimes \mathcal{B}([0,t])$$

As $X \mid_{\Omega \times [0,t]}$ is $\mathcal{F}_t \otimes \mathcal{B}([0,t])/\mathcal{S}$-measurable by progressive measurability of X, the claim follows from Lemma 2.13. The fact that $X_{\tau \wedge t}$ is $\mathcal{F}_t$-measurable for all $t \geq 0$ follows from the fact that progressive measurability implies adaptedness. $\square$

CHAPTER 10

Concentration Inequalities

LEMMA 10.1 (Markov Inequality). *Let ξ be a positive integrable random variable. Then $\boldsymbol{P}\{\xi > t\} \le \frac{E(\xi)}{t}$*

PROOF. $E(\xi) \ge E(\xi \mathbf{1}_{\{\xi > t\}}) \ge E(t \mathbf{1}_{\{\xi > t\}}) = t\mathbf{P}\{\xi > t\}$ $\square$

LEMMA 10.2 (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. $\mathbf{P}\{|\xi - \mu| > t\} = \mathbf{P}\{(\xi - \mu)^2 > t^2\} \le \frac{\mathbf{E}\left[(\xi-\mu)^2\right]}{t^2} = \frac{\sigma^2}{t^2}$ $\square$

LEMMA 10.3 (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 $c > 0$.

$$\begin{aligned} \mathbf{P}\{\xi > \lambda\} &= \mathbf{P}\{\xi + c > \lambda + c\} \\ &\le \mathbf{P}\{(\xi + c)^2 > (\lambda + c)^2\} && \text{because } \lambda + c > 0 \\ &\le \frac{\mathbf{E}\left[\xi^2\right] + c^2}{(\lambda + c)^2} \end{aligned}$$

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.

$$\begin{aligned} \frac{\mathbf{E}\left[\xi^2\right] + (\frac{\mathbf{E}\left[\xi^2\right]}{\lambda})^2}{(\lambda + \frac{\mathbf{E}\left[\xi^2\right]}{\lambda})^2} &= \frac{\mathbf{E}\left[\xi^2\right](1 + \frac{\mathbf{E}\left[\xi^2\right]}{\lambda^2})}{\lambda^2(1 + \frac{\mathbf{E}\left[\xi^2\right]}{\lambda^2})^2} \\ &= \frac{\mathbf{E}\left[\xi^2\right]}{\lambda^2 + \mathbf{E}\left[\xi^2\right]} \end{aligned}$$

Now apply the above inequality to the centered random variable $\xi - \mu$ to get the general result. $\square$

DEFINITION 10.4. 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 10.5. 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 10.6. *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 10.7. *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:

$$\begin{aligned}\mathbf{E}\left[e^{t\xi}\right] &= \int_0^\infty \mathbf{P}\{e^{t\xi} \geq \lambda\} d\lambda = \int_{-\infty}^\infty \mathbf{P}\{e^{t\xi} \geq e^{t\eta}\} te^{t\eta} d\eta \\ &= \int_{-\infty}^\infty \mathbf{P}\{\xi \geq \eta\} te^{t\eta} d\eta \leq \int_{-\infty}^\infty Cte^{t\eta - c\eta^2} d\eta = Cte^{\frac{t^2}{4c}} \int_{-\infty}^\infty e^{-\left(\sqrt{c}\eta - \frac{t}{2\sqrt{c}}\right)^2} d\eta \\ &= C' te^{\frac{t^2}{4c}} \leq C' e^{\frac{5ct^2}{4c}}\end{aligned}$$

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:

$$\mathbf{P}\{\xi \geq \lambda\} = \mathbf{P}\{e^{t\xi} \geq e^{t\lambda}\} \leq \frac{\mathbf{E}\left[e^{t\xi}\right]}{e^{t\lambda}} \leq Ce^{Ct^2 - t\lambda}$$

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 10.7 is that the constant on the right hand side is 1. With this definition, we must add the hypothesis $\mathbf{E}[\xi] = 0$ to get equivalence with the other definition.

LEMMA 10.8. *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

$$\sum_{n=0}^{\infty} \frac{t^n}{n!}\mathbf{E}[\xi^n] = \mathbf{E}\left[e^{t\xi}\right] \leq e^{\frac{c^2t^2}{2}} = \sum_{n=0}^{\infty} \frac{c^{2n}}{2^n n!}t^{2n}$$

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

$$t\mathbf{E}[\xi] + \frac{t^2}{2}\mathbf{E}\left[\xi^2\right] = \frac{c^2t^2}{2} + o(t^2) \text{ as } t \to 0$$

If we divide both sides by $t > 0$ and take the limit as $t \to 0^+$ then we get $\mathbf{E}[\xi] \leq 0$. If we divide by $t < 0$ and take the limit as $t \to 0^-$ then we get $\mathbf{E}[\xi] \geq 0$. Thus we can conclude $\mathbf{E}[\xi] = 0$. 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 10.9. *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 10.10 (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

$$\begin{aligned}
\mathbf{P}\{\sum_{i=1}^m \xi_i > \lambda\} &= \mathbf{P}\{c\sum_{i=1}^m \xi_i > c\lambda\} && \text{since } c > 0. \\
&= \mathbf{P}\{e^{c\sum_{i=1}^m \xi_i} > e^{c\lambda}\} && \text{since } e^x \text{ is increasing} \\
&\leq e^{-c\lambda}\mathbf{E}\left[e^{c\sum_{i=1}^m \xi_i}\right] && \text{by Markov's Inequality(10.1)} \\
&= e^{-c\lambda}\prod_{i=1}^m \mathbf{E}\left[e^{c\xi_i}\right] && \text{by independence and boundedness. TODO: do we really need boundedne}
\end{aligned}$$

Now we consider an individual term $\mathbf{E}\left[e^{c\xi_i}\right]$ for an almost surely bounded ξ_i with zero mean.

$$\begin{aligned}
\mathbf{E}\left[e^{c\xi_i}\right] = \mathbf{E}\left[\sum_{k=0}^\infty \frac{c^k\xi_i^k}{k!}\right] &= \sum_{k=0}^\infty \frac{c^k}{k!}\mathbf{E}\left[\xi_i^k\right] && \text{by dominated convergence} \\
&= 1 + \sum_{k=2}^\infty \frac{c^k}{k!}\mathbf{E}\left[\xi_i^k\right] && \text{by mean zero} \\
&\leq 1 + \sum_{k=2}^\infty \frac{c^k M^{k-2}\sigma_i^2}{k!} && \text{by boundedness and definition of variance} \\
&\leq e^{\sum_{k=2}^\infty \frac{c^k M^{k-2}\sigma_i^2}{k!}} && \text{since } 1 + x \leq e^x \text{(C.1)} \\
&= e^{\frac{(e^{cM}-1-cM)\sigma_i^2}{M^2}}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbf{P}\{\sum_{i=1}^m \xi_i > \lambda\} &\leq e^{-c\lambda}\prod_{i=1}^m e^{\frac{(e^{cM}-1-cM)\sigma_i^2}{M^2}} \\
&= e^{\frac{(e^{cM}-1-cM)\Sigma^2}{M^2}}
\end{aligned}$$

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

$$\begin{aligned}
\mathbf{P}\{\sum_{i=1}^m \xi_i > \lambda\} &\leq e^{-(\lambda + \frac{\Sigma^2}{M})\frac{1}{M}\ln(1+\frac{M\lambda}{\Sigma^2}) + \frac{\lambda}{M}} \\
&= e^{-\frac{\lambda}{M}\{(1+\frac{\Sigma^2}{\lambda M})\ln(1+\frac{M\lambda}{\Sigma^2})-1\}}
\end{aligned}$$

□

LEMMA 10.11 (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, $|\xi_i - \mu_i| \leq M$ almost everywhere then for every $\lambda > 0$ we have*

$$\boldsymbol{P}\{\sum_{i=1}^m [\xi_i - \mu_i] > \lambda\} \leq e^{-\{\frac{\lambda^2}{2(\Sigma^2 + \frac{1}{3}M\lambda)}\}}$$

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 10.12. 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 $\lambda > 0$,

$$\mathbf{P}\{[\xi - \mu] > \lambda\} \leq Ce^{-c\lambda^2}.$$

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 10.13. *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 10.14 (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^x on the interval $[0, s]$. For $0 \leq \theta \leq 1$,

$$e^{s\theta} = e^{\theta \cdot s + (1-\theta)\cdot 0} \leq \theta e^s + (1-\theta)e^0 = \theta e^s + (1-\theta)$$

Substituting $\theta = \xi_i$ 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 $\left(\frac{\mu_i}{\lambda}\right)^{\lambda} e^{\lambda-\mu_i}$. $\square$

1. Sanov's Theorem for Finite Alphabets

In this section we develop the simplest version of a large deviation principle for empirical measures. We assume a finite alphabet $\Sigma = \{a_1, \dots, a_{\text{card}(\Sigma)}\}$ with the powerset σ-algebra. We investigate the concentration properties of sequences $\xi_1, \xi_2, \dots$ of random elements in Σ. The empirical measure of a sequence of random elements in Σ is often called the *type* of the sequence.

TODO: Topology on the space of probability measures on a finite space. Identification with the induced topology from $\mathbb{R}^n$.

Definition 10.15. Given a finite measure space $(\Sigma, 2^{\Sigma}, \mu)$ we let

$$\Sigma_\mu = \{a \in \Sigma \mid \mu(a) > 0\}$$

Proposition 10.16. *Let μ and ν be probability measures on the finite measurable space $(\Sigma, 2^{\Sigma}, \mu)$ then the following are equivalent*

(i) $\Sigma_\nu \subset \Sigma_\mu$
(ii) $\nu \ll \mu$
(iii) $D(\nu \mid\mid \mu) < \infty$

Moreover for each fixed μ the set of ν such that $\nu \ll \mu$ is compact and $D(\nu \mid\mid \mu)$ is finite and continuous on this set.

Proof. Left to the reader. TODO: Add to exercises and provide the simple solution.

To see the compactness, we only need to see that the set of ν with $\nu \ll \mu$ is closed (it is trivially bounded). Letting $\nu = \lim_{n\to\infty} \nu_n$ with $\nu_n \ll \mu$ we see that if $\mu(a) = 0$ then it follows that $\nu_n(a) = 0$ for all $n \in \mathbb{N}$ and therefore $\nu(a) = \lim_{n\to\infty} \nu_n(a) = 0$. $\square$

Definition 10.17. Let $(\Sigma, 2^{\Sigma})$ be a finite measurable space and let $y \in \Sigma^n$

$$\mathcal{T}(y) = \frac{1}{n} \sum_{a\in\Sigma} \sum_{j=1}^{n} \mathbf{1}_a(y_j)\delta_a$$

The range $\mathcal{T}(\Sigma^n)$ is denoted $\mathcal{T}_n$. Given a probability measure $\nu \in \mathcal{T}_n$ the *type class* of ν is

$$\mathfrak{T}(\nu) = \mathcal{T})^{-1}y = \{y \in \Sigma^n \mid \mathcal{T}(y) = \nu\}$$

LEMMA 10.18. *Let $(\Sigma, 2^\Sigma)$ be a finite measure space then*

$$\operatorname{card}(\mathcal{T}_n) \leq (n+1)^{\operatorname{card}(\Sigma)}$$

PROOF. Each element $\nu \in \mathcal{T}_n$ must have probabilities $\nu(a) = j_a/n$ for some $j_a \in \{0, 1, \dots, n\}$; therefore it is uniquely determined by a tuple in $\{0, 1, \dots, n\}^\Sigma$. The size of the latter set is $(n+1)^{\operatorname{card}(\Sigma)}$. □

LEMMA 10.19. *Let μ and ν be probability measures on the finite measure space $(\Sigma, 2^\Sigma)$. Let $\xi_1, \xi_2, \dots$ be i.i.d. random elements in Σ with $\mathcal{L}(\xi_j) = \mu$ for $j \in \mathbb{N}$, then*

$$\boldsymbol{P}\{(\xi_1, \dots, \xi_n) = y\} = e^{-n(H(\mathcal{T}(y)) + D(\mathcal{T}(y)||\mu))}$$

PROOF. First suppose that $\Sigma_{\mathcal{T}(y)} \not\subset \Sigma_\mu$. In this case there exists $1 \leq j \leq n$ such that $y_j \notin \Sigma_\mu$ and therefore $\mathbf{P}\{(\xi_1, \dots, \xi_n) = y\} = 0$. Also by Proposition 10.16, $D\left(\mathcal{T}(y) \mid\mid \mu\right) = \infty$ which implies $e^{-n(H(\mathcal{T}(y)) + D(\mathcal{T}(y)||\mu))} = 0$ as well.

Thus we may assume that $\Sigma_{\mathcal{T}(y)} \subset \Sigma_\mu$. By Lemma 4.5 we have

$$\begin{aligned}
\mathbf{P}\{(\xi_1, \dots, \xi_n) = y\} &= \prod_{j=1}^n \mathbf{P}\{\xi_j = y_j\} = \prod_{a \in \Sigma_{\mathcal{T}(y)}} \mathbf{P}\{\xi_j = a_j\}^{n\mathcal{T}(y)(a_j)} \\
&= \prod_{a \in \Sigma_{\mathcal{T}(y)}} \mu(a)^{n\mathcal{T}(y)(a)} = \prod_{a \in \Sigma_{\mathcal{T}(y)}} e^{n\mathcal{T}(y)(a)\log\mu(a)} \\
&= \exp\left(n \sum_{a \in \Sigma_{\mathcal{T}(y)}} \mathcal{T}(y)(a) \log \mu(a)\right) \\
&= \exp\left(-n \sum_{a \in \Sigma_{\mathcal{T}(y)}} -\mathcal{T}(y)(a)\log\mathcal{T}(y)(a) + \mathcal{T}(y)(a)\log\mathcal{T}(y)(a) - \mathcal{T}(y)(a)\log\mu(a)\right) \\
&= \exp\left(-n \sum_{a \in \Sigma_{\mathcal{T}(y)}} -\mathcal{T}(y)(a)\log\mathcal{T}(y)(a) + \mathcal{T}(y)(a)\log(\frac{\mathcal{T}(y)(a)}{\mu(a)})\right) \\
&= \exp\left(-n \sum_{a \in \Sigma} -\mathcal{T}(y)(a)\log\mathcal{T}(y)(a) + \mathcal{T}(y)(a)\log(\frac{\mathcal{T}(y)(a)}{\mu(a)})\right) \\
&= e^{-n(H(\mathcal{T}(y)) + D(\mathcal{T}(y)||\mu))}
\end{aligned}$$

The reader is invited to make careful note of where we used the assumption $\Sigma_{\mathcal{T}(y)} \subset \Sigma_\mu$ and our conventions $0 \cdot \log 0 = 0$ and $\log \frac{0}{0} = 0$ in the calculation above. □

If we consider a probability measure $\mu \in \mathcal{T}_n$ then $(y_1, \dots, y_n) \in \mathfrak{T}(\mu)$ if and only if the number of y_j equal to $a \in \Sigma$ is equal to $n\mu(a)$; thus

$$\operatorname{card}(\mathfrak{T}(\mu)) = \frac{n!}{(n\mu(a_1))! \cdots (n\mu(a_{\operatorname{card}(\Sigma)}))!}$$

is just a multinomial coefficient. Our goal is to understand the asymptotic behavior of such a coefficient. One way to approach this is to use Stirling's approximation. We present a different approach here.

LEMMA 10.20. *For any $\nu \in \mathcal{T}_n$ we have*

$$\frac{1}{(n+1)^{\text{card}(\Sigma)}} e^{nH(\nu)} \leq \text{card}(\mathfrak{T}(\nu)) \leq e^{nH(\nu)}$$

PROOF. The upper bound is a straightforward corollary of Lemma 10.19 in the case $\mu = \nu$; let $\xi_1, \dots, \xi_n$ be i.i.d. with $\mathcal{L}(\xi_j) = \nu$ then

$$1 \geq \mathbf{P}\{(\xi_1, \dots, \xi_n) \in \mathfrak{T}(\nu)\} = \sum_{y \in \mathfrak{T}(\nu)} \mathbf{P}\{(\xi_1, \dots, \xi_n) = y\} = \text{card}(\mathfrak{T}(\nu)) e^{-nH(\nu)}$$

CLAIM 10.20.1. For all $\mu, \nu \in \mathcal{T}_n$ we have

$$\mathbf{P}\{(\xi_1, \dots, \xi_n) \in \mathfrak{T}(\mu)\} \geq \mathbf{P}\{(\xi_1, \dots, \xi_n) \in \mathfrak{T}(\nu)\}$$

If $\mathbf{P}\{(\xi_1, \dots, \xi_n) \in \mathfrak{T}(\nu)\} = 0$ there is nothing to prove. If $\mathbf{P}\{\mathcal{T}((\xi_1, \dots, \xi_n)) = \mu\} > 0$ then it is clear that $\Sigma_\mu \subset \Sigma_\nu$. Using this fact we can calculate

$$\begin{aligned}
\frac{\mathbf{P}\{(\xi_1, \dots, \xi_n) \in \mathfrak{T}(\mu)\}}{\mathbf{P}\{(\xi_1, \dots, \xi_n) \in \mathfrak{T}(\nu)\}} &= \frac{\sum_{y \in \mathfrak{T}(\mu)} \mathbf{P}\{(\xi_1, \dots, \xi_n) = y\}}{\sum_{y \in \mathfrak{T}(\nu)} \mathbf{P}\{(\xi_1, \dots, \xi_n) = y\}} = \frac{\text{card}(\mathfrak{T}(\mu)) \prod_{a \in \Sigma_\mu} \nu(a)^{n\mu(a)}}{\text{card}(\mathfrak{T}(\nu)) \prod_{a \in \Sigma_\nu} \nu(a)^{n\nu(a)}} \\
&= \frac{\text{card}(\mathfrak{T}(\mu)) \prod_{a \in \Sigma_\nu} \nu(a)^{n\mu(a)}}{\text{card}(\mathfrak{T}(\nu)) \prod_{a \in \Sigma_\nu} \nu(a)^{n\nu(a)}} = \prod_{a \in \Sigma_\nu} \frac{(n\nu(a))!}{(n\mu(a))!} \nu(a)^{n(\mu(a) - \nu(a))}
\end{aligned}$$

For arbitrary $m, l \in \mathbb{Z}$ and $n \in \mathbb{N}$ we have $\frac{m!}{l!} \geq l^{m-l}$ (if $m = l$ then both sides are 1, if $m > l$ then this follows from $m(m-1)\cdots(m-l+1) \geq l^{m-l}$ and if $m < l$ this follows from $l^{l-m} \geq l(l-1)\cdots(l-m+1)$). Applying this inequality to the last product above yields

$$\frac{\mathbf{P}\{(\xi_1, \dots, \xi_n) \in \mathfrak{T}(\mu)\}}{\mathbf{P}\{(\xi_1, \dots, \xi_n) \in \mathfrak{T}(\nu)\}} \geq \prod_{a \in \Sigma_\nu} n^{n(\nu(a) - \mu(a))} = n^{n \sum_{a \in \Sigma_\nu} (\nu(a) - \mu(a))} = 1$$

and the claim is proved.

From the claim, for $\nu \in \mathcal{T}_n$ we have by Lemma 10.19 and Lemma 10.18

$$\begin{aligned}
1 &= \sum_{\mu \in \mathcal{T}_n} \mathbf{P}\{(\xi_1, \dots, \xi_n) \in \mathfrak{T}(\mu)\} \leq \text{card}(\mathcal{T}_n) \mathbf{P}\{(\xi_1, \dots, \xi_n) \in \mathfrak{T}(\nu)\} \\
&= \text{card}(\mathcal{T}_n) \sum_{y \in \mathfrak{T}(\nu)} \mathbf{P}\{(\xi_1, \dots, \xi_n) = y\} = \text{card}(\mathcal{T}_n) \,\text{card}(\mathfrak{T}(\nu)) e^{-nH(\nu)} \\
&\leq (n+1)^{\text{card}(\Sigma)} \,\text{card}(\mathfrak{T}(\nu)) e^{-nH(\nu)}
\end{aligned}$$

□

TODO: Alternative proof using Stirling's formula

LEMMA 10.21. *Let $\xi_1, \xi_2, \dots$ be i.i.d. random elements in Σ with $\mathcal{L}(\xi_j) = \mu$ then for any $\nu \in \mathcal{T}_n$ we have*

$$(n+1)^{-\text{card}(\Sigma)} e^{-nD(\nu || \mu)} \leq \boldsymbol{P}\{(\xi_1, \dots, \xi_n) \in \mathfrak{T}(\nu)\} \leq e^{-nD(\nu || \mu)}$$

PROOF. From Lemma 10.19 we get

$$\mathbf{P}\{(\xi_1,\ldots,\xi_n)\in\mathfrak{T}(\nu)\} = \sum_{y\in\mathfrak{T}(\nu)} \mathbf{P}\{(\xi_1,\ldots,\xi_n)=y\} = \operatorname{card}(\mathfrak{T}(\nu))e^{-n(H(\nu)+D(\nu||\mu))}$$

Now by Lemma 10.20 we conclude

$$\begin{aligned}(n+1)^{-\operatorname{card}(\Sigma)}e^{-nD(\nu||\mu)} &= (n+1)^{-\operatorname{card}(\Sigma)}e^{nH(\nu)}e^{-n(H(\nu)+D(\nu||\mu))} \leq \mathbf{P}\{(\xi_1,\ldots,\xi_n)\in\mathfrak{T}(\nu)\}\\ &\leq e^{nH(\nu)}e^{-n(H(\nu)+D(\nu||\mu))} = e^{-nD(\nu||\mu)}\end{aligned}$$

$\square$

THEOREM 10.22. *Let $(\Sigma, 2^{\Sigma})$ be a finite measure space, let $\xi_1,\xi_2,\ldots$ be i.i.d. random elements in Σ with $\mathcal{L}(\xi_j)=\mu$ for all $j\in\mathbb{N}$ and let Γ be a set of probability measures on Σ then*

$$\begin{aligned}-\inf_{\nu\in\operatorname{int}(\Gamma)} D(\nu\,||\,\mu) &\leq \liminf_{n\to\infty}\frac{1}{n}\log \boldsymbol{P}\{\mathcal{T}((\xi_1,\ldots,\xi_n))\in\Gamma\}\\ &\leq \limsup_{n\to\infty}\frac{1}{n}\log \boldsymbol{P}\{\mathcal{T}((\xi_1,\ldots,\xi_n))\in\Gamma\} \leq -\inf_{\nu\in\Gamma} D(\nu\,||\,\mu)\end{aligned}$$

PROOF. First note that by Lemma 10.21 and Lemma 10.18

$$\begin{aligned}\mathbf{P}\{\mathcal{T}((\xi_1,\ldots,\xi_n))\in\Gamma\} &= \sum_{\nu\in\Gamma\cap\mathcal{T}_n}\mathbf{P}\{\mathcal{T}((\xi_1,\ldots,\xi_n))=\nu\} \leq \sum_{\nu\in\Gamma\cap\mathcal{T}_n} e^{-nD(\nu||\mu)}\\ &\leq \operatorname{card}(\Gamma\cap\mathcal{T}_n)e^{-n\inf_{\nu\in\Gamma\cap\mathcal{T}_n}D(\nu||\mu)}\\ &\leq (n+1)^{\operatorname{card}(\Sigma)}e^{-n\inf_{\nu\in\Gamma\cap\mathcal{T}_n}D(\nu||\mu)}\end{aligned}$$

and by Lemma 10.21 and the non-negativity of the terms $e^{-nD(\nu||\mu)}$ we get

$$\begin{aligned}\mathbf{P}\{\mathcal{T}((\xi_1,\ldots,\xi_n))\in\Gamma\} &= \sum_{\nu\in\Gamma\cap\mathcal{T}_n}\mathbf{P}\{\mathcal{T}((\xi_1,\ldots,\xi_n))=\nu\} \geq \sum_{\nu\in\Gamma\cap\mathcal{T}_n}(n+1)^{-\operatorname{card}(\Sigma)}e^{-nD(\nu||\mu)}\\ &\geq (n+1)^{-\operatorname{card}(\Sigma)}e^{-n\inf_{\nu\in\Gamma\cap\mathcal{T}_n}D(\nu||\mu)}\end{aligned}$$

By L'Hopital's rule

$$\lim_{n\to\infty}\frac{\log(n+1)^{\pm\operatorname{card}(\Sigma)}}{n} = \pm\operatorname{card}(\Sigma)\lim_{n\to\infty}\frac{1}{n+1}=0$$

and therefore

$$\begin{aligned}\limsup_{n\to\infty}\frac{1}{n}\log\mathbf{P}\{\mathcal{T}((\xi_1,\ldots,\xi_n))\in\Gamma\} &\leq \limsup_{n\to\infty}(-\inf_{\nu\in\Gamma\cap\mathcal{T}_n}D(\nu\,||\,\mu))\\ &\leq \limsup_{n\to\infty}(-\inf_{\nu\in\Gamma}D(\nu\,||\,\mu)) = -\inf_{\nu\in\Gamma}D(\nu\,||\,\mu)\end{aligned}$$

which gives the upper bound.

For the lower bound we start with

$$\begin{aligned}\liminf_{n\to\infty}\frac{1}{n}\log\mathbf{P}\{\mathcal{T}((\xi_1,\ldots,\xi_n))\in\Gamma\} &\geq \liminf_{n\to\infty}(-\inf_{\nu\in\Gamma\cap\mathcal{T}_n}D(\nu\,||\,\mu))\\ &= -\limsup_{n\to\infty}(\inf_{\nu\in\Gamma\cap\mathcal{T}_n}D(\nu\,||\,\mu))\end{aligned}$$

Note that if there is no $\nu\in\operatorname{int}(\Gamma)$ such that $\nu\ll\mu$ then $-\inf_{\nu\in\operatorname{int}(\Gamma)}D(\nu\,||\,\mu) = -\infty$ and the lower bound is trivially true. Thus we may assume that such a ν exists and in this case $-\inf_{\nu\in\text{int}(\Gamma)} D(\nu \mid\mid \mu) = -\inf_{\substack{\nu\in\text{int}(\Gamma)\\ \nu\ll\mu}} D(\nu \mid\mid \mu)$. Suppose we are given a $\nu \in \text{int}(\Gamma)$ such that $\nu \ll \mu$. There exists a $\delta > 0$ such that $\|\nu' - \nu\| < \delta$ implies $\nu' \in \Gamma$ (TODO: We need the fact that we are using the total variation metric on the space of probability measures; for finite alphabets it shouldn't matter). By Lemma ??? we know that we can find a sequence ν_n with $\nu_n \in \mathcal{T}_n$ and $\lim_{n\to\infty} \|\nu_n - \nu\| = 0$. In particular, eventually $\nu_n \in \text{int}(\Gamma)$. Furthermore since $D(\nu \mid\mid \mu) < \infty$ and the relative entropy is continuous with respect to the total variation we know that eventually $D(\nu_n \mid\mid \mu) < \infty$ as well.

$$-\limsup_{n\to\infty} \inf_{\nu'\in\Gamma\cap\mathcal{T}_n} D(\nu' \mid\mid \mu) \geq -\limsup_{n\to\infty} D(\nu_n \mid\mid \mu) = -D(\nu \mid\mid \mu)$$

from which it follows that

$$\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}\{\mathcal{T}((\xi_1,\dots,\xi_n)) \in \Gamma\} \geq \sup_{\substack{\nu\in\text{int}(\Gamma)\\ \nu\ll\mu}} -D(\nu \mid\mid \mu) = -\inf_{\nu\in\text{int}(\Gamma)} D(\nu \mid\mid \mu)$$

$\square$

With Sanov's Theorem in hand we can derive a large deviation principle for empirical means. We reiterate that we are deviating from the historical order in that large deviation principles for empirical means we developed prior to large deviation principles for empirical measures.

THEOREM 10.23. *Let $(\Sigma, 2^\Sigma)$ be a finite measure space and let $\xi_1, \xi_2, \dots$ be an i.i.d. sequence of random elements in Σ with $\mathcal{L}(\xi_j) = \mu$ for all $j \in \mathbb{N}$. Let $f : \Sigma \to \mathbb{R}$ be a function, let $\eta_j = f(\xi_j)$ for all $j \in \mathbb{N}$ and*

$$\hat{S}_n = \frac{1}{n}\sum_{j=1}^n \eta_j$$

If

$$I(x) = \inf\{D(\nu \mid\mid \mu) \mid \boldsymbol{E}_\nu[f] = x\}$$

then for every $A \subset \mathbb{R}$

$$-\inf_{x\in\text{int}(A)} I(x) \leq \liminf_{n\to\infty} \frac{1}{n} \log \boldsymbol{P}\{\hat{S}_n \in A\}$$
$$\leq \limsup_{n\to\infty} \frac{1}{n} \log \boldsymbol{P}\{\hat{S}_n \in A\} \leq -\inf_{x\in A} I(x)$$

Moreover, $I(x)$ is continuous on the interval $K = [\min_{a\in\Sigma} f(a), \max_{a\in\Sigma} f(a)]$ and

$$I(x) = \sup_{-\infty<\lambda<\infty} (\lambda x - \Lambda(\lambda)) \text{ for } x \in K$$

where

$$\Lambda(\lambda) = \log \boldsymbol{E}_\mu \left[e^{\lambda f}\right]$$

PROOF. Given $\mathbf{E}_\cdot[f]$ is a continuous map from the

CLAIM 10.23.1. For every $\lambda, x \in \mathbb{R}$

$$\lambda x - \Lambda(\lambda) \leq \inf_{\{\nu \mid \mathbf{E}_\nu[f]=x\}} D(\nu \mid\mid \mu) = I(x)$$

and equality holds when $x = \mathbf{E}_{\nu_\lambda}[f]$ where $\nu_\lambda(a) = \mu(a) e^{\lambda f(a) - \Lambda(\lambda)}$.

Now by the strict concavity of log and Jensen's Inequality

$$\begin{aligned}\log \sum_{a \in \Sigma} \mu(a) e^{\lambda f(a)} &= \log \sum_{a \in \Sigma} \nu(a) \frac{\mu(a)}{\nu(a)} e^{\lambda f(a)} \geq \sum_{a \in \Sigma} \nu(a) \log \frac{\mu(a)}{\nu(a)} e^{\lambda f(a)} \\ &= \lambda \mathbf{E}_\nu [f] - D(\nu \mid\mid \mu)\end{aligned}$$

and equality holds if and only if $\nu(a) > 0$ implies

$$\frac{\mu(a)}{\nu(a)} e^{\lambda f(a)} = \mathbf{E}_\nu \left[\frac{\mu(a)}{\nu(a)} e^{\lambda f(a)} \right] = \mathbf{E}_\mu \left[e^{\lambda f(a)} \right] = e^{\Lambda(\lambda)}$$

which is to say we have equality precisely at the probability measure $\nu_\lambda(a) = \mu(a) e^{\lambda f(a) - \Lambda(\lambda)}$. If we pick $-\infty < x < \infty$ then by considering the above inequality for all ν with $\mathbf{E}_\nu [f] = x$ the claim follows.

Now we need to get a handle on the conditions under which $x = \mathbf{E}_{\nu_\lambda} [f]$. First note that

$$\Lambda'(\lambda) = \frac{d}{d\lambda} \log \sum_{a \in \Sigma} \mu(a) e^{\lambda f(a)} = \frac{\sum_{a \in \Sigma} \mu(a) f(a) e^{\lambda f(a)}}{\sum_{a \in \Sigma} \mu(a) e^{\lambda f(a)}} = \sum_{a \in \Sigma} \mu(a) f(a) e^{\lambda f(a) - \Lambda(\lambda)} = \mathbf{E}_{\nu_\lambda} [f]$$

so by the prior claim we have $\lambda x - \Lambda(\lambda) = I(x)$ for all $x \in \{\Lambda'(\lambda) \mid -\infty < \lambda < \infty\}$. Observe that

$$\Lambda''(\lambda) = \frac{\mathbf{E}_\mu \left[e^{\lambda f}\right] \mathbf{E}_\mu \left[f^2 e^{\lambda f}\right] - \mathbf{E}_\mu \left[f e^{\lambda f}\right]^2}{\mathbf{E}_\mu \left[e^{\lambda f}\right]^2}$$

and by Cauchy-Schwartz

$$\mathbf{E}_\mu \left[f e^{\lambda f}\right]^2 \leq \mathbf{E}_\mu \left[|f| e^{\lambda f}\right]^2 \leq \mathbf{E}_\mu \left[f^2 e^{\lambda f}\right] \mathbf{E}_\mu \left[e^{\lambda f}\right]$$

which shows that $\Lambda'(\lambda)$ is non-decreasing. Let $a_- = \arg\min_{a \in \Sigma} f(a)$ and $a_+ = \arg\max_{a \in \Sigma} f(a)$ and that

$$\lim_{\lambda \to \infty} \nu_\lambda(a) = \mu(a) \lim_{\lambda \to \infty} \frac{1}{\sum_{a' \in \Sigma} \mu(a') e^{\lambda (f(a') - f(a))}} = \delta_{a_+}$$

and

$$\lim_{\lambda \to -\infty} \nu_\lambda(a) = \mu(a) \lim_{\lambda \to -\infty} \frac{1}{\sum_{a' \in \Sigma} \mu(a') e^{\lambda (f(a') - f(a))}} = \delta_{a_-}$$

hence $\lim_{\lambda \to \pm\infty} \Lambda'(\lambda) = \lim_{\lambda \to \pm\infty} \mathbf{E}_{\nu_\lambda} [f] = f(a_\pm)$. Since Λ' is continuous and non-decreasing it follows that $(f(a_-), f(a_+)) \subset \{\Lambda'(\lambda) \mid -\infty < \lambda < \infty\}$ and thus $\lambda x - \Lambda(\lambda) = I(x)$ on $(f(a_-), f(a_+))$. Consider the endpoints $a_\pm$ and compute using $\mathbf{E}_{\delta_{a_\pm}} [f] = f(a_\pm)$ and the claim to see

$$\begin{aligned}-\log \mu(a_\pm) = D\left(\delta_{a_\pm} \mid\mid \mu\right) &\geq I(f(a_\pm)) \geq \sup_{-\infty < \lambda < \infty} \lambda f(a_\pm) - \Lambda(\lambda) \\ &\geq \lim_{\lambda \to \pm\infty} \lambda f(a_\pm) - \Lambda(\lambda) = \lim_{\lambda \to \pm\infty} \log \frac{e^{\lambda f(a_\pm)}}{\mathbf{E}_\mu \left[e^{\lambda f}\right]} \\ &= \lim_{\lambda \to \pm\infty} \log \frac{1}{\sum_{a \in \Sigma} \mu(a) e^{\lambda (f(a) - f(a_\pm))}} = -\log \mu(a_\pm)\end{aligned}$$

$\square$

2. Cramer's Theorem

We now extend large deviations to random vectors. We first handle the case of large deviations of empirical means of a random variable: a result known as Cramer's Theorem.

LEMMA 10.24. *Let ξ be a random variable, let $\Lambda_\xi(\lambda) = \log \boldsymbol{E}\left[e^{\lambda\xi}\right]$ be the cumulant generating function and let $\Lambda^*_\xi(x) = \sup_{\lambda\in\mathbb{R}}\{\lambda x - \Lambda_\xi(\lambda)\}$ be the Legendre-Fenchel transform of the cumulant generating function. The*

(i) *The range of Λ_ξ is $(-\infty, \infty]$, $\Lambda_\xi(0) = 0$ and Λ_ξ is a convex function. Moreover $\Lambda_\xi(\lambda) = \Lambda_{-\xi}(-\lambda)$.*
(ii) *Λ^*_ξ is convex, $0 \le \Lambda^*_\xi(x) \le \infty$ and Λ^*_ξ is lower semicontinuous. Moreover $\Lambda^*_\xi(x) = \Lambda^*_{-\xi}(-x)$.*
(iii) *If $\Lambda_\xi(\lambda) = \infty$ for all $\lambda \ne 0$ then $\Lambda^*_\xi \equiv 0$.*
(iv) *If $\Lambda_\xi(\lambda) < \infty$ for some $\lambda > 0$ then $-\infty \le \boldsymbol{E}[\xi] < \infty$. When $-\infty \le \boldsymbol{E}[\xi] < \infty$ we have*

$$\Lambda^*_\xi(x) = \sup_{0\le\lambda<\infty}\{\lambda x - \Lambda_\xi(\lambda)\} \text{ for all } x \ge \boldsymbol{E}[\xi]$$

and moreover Λ^*_ξ is non-decreasing on $(\boldsymbol{E}[\xi], \infty)$.
(v) *If $\Lambda_\xi(\lambda) < \infty$ for some $\lambda < 0$ then $-\infty < \boldsymbol{E}[\xi] \le \infty$. When $-\infty < \boldsymbol{E}[\xi] \le \infty$ we have*

$$\Lambda^*_\xi(x) = \sup_{-\infty<\lambda\le 0}\{\lambda x - \Lambda_\xi(\lambda)\} \text{ for all } x \le \boldsymbol{E}[\xi]$$

and moreover Λ^*_ξ is non-increasing on $(-\infty, \boldsymbol{E}[\xi])$.
(vi) *When $-\infty < \boldsymbol{E}[\xi] < \infty$ then $\Lambda^*_\xi(\boldsymbol{E}[\xi]) = 0$ and in all cases $\inf_{-\infty<x<\infty}\Lambda^*_\xi(x) = 0$.*
(vii) *Λ_ξ is differentiable on the interior of $\mathcal{D}(\Lambda_\xi) = \{\lambda \mid \Lambda_\xi(\lambda) < \infty\}$ and*

$$\Lambda'_\xi(\lambda) = \frac{\boldsymbol{E}\left[\xi e^{\lambda\xi}\right]}{\boldsymbol{E}[e^{\lambda\xi}]}$$

moreover $\Lambda'_\xi(\lambda) = y$ implies $\Lambda^*_\xi(y) = \lambda y - \Lambda_\xi(\lambda)$.

PROOF. To see (i) we note that $0 < \mathbf{E}\left[e^{\lambda\xi}\right] \le \infty$ and therefore $-\infty < \Lambda_\xi(\lambda) \le \infty$. Clearly, $\Lambda_\xi(0) = \log \mathbf{E}[1] = 0$ and given $\lambda, \eta \in \mathbb{R}$, $0 < \theta < 1$ by Hölder's inequality with $p = \theta^{-1}$ and $q = (1-\theta)^{-1}$ we get

$$\begin{aligned}
\Lambda_\xi(\theta\lambda + (1-\theta)\eta) &= \log \mathbf{E}\left[e^{(\theta\lambda+(1-\theta)\eta)\xi}\right] = \log \mathbf{E}\left[(e^{\lambda\xi})^\theta (e^{\eta\xi})^{1-\theta}\right] \\
&\le \log \mathbf{E}\left[e^{\lambda\xi}\right]^\theta \mathbf{E}\left[e^{\eta\xi}\right]^{1-\theta} = \log\theta \mathbf{E}\left[e^{\lambda\xi}\right] + (1-\theta)\log \mathbf{E}\left[e^{\eta\xi}\right] \\
&= \theta\Lambda_\xi(\lambda) + (1-\theta)\Lambda_\xi(\eta)
\end{aligned}$$

and therefore Λ_ξ is convex. Trivially,

$$\Lambda_\xi(\lambda) = \log \mathbf{E}\left[e^{\lambda\xi}\right] = \log \mathbf{E}\left[e^{(-\lambda)(-\xi)}\right] = \Lambda_{-\xi}(-\lambda)$$

To see (ii) we first note that Λ^*_ξ is convex letting $x, y \in \mathbb{R}$ and $0 < \theta < 1$

$$\begin{aligned}\Lambda^*_\xi(\theta x + (1-\theta)y) &= \sup_{-\infty<\lambda<\infty} \{\lambda(\theta x + (1-\theta)y) - \Lambda_\xi(\lambda)\} \\ &= \sup_{-\infty<\lambda<\infty} \{\theta(\lambda x - \Lambda_\xi(\lambda)) + (1-\theta)(\lambda y - \Lambda_\xi(\lambda))\} \\ &\leq \theta \sup_{-\infty<\lambda<\infty} \{\lambda x - \Lambda_\xi(\lambda)\} + (1-\theta) \sup_{-\infty<\lambda<\infty} \{\lambda y - \Lambda_\xi(\lambda)\} \\ &= \theta \Lambda^*_\xi(x) + (1-\theta)\Lambda^*_\xi(y)\end{aligned}$$

As we have shown $\Lambda_\xi(0) = 0$ therefore for all $-\infty < x < \infty$, $\Lambda^*_\xi(x) \geq 0 \cdot x - \Lambda_\xi(0) = 0$. To see the lower semicontinuity of Λ^*_ξ suppose that $\lim_{n\to\infty} x_n = x$ and compute for every $-\infty < \lambda < \infty$,

$$\begin{aligned}\liminf_{n\to\infty} \Lambda^*_\xi(x_n) &= \liminf_{n\to\infty} \sup_{-\infty<\eta<\infty} \{\eta x_n - \Lambda_\xi(\eta)\} \geq \liminf_{n\to\infty} \lambda x_n - \Lambda_\xi(\lambda) \\ &= \lambda x - \Lambda_\xi(\lambda)\end{aligned}$$

Now take the supremum over all λ to see that

$$\liminf_{n\to\infty} \Lambda^*_\xi(x_n) \geq \sup_{-\infty<\lambda<\infty} \{\lambda x - \Lambda_\xi(\lambda)\} = \Lambda^*_\xi(x)$$

Lastly we see that

$$\begin{aligned}\Lambda^*_{-\xi}(-x) &= \sup_{-\infty<\lambda<\infty} \{-\lambda x - \Lambda_{-\xi}(\lambda)\} = \sup_{-\infty<\lambda<\infty} \{-\lambda x - \Lambda_\xi(-\lambda)\} \\ &= \sup_{-\infty<\lambda<\infty} \{\lambda x - \Lambda_\xi(\lambda)\} = \Lambda^*_\xi(x)\end{aligned}$$

To see (iii) suppose that $\Lambda_\xi(\lambda) = \infty$ for all $\lambda \neq 0$ then for all $-\infty < x < \infty$, $\{\lambda x - \Lambda_\xi(\lambda) \mid -\infty < \lambda < \infty\} = \{-\infty, 0\}$ and it follows that $\Lambda^*_\xi(x) = 0$.

To see (iv), suppose that $\Lambda_\xi(\lambda) < \infty$ and $\lambda > 0$. Letting $\xi_+ = \xi \vee 0$ then by Lemma 3.8 and Markov's Inequality 10.1

$$\begin{aligned}\mathbf{E}[\xi_+] &= \int_0^\infty \mathbf{P}\{\xi_+ > \eta\}\, d\eta = \int_0^\infty \mathbf{P}\{\xi > \eta\}\, d\eta \\ &= \int_0^\infty \mathbf{P}\{e^{\lambda\xi} > e^{\lambda\eta}\}\, d\eta \leq \mathbf{E}\left[e^{\lambda\xi}\right] \int_0^\infty e^{-\lambda\eta}\, d\eta = \lambda^{-1}\mathbf{E}\left[e^{\lambda\xi}\right] < \infty\end{aligned}$$

which shows that $-\infty \leq \mathbf{E}[\xi] < \infty$. If we assume that $\mathbf{E}\left[e^{\lambda\xi}\right] < \infty$ then by Jensen's Inequality (Theorem 3.17) we get

$$\Lambda_\xi(\lambda) = \log \mathbf{E}\left[e^{\lambda\xi}\right] \geq \mathbf{E}\left[\log e^{\lambda\xi}\right] = \lambda \mathbf{E}[\xi] \tag{7}$$

The inequality (7) trivially holds when $\mathbf{E}\left[e^{\lambda\xi}\right] = \infty$ and therefore holds for all $-\infty < \lambda < \infty$. We now jump ahead and verify the first part of (vi); if we assume $-\infty < \mathbf{E}[\xi] < \infty$ then (7) implies

$$\Lambda^*_\xi(\mathbf{E}[\xi]) = \sup_{-\infty<\lambda<\infty} \{\lambda \mathbf{E}[\xi] - \Lambda_\xi(\lambda)\} \leq 0$$

from which we conclude $\Lambda^*_\xi(\mathbf{E}[\xi]) = 0$.

We return to (iv) and assume $-\infty \leq \mathbf{E}[\xi] < \infty$. If $-\infty = \mathbf{E}[\xi]$ then for every $\lambda < 0$ we have

$$\infty = \mathbf{E}[\lambda\xi] \leq \mathbf{E}\left[e^{\lambda\xi} - 1\right] = \Lambda(\lambda) - 1$$

and therefore we get $\sup_{-\infty<\lambda<0}\{\lambda x - \Lambda(\lambda)\} = -\infty$. If $-\infty < \mathbf{E}[\xi] < \infty$ then for $x \geq \mathbf{E}[\xi]$ and $\lambda < 0$

$$\lambda x - \Lambda_\xi(\lambda) \leq \lambda \mathbf{E}[\xi] - \Lambda_\xi(\lambda) \leq \Lambda_\xi^*(\mathbf{E}[\xi]) = 0$$

so that $\sup_{-\infty<\lambda<0}\{\lambda x - \Lambda_\xi(\lambda)\} \leq 0$. In both cases, by the fact that $0 \cdot x - \Lambda_\xi(0) = 0$ for every $-\infty < x < \infty$ we have $\sup_{0\leq\lambda<\infty}\{\lambda x - \Lambda_\xi(\lambda)\} \geq 0$ and therefore we get

$$\begin{aligned}\Lambda_\xi^*(x) &= 0 \vee \Lambda_\xi^*(x) = 0 \vee \sup_{-\infty<\lambda<0}\{\lambda x - \Lambda_\xi(\lambda)\} \vee \sup_{0\leq\lambda<\infty}\{\lambda x - \Lambda_\xi(\lambda)\} \\ &= \sup_{0\leq\lambda<\infty}\{\lambda x - \Lambda_\xi(\lambda)\}\end{aligned}$$

If $\mathbf{E}[\xi] < x \leq y < \infty$ then

$$\Lambda_\xi^*(x) = \sup_{0\leq\lambda<\infty}\{\lambda x - \Lambda_\xi(\lambda)\} \leq \sup_{0\leq\lambda<\infty}\{\lambda y - \Lambda_\xi(\lambda)\} = \Lambda_\xi^*(y)$$

To see (v) note that if $\Lambda_\xi(\lambda) < \infty$ for $\lambda < 0$ then $\Lambda_{-\xi}(-\lambda) < \infty$ for $-\lambda > 0$ hence applying (iv) to $-\xi$ we get for $x \leq \mathbf{E}[\xi]$

$$\begin{aligned}\Lambda_\xi^*(x) = \Lambda_{-\xi}^*(-x) &= \sup_{0\leq\lambda<\infty}\{\lambda(-x) - \Lambda_{-\xi}(\lambda)\} = \sup_{0\leq\lambda<\infty}\{-\lambda x - \Lambda_\xi(-\lambda)\} \\ &= \sup_{-\infty<\lambda\leq 0}\{\lambda x - \Lambda_\xi(\lambda)\}\end{aligned}$$

and $\Lambda_{-\xi}^*(x) = \Lambda_\xi^*(-x)$ is non-decreasing on $(-\mathbf{E}[\xi], \infty)$ which is equivalent to $\Lambda_\xi^*(x)$ is non-increasing on $(-\infty, \mathbf{E}[\xi])$.

To see (vi) we have already shown that $-\infty < \mathbf{E}[\xi] < \infty$ implies $\Lambda^*(\mathbf{E}[\xi]) = 0$; to see that $\inf_{-\infty<x<\infty} \Lambda^*(x) = 0$ in general we consider a few cases. Firstly if $\Lambda(\lambda) = \infty$ for all $\lambda \neq 0$ then we know that $\Lambda^* \equiv= 0$ so the result holds in this case. It remains to consider the cases $\Lambda(\lambda) < \infty$ and $\mathbf{E}[\xi] = \pm\infty$. Thus assume that $\mathbf{E}[\xi] = -\infty$ and that $\Lambda(\lambda) < \infty$ for some $\lambda \neq 0$. Note first that we may assume that $\Lambda(\lambda) < \infty$ for $\lambda > 0$ since by (v) we know that $\Lambda(\lambda) < \infty$ for $\lambda < 0$ implies $\mathbf{E}[\xi] > -\infty$. For every $\lambda > 0$ we compute using the Markov Inequality

$$\log \mathbf{P}\{\xi \geq x\} = \log \mathbf{P}\{e^{\lambda\xi} \geq e^{\lambda x}\} \leq \log \mathbf{E}\left[e^{\lambda\xi}\right] - \lambda x = \Lambda(\lambda) - \lambda x$$

Take the infimum over all $\lambda > 0$ hence

$$\begin{aligned}\log \mathbf{P}\{\xi \geq x\} &\geq \inf\{\Lambda(\lambda) - \lambda x \mid 0 < \lambda < \infty\} \geq \inf\{\Lambda(\lambda) - \lambda x \mid 0 \leq \lambda < \infty\} \\ &= -\sup\{\lambda x - \Lambda(\lambda) \mid 0 \leq \lambda < \infty\} = -\Lambda^*(x)\end{aligned}$$

and therefore by the fact that $\Lambda^* \geq 0$ and continuity of measure (Lemma 2.30) we get

$$0 \leq \lim_{x\to-\infty} \Lambda^*(x) \leq -\lim_{x\to-\infty} \log \mathbf{P}\{\xi \geq x\} = 0$$

To prove the first part of (vii) amounts to justifying exchanging the order of integration and differentiation. Let $\lambda \in \operatorname{int}(\mathcal{D}(\Lambda_\xi))$. Pick a $\delta > 0$ such that $[\lambda-\delta, \lambda+\delta] \subset \operatorname{int}(\mathcal{D}(\Lambda_\xi))$. By elementary calculus we have pointwise $\lim_{h\to 0} \frac{e^{(\lambda+h)\xi} - e^{\lambda\xi}}{h} = \xi e^{\lambda\xi}$. Writing

$$\left|\frac{e^{(\lambda+h)\xi} - e^{\lambda\xi}}{h}\right| = e^{\lambda\xi}\left|\frac{e^{h\xi} - 1}{h}\right|$$

and by differentiation and the estimate $1 + x \leq e^x$

$$\frac{d}{dh}\frac{e^{h\xi} - 1}{h} = \frac{h\xi e^{h\xi} - e^{h\xi} + 1}{h^2} \geq \frac{(h\xi - 1)(h\xi + 1) + 1}{h^2} = \xi^2$$

and therefore $\frac{e^{h\xi}-1}{h}$ is non-decreasing and therefore

$$\sup_{-\delta \leq h \leq \delta} \left| \frac{e^{h\xi}-1}{h} \right| = \left| \frac{e^{-\delta\xi}-1}{\delta} \right| \vee \left| \frac{e^{\delta\xi}-1}{\delta} \right| = \left| \frac{e^{\delta|\xi|}-1}{\delta} \right|$$

We also have

$$\begin{aligned} \mathbf{E}\left[e^{\lambda\xi} \left| \frac{e^{\delta|\xi|}-1}{\delta} \right| \right] &= \mathbf{E}\left[e^{\lambda\xi} \left(\frac{e^{\delta|\xi|}-1}{\delta} \right) \right] = \frac{1}{\delta}\mathbf{E}\left[e^{\lambda\xi+\delta|\xi|} - e^{\lambda\xi} \right] \\ &= \frac{1}{\delta}\left(\mathbf{E}\left[e^{(\lambda+\delta)\xi}; \xi \geq 0 \right] + \mathbf{E}\left[e^{(\lambda-\delta)\xi}; \xi < 0 \right] - \mathbf{E}\left[e^{\lambda\xi} \right] \right) \\ &\leq \frac{1}{\delta}\left(\mathbf{E}\left[e^{(\lambda+\delta)\xi} \right] + \mathbf{E}\left[e^{(\lambda-\delta)\xi} \right] - \mathbf{E}\left[e^{\lambda\xi} \right] \right) < \infty \end{aligned}$$

Now by Dominated Convergence applies and we conclude that $\frac{d}{d\lambda}\mathbf{E}\left[e^{\lambda\xi}\right] = \mathbf{E}\left[\xi e^{\lambda\xi}\right]$; $\Lambda'_\xi(\lambda) = \frac{\mathbf{E}\left[\xi e^{\lambda\xi}\right]}{\mathbf{E}\left[e^{\lambda\xi}\right]}$ follows from the Chain Rule.

Suppose that $\Lambda'_\xi(\lambda) = y$. The function $g(\eta) = \eta y - \Lambda_\xi(\eta)$ is concave and satisfies $g'(\lambda) = 0$; it follows that $g(\lambda)$ is a global maximum (TODO: Where do we show this???) and therefore $\lambda y - \Lambda_\xi(\lambda) = \sup_{-\infty<\eta<\infty}\{\lambda\eta - \Lambda_\xi(\eta)\} = \Lambda^*_\xi(y)$. □

TODO: The proof of (i) here provides some motivation for the form of the rate function (see discussion in Stroock).

THEOREM 10.25 (Cramér's Theorem). *Let $\xi, \xi_1, \xi_2, \dots$ be i.i.d. random variables then*

(i) *For every closed set $F \subset \mathbb{R}$ we have*

$$\limsup_{n\to\infty} \frac{1}{n} \log \boldsymbol{E}\left[\frac{1}{n}\sum_{j=1}^n \xi_j \in F \right] \leq -\inf_{x\in F} \Lambda^*_\xi(x)$$

(i) *For every open set $G \subset \mathbb{R}$ we have*

$$\liminf_{n\to\infty} \frac{1}{n} \log \boldsymbol{E}\left[\frac{1}{n}\sum_{j=1}^n \xi_j \in G \right] \geq -\inf_{x\in G} \Lambda^*_\xi(x)$$

PROOF. We start with (i). Note that if either F is empty or $\inf_{x\in F}\Lambda^*_\xi(x) = 0$ then (i) holds trivially so we may assume F is non-empty and that $\inf_{x\in F}\Lambda^*_\xi(x) > 0$. From Lemma 10.24 it follows that

- Λ^*_ξ is not identically zero
- $\mathcal{D}(\Lambda_\xi) \neq \emptyset$
- Either $\mathbf{E}[\xi_+] < \infty$ and $\Lambda^*_\xi(x) = \sup_{\lambda\leq 0}\{\lambda x - \Lambda(\lambda)\}$ or $\mathbf{E}[\xi_-] < \infty$ and $\Lambda^*_\xi(x) = \sup_{\lambda\geq 0}\{\lambda x - \Lambda(\lambda)\}$

Let $-\infty < x < \infty$ and $\lambda \geq 0$ be given and suppose that $\mathbf{E}[\xi] < \infty$ then by Chernoff bounding and the independence of ξ_n we get

$$\begin{aligned} \mathbf{P}\{\frac{1}{n}\sum_{j=1}^n \xi_j \geq x\} = \mathbf{P}\{e^{\sum_{j=1}^n \xi_j} \geq e^{n\lambda x}\} &\leq e^{-n\lambda x}\mathbf{E}\left[e^{\lambda \sum_{j=1}^n \xi_j} \right] \\ &= e^{-n\lambda x}\prod_{j=1}^n \mathbf{E}\left[e^{\lambda\xi_j} \right] = \mathbf{E}\left[e^{-n(\lambda x - \lambda\xi)} \right] \end{aligned}$$

Taking the infimum of the right hand side over all $\lambda \geq 0$ we get

$$\mathbf{P}\{\frac{1}{n}\sum_{j=1}^n \xi_j \geq x\} \leq e^{-n\Lambda_\xi^*(x)}$$

If assume that $\mathbf{E}\left[\xi\right] > -\infty$ and $\lambda \leq 0$ the the same argument yields $\mathbf{P}\{\frac{1}{n}\sum_{j=1}^n \xi_j \leq x\} \leq \mathbf{E}\left[e^{-n(\lambda x - \lambda\xi)}\right]$ and

$$\mathbf{P}\{\frac{1}{n}\sum_{j=1}^n \xi_j \leq x\} \leq e^{-n\Lambda_\xi^*(x)}$$

Now assume that $-\infty < \mathbf{E}\left[\xi\right] < \infty$. From Lemma 10.24 we know that $\Lambda_\xi^*(\mathbf{E}\left[\xi\right]) = 0$ and therefore $\mathbf{E}\left[\xi\right] \notin F$. Let (x_-, x_+) be the union of open intervals $(a,b) \subset F^c$ that contain $\mathbf{E}\left[\xi\right]$. Since F is nonempty either x_- or x_+ is finite. If $x_\pm$ is finite it also follows that $x_\pm \in F$ for otherwise since F^c is open we could find a bigger open interval containing $\mathbf{E}\left[\xi\right]$. In either case we have $\Lambda_\xi^*(x_\pm) \geq \inf_{x\in F}\Lambda_\xi^*(x)$ and therefore by a union bound

$$\begin{aligned}\mathbf{P}\{\frac{1}{n}\sum_{j=1}^n \xi_j \in F\} &\leq \mathbf{P}\{\frac{1}{n}\sum_{j=1}^n \xi_j \geq x_+\} + \mathbf{P}\{\frac{1}{n}\sum_{j=1}^n \xi_j \leq x_-\} \\ &\leq e^{-n\Lambda_\xi^*(x_+)} + e^{-n\Lambda_\xi^*(x_-)} \leq 2e^{-n\inf_{x\in F}\Lambda_\xi^*(x)}\end{aligned}$$

and

$$\limsup_{n\to\infty}\frac{1}{n}\log\mathbf{P}\{\frac{1}{n}\sum_{j=1}^n \xi_j \in F\} \leq \limsup_{n\to\infty}\frac{1}{n}(\log 2 - n\inf_{x\in F}\Lambda_\xi^*(x)) = -\inf_{x\in F}\Lambda_\xi^*(x)$$

If $\mathbf{E}\left[\xi\right] = -\infty$ then Lemma 10.24 says that Λ_ξ^* is nondecreasing and therefore $\lim_{x\to-\infty}\Lambda_\xi^*(x) = 0$. Let $x_+ = \inf\{x \mid x \in F\}$ and observe that if $x_+ = -\infty$ it follows that $\inf_{x\in F}\Lambda_\xi^*(x) = 0$ which is contradiction. Thus since F is non-empty $-\infty < x_+ < \infty$ and since F is closed $x_+ \in F$ and $\Lambda_\xi^*(x_+) \geq \inf_{x\in F}\Lambda_\xi^*(x)$.

$$\mathbf{P}\{\frac{1}{n}\sum_{j=1}^n \xi_j \in F\} \leq \mathbf{P}\{\frac{1}{n}\sum_{j=1}^n \xi_j \geq x_+\} \leq e^{-n\Lambda_\xi^*(x_+)} \leq e^{-n\inf_{x\in F}\Lambda_\xi^*(x)}$$

and the upper bound follows. If $\mathbf{E}\left[\xi\right] = \infty$ argue similarly.

TODO: Tie in the mechanism below with the technique of importance sampling discussed in Steele's development of Girsanov Theory.

Claim 10.25.1. For every $\delta > 0$ we have

$$\liminf_{n\to\infty}\frac{1}{n}\log\mathbf{P}\{-\delta < \frac{1}{n}\sum_{j=1}^n \xi_j < \delta\} \geq \inf_{-\infty<\lambda<\infty}\Lambda_\xi(\lambda) = -\Lambda_\xi^*(0)$$

First we suppose that $\mathbf{P}\{\xi > 0\} > 0$, $\mathbf{P}\{\xi < 0\} > 0$ and there exists an N such that $\mathbf{P}\{-N \leq \xi \leq N\} = 1$. By continuity of measure there exists $\epsilon > 0$ such that $\mathbf{P}\{\xi > \epsilon\} > 0$ and $\mathbf{P}\{\xi < -\epsilon\} > 0$ and therefore

$$\lim_{\lambda\to\infty}\Lambda_\xi(\lambda) = \lim_{\lambda\to\infty}\mathbf{E}\left[e^{\lambda\xi}\right] \geq \lim_{\lambda\to\infty}\mathbf{E}\left[e^{\lambda\xi}; \xi > \epsilon\right] \geq \mathbf{P}\{\xi > \epsilon\}\lim_{\lambda\to\infty}e^{\lambda\epsilon} = \infty$$

and

$$\lim_{\lambda\to-\infty}\Lambda_\xi(\lambda) = \lim_{\lambda\to-\infty}\mathbf{E}\left[e^{\lambda\xi}\right] \geq \lim_{\lambda\to-\infty}\mathbf{E}\left[e^{\lambda\xi}; \xi < -\epsilon\right] \geq \mathbf{P}\{\xi < -\epsilon\}\lim_{\lambda\to-\infty}e^{-\lambda\epsilon} = \infty$$

Since $\mathbf{P}\{|\xi| \leq N\} = 1$ we have for all $-\infty < \lambda < \infty$

$$\Lambda(\lambda) = \mathbf{E}\left[e^{\lambda\xi}\right] = \mathbf{E}\left[e^{\lambda\xi}; |\xi| \leq N\right] \leq e^{|\lambda| N} < \infty$$

Thus $\mathcal{D}(\Lambda) = \mathbb{R}$ and applying Lemma 10.24 we see that Λ is continuous and differentiable on all of $\mathbb{R}$ and $\Lambda'(\lambda) = \mathbf{E}\left[\xi e^{\lambda\xi}\right] / \mathbf{E}\left[e^{\lambda\xi}\right]$. By continuity of Λ and the fact that $\lim_{\lambda \to \pm\infty} \Lambda(\lambda) = \infty$ we know there exists λ_0 such that $\Lambda(\lambda_0) = \inf_{-\infty<\lambda<\infty} \Lambda(\lambda)$ and the fact that Λ is differentiable at λ_0 implies that $\Lambda'(\lambda_0) = 0$.

Let $\mu = \mathcal{L}(\xi)$ and define $\tilde{\mu} = e^{\lambda_0 x - \Lambda_\xi(\lambda_0)} \cdot \mu$ so that

$$\tilde{\mu}(A) = \int_A e^{\lambda_0 x - \Lambda_\xi(\lambda_0)}\, \mu(dx)$$

Since $\int e^{\lambda_0 x}\, \mu(dx) = \mathbf{E}\left[e^{\lambda_0 \xi}\right] = \Lambda_\xi(\lambda_0)$ it follows that $\tilde{\mu}$ is a probability measure. Let $\eta, \eta_1, \eta_2, \dots$ be i.i.d. with $\mathcal{L}(\eta) = \tilde{\mu}$ and observe that

$$\mathbf{E}\left[\eta\right] = \int x e^{\lambda_0 x - \Lambda_\xi(\lambda_0)}\, \mu(dx) = e^{-\Lambda_\xi(\lambda_0)} \mathbf{E}\left[\xi e^{\lambda_0 \xi}\right] = \Lambda'(\lambda_0) = 0$$

and η is integrable since

$$\mathbf{E}\left[|\eta|\right] \leq e^{-\Lambda_\xi(\lambda_0)} \mathbf{E}\left[|\xi|\, e^{\lambda_0 \xi}\right] \leq e^{-\Lambda_\xi(\lambda_0)} N e^{|\lambda_0| N} < \infty$$

Thus we can apply the Weak Law of Large Numbers Theorem 5.15 to η to conclude

$$\lim_{n \to \infty} \mathbf{P}\{-\epsilon < \frac{1}{n} \sum_{j=1}^n \eta_j < \epsilon\} = 1 \text{ for every } \epsilon > 0$$

It remains to see what this implies about the normalized sums $\frac{1}{n} \sum_{j=1}^n \xi_j$. We compute using the i.i.d. property of the ξ_n and the fact that $-n\epsilon < y < n\epsilon$ implies $\lambda_0 y \leq |\lambda_0|\, n\epsilon$ to see

$$\begin{aligned}
\mathbf{P}\{-\epsilon < \frac{1}{n} \sum_{j=1}^n \xi_j < \epsilon\} &= \int_{\left|\sum_{j=1}^n x_j\right| < n\epsilon} \mu(dx_1) \cdots \mu(dx_n) \\
&\geq \int_{\left|\sum_{j=1}^n x_j\right| < n\epsilon} e^{-|\lambda_0| n\epsilon + \lambda_0 \sum_{j=1}^n x_j} \mu(dx_1) \cdots \mu(dx_n) \\
&= e^{-|\lambda_0| n\epsilon + n\Lambda_\xi(\lambda_0)} \int_{\left|\sum_{j=1}^n x_j\right| < n\epsilon} e^{\sum_{j=1}^n (\lambda_0 x_j - \Lambda_\xi(\lambda_0))}\, \mu(dx_1) \cdots \mu(dx_n) \\
&= e^{-|\lambda_0| n\epsilon + n\Lambda_\xi(\lambda_0)} \int_{\left|\sum_{j=1}^n x_j\right| < n\epsilon} \tilde{\mu}(dx_1) \cdots \tilde{\mu}(dx_n) \\
&= e^{-n(|\lambda_0|\epsilon + \Lambda_\xi(\lambda_0))} \mathbf{P}\{-\epsilon < \frac{1}{n} \sum_{j=1}^n \eta_j < \epsilon\}
\end{aligned}$$

Therefore for every $0 < \epsilon < \delta$ we have

$$\begin{aligned}\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}\{-\delta < \frac{1}{n}\sum_{j=1}^n \xi_j < \delta\} &\geq \liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}\{-\epsilon < \frac{1}{n}\sum_{j=1}^n \xi_j < \epsilon\} \\ &\geq \liminf_{n\to\infty} \frac{1}{n} \log e^{-n(|\lambda_0|\epsilon + \Lambda_\xi(\lambda_0))} \mathbf{P}\{-\epsilon < \frac{1}{n}\sum_{j=1}^n \eta_j < \epsilon\} \\ &= -\left|\lambda_0\right|\epsilon + \Lambda_\xi(\lambda_0) + \liminf_{n\to\infty} \mathbf{P}\{-\epsilon < \frac{1}{n}\sum_{j=1}^n \eta_j < \epsilon\} \\ &= -\left|\lambda_0\right|\epsilon + \Lambda_\xi(\lambda_0)\end{aligned}$$

Taking the limit as $\epsilon \to 0$ we see that $\liminf_{n\to\infty} \frac{1}{n} \log \mathbf{P}\{-\delta < \frac{1}{n}\sum_{j=1}^n \xi_j < \delta\} = \Lambda_\xi(\lambda_0)$ and for this special case the claim is proven.

TODO: Show the claim implies the lower bound and complete the proof of the claim. ☐

Before attacking the generalization of Cramér's Theorem to $\mathbb{R}^d$ we need to spend some time to arm ourselves with a few technical tools.

Unboundedness of rate functions can be a technical impediment in certain proof scenarios. The following proposition shows that for proving upper bounds we can reduce to the bounded case by giving up positivity; this is often a good tradeoff.

Proposition 10.26. $\lim_{\delta\to 0} \inf_{x\in\Gamma}(I(x) - \delta) \wedge \frac{1}{\delta} = \inf_{x\in\Gamma} I(x)$.

Proof. Suppose $\inf_{x\in\Gamma} I(x) = \infty$ (i.e. $I \equiv \infty$) then if follows that $\inf_{x\in\Gamma}(I(x) - \delta) \wedge \frac{1}{\delta} = \frac{1}{\delta}$ and clearly the result holds. Similarly if $\inf_{x\in\Gamma} I(x) = -\infty$ then we may find $x_n \in \Gamma$ such that $I(x_n) < -n$ for $n \in \mathbb{N}$ and therefore $\inf_{x\in\Gamma}(I(x)-\delta)\wedge\frac{1}{\delta} = -\infty$ for all $\delta > 0$.

So we may assume $-\infty < \inf_{x\in\Gamma} I(x) < \infty$ which implies $\inf_{x\in\Gamma}(I(x)-\delta)\wedge\frac{1}{\delta}$ is finite for all $\delta > 0$. From the fact that $(I(x)-\delta)\wedge\frac{1}{\delta} < I(x)$ for all $\delta > 0$ it follows that $\inf_{x\in\Gamma}(I(x)-\delta)\wedge\frac{1}{\delta} \leq \inf_{x\in\Gamma} I(x)$ for all $\delta > 0$ and thus $\lim_{\delta\to 0}\inf_{x\in\Gamma}(I(x)-\delta)\wedge\frac{1}{\delta} \leq \inf_{x\in\Gamma} I(x)$. Also $(I(x)-\delta)\wedge\frac{1}{\delta}$ is a non-decreasing function of $\delta > 0$ for each $x \in \Gamma$ thus $C = \lim_{\delta\to 0}\inf_{x\in\Gamma}(I(x)-\delta)\wedge\frac{1}{\delta}$ is defined, bounded above by $\inf_{x\in\Gamma} I(x)$ hence finite. For every $0 < \epsilon < 1$ we may pick $0 < \delta < \epsilon/2 \min /frac1C + 1$ such that $C - \epsilon/2 < \inf_{x\in\Gamma}(I(x)-\delta)\wedge\frac{1}{\delta} \leq C$. Now pick $x \in \Gamma$ such that $C-\epsilon/2 < (I(x)-\delta)\wedge\frac{1}{\delta} < C + \epsilon/2$ which by our choice of δ is equivalent to $C - \epsilon/2 < I(x) - \delta < C + \epsilon/2$. Thus $I(x) < C + \epsilon$ hence $\inf_{x\in\Gamma} < C + \epsilon$. Now let $\epsilon \to 0$. ☐

The most primitive tool for bounding probabilities is the union bound. When proving large deviation results we often need the following logarithmic union bound.

Proposition 10.27. *Let $(\Omega, \mathcal{A})$ be a measurable space, μ_ϵ be a family of probability measures and let $A_1, A_2, \dots, A_n$ be measurable sets then*

$$\limsup_{\epsilon\to 0} \epsilon \log \mu_\epsilon(\cup_{j=1}^n A_j) \leq \max_{1\leq j\leq n} \limsup_{\epsilon\to 0} \epsilon \log \mu_\epsilon(A_j)$$

Proof. By a union bound and the fact that log is increasing we have $\limsup_{\epsilon\to 0} \epsilon \log \mu_\epsilon(\cup_{j=1}^n A_j) \leq \limsup_{\epsilon\to 0} \epsilon \log \sum_{j=1}^n \mu_\epsilon(A_j)$; we just need to bound the right hand side. Note that for every $1 \leq k \leq n$ since log is increasing

$$\begin{aligned}\log \mu_\epsilon(A_k) &\leq \log \sum_{j=1}^{n} \mu_\epsilon(A_j) \leq \log n \max_{1 \leq j \leq n} \mu_\epsilon(A_j) \\ &= \log n + \max_{1 \leq j \leq n} \log \mu_\epsilon(A_j)\end{aligned}$$

Multiplying by ϵ, taking the limit and using the fact that $\limsup(f \vee g) = \limsup f \vee \limsup g$ we get

$$\max_{1 \leq j \leq n} \limsup_{\epsilon \to 0} \epsilon \log \mu_\epsilon(A_j) = \limsup_{\epsilon \to 0} \epsilon \max_{1 \leq j \leq n} \log \mu_\epsilon(A_j) = \limsup_{\epsilon \to 0} \epsilon \log \sum_{j=1}^{n} \mu_\epsilon(A_j)$$

$\square$

TODO: Example of weak LDP for which there is no LDP

Definition 10.28. Let $(\Omega, \mathcal{A})$ be a topological measurable space such that every compact set is measurable and μ_ϵ be a family of probability measures. We say that μ_ϵ is *exponentially tight* if for every $0 < \alpha < \infty$ there exists a compact set K_α such that

$$\limsup_{\epsilon \to 0} \epsilon \log \mu_\epsilon(K_\alpha^c) < -\alpha$$

CHAPTER 11

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 11.1. Suppose μ and ν such that $\mu \ll \nu$. The the *Kullback-Liebler divergence* or *relative entropy* of μ and ν is defined as

$$D\left(\mu \mid\mid \nu\right) = \mathbf{E}_\mu \left[\log \frac{d\mu}{d\nu} \right]$$

If μ is not absolutely continuous with respect to ν then by convention $D\left(\mu \mid\mid \nu\right) = \infty$.

Example 11.2. Suppose μ and ν are probability measures that are both absolutely continuous with respect to a third measure λ and furthermore $\mu \ll \nu$. 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

$$D\left(\mu \mid\mid \nu\right) = \int \log(\frac{f}{g}) f \, d\lambda$$

that the user may have encountered before.

EXAMPLE 11.3. Suppose μ and ν are probability measures on a finite state space $\Omega = \{a_1, \dots, a_n\}$. Let $\mu_j = \mu(a_j)$ and $\nu_j = \nu(a_j)$ for $j = 1, \dots, n$. Both μ and ν are absoutely continuous with respect to the counting measure on Ω. Furthermore $\mu \ll \nu$ if and only if $\nu_j = 0$ implies $\mu_j = 0$. From the previous example

$$D(\mu \,||\, \nu) = \sum_{j=1}^{n} \mu_j \log(\frac{\mu_j}{\nu_j})$$

where again we have the convention that $\log(\frac{0}{0}) = 0$.

EXAMPLE 11.4. 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 $\mu \ll \nu$ and using the formula for relative entropy in terms of densities in the previous example we compute

$$D(\mu \,||\, \nu) = \log(\frac{1}{1/2}) \cdot 1 + \log(\frac{0}{1/2}) \cdot 0 = \log 2 = 1$$

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

$$D(\nu \,||\, \mu) = \log(\frac{1/2}{1}) \cdot \frac{1}{2} + \log(\frac{1/2}{0}) \cdot \frac{1}{2} = \infty$$

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 11.5 (Gibbs Inequality). *For all probability distributions μ and ν, $D(\mu \,||\, \nu) \geq 0$ 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 $\nu \ll \mu$. 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

$$D(\mu \,||\, \nu) = \mathbf{E}_\mu \left[\log \frac{d\mu}{d\nu}\right] = \mathbf{E}_\nu \left[\frac{d\mu}{d\nu} \log \frac{d\mu}{d\nu}\right] \geq \mathbf{E}_\nu \left[\frac{d\mu}{d\nu}\right] \log \mathbf{E}_\nu \left[\frac{d\mu}{d\mu}\right] = \mathbf{E}_\mu [1] \log \mathbf{E}_\mu [1] = 0$$

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. $\square$

EXAMPLE 11.6. 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 11.5 suggests a potential path. We know from the previous example that

$$D(\theta_0 \mid\mid \theta) = \mathbf{E}_{\theta_0}\left[\log(\frac{f(x \mid \theta_0)}{f(x \mid \theta)})\right] = \mathbf{E}_{\theta_0}\left[\log(f(x \mid \theta_0))\right] - \mathbf{E}_{\theta_0}\left[\log(f(x \mid \theta))\right] \geq 0$$

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 $\mathbf{E}_{\theta_0}\left[\log(f(x \mid \theta))\right]$ 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...

1. The Delta Method

TODO: Move the discussion of tightness into the convergence chapter.

DEFINITION 11.7. 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 relatively 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 11.8. *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 $M = \max(M_2, M'_1, \dots, M'_{n-1})$ then we get $\sup_n \mathbf{P}\{|\xi_n| > M\} < \epsilon$ and tightness is shown. ☐

Lemma 11.9. *Suppose r_n is a sequence of real numbers such that $\lim_{n\to\infty} |r_n| = \infty$ and $\eta, \xi, \xi_1, \xi_2, \dots$ 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 11.8). 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 11.10. *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 5.30) the therefore we know ξ_n is tight by Lemma 11.8.

The statement of (ii) is a corollary to the Continuous Mapping Theorem (Corollary 5.14).

TODO: Finish... □

Here is a slightly more involved fact that we shall use in the sequel.

LEMMA 11.11. *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 5.11) 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

$$\begin{aligned}
(1-A)^{-1} &= \sum_{n=0}^{\infty} A^n && \text{converges absolutely} \\
\left\|(1-A)^{-1}\right\| &\leq \sum_{n=0}^{\infty} \|A^n\| \leq \sum_{n=0}^{\infty} \|A\|^n = (1-\|A\|)^{-1} \\
(1-A)(1-A)^{-1} &= \sum_{n=0}^{\infty} A^n - \sum_{n=1}^{\infty} A^n = 1 \\
(1-A)^{-1}(1-A) &= \sum_{n=0}^{\infty} A^n - \sum_{n=1}^{\infty} A^n = 1
\end{aligned}$$

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 $\left\|T^{-1}A\right\| \leq \left\|T^{-1}\right\| \|A\| < 1$ 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

$$\left\|(T - A)^{-1}\right\| \leq \|T\| \left\|(1 - T^{-1}A)^{-1}\right\| \leq \frac{\|T\|}{1 - \|T^{-1}A\|} \leq \frac{\|T\|}{1 - \|T^{-1}\| \|A\|}$$

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 $\left\|\Psi_n^{-1}\right\| < 2\left\|\Psi^{-1}\right\|$. 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

$$\sup_{n > N} \mathbf{P}\{\|\xi_n\| > 2M \left\|\Psi^{-1}\right\|\} < \delta$$

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 11.12. 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 11.13 (Delta Method). *Let $\phi : D \subset \mathbb{R}^k \to \mathbb{R}^m$ be Frechet differentiable at $\theta \in D$. Let $\xi, \xi_1, \xi_2, \dots$ 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 11.9 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 5.45) 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 5.45) to see that $D\phi(\theta) r_n(\xi_n - \theta) \xrightarrow{d} D\phi(\theta)\xi$ and Slutsky's Lemma (Lemma 5.46)) and the part of this Theorem just proven to conclude $r_n(\phi(\xi_n) - \phi(\theta)) \xrightarrow{d} D\phi(\theta)\xi$. □

Example 11.14. 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 11.13) 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 11.15. 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 $i = 1, \dots, n$, every $\theta \in \Theta$ and every $x \in B$.

(ii) For every $k = 1, \dots, n$,

$$\frac{\partial}{\partial \theta_i} \int f(x \mid \theta)\, d\mu(x) = \int \frac{\partial}{\partial \theta_i} f(x \mid \theta)\, d\mu(x)$$

(iii) The set $C = \{x \in X \mid f(x \mid \theta) > 0$ does not depend on θ.

Definition 11.16. 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

$$U(\xi \mid \theta) = \left(\frac{\partial}{\partial \theta_1} \log f(\xi \mid \theta), \dots, \frac{\partial}{\partial \theta_n} \log f(\xi \mid \theta) \right)$$

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

$$\begin{aligned}
\mathbf{E}_\theta \left[\frac{\partial}{\partial \theta_i} \log f(\xi \mid \theta) \right] &= \int \frac{\frac{\partial}{\partial \theta_i} f(x \mid \theta)}{f(x \mid \theta)} f(x \mid \theta)\, d\mu(x) \\
&= \int \frac{\partial}{\partial \theta_i} f(x \mid \theta)\, d\mu(x) \\
&= \frac{\partial}{\partial \theta_i} \int f(x \mid \theta)\, d\mu(x) = \frac{\partial}{\partial \theta_i} 1 = 0
\end{aligned}$$

and therefore $\mathbf{E}_\theta \left[U(\xi \mid \theta) \right] = 0$ under the FI regularity constraints.

If we differentiate both side of this latter equality

$$\begin{aligned}
0 &= \frac{\partial}{\partial \theta_j} \int \frac{\partial}{\partial \theta_i} \log f(x \mid \theta) f(x \mid \theta)\, d\mu(x) \\
&= \int \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(x \mid \theta) f(x \mid \theta) + \frac{\partial}{\partial \theta_i} \log f(x \mid \theta) \frac{\partial}{\partial \theta_j} f(x \mid \theta)\, d\mu(x) \\
&= \int \left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(x \mid \theta) + \frac{\partial}{\partial \theta_i} \log f(x \mid \theta) \frac{\partial}{\partial \theta_j} \log f(x \mid \theta) \right) f(x \mid \theta)\, d\mu(x)
\end{aligned}$$

which shows that when ξ has density $f(x \mid \theta)$, we have the identity

$$-\mathbf{E}_\theta \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log f(\xi \mid \theta) \right] = \mathbf{E}_\theta \left[\frac{\partial}{\partial \theta_i} \log f(\xi \mid \theta) \frac{\partial}{\partial \theta_j} \log f(\xi \mid \theta) \right]$$

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 11.17. 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 11.18. 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 $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)$ 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 11.5) 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 $\frac{1}{n}\sum_{i=1}^n \log f(\xi_i \mid \theta) \stackrel{a.s.}{\to} \mathbf{E}_{\theta_0}[\log f(x \mid \theta)]$ 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 11.19. 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 11.20. *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 11.19 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 11.21 (Asymptotic Consistency of MLE). *Let $\xi, \xi_1, \xi_2, \dots$ 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 5.4, 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 $U_1, \dots, U_{m-1}$ 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 $U_1, \dots, U_m$ 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 $1 \leq j \leq m$,

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n Z(U_j, \xi_i) = \mathbf{P}_{\theta_0}\{Z(U_j, \xi)\} > 0$$

$\square$

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). TODO: Check with Charles Geyer's notes on MLE; I think he addresses this issue.

EXAMPLE 11.22. 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 11.21 we need to show $\mathbf{E}_{\theta_0}\left[Z(U, \theta)\right] > 0$ 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 $\mathbf{E}_{\theta_0}\left[Z(U_\theta, x)\right] > 0$.

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 $\mathbf{E}_{\theta_0}\left[Z(U_\theta, x)\right] = +\infty$.

Lastly we have to find a compact set K such that $\mathbf{E}_{\theta_0}\left[Z(\mathbb{R}_+ \setminus K, x)\right] > 0$. 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,

$$\begin{aligned} \mathbf{E}_{\theta_0}\left[Z(\mathbb{R}_+ \setminus [\theta_0/a, a\theta_0], x)\right] &= \frac{1}{\theta_0} \int_0^{\frac{\theta_0}{a}} \log \frac{x}{\theta_0}\, dx + \frac{\theta_0 - \frac{\theta_0}{a}}{\theta_0} \log a \\ &= (\frac{1}{a} \log \frac{1}{a} - \frac{\theta_0}{a}) + \frac{\theta_0 - \frac{\theta_0}{a}}{\theta_0} \log a \end{aligned}$$

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 $\mathbf{E}_{\theta_0}\left[Z(\mathbb{R}_+ \setminus [\theta_0/a, a\theta_0], x)\right] > 0$.

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 $\mathbf{E}_{\theta_0}\left[Z(U, x)\right] < 0$ 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 11.23. 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 $\mathbf{E}_{\theta_0}\left[Z(K^c, x)\right] = -\infty$ and therefore Theorem 11.21 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 11.24. *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 $(X, \mathcal{X}, \mu)$. 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 $\boldsymbol{E}_{\theta_0}\left[Z(U_\theta, \xi)\right] > -\infty$. 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 $\boldsymbol{E}_{\theta_0}\left[Z(\Theta \setminus K, \xi)\right] > 0$. 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 $\mathbf{E}_{\theta_0}\left[Z(U_\theta, \xi)\right] > 0$ and then apply the previous Theorem 11.21.

Pick $\theta \neq \theta_0$ and assume that we have an open neighborhood U_θ with $\theta \in U_\theta$ and $\mathbf{E}_{\theta_0}\left[Z(U_\theta, \xi)\right] > -\infty$. If $\mathbf{E}_{\theta_0}\left[Z(U_\theta, \xi)\right] > 0$ then have found a suitable neighborhood so we may assume $\mathbf{E}_{\theta_0}\left[Z(U_\theta, \xi)\right] \leq 0$ 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 $\mathbf{E}_{\theta_0}\left[Z(U_\theta, \xi)\right]$ is finite, we may apply Fatou's Lemma (Theorem 2.45)

$$\begin{aligned}
\liminf_{n\to\infty} \mathbf{E}_{\theta_0}\left[Z(U_\theta^n, x)\right] - \mathbf{E}_{\theta_0}\left[Z(U_\theta, x)\right] &= \liminf_{n\to\infty} \mathbf{E}_{\theta_0}\left[Z(U_\theta^n, x) - Z(U_\theta, x)\right] \\
&\geq \mathbf{E}_{\theta_0}\left[\lim_{n\to\infty} \left(Z(U_\theta^n, x) - Z(U_\theta, x)\right)\right] \\
&= \mathbf{E}_{\theta_0}\left[\log \frac{f(x \mid \theta_0)}{f(x \mid \theta)}\right] - \mathbf{E}_{\theta_0}\left[Z(U_\theta, x)\right]
\end{aligned}$$

Cancelling the (finite) common term $\mathbf{E}_{\theta_0}\left[Z(U_\theta, x)\right]$ we get

$$\liminf_{n\to\infty} \mathbf{E}_{\theta_0}\left[Z(U_\theta^n, x)\right] \geq \mathbf{E}_{\theta_0}\left[\log \frac{f(x \mid \theta_0)}{f(x \mid \theta)}\right] > 0$$

where the last inequality follows from the positivity of relative entropy (Lemma 11.5). Now by this inequality we can find an $N > 0$ such that $\mathbf{E}_{\theta_0}\left[Z(U_\theta^n, x)\right] > 0$ 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 $i(\theta_0) = \mathbf{E}_{\theta_0} \left[\log^2 f(\xi \mid \theta_0)\right]$ 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

$$-\frac{1}{n} \frac{\partial^2}{\partial \theta^2} \ell_n(\xi \mid \theta_0) = -\frac{1}{n} \sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \ell_n(\xi_i \mid \theta_0) \xrightarrow{P} \mathbf{E}_{\theta_0} \left[-\frac{\partial^2}{\partial \theta^2} \log f(\xi \mid \theta_0) \right] = i(\theta_0)$$

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 11.25. *Let $\xi, \xi_1, \xi_2, \dots$ 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 $\left\| \hat{\theta}_n - \theta_0 \right\| < R$ 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 $\left\| \hat{\theta}_n - \theta_0 \right\| \geq R$ won't matter since it is an event that occurs with vanishingly small probability as n grows. Accordingly, we define

$$\Delta_n = \begin{cases} D^2_{\theta_n^*} \ell_n(\boldsymbol{\xi} \mid \theta) & \text{when } \left\| \hat{\theta}_n - \theta_0 \right\| < R \\ 0 & \text{when } \left\| \hat{\theta}_n - \theta_0 \right\| \geq R \end{cases}$$

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

$$\frac{1}{\sqrt{n}}(D_{\theta_0} \ell_n(\boldsymbol{\xi} \mid \theta) + \Delta_n \cdot (\hat{\theta}_n - \theta_0)) = \begin{cases} \frac{1}{\sqrt{n}}(D_{\hat{\theta}_n} \ell_n(\boldsymbol{\xi} \mid \theta) & \text{when } \left\| \hat{\theta}_n - \theta_0 \right\| < R \\ \frac{1}{\sqrt{n}}(D_{\theta_0} \ell_n(\boldsymbol{\xi} \mid \theta) & \text{when } \left\| \hat{\theta}_n - \theta_0 \right\| \geq R \end{cases}$$

and therefore

$$\begin{aligned}&\lim_{n \to \infty} \mathbf{P}_{\theta_0}\{\frac{1}{\sqrt{n}}(D_{\theta_0} \ell_n(\boldsymbol{\xi} \mid \theta) + \Delta_n \cdot (\hat{\theta}_n - \theta_0)) > \epsilon\} \\ &= \lim_{n \to \infty} \mathbf{P}_{\theta_0}\{\frac{1}{\sqrt{n}} D_{\hat{\theta}_n} \ell_n(\boldsymbol{\xi} \mid \theta) > \epsilon; \left\| \hat{\theta}_n - \theta_0 \right\| < R\} \\ &+ \lim_{n \to \infty} \mathbf{P}_{\theta_0}\{\frac{1}{\sqrt{n}} D_{\theta_0} \ell_n(\boldsymbol{\xi} \mid \theta) > \epsilon; \left\| \hat{\theta}_n - \theta_0 \right\| \geq R\} \\ &\leq \lim_{n \to \infty} \mathbf{P}_{\theta_0}\{\frac{1}{\sqrt{n}} D_{\hat{\theta}_n} \ell_n(\boldsymbol{\xi} \mid \theta) > \epsilon\} + \lim_{n \to \infty} \mathbf{P}_{\theta_0}\{\left\| \hat{\theta}_n - \theta_0 \right\| \geq R\} = 0\end{aligned}$$

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

$$\frac{1}{n}D^2_{\theta_0}\ell_n(\boldsymbol{\xi} \mid \theta) = \frac{1}{n}\sum_{i=1}^n D^2_{\theta_0}\log f(\xi_i \mid \theta) \xrightarrow{P} \mathbf{E}_{\theta_0}\left[D^2_{\theta_0}\log f(\xi \mid \theta)\right] = -\mathcal{I}_\xi(\theta_0)$$

and therefore by Corollary 5.14 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 $\mathbf{E}_{\theta_0}\left[H_r(\xi, \theta_0)\right] < \frac{\epsilon}{2}$. Again applying the Weak Law of Large Numbers

$$\frac{1}{n}\sum_{i=1}^n H_r(\xi_i, \theta_0) \xrightarrow{P} \mathbf{E}_{\theta_0}\left[H_r(\xi, \theta_0)\right] < \frac{\epsilon}{2}$$

and therefore

$$\lim_{n\to\infty}\mathbf{P}_{\theta_0}\{\frac{1}{n}\sum_{i=1}^n H_r(\xi_i, \theta_0) < \epsilon\} \leq \lim_{n\to\infty}\mathbf{P}_{\theta_0}\{\left|\frac{1}{n}\sum_{i=1}^n H_r(\xi_i, \theta_0) - \mathbf{E}_{\theta_0}\left[H_r(\xi, \theta_0)\right]\right| < \frac{\epsilon}{2}\} = 0$$

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

$$\frac{1}{n}D_{\theta_0}\ell_n(\boldsymbol{\xi} \mid \theta) = \frac{1}{n}\sum_{i=1}^n D_{\theta_0}\log f(\xi_i \mid \theta) \xrightarrow{P} \mathbf{E}_{\theta_0}\left[D_{\theta_0}\log f(\xi \mid \theta)\right]$$

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

$$\mathbf{E}_{\theta_0}\left[\frac{\partial}{\partial\theta_i}\log f(\xi_i \mid \theta)\right] = \int \frac{\partial}{\partial\theta_i}\log f(x \mid \theta) f(x \mid \theta_0) dx = \int \frac{\partial}{\partial\theta_i} f(x \mid \theta_0)\, dx = \frac{\partial}{\partial\theta_i}\int f(x \mid \theta_0)\, dx = 0$$

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 11.8, 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 11.10). Claim 3 and the invertibility of $\mathcal{I}_\xi(\theta_0)$ allows us to apply Lemma 11.11 to conclude that $\frac{1}{\sqrt{n}}\Delta_n\cdot(\hat{\theta}_n - \theta_0)$ is tight. Now by Claim 3 and Lemma 11.10 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 5.46) 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 11.26 (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.

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.

$$\binom{n}{m} p^m (1-p)^{n-m} = e^{\ln(\binom{n}{m})} e^{\ln(p^m)} e^{\ln((1-p)^{n-m}} \tag{8}$$

$$= e^{\ln(\binom{n}{m})+m\ln(p/(1-p))+n\ln(1-p)} \tag{9}$$

$$= e^{\ln(\binom{n}{m})+m\ln(p/(1-p))-n\ln(1+p/(1-p))} \tag{10}$$

$$= e^{\ln(\binom{n}{m})+m\theta-n\ln(1+e^\theta)} \tag{11}$$

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

$$s(\theta; y) = \frac{\partial}{\partial\theta} l(\theta; y) = y - \frac{ne^\theta}{1+e^\theta} = y - np \tag{12}$$

$$i(\theta; y) = -\frac{\partial}{\partial\theta} s(\theta; y) = np(1-p) \tag{13}$$

3. Bayesian Models

Here are some simple examples of Bayesian updating for models in which conjugate priors exist so that we have closed form solutions.

Example 11.27. Suppose we have a normal population $N(\mu, \sigma^2)$ with σ^2 assumed known and μ assumed to be distributed $N(\mu_0, \sigma_0^2)$ with μ_0 and σ_0^2 known. If we are given independent observations $x_1, \dots, x_n$ then we have a likelihood function

$$p(\mathbf{x} \mid \mu) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i-\mu)^2/2\sigma^2} = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2}$$

and by Bayes' Theorem

$$p(\mu \mid \mathbf{x}) \propto \frac{1}{\sqrt{2\pi}\sigma_0} e^{-(\mu-\mu_0)^2/2\sigma_0^2} \cdot \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2}$$

A generalization of the above to a linear regression scenario is

Example 11.28. Here we have a model $y = X^t\beta + \epsilon$ with

(i) ϵ is $N(0, \sigma^2)$
(ii) $1/\sigma^2$ is $\Gamma(\alpha, \beta)$ with α and β known.
(iii) β is $N(\mu_0, \sigma^2\Lambda_0^{-1})$ with μ_0 and Λ_0 known

We suppose that we are given independent observations $X_1, \dots, X_n$ which we assemble into an observation matrix X. TODO: Finish

CHAPTER 12

Brownian Motion

We begin by studying the one dimensional version of Brownian motion.

Definition 12.1. 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 12.2. *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 4.34).

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,

$$\begin{aligned} \mathbf{Var}\,(B_1(1/4)) &= \frac{1}{16} + \frac{1}{16} = 1/8 = 1/4 - 1/8 \\ \mathbf{Var}\,(B_1(3/4)) &= \frac{9}{16} + \frac{1}{16} = 5/8 = 3/4 - 1/8 \end{aligned}$$

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

$$
\begin{aligned}
B_n(x) &= B_0(x) + \sum_{j=0}^{n-1} F_j(x) \\
B(x) &= B_0(x) + \sum_{j=0}^{\infty} F_j(x)
\end{aligned}
$$

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 12.3. *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 (Proposition 7.21; TODO: Show that they are jointly Gaussian) which is a simple computation using the formulae above and the induction hypothesis

$$\begin{aligned}
&\mathbf{E}\left[(B((2k+1)/2^n) - B(2k/2^n))(B((2k+2)/2^n) - B((2k+1)/2^n))\right] \\
&= \mathbf{E}\left[\left(\frac{B((2k+2)/2^n) - B(2k/2^n)}{2} + \frac{1}{\sqrt{2^{n+1}}} Z_{(2k+1)/2^n}\right)\left(\frac{B((2k+2)/2^n) - B(2k/2^n)}{2} - \frac{1}{\sqrt{2^{n+1}}} Z_{(2k+1)/2^n}\right)\right] \\
&= \frac{1}{4}\mathbf{E}\left[(B((2k+2)/2^n) - B(2k/2^n))^2\right] - \frac{1}{2^{n+1}} \\
&= \frac{1}{4}\frac{1}{2^{n-1}} - \frac{1}{2^{n+1}} = 0
\end{aligned}$$

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 $0 \leq j < k \leq 2^n$. In a similar way to the case above we know that by using the result (ii) we can see that for $0 \leq k < 2^n$,

$$B((k+1)/2^n) - B(k/2^n) = \begin{cases} \frac{B((k+1)/2^n)-B((k-1)/2^n)}{2} - \frac{1}{\sqrt{2^{n+1}}} Z_{k/2^n} & k \text{ is odd} \\ \frac{B((k+2)/2^n)-B(k/2^n)}{2} + \frac{1}{\sqrt{2^{n+1}}} Z_{(k+1)/2^n} & k \text{ is even} \end{cases}$$

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 $\frac{k}{2^n} \leq x \leq \frac{k+1}{2^n}$ and $0 \leq k < 2^n$, 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

$$\begin{aligned}\mathbf{Var}\,(B_{n+1}(x)) &= \mathbf{Var}\,(B_n(x)) + 2^n(x-\frac{k}{2^n})^2\\ &= 2^n(x-\frac{k}{2^n})^2 + \frac{k}{2^n} + 2^n(x-\frac{k}{2^n})^2\\ &= 2^{n+1}(x-\frac{k}{2^n})^2 + \frac{k}{2^n}\end{aligned}$$

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

$$\begin{aligned}\mathbf{Var}\,(B_{n+1}(x)) &= \mathbf{Var}\,(B_n(x)) + 2^n(x-\frac{k+1}{2^n})^2\\ &= 2^n(x-\frac{k}{2^n})^2 + \frac{k}{2^n} + 2^n(x-\frac{k+1}{2^n})^2\\ &= 2^n\left[(x-\frac{2k+1}{2^{n+1}})^2 + \frac{1}{2^{n+1}}(x-\frac{2k+1}{2^{n+1}}) + \frac{1}{2^{2n+2}}\right] +\\ &\quad 2^n\left[(x-\frac{2k+1}{2^{n+1}})^2 - \frac{1}{2^{n+1}}(x-\frac{2k+1}{2^{n+1}}) + \frac{1}{2^{2n+2}}\right] + \frac{k}{2^n}\\ &= 2^{n+1}(x-\frac{2k+1}{2^{n+1}})^2 + \frac{2k+1}{2^{n+1}}\end{aligned}$$

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 $x = 0, \frac{1}{2^n}, \dots, 1$, 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.

$$\begin{aligned}\mathbf{P}\{|Z_d| \ge \lambda\} &= \frac{2}{\sqrt{2\pi}}\int_\lambda^\infty e^{\frac{-u^2}{2}}\,du\\ &\le \frac{2}{\sqrt{2\pi}}\int_\lambda^\infty \frac{u}{\lambda}e^{\frac{-u^2}{2}}\,du\\ &= \frac{1}{\lambda\sqrt{2\pi}}e^{\frac{-\lambda^2}{2}}\end{aligned}$$

so if we pick any constant $c > 1$ and $n > 0$, then

$$\mathbf{P}\{|Z_d| \ge c\sqrt{n}\} \le \frac{1}{c\sqrt{2\pi n}}e^{\frac{-c^2 n}{2}} \le e^{\frac{-c^2 n}{2}}$$

Now using this bound, we see that

$$\sum_{n=0}^{\infty} \mathbf{P}\{\text{there exists } d \in \mathcal{D}_n \text{ such that } |Z_d| \geq c\sqrt{n}\} \leq \sum_{n=0}^{\infty} \sum_{d \in \mathcal{D}_n} \mathbf{P}\{|Z_d| \geq c\sqrt{n}\}$$
$$\leq \sum_{n=0}^{\infty} 2^n e^{\frac{-c^2 n}{2}}$$
$$= \sum_{n=0}^{\infty} e^{-n(c^2 - 2\ln 2)/2}$$

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 $d \in \mathcal{D}_n$. Using this result with the definition of $F_n(x) = \sum_{k=0}^{2^n-1} \frac{1}{\sqrt{2^{n+2}}} Z_{\frac{2k+1}{2^{n+1}}} \Delta_{n,k}(x)$, the disjointness of the support of $\Delta_{n,k}(x)$ for fixed n and the fact that $|\Delta_{n,k}(x)| \leq 1$ we have $\|F_n\|_\infty \leq 2^{-(n+2)/2} c\sqrt{n+1}$ 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 1.39).

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

$$\lim_{m\to\infty} \mathbf{Cov}\left(B(x_{j,m}) - B(x_{j-1,m}), B(x_{i,m}) - B(x_{i-1,m})\right) = \delta_{i,j} \lim_{m\to\infty} (x_{i,m} - x_{i-1,m})$$
$$= \delta_{i,j}(x_i - x_{i-1})$$

and therefore by Lemma 7.22 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.

Note that we have ignored measurability considerations up to this point and it is worth filling in that gap so that we have verified our construction defines a proper stochastic process. Since we have defined B as an almost sure limit of the B_n it suffices to show that each B_n is measurable (Lemma 2.14). Now each B_n is a sum of terms each of which is a random variable times a deterministic function so by Lemma 2.19 it suffices to show each such term is measurable. So let ξ be a random variable and let $g(x)$ an element of $\mathbb{R}^{[0,1]}$. If we pick $0 \leq x \leq 1$ and $A \in \mathcal{B}(\mathbb{R})$ then $\{\xi g(x) \in A\} = \{\xi \in A \cdot 1/g(x)\}$ which is measurable because ξ is (here we have ignored the case in which $g(x) = 0$; in that case the set is either $\emptyset$ or Ω so is measurable). Since sets of the form $\{f(t) \in A\}$ generate the σ-algebra on $\mathbb{R}^{[0,1]}$ we have shown that $\xi g(x)$ is measurable (Lemma 2.12). $\square$

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: 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 (either just assume that the σ-algebra on $C[0,\infty)$ is induced from the product $\mathbb{R}^{[0,\infty)}$ or note that the Borel σ-algebra on $C[0,\infty)$ is generated by the projections; in either case this follows from Lemma 9.6).

Definition 12.4. A function $f : (S,d) \to (T,d')$ between metric spaces is said to be *Hölder continuous* with exponent α if there exists a constant $C > 0$ such that $d'(f(x), f(y)) \leq C d(x,y)^\alpha$ for all $x, y \in S$.

Lemma 12.5. *Let $f : [0,1] \to \mathbb{R}$ be continuous with $f(x) = c_0 + \sum_{n=0}^\infty \sum_{k=0}^{2^n-1} c_{n,k} \Delta_{n,k}(x)$. Suppose $|c_{n,k}| \leq 2^{-\alpha n}$ for some $0 < \alpha < 1$ then $f \in C^\alpha[0,1]$.*

Proof. Since the condition for Hölder continuity only depends on differences between a function we may assume that $c_0 = 0$. Pick $s, t \in [0,1]$ and use the triangle inequality to conclude

$$|f(s) - f(t)| \leq \sum_{n=0}^\infty \left| \sum_{k=0}^{2^n-1} c_{n,k} \left(\Delta_{n,k}(s) - \Delta_{n,k}(t) \right) \right|$$

To clean up our notation a bit we define

$$D_n(s,t) = \sum_{k=0}^{2^n-1} c_{n,k} \left(\Delta_{n,k}(s) - \Delta_{n,k}(t) \right)$$

for $n \geq 0$ and we work on getting a bound on $|D_n|$. Since we have a very concrete description of the $\Delta_{n,k}$ elementary (but detailed) tools can be used. Because the support of $\Delta_{n,k}$ for fixed n are disjoint $\Delta_{n,k}(s)$ is non-zero for at most one k and similarly with $\Delta_{n,k}(t)$. Let $0 \leq k_s < 2^n$ be an integer such that $k_s/2^n \leq s \leq (k_s+1)2^n$ and similarly with k_t (there is ambiguity in the choice for $s, t = k/2^n$ but it doesn't matter since the $\Delta_{n,k}$ all vanish at such points); with these choices, $D_n(s,t) = c_{n,k_s} \Delta_{n,k_s}(s) - c_{n,k_t} \Delta_{n,k_t}(t)$. Each function $\Delta_{n,k}$ is piecewise linear and comprises two line segments with slope $\pm 2^{n+1}$ and it is geometrically clear that $\Delta_{n,k_s}(s)$ and $\Delta_{n,k_t}(t)$ can be no farther than if they are on the same such line : hence $|\Delta_{n,k_s}(s) - \Delta_{n,k_t}(t)| \leq |s-t| \, 2^{n+1}$ and by the bounds we have on the coefficients $c_{n,k}$ we get

$$|D_n(s,t)| \leq (|c_{n,k_s}| \vee |c_{n,k_t}|) \, |\Delta_{n,k_s}(s) - \Delta_{n,k_t}(t)| \leq 2^{-\alpha n} \, |s-t| \, 2^{n+1}$$

This is a good bound when s, t are close (in fact it is a tight bound when $k_s = k_t$ and s, t are on the same line segment). However, as s, t get farther apart we can do better just by using the fact that $0 \leq \Delta_{n,k} \leq 1$. Indeed by the triangle inequality

$$|D_n(s,t)| = |c_{n,k_s}\Delta_{n,k_s}(s)| + |c_{n,k_t}\Delta_{n,k_t}(t)| \leq |c_{n,k_s}| + |c_{n,k_t}| \leq 2^{-\alpha n+1}$$

and therefore we have the two bounds

$$D_n(s,t) \leq 2^{-\alpha n}\,|s-t|\,2^{n+1} \wedge 2^{-\alpha n+1}$$

As mentioned, the first of these bounds is a better estimate when s, t are closer that 2^{-n} and the latter is better otherwise. So with s, t given pick $N \geq 0$ such that $2^{-N-1} \leq |s-t| < 2^{-N}$ and use the appropriate mix of the two estimates

$$\begin{aligned}
|f(s) - f(t)| &\leq \sum_{n=0}^{N} \left| \sum_{k=0}^{2^n-1} c_{n,k}\left(\Delta_{n,k}(s) - \Delta_{n,k}(t)\right)\right| + \sum_{n=N+1}^{\infty} \left| \sum_{k=0}^{2^n-1} c_{n,k}\left(\Delta_{n,k}(s) - \Delta_{n,k}(t)\right)\right| \\
&\leq \sum_{n=0}^{N} 2^{-\alpha n}\,|s-t|\,2^{n+1} + \sum_{n=N+1}^{\infty} 2^{-\alpha n+1} \\
&= 2\,|s-t|\,\frac{2^{(1-\alpha)(N+1)}-1}{2^{1-\alpha}-1} + 2\cdot 2^{-\alpha(N+1)}\cdot\frac{1}{1-2^{-\alpha}} \\
&\leq \frac{2}{2^{1-\alpha}-1}\,|s-t|^{\alpha} - \frac{2}{2^{1-\alpha}-1}\,|s-t| + \frac{2}{1-2^{-\alpha}}\,|s-t|^{\alpha} \\
&\leq \left(\frac{2}{2^{1-\alpha}-1} + \frac{2}{1-2^{-\alpha}}\right)|s-t|^{\alpha}
\end{aligned}$$

where we have used the assumption that $0 < \alpha < 1$ to determine the sign of coefficients in the estimates (e.g. to conclude that $\frac{2}{2^{1-\alpha}-1}\,|s-t| > 0$ so that this term may be dropped from the estimate). ☐

A corollary of this result and our construction of Brownian motion is the fact that Brownian paths are Hölder continuous with any exponent less that 1/2.

THEOREM 12.6 (Hölder Continuity of Brownian Paths). *Let B_t be a standard Brownian motion then almost surely B_t is Hölder continuous for any exponent $\alpha < 1/2$. Furthermore there exists a constant $C > 0$ (independent of ω) such that almost surely there exists a constant $\epsilon > 0$ (depending on ω) such that for all $0 \leq h \leq \epsilon$ and $0 \leq t \leq 1-h$ we have*

$$|B_{t+h} - B_t| \leq C\sqrt{h\log(1/h)}$$

PROOF. From our construction of Brownian motion recall that we had the representation

$$B_t = tZ_0 + \sum_{n=0}^{\infty} \frac{1}{\sqrt{2^{n+2}}} \sum_{k=0}^{2^n-1} \Delta_{n,k}(t) Z_{\frac{2k+1}{2^{n+1}}}$$

and moreover we have shown during the construction of Brownian motion for $c > \sqrt{2\ln 2}$ almost surely there is an $N > 0$ such that

$$\left| Z_{\frac{2k+1}{2^{n+1}}} \right| \leq c\sqrt{n+1}$$

for all $n \geq N$. Note that we can ignore the leading term tZ_0 since is clearly Hölder continuous, so to apply Lemma 12.5 it suffices to observe that we have coefficients $c_{n,k} = \frac{1}{\sqrt{2^{n+2}}} Z_{\frac{2k+1}{2^{n+1}}}$ with the bound

$$|c_{n,k}| \leq \frac{c\sqrt{n+1}}{\sqrt{2^{n+2}}} \leq 2^{-\alpha n}$$

for n sufficiently large. TODO: In the previous Lemma we need to rephrase things to note that it suffices to have the bound hold eventually.

TODO: Extend the estimates from the prior Lemma to yield the simple upper bound for modulus of continuity. Following the proof of the prior Lemma and using our estimate on the $c_{n,k}$ directly instead of the derived bound $|c_{n,k}| \leq 2^{-\alpha n}$ we get by picking $2^{-M-2} < |s-t| \leq 2^{-M-1}$ (so that $M+1 \leq \log_2(1/|s-t|)$)

$$|B_s - B_t| \leq \sum_{n=0}^{N-1} \max_{0 \leq k < 2^n} |c_{n,k}| \, |s-t| \, 2^{n+1} + \sum_{n=N}^{M} |s-t| \, 2^{n+1} \frac{c\sqrt{n+1}}{2^{(n+2)/2}} + 2 \sum_{n=M+1}^{\infty} \frac{c\sqrt{n+1}}{2^{(n+2)/2}}$$

For the first term, we use the fact that $\lim_{\epsilon \to 0^+} \epsilon / \sqrt{\epsilon \log(1/\epsilon)} = 0$ to find ϵ (depending on ω) sufficiently small so that provided $|s-t| \leq \epsilon$ we have

$$\sum_{n=0}^{N-1} \max_{0 \leq k < 2^n} |c_{n,k}| \, |s-t| \, 2^{n+1} \leq \sqrt{|s-t| \log(1/|s-t|)}$$

For the second term, by choice of M we get

$$\begin{aligned}
\sum_{n=N}^{M} |s-t| \, 2^{n+1} \frac{c\sqrt{n+1}}{2^{(n+2)/2}} &\leq c\,|s-t| \sum_{n=0}^{M} 2^{n/2} \sqrt{n+1} \\
&\leq c\,|s-t| \sqrt{M+1} \frac{2^{(M+1)/2} - 1}{\sqrt{2} - 1} \\
&\leq \frac{c}{\sqrt{2}-1} \sqrt{|s-t| \log_2(1/|s-t|)}
\end{aligned}$$

For the third term by choice of M we get

$$\begin{aligned}
2 \sum_{n=M+1}^{\infty} \frac{c\sqrt{n+1}}{2^{(n+2)/2}} &\leq \sqrt{M+1} \frac{c}{2^{(M+1)/2}} \sum_{n=0}^{\infty} \sqrt{\frac{n+M+1}{M+1}} \frac{1}{2^{n/2}} \\
&\leq \sqrt{M+1} \frac{c}{2^{(M+1)/2}} \sum_{n=0}^{\infty} \sqrt{n+1} \frac{1}{2^{n/2}} \\
&\leq C_2 \sqrt{|s-t| \log_2(1/|s-t|)}
\end{aligned}$$

where the constant C_2 depends only on the value of the convergent series and the choice of c. ∎

TODO: Levy's modulus of continuity Lemmas

Theorem 12.7. *Almost surely*

$$\limsup_{h \downarrow 0} \sup_{0 \leq t \leq 1-h} \frac{|B_{t+h} - B_t|}{\sqrt{2h \log(1/h)}} = 1$$

(TODO: Is this $\log_e$ *or* $\log_2$*?)*

PROOF. My notes on the proof from Peres and Morters Fix a $c > \sqrt{2}$ and pick $0 < \epsilon < 1/2$. For this ϵ, by Lemma ? we pick $m > 0$ such that for every $[s,t] \subset [0,1]$ we get $[s',t'] \in \Lambda(m)$ such that $|t - t'| < \epsilon(t-s)$ and $|s - s'| < \epsilon(t-s)$. Now by Lemma ? we choose $N > 0$ such that for all $n \geq N$, almost surely for every $[s',t'] \in \Lambda_n(m)$

$$|B_{t'} - B_{s'}| \leq c\sqrt{(t'-s')\log(1/(t'-s'))}$$

(we want this to be true for the approximating $[s',t']$: how do we know that $[s',t'] \in \Lambda_n(m)$ for sufficiently large n; I think it is true that $\Lambda_n(m) \subset \Lambda_{2n}(m)$? No I think we make the assumption that $t-s < 2^{-N}$). But we also have Theorem 12.6 (TODO: Does this work; this result gives the bound for h smaller than a *random* constant but here it seems we are assuming that it is not random) so we can estimate

$$\begin{aligned}|B_t - B_s| &\leq |B_t - B_{t'}| + |B_{t'} - B_{s'}| + |B_{s'} - B_s| \\ &\leq C\sqrt{|t-t'|\log(1/|t-t'|)} + c\sqrt{(t'-s')\log(1/(t'-s'))} + C\sqrt{|s-s'|\log(1/|s-s'|)}\end{aligned}$$

The function $x\log(1/x)$ is increasing for $0 \leq x \leq 1/2$ (here we are using $\log_2$; otherwise $1/e$) therefore if we assume $t - s < \epsilon$ then $|t-t'| < \epsilon(t-s) < 1/4$ so we get the estimate

$$\begin{aligned}C\sqrt{|t-t'|\log(1/|t-t'|)} &\leq C\sqrt{\epsilon(t-s)\log(1/\epsilon(t-s))} \\ &\leq C\sqrt{\epsilon(t-s)\log(1/(t-s)^2)} \\ &= \sqrt{2\epsilon}C\sqrt{(t-s)\log(1/(t-s))}\end{aligned}$$

and similarly with the term involving $|s-s'|$. As for the middle term, we have by choice of $[s',t']$ that $(1-2\epsilon)(t-s) \leq (t'-s') \leq (1+2\epsilon)(t-s)$ and by assumption $\log(1/(t-s)) > 1$ therefore

$$\begin{aligned}c\sqrt{(t'-s')\log\frac{1}{t'-s'}} &\leq c\sqrt{(1+2\epsilon)(t-s)\log\frac{1}{(1-2\epsilon)(t-s)}} \\ &= c\sqrt{(1+2\epsilon)(t-s)(\log\frac{1}{(1-2\epsilon)} + \log\frac{1}{(t-s)})} \\ &\leq c\sqrt{(1+2\epsilon)(t-s)\log\frac{1}{(t-s)}(1-\log(1-2\epsilon))}\end{aligned}$$

Now since $\epsilon > 0$ was arbitrary, we can put all three estimates together conclude for any $0 < h < \epsilon$,

$$\sup_{0\leq t\leq 1-h} |B_{t+h} - B_t| \leq \left(2\sqrt{2\epsilon}C + c\sqrt{(1+2\epsilon)(1-\log(1-2\epsilon))}\right)\sqrt{h\log(1/h)}$$

and thus

$$\limsup_{h\downarrow 0}\sup_{0\leq t\leq 1-h}\frac{|B_{t+h}-B_t|}{\sqrt{h\log(1/h)}} \leq 2\sqrt{2\epsilon}C + c\sqrt{(1+2\epsilon)(1-\log(1-2\epsilon))}$$

Now since $0 < \epsilon < 1/2$ was arbitrary and $c > \sqrt{2}$ was arbitrary we can let $\epsilon \downarrow 0$ and then $c \downarrow \sqrt{2}$ to conclude the result. $\square$

The approach above to studying the sample path properties of Brownian motion is based on examing the (random) coefficients of the expression of the Brownian motion in the Schauder basis. This has advantages and disadvantages. The obvious advantage is a certain concreteness that is appealing. The disadvantage is that the analysis is less general than it can be. Here we provide a classical alternative to the construction of Brownian motion and the analysis of sample paths that relies on tools that are more general. It is critical to have these more general tools at hand when discussing larger classes of stochastic process.

THEOREM 12.8 (Kolmogorov-Centsov). *Let X_t be a stochastic process on $[0,T]^d$ with values in a complete metric space (S,d) and suppose that there exist constant C, α, β such that*

$$\boldsymbol{E}[d(X_s, X_t)^\alpha] \leq |s-t|^{d+\beta} \text{ for all } s,t \in \mathbb{R}^d$$

then X_t has a continuous modification $\tilde{X}_t$ and furthermore the paths of $\tilde{X}_t$ are almost surely Hölder continuous with exponent γ for every $0 < \gamma < \beta/\alpha$.

PROOF. We do the proof with $T = 1$ and $d = 1$.

The basic idea of the proof is that via Markov bounding, the moment condition controls the variations of X_t pointwise; furthermore by careful selection of constants we can extend this to uniform continuity of X_t on a countable subset of $[0,T]^d$. By chosing a countable dense subset of $[0,T]^d$ we will then be in position to create the modification.

For each $n \geq 0$, let $\mathcal{D}_n = \{k/2^n \mid 0 \leq k \leq 2^n\}$ be the dyadic rationals with scale n and consider the behavior of X_t on the grid $\mathcal{D}_n^d \subset [0,1]^d$. To begin bound the variation on adjacent points in the grid using a union bound and a Markov bound (TODO: Fix up the sum below for the case $d > 1$)

$$\begin{aligned}\mathbf{P}\{\max_{0<k\leq 2^n} d(X_{k/2^n}, X_{(k-1)/2^n}) \geq \epsilon\} &\leq \sum_{k=1}^{2^n} \mathbf{P}\{d(X_{k/2^n}, X_{(k-1)/2^n}) \geq \epsilon\} \\ &\leq \sum_{k=1}^{2^n} 2^{-n(d+\beta)}/\epsilon^\alpha = 2^{-n\beta}\epsilon^{-\alpha}\end{aligned}$$

So if we pick $0 < \gamma < \beta/\alpha$ and $\epsilon = 2^{-n\gamma}$ then we have the bound

$$\sum_{n=0}^\infty \mathbf{P}\{\max_{0<k\leq 2^n} d(X_{k/2^n}, X_{(k-1)/2^n}) \geq 2^{-n\gamma}\} \leq \sum_{n=1}^\infty 2^{-n(\beta-\gamma\alpha)} < \infty$$

and Borel Cantelli tells us that there is an event $A \subset \Omega$ with $\mathbf{P}\{A\} = 1$ and for each $\omega \in A$ there exists an $N(\omega)$ such that

$$d(X_{k/2^n}(\omega), X_{(k-1)/2^n}(\omega)) < 2^{-n\gamma} \text{ for all } n \geq N(\omega) \text{ and } 0 < k \leq 2^n$$

We have gained some control on the behavior of X_t on a sequence of successively finer dyadic grids but what we need is to translate this into control of X_t simultaneously over the union of all grids (to see what we are lacking at this point realise that we have an almost sure bound on a term like $d(X_{k/2^n}, X_{(k-1)/2^n})$ with $k/2^n - (k-1)/2^n = 1/2^n$ but we don't yet have a bound on a term like $d(X_{(2k+1)/2^n}, X_{(2k-1)/2^n})$ with $(2k+1)/2^{n+1} - (2k-1)/2^{n+1} = 1/2^n$).

Claim: For every $n \geq N(\omega)$ and every $m > n$ we have

$$d(X_t(\omega), X_s(\omega)) \leq 2\sum_{k=n+1}^m 2^{-k\gamma} \text{ for } s,t \in \mathcal{D}_m \text{ with } 0 < |s-t| < 2^{-n}$$

The proof of the claim is by induction. For $m = n+1$ the only way for $0 < |s-t| < 2^{-n}$ when $s, t \in \mathcal{D}_{n+1}$ is when $s = (k-1)/2^{n+1}$ and $t = k/2^{n+1}$ and therefore by what we have already shown $d(X_t(\omega), X_s(\omega)) \leq 2^{-(n+1)\gamma}$ so the result holds in this case. Now assume that the result holds for all $n+1, \ldots, m$ and we show it for $m+1$. Assume without loss of generality that $s < t$ define $s^* = \lceil 2^m s \rceil / 2^m$ and $t^* = \lfloor 2^m t \rfloor / 2^m$ (that is to say round s up to nearest point on the grid $\mathcal{D}_m$ and round t down to the nearest point on the grid $\mathcal{D}_m$). Then the following are easily seen to be true

(i) $s^*, t^* \in \mathcal{D}_m$
(ii) $s \leq s^* \leq t^* \leq t$
(iii) $0 \leq s^* - s \leq 1/2^{m+1}$
(iv) $0 \leq t - t^* \leq 1/2^{m+1}$
(v) $0 \leq t^* - s^* < 1/2^n$

Now by the triangle inequality, the induction hypothesis and the result for adjacent points in the grid $\mathcal{D}_{m+1}$ we get

$$\begin{aligned} d(X_t, X_s) &\leq d(X_t, X_{t^*}) + d(X_{t^*}, X_{s^*}) + d(X_{s^*}, X_s) \\ &\leq 2^{-(m+1)\gamma} + 2 \sum_{k=n+1}^{m} 2^{-k\gamma} + 2^{-(m+1)\gamma} = 2 \sum_{k=n+1}^{m+1} 2^{-k\gamma} \end{aligned}$$

and we are done with the claim.

The claim establishes the local Hölder continuity of $X_t(\omega)$ on $\mathcal{D} = \cup_{n=1}^{\infty} \mathcal{D}_n$ (hence uniform continuity). To see this, pick $s, t \in \mathcal{D}$ such that $|s-t| < 2^{-N(\omega)}$ and find $n > N(\omega)$ such that $2^{-(n+1)} \leq |s-t| < 2^{-n}$, then $s, t \in \mathcal{D}_m$ for all m large enough and so

$$d(X_t(\omega), X_s(\omega)) \leq 2 \sum_{k=n+1}^{m} 2^{-k\gamma} \leq 2^{-(n+1)\gamma} \frac{2}{1-2^{-\gamma}} \leq |s-t|^{\gamma} \frac{2}{1-2^{-\gamma}}$$

Since X_t is almost surely Hölder continuous on $\mathcal{D}^d$ which is a dense subset of $[0,1]^d$ we know that X_t has a unique extension $\tilde{X}_t$ to a continuous function on $[0,1]^d$ and that the extension is Hölder continuous with the same exponent and constant. Define $\tilde{X}_t = 0$ for $\omega \notin A$.

It remains to show that $\tilde{X}_t$ defined in this way is a modification of X_t. Assume $\epsilon > 0$ and apply a Markov bound

$$\mathbf{P}\{d(X_t, X_s) > \epsilon\} \leq \frac{\mathbf{E}\left[d(X_t, X_s)^{\alpha}\right]}{\epsilon^{\alpha}} \leq \frac{|s-t|^{d+\beta}}{\epsilon^{\alpha}}$$

which shows that for every $s \in [0,1]^d$ we have $X_t \xrightarrow{P} X_s$ as $t \to s$.

TODO: Finish the argument that this is a modification. □

The flip side of the positive results showing that Brownian paths are Hölder continuous is the following result showing that a sea change occurs at $\alpha = 1/2$. As we'll note, in particular this shows that Brownian paths are almost surely nowhere differentiable.

THEOREM 12.9. *For every $\alpha > 1/2$ almost surely a Brownian path has no point that is locally Hölder continuous with exponent α.*

PROOF. Pick an $\alpha > 1/2$, $C > 0$, $\epsilon > 0$ and define

$$G(\alpha, C, \epsilon) = \{\omega \mid \text{ there exists } s \in [0,1] \text{ such that } |B_t(\omega) - B_s(\omega)| < C\,|t-s| \text{ for every } t \in [0,1] \text{ with } |t-s| <$$

The set $G(\alpha, C, \epsilon)$ is not necessarily measurable so it doesn't make sense to show that it has measure zero; however we will show that it is contained in a set of a measure zero. The trick to doing this is the observation that the α-Hölder continuity of $B_s(\omega)$ from the definition of $G(\alpha, C, \epsilon)$ implies an arbitrarily large number of independent increments to be small. By the Gaussian nature of the increments and a very crude tail probability estimate we'll be able to conclude that the probability of the increments all being small can be sent to zero. At the risk of being pedantic, note that while the positive results on Hölder continuity relied on bounds showing it is unlikely that a collection of independent Gaussians will simulaneously be large, this result requires a bound showing it is unlikely that a collection of independent Gaussians will simultaneously be small.

To make this precise, pick an $\omega \in G(\alpha, C, \epsilon)$ and let $s \in [0,1]$ be an appropriate Hölder continuous point. Now define $U = [0,1] \cap (s-\epsilon, s+\epsilon)$ so that the diameter is at least ϵ. Now for any $m > 0$ there is an $N_{m,\epsilon}$ (roughly speaking $N_{m,\epsilon} = 2m/\epsilon$) such that for all $n \geq N_{m,\epsilon}$ there exists a k with $0 \leq k < n-m$ such that for all $0 \leq i < m$, $[\frac{k+i}{n}, \frac{k+i+1}{n}] \subset U$ and either $s \in [\frac{k}{n}, \frac{k+1}{n}]$ or $s \in [\frac{k+m-1}{n}, \frac{k+m}{n}]$ (we only need the last option when $s = 1$). Now using the fact that the diameter of U is less that ϵ, the triangle inequality and the Hölder continuity at s we see for every $0 \leq i < m$,

$$\left|B_{\frac{k+i+1}{n}}(\omega) - B_{\frac{k+i}{n}}(\omega)\right| \leq \left|B_{\frac{k+i+1}{n}}(\omega) - B_s(\omega)\right| + \left|B_s(\omega) - B_{\frac{k+i}{n}}(\omega)\right| \leq 2C\left(\frac{m}{n}\right)^{\alpha}$$

From this argument we conclude that for every $m > 0$ and every $n \geq N_{m,\epsilon}$

$$G(\alpha, C, \epsilon) \subset \cup_{k=0}^{n-m-1} \cap_{i=0}^{m-1} \{\omega \mid \left|B_{\frac{k+i+1}{n}}(\omega) - B_{\frac{k+i}{n}}(\omega)\right| \leq 2C\left(\frac{m}{n}\right)^{\alpha}\}$$

We know that each increment $B_{\frac{k+i+1}{n}}(\omega) - B_{\frac{k+i}{n}}(\omega)$ is Gaussian with variance $1/n$. Thus we can apply the simple bound for a $N(0,1)$ random variable Z,

$$\mathbf{P}\{|Z| \leq \lambda\} = \frac{1}{\sqrt{2\pi}} \int_{-\lambda}^{\lambda} e^{-x^2/2}\,dx \leq \frac{1}{\sqrt{2\pi}} \int_{-\lambda}^{\lambda} dx = \frac{2\lambda}{\sqrt{2\pi}}$$

to conclude

$$\mathbf{P}\{\left|B_{\frac{k+i+1}{n}}(\omega) - B_{\frac{k+i}{n}}(\omega)\right| \leq 2C\left(\frac{m}{n}\right)^{\alpha}\} \leq \frac{4C\sqrt{n}}{\sqrt{2\pi}}\left(\frac{m}{n}\right)^{\alpha}$$

By a union bound and the independence of Brownian increments we know that

$$\begin{aligned}
&\mathbf{P}\{\cup_{k=0}^{n-m-1} \cap_{i=0}^{m-1} \{\omega \mid \left|B_{\frac{k+i+1}{n}}(\omega) - B_{\frac{k+i}{n}}(\omega)\right| \leq 2C\left(\frac{m}{n}\right)^{\alpha}\}\} \\
&\leq n\left(\frac{4C\sqrt{n}}{\sqrt{2\pi}}\left(\frac{m}{n}\right)^{\alpha}\right)^m = \left(\frac{4Cm^{\alpha}}{\sqrt{2\pi}}\right)^m n^{1+(\frac{1}{2}-\alpha)m}
\end{aligned}$$

The important point is if we choose any value of $m > \frac{1}{\alpha - 1/2}$ (possible since $\alpha > 1/2$) then the exponent $1 + (\frac{1}{2} - \alpha)m < 0$ and taking the limit as $n \to \infty$ we see that $G(\alpha, C, \epsilon)$ is contained in a set of measure zero.

The proof of the Theorem is completed by taking the countable union over all rational C and rational ϵ and noting that this is also contained in a set of measure zero. $\square$

COROLLARY 12.10 (Nondifferentiability of Brownian Motion). *Almost sure a Brownian path is nowhere differentiable.*

PROOF. Take $\alpha = 1$ in the Theorem 12.9 $\square$

THEOREM 12.11 (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 4.17 we know that it is sufficient to show independence for finite sets $\{B_{t_1+s} - B_s, \dots, B_{t_n+s} - B_s\}$ and $\{B_{s_1}, \dots, B_{s_m}\}$ 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 $\xi_1, \dots, \xi_n$ we have $\sigma(\xi_1, \xi_2 - \xi_1, \dots, \xi_n - \xi_1) = \sigma(\xi_1, \xi_2 - \xi_1, \dots, \xi_n - \xi_{n-1})$ (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 4.14 we know that $\sigma(B_{t_1+s} - B_s, \dots, B_{t_n+s} - B_{t_{n-1}})$ and $\sigma(B_{s_1} - B_0, \dots, B_{s_m} - B_{s_{m-1}})$ are independent which establishes the result by applying the previous observation. $\square$

1. Skorohod Embedding and Donsker's Theorem

TODO: Clarify what we mean when we say a Brownian motion is independent of a σ-algebra. ANSWER: Independence of a Brownian motion and σ-algebra is interpreted by thinking of the Brownian motion as a stochastic process. Because the σ-algebra on $\mathbb{R}^{\mathbb{R}_+}$ (or $\mathbb{R}^{[0,1]}$) is generated by projections/evaluation maps $\pi_t(f) = f(t)$ we can check independence by checking independence on finite dimensional projections $\{(B_{t_1}, \dots, B_{t_n}) \in A\}$ by monotone classes. Up till this point we have been treating independence in a slightly different (though I expect equivalent) way of saying that the σ-algebra $\sigma(B_t)$ is the basis of independence.

TODO: Introduce the right continuous filtration $\mathcal{F}_t^+$

TODO: Extend the Markov property of Brownian motion to the filtration $\mathcal{F}_t^+$.

TODO: Define $\mathcal{F}$-Brownian motion

THEOREM 12.12 (Markov Property). *Let B_t be a Brownian motion then for any $s \geq 0$ the process $\tilde{B}_t = B_{t+s} - B_s$ 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 $\tilde{B}_0 = B_s - B_s = 0$.

(ii) For any $0 \leq t_1 \leq \cdots \leq t_n$ the increment $\tilde{B}_{t_j} - \tilde{B}_{t_{j-1}} = \tilde{B}_{s+t_j} - \tilde{B}_{s+t_{j-1}}$ therefore the independence of the increments $\tilde{B}_{t_2} - \tilde{B}_{t_1}, \dots, \tilde{B}_{t_n} - \tilde{B}_{t_{n-1}}$ 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 $\tilde{B}_{t_2} - \tilde{B}_{t_1} = B_{s+t_2} - B_{s+t_1}$ is normallly distributed with mean 0 and variance $(s+t_2) - (s+t_1) = t_2 - t_1$.

(iv) The paths $\tilde{B}_t = B_{s+t}$ 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 □

DEFINITION 12.13. A process X_t on a time scale T is called a *Gaussian process* if $c_1 X_{t_1} + \cdots + c_n X_{t_n}$ is a Gaussian random variable for all $n \in \mathbb{N}$, $(t_1, \ldots, t_n) \in T^n$ and all $(c_1, \ldots, c_n) \in \mathbb{R}^n$. The process is said to be a *centered Gaussian process* if in addition $\mathbf{E}[X_t] = 0$ for all $t \in T$.

Just as the distribution of a Gaussian random variable or vector is characterised by its first two moments, so to with a Gaussian process.

LEMMA 12.14. *Let X_t be a Gaussian process on a time scale T, then the distribution of X is determined by the values $\boldsymbol{E}[X_t]$ for all $t \in T$ and $\boldsymbol{E}[X_s X_t]$ for all $s, t \in T$.*

PROOF. Suppose we have Gaussian processes X and Y with same first two moments in the sense of the hypothesis of the lemma. If we pick $n \in naturals$, $(t_1, \ldots, t_n) \in T^n$ and $(c_1, \ldots, c_n) \in \mathbb{R}^n$ then each of $c_1 X_{t_1} + \cdots + c_n X_{t_n}$ and $c_1 Y_{t_1} + \cdots + c_n Y_{t_n}$ is Gaussian and has the same mean and variance and therefore are equal in distribution. By the Cramer-Wold Device (Corollary 7.15) this tells us that $(X_{t_1}, \ldots, X_{t_n}) \overset{d}{=} (Y_{t_1}, \ldots, Y_{t_n})$ and therefore $X \overset{d}{=} Y$ by Lemma 9.6. □

LEMMA 12.15 (Brownian Time Inversion). *Let B_t be a Brownian motion starting at $x \in \mathbb{R}$ and define*

$$X_t = \begin{cases} 0 & \text{if } t = 0 \\ t B_{1/t} & \text{if } t > 0 \end{cases}$$

then X_t is also a standard Brownian motion.

PROOF. Clearly, X_t is a centered Gaussian process (remember constants are Gaussian with variance 0) and therefore its distribution is determined by the covariance function by Lemma 12.14. It is straightforward to see that

$$\mathbf{E}[X_s X_t] = st\mathbf{E}\left[B_{1/s} B_{1/t}\right] = st(1/s \wedge 1/t) = s \wedge t$$

and therefore X has the distribution of a Brownian motion. It remains to show that X has continuous sample paths almost surely. In fact we know that the sample paths of X are almost surely continuous on $(0, \infty)$ since those of B are; we only need to show almost sure continuity at 0. First note that when restricted to $\mathbb{Q}$ we have the event

$$A = \{f : [0, \infty) \to \mathbb{R} \mid \lim_{\substack{q \downarrow 0 \\ q \in \mathbb{Q}}} f(q) = 0\} = \cap_{n=1}^{\infty} \cup_{\substack{q > 0 \\ q \in \mathbb{Q}}} \cap_{\substack{0 < p < q \\ p \in \mathbb{Q}}} \{-1/n < f(p) < 1/n\}$$

is measurable and as B and X have the same distribution and B is almost surely continuous at 0 we know that $\mathbf{P}\{B \in A\} = \mathbf{P}\{X \in A\} = 1$; that is to say, $\lim_{\substack{q \downarrow 0 \\ q \in \mathbb{Q}}} X_q = 0$ a.s. On the other hand, as X is almost surely continuous on $(0, \infty)$ we know that $\lim_{t \downarrow 0} X_t = \lim_{\substack{q \downarrow 0 \\ q \in \mathbb{Q}}} X_q$ almost surely so $\lim_{t \downarrow 0} X_t = 0$ a.s. □

The Markov property of Brownian motion can be extended to a slightly stronger statement. TODO: Should we remove this statement as it is subsumed by the Strong Markov property proved next.

THEOREM 12.16. *Let B_t be a Brownian motion then for any $s \geq 0$ the process $\tilde{B}_t = B_{t+s} - B_s$ is a standard Brownian motion independent of $\mathcal{F}_s^+$.*

PROOF. Suppose $s \geq 0$ is chosen. We have already shown that $\tilde{B}_t$ is a standard Brownian motion independent of $\mathcal{F}_s^0$; we only need to extend the independence statement to the larger filtration $\mathcal{F}_s^+ = \cap_{t>s}\mathcal{F}_t^0$. Let s_n be a sequence of real numbers such that $s_n \downarrow s$ and define for each $n \in \mathbb{N}$ the process $B_t^n = B_{t+s_n} - B_{s_n}$ which by the Markov Property Theorem 12.12 is a Brownian motion independent of $\mathcal{F}_{s_n}^0 \supset \mathcal{F}_s^+$. By almost sure continuity of B_t we know that almost surely $\lim_{n\to\infty} B_t^n = \tilde{B}_t$ and therefore $\tilde{B}_t$ is also independent of $\mathcal{F}_s^+$. TODO: We make this last argument several times in this section (see proof of the Strong Markov Property below) so we should factor it out for easy reference. $\square$

DEFINITION 12.17. Let B_t be a Brownian motion starting at x then the σ-algebra $\mathcal{F}_0^+ = \cap_{t>0} \vee_{0\leq s\leq t} \sigma(B_s)$ is called the *germ σ-algebra* and the σ-algebra $\mathcal{T} = \cap_{t>0} \vee_{s\geq t} \sigma(B_s)$ is called the *tail σ-algebra*.

LEMMA 12.18 (Blumenthal 0-1 Law). *Let B_t be a Brownian motion then the germ σ-algebra $\mathcal{F}_0^+$ and the tail σ-algebra $\mathcal{T}$ are both trivial.*

PROOF. By Theorem 12.16 we know that for any $t > 0$, B is independent of $\mathcal{F}_0^+$. However we know that $\mathcal{F}_0^+ \subset \sigma(B)$ it follows that $\mathcal{F}_0^+$ is independent of itself and therefore for any $A \in \mathcal{F}_0^+$ we know that $\mathbf{P}\{()A\} = \mathbf{P}\{()A \cap A\} = \mathbf{P}\{()A\}^2$ which implies $\mathbf{P}\{()A\} \in \{0, 1\}$ which shows that the germ σ-algebra is trivial.

Triviality of the tail σ-algebra follows by noting that

$$\mathcal{T} = \cap_{t>0} \vee_{s\geq t} \sigma(B_s) = \cap_{1/t>0} \vee_{1/s\geq 1/t} \sigma(B_{1/s}) = \cap_{t>0} \vee_{s\leq t} \sigma(B_{1/s})$$

so we see that the tail σ-algebra for the Brownian motion B coincides with the germ σ-algebra for the time inverted Brownian motion $tB_{1/t}$. As $tB_{1/t}$ is a Brownian motion starting at 0 we see that that the tail $\mathcal{T}$ is trivial. $\square$

THEOREM 12.19 (Strong Markov Property). *Let B_t be a Brownian motion and let τ be an almost surely finite $\mathcal{F}^+$-optional time, then $\tilde{B}_t = B_{\tau+t} - B_\tau$ is a standard Brownian motion independent of $\mathcal{F}_\tau^+$.*

PROOF. First suppose that τ has a countable range $S \subset \mathbb{R}_+$ and for each $s \in S$ define $B_t^s = B_{s+t} - B_s$. Let A be a measurable subset of $\mathbb{R}^{[0,\infty)}$ and let $E \in \mathcal{F}_\tau^+$. Now using the fact that τ is $\mathcal{F}^+$-optional and the Markov property of B_t^s (Theorem 12.16) we get

$$\begin{aligned}
\mathbf{P}\{\{\tilde{B}_t \in A\} \cap E\} &= \sum_{s\in S} \mathbf{P}\{\{B_t^s \in A\} \cap E \cap \{\tau = s\}\} \\
&= \sum_{s\in S} \mathbf{P}\{B_t^s \in A\}\mathbf{P}\{E \cap \{\tau = s\}\} \\
&= \mathbf{P}\{B_t \in A\} \sum_{s\in S} \mathbf{P}\{E \cap \{\tau = s\}\} \\
&= \mathbf{P}\{B_t \in A\}\mathbf{P}\{E\}
\end{aligned}$$

which shows that the process $\tilde{B}_t$ is independent of $\mathcal{F}^+$. It is clear that $\tilde{B}_t$ is almost surely continuous and $\tilde{B}_0 = 0$; furthermore taking $E = \Omega$ and $A = (\pi_{t_1}, \dots, \pi_{t_d})^{-1}(C)$ for some $C \in \mathcal{B}(\mathbb{R}^d)$ in the above calculation and we see that $\tilde{B}_t$ has independent Gaussian increments, thus $\tilde{B}_t$ is a standard Brownian motion. TODO: Is there anything that needs to be done to show that $\tilde{B}_t$ is measurable?

It remains to extend the result to arbitrary $\mathcal{F}^+$-optional times. It is clear that $\tilde{B}_t$ is almost surely continuous and $\tilde{B}_0 = 0$. Given τ an $\mathcal{F}^+$-optional time let $\tau_n = \frac{1}{2^n}\lfloor 2^n\tau + 1 \rfloor$ so that $\tau_n \downarrow \tau$ (Lemma 9.71) and by definition each τ_n is $\mathcal{F}^+$-optional. For each $n \in \mathbb{N}$ define $B^n_t = B_{\tau_n + t} - B_{\tau_n}$ and apply the result for countably valued optional times to conclude B^n_t is a standard Brownian motion independent of $\mathcal{F}^+_{\tau_n} \supset \mathcal{F}^+_\tau$. Since B_t is almost surely continuous we know that almost surely $\tilde{B}_t = \lim_{n\to\infty} B^n_t$. Now by Dominated Convergence and the fact that B^n_t is independent of $\mathcal{F}^+_\tau$ for all n we get for all $E \in \mathcal{F}^+_\tau$, all bounded continuous functions $f : \mathbb{R}^d \to \mathbb{R}$ and $0 \leq t_1 \leq \dots \leq t_d < \infty$,

$$\begin{aligned}\mathbf{E}\left[f(\tilde{B}_{t_1}, \dots, \tilde{B}_{t_d}); E\right] &= \lim_{n\to\infty} \mathbf{E}\left[f(B^n_{t_1}, \dots, B^n_{t_d}); E\right] \\ &= \lim_{n\to\infty} \mathbf{E}\left[f(B^n_{t_1}, \dots, B^n_{t_d})\right] \mathbf{P}\{E\} = \mathbf{E}\left[f(\tilde{B}_{t_1}, \dots, \tilde{B}_{t_d})\right] \mathbf{P}\{E\}\end{aligned}$$

Given an arbitrary open set $U \subset \mathbb{R}^d$ we define $f_n(x) = nd(x, U^c) \wedge 1$ so that f_n is bounded and continuous and $f_n \downarrow \mathbf{1}_U$ so that by Monotone Convergence we get

$$\mathbf{P}\{\{(\tilde{B}_{t_1}, \dots, \tilde{B}_{t_d}) \in U\} \cap E\} = \mathbf{P}\{(\tilde{B}_{t_1}, \dots, \tilde{B}_{t_d}) \in U\}\mathbf{P}\{E\}$$

and because sets of the form $\{(\tilde{B}_{t_1}, \dots, \tilde{B}_{t_d}) \in U\}$ are a π-system generating $\tilde{B}_t \in E$, we get

$$\mathbf{P}\{\{\tilde{B}_t \in A\} \cap E\} = \mathbf{P}\{\tilde{B}_t \in A\}\mathbf{P}\{E\}$$

Now arguing exactly as in the countable case this shows that $\tilde{B}_t$ is a standard Brownian motion independent of $\mathcal{F}^+_\tau$. ☐

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 12.20 (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. First assume that τ is almost surely finite. Define $B^\tau_t = B_{\tau\wedge t}$ and $\tilde{B}_t = B_{\tau+t} - B_\tau$. Because B_t is continuous we know that B_t is progressively measurable (Lemma 9.89) and therefore B^τ is $\mathcal{F}_\tau$-measurable (Lemma 9.90). By the Strong Markov Property (Theorem 12.19) we know that $\tilde{B}$ is a standard Brownian motion independent of $\mathcal{F}^+_\tau$ hence independent of τ and B^τ; the same is true of $-\tilde{B}$.

Combining independence and the equality of the marginal distributions we know $(\tau, B^\tau, B) \overset{d}{=} (\tau, B^\tau, -\tilde{B})$ (Lemma 4.5). Now define $G : \mathbb{R} \times \mathbb{R}^{[0,\infty)} \times \mathbb{R}^{[0,\infty)} \to \mathbb{R}^{[0,\infty)}$ by $G(t, f, g)(s) = f(s) + g((s-t)_+)$ and note that $B_t = G(\tau, B^\tau, \tilde{B})$ and $B_t^\prime = G(\tau, B^\tau, -\tilde{B})$ so the result follows once we verify that G is measurable.

Unfortunately in the generality we've defined it, G is not measurable. However all we really need is the fact that restriction of G to $C([0,\infty), \mathbb{R})$ is measurable. Here we peek ahead to use the fact that the σ-algebra induced on $C([0,\infty), \mathbb{R})$ from the product σ-algebra is the Borel σ-algebra coresponding to the topology of uniform convergence on compact sets (see Lemma 15.29). It is easy to see that G is continuous hence measurable.

TODO: Can we avoid appealing to the continuity argument and see the measurability directly? □

LEMMA 12.21. *Let B_t be a standard Brownian motion, $0 \leq x < b$ and $\tau = \inf\{t \mid B_t \geq b\}$. Then*

$$\boldsymbol{P}\{\sup_{0 \leq s \leq t} B_t \geq b; B_t < x\} = \boldsymbol{P}\{B_t > 2b - x\}$$

PROOF. A general fact seems to be that many of the consequences of the Strong Markov Property can be shown without making a direct appeal to the Strong Markov Property. Here is a proof of the reflection principle that doesn't use the Strong Markov Property directly but instead replays key parts of the proof Strong Markov Property for Brownian motion.

Define $\tau_n = \frac{1}{2^n}\lfloor 2^n \tau + 1 \rfloor$ so that $\tau_n \downarrow \tau$ (Lemma 9.71). First consider

$$\begin{aligned}
\mathbf{P}\{\tau_n \leq t; B_t - B_{\tau_n} < x - b\} &= \sum_{k=0}^{\lfloor 2^n t \rfloor} \mathbf{P}\{\tau_n = k/2^n; B_t - B_{k/2^n} < x - b\} \\
&= \sum_{k=0}^{\lfloor 2^n t \rfloor} \mathbf{P}\{\tau_n = k/2^n\}\mathbf{P}\{B_t - B_{k/2^n} < x - b\} \\
&= \sum_{k=0}^{\lfloor 2^n t \rfloor} \mathbf{P}\{\tau_n = k/2^n\}\mathbf{P}\{B_t - B_{k/2^n} > b - x\} \\
&= \mathbf{P}\{\tau_n \leq t; B_t - B_{\tau_n} > b - x\}
\end{aligned}$$

where we have used the fact that $B_t - B_{k/2^n}$ is independent of $\mathcal{F}_{k/2^n}$ and $\tau_n = k/2^n \in \mathcal{F}_{k/2^n}$ since τ_n is an optional time and the fact that a Gaussian distribution is symmetric about 0.

Now because B_t is almost surely continuous we have that $B_\tau = b$ almost surely and therefore $\{\tau = t\} \subset \{B_t = b\}$ and because B_t is a Gaussian random variable we know that $\mathbf{P}\{\tau = t\} = \mathbf{P}\{B_t = b\} = 0$. Similarly, because the increment $B_t - B_\tau$ is Gaussian we know that $\mathbf{P}\{B_t - B_\tau = b - x\} = \mathbf{P}\{B_t - B_\tau = x - b\} = 0$. Therefore both $(-\infty, t] \times (b - x, \infty)$ and $(-\infty, t] \times (-\infty, x - b)$ are $\mathcal{L}(\tau, B_t - B_\tau)$-continuity sets and by the Portmanteau Theorem 5.43 we get

$$\begin{aligned}\mathbf{P}\{\tau \leq t; B_t - B_\tau < x - b\} &= \lim_{n\to\infty} \mathbf{P}\{\tau_n \leq t; B_t - B_{\tau_n} < x - b\} \\ &= \lim_{n\to\infty} \mathbf{P}\{\tau_n \leq t; B_t - B_{\tau_n} > b - x\} \\ &= \mathbf{P}\{\tau \leq t; B_t - B_\tau > b - x\}\end{aligned}$$

Using the fact that $B_\tau = b$ we can rewrite the equality as

$$\mathbf{P}\{\tau \leq t; B_t < x\} = \mathbf{P}\{\tau \leq t; B_t > 2b - x\}$$

and by the continuity of B_t we know that $\{\tau \leq t\} = \{\sup_{0\leq s\leq t} B_s \geq b\}$ and $\{\sup_{0\leq s\leq t} B_s \geq b\} \subset \{B_t > 2b - x\}$ and therefore we get

$$\mathbf{P}\{\sup_{0\leq s\leq t} B_s \geq b; B_t < x\} = \mathbf{P}\{B_t > 2b - x\}$$

□

LEMMA 12.22. *Let $M_t = \sup_{0\leq s\leq t} B_s$ be the maximal process associated with a standard Brownian motion then $M_t \overset{d}{=} |B_t|$.*

PROOF. Suppose $x > 0$ then we can calculate using continuity of measure, the Reflection Principle and the fact that $\mathbf{P}\{B_t = x\} = 0$,

$$\begin{aligned}\mathbf{P}\{\sup_{0\leq s\leq t} B_s \geq x\} &= \mathbf{P}\{\sup_{0\leq s\leq t} B_s \geq x, B_t < x\} + \mathbf{P}\{\sup_{0\leq s\leq t} B_s \geq x, B_t \geq x\} \\ &= \lim_{n\to\infty} \mathbf{P}\{\sup_{0\leq s\leq t} B_s \geq x, B_t < x - 1/n\} + \mathbf{P}\{B_t \geq x\} \\ &= \lim_{n\to\infty} \mathbf{P}\{B_t > x + 1/n\} + \mathbf{P}\{B_t \geq x\} \\ &= \mathbf{P}\{B_t > x\} + \mathbf{P}\{B_t \geq x\} = \mathbf{P}\{|B_t| \geq x\}\end{aligned}$$

From this it follows that $\mathbf{P}\{\sup_{0\leq s\leq t} B_s \geq 0\} = 1$ so in addition we have $\mathbf{P}\{\sup_{0\leq s\leq t} B_s \leq x\} = 0$ for all $x \leq 0$ and the result is shown. □

LEMMA 12.23. *Let the B_t be a standard Brownian motion the $B_t^2 - t$ and $B_t^4 - 6tB_t^2 + 3t^2$ are both martingales.*

PROOF. TODO: Prove using independent increments. □

TODO: Wald's Lemma seems to apply only to specific filtrations. Clarify this in the statement.

LEMMA 12.24 (Wald's Lemma). *Let B_t be a standard Brownian motion and let τ be an $\mathcal{F}$-optional time such that B_τ is bounded then*

(i) $\boldsymbol{E}[B_\tau] = 0$
(ii) $\boldsymbol{E}\left[B_\tau^2\right] = \boldsymbol{E}[\tau]$
(iii) $\boldsymbol{E}\left[\tau^2\right] \leq 4\boldsymbol{E}\left[B_\tau^4\right]$

PROOF. The idea is that the first two results are consequences of optional stopping (e.g. to get (i) let $\sigma = 0$ then apply Optional Stopping to conclude $\mathbf{E}[B_\tau] = \mathbf{E}[\mathbf{E}[B_\tau \mid \mathcal{F}_0]] = B_0 = 0$; to get (ii) one argues using the martingale $B_t^2 - t$ and to get (iii) one argues using the martingale $B_t^4 - 6tB_t^2 + 3t^2$). The trick is that τ is not assumed bounded so we cannot apply Theorem 9.82. To fix this, pick an arbitrary $T > 0$ and argue as above to conclude that $\mathbf{E}[B_{\tau\wedge T}] = 0$,

$\mathbf{E}\left[B^2_{\tau\wedge T}\right] = \mathbf{E}\left[\tau \wedge T\right]$ and $\mathbf{E}\left[B^4_{\tau\wedge T}\right] + 3\mathbf{E}\left[(\tau\wedge T)^2\right] = 6\mathbf{E}\left[(\tau\wedge T)B^2_{\tau\wedge T}\right]$. Now by the boundedness of B_τ, we know that $B_{\tau\wedge T}$ is bounded so we may apply Dominated Convergence to conclude $\mathbf{E}\left[B_\tau\right] = \lim_{T\to\infty}\mathbf{E}\left[B_{\tau\wedge T}\right] = 0$ and $0 \leq \tau\wedge T \uparrow \tau$ so by Dominated Convergence and Monotone Convergence we have

$$\mathbf{E}\left[B^2_\tau\right] = \lim_{T\to\infty}\mathbf{E}\left[B^2_{\tau\wedge T}\right] = \lim_{T\to\infty}\mathbf{E}\left[\tau\wedge T\right] = \mathbf{E}\left[\tau\right]$$

As a consequence of (ii) and the boundedness of B_τ we now that $\mathbf{E}\left[\tau\right] < \infty$ and therefore $(\tau\wedge T)B^2_{\tau\wedge T}$ is bounded by the integrable function $C^2\tau$ where $C = \sup_{0\leq t<\infty}|B_t|$ and we can use Dominated Convergence and Monontone Convergence to take limits and conclude

$$\begin{aligned}
\mathbf{E}\left[B^4_\tau\right] + 3\mathbf{E}\left[\tau^2\right] &= \lim_{T\to\infty}\mathbf{E}\left[B^4_{\tau\wedge T}\right] + 3\lim_{T\to\infty}\mathbf{E}\left[(\tau\wedge T)^2\right] \\
&= 6\lim_{T\to\infty}\mathbf{E}\left[(\tau\wedge T)B^2_{\tau\wedge T}\right] = 6\mathbf{E}\left[\tau B^2_\tau\right] \\
&\leq 6(\mathbf{E}\left[\tau^2\right]\mathbf{E}\left[B^4_\tau\right])^{1/2}
\end{aligned}$$

where in the last line we have used Cauchy Schwartz (Lemma 3.9). If we divide by $\mathbf{E}\left[B^4_\tau\right]$ and write $r = (\mathbf{E}\left[\tau^2\right]/\mathbf{E}\left[B^4_\tau\right])^{1/2}$ the inequality we have proven is $1 + 3r^2 \leq 6r$. Now simple algebra shows $3(r-1)^2 = 3r^2 - 6r + 3 \leq 2$ and therefore $r \leq 1 + \sqrt{2/3} < 2$. Upon backsubstituting the definition of r the inequality (iii) is proven. □

As a small step toward Skorohod embedding we first let $x \leq 0 \leq y$ be two real numbers and consider the hitting time $\tau_{x,y} = \inf\{t \geq 0 \mid B_t = x \text{ or } B_t = y\}$. By continuity of B_t and the closedness of the set $\{x, y\}$ we know from Lemma 9.70 that $\tau_{x,y}$ is an optional time and by Wald's Lemma just proven (TODO: Show that $\tau_{x,y}$ is almost surely finite; from this it follows trivially from the definition of $\tau_{x,y}$ that $B_{\tau_{x,y}}$ is almost surely bounded by $-x \vee y$) we know that $\mathbf{E}\left[B_{\tau_{x,y}}\right] = 0$. The point to bring out is that since the distribution of $B_{\tau_{x,y}}$ is supported on the two points $\{x, y\}$ by definition the condition $\mathbf{E}\left[B_{\tau_{x,y}}\right] = 0$ uniquely determines the distribution to be $\mathcal{L}(B_{\tau_{x,y}}) = \frac{y\delta_x - x\delta_y}{y-x}$ and moreover tells us that every mean zero measure supported on two points $\{x, y\}$ may be represented as a $B_{\tau_{x,y}}$. Given the tools we have developed, this fact was quite easy to see but what is less clear is that it can pushed further to represent an arbitrary mean zero random variable as a stopped Brownian motion.

TODO: Have we shown that the integral of the measures $\nu_{a,b}$ is a well defined object?

Lemma 12.25. *Let μ a Borel measure on $\mathbb{R}$ such that $\int x\,d\mu = 0$ and for all $a \leq 0 \leq b \in \mathbb{R}$ define the measure*

$$\nu_{a,b} = \begin{cases} \delta_0 & \text{if } a = 0 \text{ or } b = 0 \\ \frac{b\delta_a - a\delta_b}{b-a} & \text{if } a < 0 < b \end{cases}$$

then $\nu_{a,b} : \mathbb{R}_- \times \mathbb{R}_+ \to \mathcal{P}(\mathbb{R})$ is a probability kernel and there exists a measure $\tilde{\mu}$ on $\mathbb{R}_- \times \mathbb{R}_+$ such that

$$\mu(A) = \int \nu_{a,b}(A)\,d\tilde{\mu}(a,b) \text{ for all } A \in \mathcal{B}(\mathbb{R})$$

Proof. To see that $\nu_{a,b}$ is a kernel, it is immediate from the definition that for fixed $(a, b) \in \mathbb{R}_- \times \mathbb{R}_+$ $\nu_{a,b}$ is a probability measure. For fixed $A \in \mathcal{B}(\mathbb{R})$ we have

$$\nu_{a,b}(A) = \begin{cases} \mathbf{1}_A(0) & \text{if } a = 0 \text{ or } b = 0 \\ \frac{b}{b-a}\mathbf{1}_A(a) - \frac{a}{b-a}\mathbf{1}_A(b) & \text{if } a < 0 < b \end{cases}$$

which is a measurable function of (a, b) by measurability of the sets A, $\{(a, b) \in \mathbb{R}^2 \mid a = 0 \text{ or } b = 0\}$ and $\{(a, b) \in \mathbb{R}^2 \mid a < 0 < b\}$.

Denote by μ_+ the restriction $\mu_{(0,\infty)}$ of μ to the interval $(0, \infty)$ and by μ_- the restriction $\mu_{(-\infty,0)}$ of μ to the interval $(-\infty, 0)$. Define $c = \int x\, d\mu_+$ and by the condition $\int x\, d\mu$ note that $c = -\int x\, d\mu_-$. Now let $f : \mathbb{R} \to \mathbb{R}_+$ be a non-negative Borel measurable function and calculate using Tonelli's Theorem (Theorem 2.88)

$$\begin{aligned} c\int f(y)\, d\mu(y) &= c\mu(\{0\})f(0) + c\int f(y)\, d\mu_+(y) + c\int f(y)\, d\mu_-(y) \\ &= c\mu(\{0\})f(0) - \int x\, d\mu_-(x) \int f(y)\, d\mu_+(y) + \int x\, d\mu_+(x) \int f(y)\, d\mu_-(y) \\ &= c\mu(\{0\})f(0) + \int (yf(x) - xf(y))\; d(\mu_- \otimes \mu_+)(x, y) \\ &= c\mu(\{0\})f(0) + \int (y - x) \left[\int f(z)\, d\nu_{x,y}\right] d(\mu_- \otimes \mu_+)(x, y) \end{aligned}$$

where in the last line we have used the direct calculation $\int f\, d\nu_{x,y} = \frac{yf(x)-xf(y)}{y-x}$.

We can also compute for measurable f

$$\int \left[\int f\, d\nu_{x,y}\right] d\delta_{0,0}(x, y) = \int f\, d\nu_{0,0} = f(0)$$

Thus if for every μ we define

$$\tilde{\mu} = \mu(\{0\})\delta_{0,0} + \frac{y - x}{\int x\, d\mu_+(x)}\mu_- \otimes \mu_+$$

we have for all non-negative Borel measurable f, $\int f\, d\mu = \int \left[\int f\, d\nu_{a,b}\right] d\tilde{\mu}(a, b)$. In particular this holds for indicator functions.

The measurability of the map $\mu \to \tilde{\mu}$ follows by noting it is a composition of a number of measurable maps; indeed by definition of the σ-algebra on the space of measures we know that $\mu \to \mu(\{0\})$ is measurable and Lemma 8.26 shows that the restrictions $\mu \to \mu_\pm$, the integral $\mu \to \int x\, d\mu_+(x)$ and the product measure $(\mu_+, \mu_-) \to \mu_- \otimes \mu_+$ are all measurable mappings of measures. ☐

We are now ready to show that one can represent any mean zero Borel probability measure in the form B_τ for an appropriate optional time τ.

Lemma 12.26. *Let*

(i) *μ be a Borel probability measure on $\mathbb{R}$ with $\int x\, d\mu = 0$*
(ii) *$\tilde{\mu}$ is a Borel measure on $\mathbb{R}_- \times \mathbb{R}_+$ such that $\mu(A) = \int \nu_{x,y}(A)\, d\tilde{\mu}(x, y)$ for all $A \in \mathcal{B}(\mathbb{R})$*
(iii) *(α, β) be a random element in $\mathbb{R}_- \times \mathbb{R}_+$ such that $\mathcal{L}(\alpha, \beta) = \tilde{\mu}$*
(iv) *B_t be an independent Brownian motion*
(v) *$\mathcal{F}$ be the filtration defined by $\mathcal{F}_t = \sigma(\cup_{s\leq t}\sigma(B_s) \cup \sigma(\alpha) \cup \sigma(\beta))$*

Then

$$\tau = \inf\{t \geq 0 \mid B_t = \alpha \text{ or } B_t = \beta\}$$

is an $\mathcal{F}$-optional time and

$$\mathcal{L}(B_\tau) = \mu \qquad \boldsymbol{E}[\tau] = \int x^2 \, d\mu(x) \qquad \boldsymbol{E}\left[\tau^2\right] \leq 4 \int x^4 \, d\mu(x)$$

Proof. First note that by independence of the B_t and (α, β) we also know that $B_t - B_s$ is independent of (α, β) for all $s \leq t$ and therefore B_t is an $\mathcal{F}$-Brownian motion. To see that τ is $\mathcal{F}$-optional we recast the definition of τ slightly to make it clear that it is (almost) a hitting time $\tau = t \geq 0 \mid \{\frac{B_t - \alpha}{\beta} \in \{0, 1\}\}$ (it is not a hitting time since we have the condition $t \geq 0$ rather than $t > 0$). Pick $0 \leq t < \infty$, there is an analogous but simpler argument to that in Lemma 9.70 using the the continuity of B_t, closedness of $\{0, 1\}$ to conclude

$$\begin{aligned}\{\tau \leq t\} = \{\alpha \neq \beta\} \cap \cap_{n=1}^\infty \cup_{\substack{0 \leq q \leq t \\ q \in \mathbb{Q}}} \{\frac{B_t - \alpha}{\beta - \alpha} \in (-1/n, 1/n) \cup (1 - 1/n, 1 + 1/n)\} \\ \cup \{\alpha = \beta\} \cap \cap_{n=1}^\infty \cup_{\substack{0 \leq q \leq t \\ q \in \mathbb{Q}}} \{B_t - \alpha \in (-1/n, 1/n)\}\end{aligned}$$

and therefore $\mathcal{F}_t$ measurability follows from the adaptedness of B_t.

To calculate the distribution of B_τ we let $A \in \mathcal{B}(\mathbb{R})$ and consider B_τ as a function of the independent random elements B_t and (α, β) applying Fubini (specifically Lemma 4.6), the Expectation Rule (Lemma 3.7) and the definition of $\tilde{\mu}$ to get

$$\mathbf{P}\{B_\tau \in A\} = \mathbf{E}\left[\mathbf{P}\{B_{\tau_{x,y}} \in A\} \mid_{(x,y)=(\alpha,\beta)}\right] = \mathbf{E}\left[\nu_{\alpha,\beta}(A)\right] = \int \nu_{x,y}(A) \, d\tilde{\mu}(x, y) = \mu(A)$$

Now we compute using Lemma 4.6 , Wald's Lemma 12.24, the Expectation Rule and the just proven fact that the distribution of B_τ is μ to see

$$\mathbf{E}[\tau] = \mathbf{E}\left[\mathbf{E}[\tau_{x,y}] \mid_{(x,y)=(\alpha,\beta)}\right] = \mathbf{E}\left[B_\tau^2\right] = \int x^2 \, d\mu(x)$$

and in the same way

$$\mathbf{E}\left[\tau^2\right] = \mathbf{E}\left[\mathbf{E}\left[\tau_{x,y}^2\right] \mid_{(x,y)=(\alpha,\beta)}\right] \leq 4\mathbf{E}\left[B_\tau^4\right] = 4 \int x^4 \, d\mu(x)$$

$\square$

We can now complete the embedding of a random walk in a suitable Brownian motion.

Theorem 12.27 (Skorohod Embedding). *Let $\xi, \xi_1, \xi_2, \dots$ be i.i.d. random variables such that $\boldsymbol{E}[\xi] = 0$. Define $S_n = \xi_1 + \cdots + \xi_n$, there exists a probability space $(\Omega, \mathcal{A}, P)$ and a filtration $\mathcal{F}$ with a Brownian motion B_t and $\mathcal{F}$-optional times $0 = \tau_0 \leq \tau_1 \leq \cdots$ such that $(B_{\tau_1}, B_{\tau_2}, \dots) \overset{d}{=} (S_1, S_2, \dots)$ and the differences $\Delta\tau_n = \tau_n - \tau_{n-1}$ are i.i.d. and satisfy $\boldsymbol{E}[\Delta\tau_n] = \boldsymbol{E}\left[\xi^2\right]$ and $\boldsymbol{E}\left[(\Delta\tau_n)^2\right] \leq 4\boldsymbol{E}\left[\xi^4\right]$.*

Proof. Let μ be distribution of ξ and let B_t be a standard Brownian motion. Because $\mathbf{E}[\xi] = 0$ by Lemma 12.25 we know there is a $\tilde{\mu}$ such that $\mu(A) = \int \nu_{x,y}(A) \, d\tilde{\mu}$ for all $A \in \mathcal{B}(\mathbb{R})$. Potentially extending the probability space of B_t we can assume that there are i.i.d random vectors $(\alpha_1, \beta_1), (\alpha_2, \beta_2), \dots$ with distribution $\tilde{\mu}$ that are independent of B_t (Theorem 4.34 TODO: This theorem is stated for random variables; what is necessary to extend to random vectors?). Define the filtrations

$$\begin{aligned}\mathcal{F}^n_t &= \sigma(\alpha_k, \beta_k, k \leq n, B_s, 0 \leq s \leq t) \text{ for } n > 0 \\ \mathcal{G}_n &= \sigma(\alpha_k, \beta_k, k \leq n, B) \\ \mathcal{F}_t &= \sigma(\alpha_n, \beta_n, n > 0, B_s, 0 \leq s \leq t)\end{aligned}$$

Claim: B_t is an $\mathcal{F}$-Brownian motion.

This follows from independence of B_t and the (α_n, β_n) (TODO: more detail presumably referencing Lemma 4.14).

Define the sequence of random times $0 = \tau_0 \leq \tau_1 \leq \cdots$ recursively by the formula for $n \geq 1$

$$\begin{aligned}B^{n-1}_t &= B_{\tau_{n-1}+t} - B_{\tau_{n-1}} \\ \tau_n &= \inf\{t \geq \tau_{n-1} \mid B_t - B_{\tau_{n-1}} \in \{\alpha_n, \beta_n\}\} \\ &= \tau_{n-1} + \inf\{t \geq 0 \mid B^{n-1}_t \in \{\alpha_n, \beta_n\}\}\end{aligned}$$

Claim: τ_n is $\mathcal{F}$-optional and $\mathcal{G}_n$-measurable for all $n \geq 0$.

In fact we shall show the stronger result that τ_n is $\mathcal{F}^n$-optional. This follows using induction and the explicit formula

$$\begin{aligned}\{\tau_n \leq t\} &= \cap_{n=1}^\infty \cup_{\substack{0 \leq q \leq t \\ q \in \mathbb{Q}}} \{\tau_{n-1} \leq q\} \cap \{\frac{B_q - B_{\tau_{n-1}} - \alpha_n}{\beta_n - \alpha_n} \in (-1/n, 1/n) \cup (1 - 1/n, 1 + 1/n)\} \\ &= \cap_{n=1}^\infty \cup_{\substack{0 \leq q \leq t \\ q \in \mathbb{Q}}} \{\tau_{n-1} \leq q\} \cap \{\frac{B_q - B_{\tau_{n-1} \wedge t} - \alpha_n}{\beta_n - \alpha_n} \in (-1/n, 1/n) \cup (1 - 1/n, 1 + 1/n)\}\end{aligned}$$

The fact that B_t has continuous sample paths implies that it is progressively measurable (Lemma 9.89) and therefore since τ_{n-1} is $\mathcal{F}^{n-1}$-optional (a fortiori $\mathcal{F}^n$-optional) we know that $B_{\tau_{n-1} \wedge t}$ is $\mathcal{F}^n_t$-measurable (Lemma 9.90). Since α_n and β_n are $\mathcal{F}^n_0$-measurable and B_q is $\mathcal{F}^n_t$-measurable for all $q \leq t$, optionality is shown. Now since $\mathcal{F}^n_t \subset \mathcal{G}_n$ for all $t \geq 0$, we have $\{\tau_n \leq t\} \in \mathcal{G}_n$ for all $t \geq 0$ and $\mathcal{G}_n$-measurability of τ_n follows (Lemma 2.6 and Lemma 2.12).

Claim: $\Delta\tau_n$ are i.i.d., $\mathbf{E}\left[\Delta\tau_n\right] = \mathbf{E}\left[\xi^2\right]$ and $\mathbf{E}\left[(\Delta\tau_n)^2\right] = \mathbf{E}\left[\xi^4\right]$

On the subset $C([0, \infty), \mathbb{R}) \times \mathbb{R} \times \mathbb{R}$, we claim that the function

$$\Psi(f, a, b) = \inf\{t \geq 0 \mid f(t) \in \{a, b\}\}$$

is measurable. First, define the mapping $\tau_F(f) = \inf\{t \geq 0 \mid f(t) \in F\}$ for continuous f and closed F. The often used formula

$$\{\tau_F(f) \leq t\} = \cap_{n=1}^\infty \cup_{\substack{0 \leq q \leq t \\ q \in \mathbb{Q}}} \{d(f(q), F) < 1/n\} \text{ for } f \text{ continuous and } F \text{ closed}$$

shows that τ_F is measurable on $C([0, \infty), \mathbb{R}) \cap \mathbb{R}^{[0,\infty)}$. Then factoring

$$\Psi(f, a, b) = \mathbf{1}_{a \neq b} \tau_{\{0,1\}}((f - a)/(b - a)) + \mathbf{1}_{a=b} \tau_{\{0\}}(f - a)$$

and using measurability of group operations on set functions (Lemma 9.3) we get the measurability of Ψ. We can write $\Delta\tau_n = \Psi(B_{\tau_{n-1}+t} - B_{\tau_{n-1}}, \alpha_n, \beta_n)$ and by the Strong Markov Property we know that for all $n \geq 0$, $B_{\tau_{n-1}+t} - B_{\tau_{n-1}}$ is a standard Brownian motion independent of $\mathcal{F}_{\tau_{n-1}}$ (hence independent of (α_n, β_n)) and by construction (α_n, β_n) is i.i.d. with distribution $\tilde{\mu}$. Therefore $\mathcal{L}(B_{\tau_{n-1}+t} - B_{\tau_{n-1}}, \alpha_n, \beta_n) = \mathcal{L}(B_t) \otimes \tilde{\mu}$. Since we have expressed $\Delta\tau_n$ as a function $\Psi(B_{\tau_{n-1}+t} - B_{\tau_{n-1}}, \alpha_n, \beta_n)$ it follows from the Expectation Rule (Lemma 3.7) that $\mathbf{P}\{\Delta\tau_n \in A\} = \int \Psi(x,y,z)\, d\mathcal{L}(B_t)(x) \otimes \tilde{\mu}(y,z)$ and is the same for all $n \geq 0$. Independence follows in a similar way. By Lemma 4.15 it suffices for us to show that $(\Delta\tau_0, \dots, \Delta\tau_n) \perp\!\!\!\perp \Delta\tau_{n+1}$ for all $n \geq 0$. In fact we shall prove something a bit stronger. Define $\mathcal{H}_n = \sigma(\tau_k, B_{\tau_k}, k \leq n)$; we shall show $\mathcal{H}_n \perp\!\!\!\perp \Delta\tau_{n+1}$. Applying Lemma 4.15 to the sequence of σ-algebras $\sigma(B), \sigma(\alpha_1, \beta_1), \dots$ we know that $(\alpha_{n+1}, \beta_{n+1}) \perp\!\!\!\perp \mathcal{G}_n$ for all $n \geq 1$. We have shown that τ_n is $\mathcal{G}_n$-measurable and moreover τ_n is $\mathcal{F}^n$-optional therefore B_{τ_n} is $\mathcal{F}^n_{\tau_n}$-measurable hence $\mathcal{G}_n$-measurable. Therefore $\sigma(B^n, \mathcal{H}_n) \subset \mathcal{G}_n$ and we conclude $(\alpha_{n+1}, \beta_{n+1}) \perp\!\!\!\perp (B^n, \mathcal{H}_n)$ for all $n \geq 1$. Now on the other hand, by the Strong Markov Property Theorem 12.19 we know that B^n is independent of $\mathcal{F}_{\tau_n}$. By $\mathcal{F}_{\tau_k}$-measurability of τ_k and B_{τ_k} and the fact that $\tau_k \leq \tau_n$ for $k \leq n$ we know that $\mathcal{H}_n \subset \mathcal{F}_{\tau_n}$ and we conclude that $\mathcal{H}_n \perp\!\!\!\perp B^n$ and therefore $(\alpha_{n+1}, \beta_{n+1}, B^n) \perp\!\!\!\perp \mathcal{H}_n$ for all $n \geq 1$ by Lemma 8.21. Now we have expressed $\Delta\tau_{n+1} = \Psi(B^n, \alpha_{n+1}, \beta_{n+1})$ and therefore $\Delta\tau_{n+1} \perp\!\!\!\perp \mathcal{H}_n$ for all $n \geq 1$ by Lemma 4.16. The fact that $\mathbf{E}\left[\Delta\tau_n\right] = \mathbf{E}\left[\xi^2\right]$ and $\mathbf{E}\left[(\Delta\tau_n)^2\right] = \mathbf{E}\left[\xi^4\right]$ follows from Lemma 12.26 applied to the standard Brownian motion B^{n-1}.

Claim: The $B^n_{\Delta\tau_{n+1}}$ are i.i.d. with $B^n_{\Delta\tau_{n+1}} \overset{d}{=} \xi$.

The fact that $B^n_{\Delta\tau_{n+1}} \overset{d}{=} \xi$ follows from Lemma 12.26 applied to the standard Brownian motion B^n_t using the facts that $\Delta\tau_{n+1} = \inf\{t \geq 0 \mid B^n_t \in \{\alpha_{n+1}, \beta_{n+1}\}\}$ and $\mathcal{L}(\alpha_{n+1}, \beta_{n+1}) = \tilde{\mu}$. To see that the $B^n_{\Delta\tau_{n+1}}$ are independent it suffices to show that $B^n_{\Delta\tau_{n+1}} \perp\!\!\!\perp (B^0_{\Delta\tau_1}, \dots, B^{n-1}_{\Delta\tau_n})$ for each $n > 0$ (Lemma 4.15). It follows from Lemma 12.26 that $\Delta\tau_n$ is $\sigma(\alpha_n, \beta_n, B^{n-1}_s; s \leq t)$-optional and therefore by Lemma 9.90 we know that $B^n_{\Delta\tau_{n+1} \wedge t}$ is $\sigma(\alpha_{n+1}, \beta_{n+1}, B^n)$-measurable for all $t \geq 0$. Taking the limit as t goes to infinity we conclude that $B^n_{\Delta\tau_{n+1}}$ is $\sigma(\alpha_{n+1}, \beta_{n+1}, B^n)$-measurable. On the other hand, since $B^{n-1}_{\Delta\tau_n} = B_{\tau_n} - B_{\tau_{n-1}}$ we know that $B^{k-1}_{\Delta\tau_k}$ is $\sigma(\tau_k, B_{\tau_k}, k \leq n) = \mathcal{H}_n$-measurable for all $k \leq n$. Having shown that $(\alpha_{n+1}, \beta_{n+1}, B^n) \perp\!\!\!\perp \mathcal{H}_n$ in the proof of the prior claim we are done with this claim.

The last part of the result to show is that $(B_{\tau_1}, B_{\tau_2}, \dots) \overset{d}{=} (S_1, S_2, \dots)$; this follows from writing B_{τ_n} as a telescoping sum

$$B_{\tau_n} = \sum_{k=1}^n B_{\tau_k} - B_{\tau_{k-1}} = \sum_{k=1}^n B^{k-1}_{\Delta\tau_k}$$

and using the previous claim to see that

$$(B_{\tau_1}, B^1_{\Delta\tau_2}, B^2_{\Delta\tau_3}, \dots) \overset{d}{=} (\xi_1, \xi_2, \xi_3, \dots)$$

and then applying the measurable mapping $g(t_1, t_2, t_3, \dots) = (t_1, t_1 + t_2, t_1 + t_2 + t_3, \dots)$. ∎

The Skorohod Embedding shows that a Brownian motion and associated optional times can be constructed to represent any unbiased random walk up to distribution. However it says a bit more than that in that it shows the optional times used in the embedding are i.i.d. sums. By the Law of Large Numbers we should therefore expect that almost surely in the large time limit the optional times should approach deterministic times so that, up to some error terms, if we sample the Brownian motion at these deterministic times it should be a random walk and we can dispense with the optional times altogether. This intuition turns out to be true and in fact a bit more is true; once we consider approximating a random walk with a Brownian motion sampled at deterministic times we are also in a position to get an almost sure approximation (as opposed to an approximation in distribution only).

To begin with we need the following result that is really a corollary of the proof of the Law Of Iterated Logarithm (Theorem 12.33).

LEMMA 12.28. *Let B_t be a standard Brownian motion then*

$$\lim_{r \downarrow 1} \limsup_{t \to \infty} \sup_{t \le u \le rt} \frac{|B_u - B_t|}{\sqrt{2t \log \log t}} = 0 \ a.s.$$

PROOF. To clean up the notation a little define $\psi(t) = \sqrt{2t \log \log t}$.

First note that $\limsup_{t \to \infty} \sup_{t \le u \le rt} \frac{|B_u - B_t|}{\sqrt{2t \log \log t}}$ is a decreasing function of r and therefore to show the result it suffices to restrict ourselves to showing

$$\lim_{n \to \infty} \limsup_{t \to \infty} \sup_{t \le u \le r_n t} \frac{|B_u - B_t|}{\sqrt{2t \log \log t}} = 0 \text{ a.s.}$$

where r_n is any sequence such that $r_n \downarrow 1$. In particular the result holds if we restrict $r \in \mathbb{Q}$ and interpret the limit in r as being over rational $r > 1$.

Claim: It suffices to show

$$\lim_{r \downarrow 1} \limsup_{n \to \infty} \sup_{r^n \le u \le r^{n+1}} \frac{|B_u - B_{r^n}|}{\psi(r^n)} = 0 \text{ a.s.}$$

This is a general fact; if $f(t)$ is a function with $t \ge 0$ then given any sequence $t_n \to \infty$ we have $\limsup_{n \to \infty} f(t_n) \le \limsup_{t \to \infty} f(t)$ (any limit point of $f(t_n)$ is also a limit point of $f(t)$).

Now let $r > 1$, $n > 0$ and $c > 0$ be fixed for the moment and define

$$A_n = \{ \sup_{r^n \le u \le r^{n+1}} |B_u - B_{r^n}| \ge c\psi(r^n) \}$$

Now applying the Markov Property of Brownian motion to conclude that $B_u - B_{r^n}$ is a standard Brownian motion, applying Lemma 12.22 to get the distribution of the maximal process, normalizing to a standard normal variable Z and using the tail bound Lemma 7.23 and some algebra we get

$$\begin{aligned}
\mathbf{P}\{A_n\} &= \mathbf{P}\{ \sup_{0 \le u \le r^{n+1} - r^n} |B_u| \ge c\psi(r^n) \} \\
&= \mathbf{P}\{ |B_{r^{n+1} - r^n}| \ge c\psi(r^n) \} \\
&= \mathbf{P}\{ |Z| \ge \frac{c\psi(r^n)}{\sqrt{r^{n+1} - r^n}} \} \\
&\le \frac{\sqrt{r^{n+1} - r^n}}{\sqrt{2\pi} c\psi(r^n)} e^{-c^2 \psi^2(r^n)/2(r^{n+1} - r^n)} \\
&= \frac{\sqrt{r-1}}{c\sqrt{4\pi(\log n + \log \log r)}} e^{-c^2 \log \log r^n / 2(r-1)} \\
&= \frac{\sqrt{r-1}}{c\sqrt{4\pi(\log n + \log \log r)}} (n \log r)^{-c^2/2(r-1)} \\
&\le C_{r,c} n^{-c^2/2(r-1)}
\end{aligned}$$

where we have selected a constant $C_{r,c}$ depending only on $r > 1$ and $c > 0$. Now for any $c > \sqrt{2(r-1)}$ we see that $\sum_{n=1}^{\infty} \mathbf{P}\{A_n\} < \infty$ and therefore we may apply Borel

Cantelli (Theorem 4.23) to conclude that almost surely there exists a random N (depending on r and c) such that $\sup_{r^n \leq u \leq r^{n+1}} |B_u - B_{r^n}| < c\psi(r^n)$ for all $n \geq N$. Therefore in particular if we choose $c = 2\sqrt{r-1}$ we conclude that almost surely $\limsup_{n\to\infty} \sup_{r^n \leq u \leq r^{n+1}} \frac{|B_u - B_{r^n}|}{\psi(r^n)} \leq 2\sqrt{r-1}$. Taking the countable intersection of events of probability 1 we get the bound almost surely for all $r \in \mathbb{Q}$. Let $r \downarrow 1$ over $r \in \mathbb{Q}$ and we are done. ☐

Now we are ready to state the results that an Brownian motion asymptotically approximates a random walk.

THEOREM 12.29. *Let $\xi, \xi_1, \xi_2, \dots$ be an i.i.d. sequence of random variables with $\boldsymbol{E}[\xi] = 0$ and $\boldsymbol{E}\left[\xi^2\right] = 1$ and let $S_n = \xi_1 + \cdots + \xi_n$. There exists a Brownian motion B such that*

$$\frac{1}{\sqrt{t}} \sup_{0 \leq s \leq t} \left|S_{\lfloor s \rfloor} - B_s\right| \xrightarrow{P} 0 \text{ as } t \to \infty$$
$$\lim_{t\to\infty} \frac{\left|S_{\lfloor t \rfloor} - B_t\right|}{\sqrt{2t \log\log t}} = 0 \text{ a.s.}$$

PROOF. The first order of business is to observe how the Skorohod embedding may be modified to get an almost sure representation of the random walk. Applying Theorem 12.27 we can conclude that there exists a Brownian motion $\tilde{B}$ and optional times $\tilde{\tau}_n$ such that $\tilde{B}_{\tilde{\tau}_n} \stackrel{d}{=} S_n$ and the $\Delta\tilde{\tau}_n = \tilde{\tau}_n - \tilde{\tau}_{n-1}$ are an i.i.d. sequence with $\mathbf{E}[\Delta\tilde{\tau}_n] = 1$. Now if we define $(\tilde{B}, \Delta\tilde{\tau}_1, \Delta\tilde{\tau}_2, \dots)$ as a random element in $C([0,\infty);\mathbb{R}) \times \mathbb{R}_+^\infty$ and define $g : C([0,\infty);\mathbb{R}) \times \mathbb{R}_+^\infty \to \mathbb{R}^\infty$ by $g(f, t_1, t_2, \dots) = (f(t_1), f(t_1 + t_2), f(t_1 + t_2 + t_3), \dots)$ then we have on the one hand

$$g(\tilde{B}, \Delta\tilde{\tau}_1, \Delta\tilde{\tau}_2, \dots) \stackrel{d}{=} (S_1, S_2, \dots)$$

Now we can apply Lemma 8.41 to conclude that there is a random element $(B, \Delta\tau_1, \Delta\tau_2, \dots)$ such that

$$(B, \Delta\tau_1, \Delta\tau_2, \dots) \stackrel{d}{=} (\tilde{B}, \Delta\tilde{\tau}_1, \Delta\tilde{\tau}_2, \dots)$$

and

$$(B_{\Delta\tau_1}, B_{\Delta\tau_1+\Delta\tau_2}, \dots) = (S_1, S_2, \dots) \text{ a.s.}$$

In particular by taking marginals we know that $B \stackrel{d}{=} \tilde{B}$ which shows B is a Brownian motion and the $\Delta\tau_n$ are i.i.d. with $\mathbf{E}[\Delta\tau_n] = 1$ and if we define random times $\tau_n = \sum_{k=1}^n \Delta\tau_k$ then we have $B_{\tau_n} = S_n$ a.s. for all $n > 0$. (TODO: Show g is measurable and justify that spaces are Borel). Note that while $\tilde{\tau}_n$ are optional times the τ_n are not nor do we need them to be; what matters here is only that τ_n is a sum of i.i.d. random variables $\Delta\tau_n$. By the Strong Law of Large Numbers (Theorem 5.22) we know that $\lim_{n\to\infty} \frac{\tau_n}{n} = 1$ and furthermore we know that $\lim_{t\to\infty} \frac{\tau_{\lfloor t \rfloor}}{t} = \lim_{t\to\infty} \frac{\tau_{\lfloor t \rfloor}}{\lfloor t \rfloor} \frac{\lfloor t \rfloor}{t} = 1$ a.s.

Claim: $\lim_{t\to\infty} \frac{\left|S_{\lfloor t \rfloor} - B_t\right|}{\sqrt{2t\log\log t}} = 0$ a.s.

As usual we define $\psi(t) = \sqrt{2t\log\log t}$. Since $\lim_{t\to\infty} \frac{\tau_{\lfloor t \rfloor}}{t} = 1$ a.s. we know that almost surely for every $r > 1$ there is a random $T > 0$ such that $1/r < \frac{\tau_{\lfloor t \rfloor}}{t} < r$

for all $t \geq T$. This implies that either $t \leq \tau_{\lfloor t \rfloor} \leq rt$ or $\tau_{\lfloor t \rfloor} \leq t \leq r\tau_{\lfloor t \rfloor}$ and consequently for every $r > 1$

$$\begin{aligned}\frac{\left|S_{\lfloor t \rfloor} - B_t\right|}{\psi(t)} &= \frac{\left|B_{\tau_{\lfloor t \rfloor}} - B_t\right|}{\psi(t)} \\ &\leq \sup_{t \leq u \leq rt} \frac{|B_u - B_t|}{\psi(t)} \wedge \sup_{\tau_{\lfloor t \rfloor} \leq u \leq r\tau_{\lfloor t \rfloor}} \frac{\left|B_u - B_{\tau_{\lfloor t \rfloor}}\right|}{\psi(t)}\end{aligned}$$

for sufficiently large $t > 0$ (depending on r). Now taking the limit as $t \to \infty$ and using the fact that $\lim_{t\to\infty} \tau_{\lfloor t \rfloor}/t = 1$ implies $\lim_{t\to\infty} \psi(\tau_{\lfloor t \rfloor})/\psi(t) - 1$ and the general fact that if $\lim_{t\to\infty} g(t)/t = 1$ implies $\limsup_{t\to\infty} f(g(t)) \leq \limsup_{t\to\infty} f(t)$ we get that almost surely for every $r > 1$,

$$\limsup_{t\to\infty} \frac{\left|S_{\lfloor t \rfloor} - B_t\right|}{\psi(t)} \leq \limsup_{t\to\infty} \sup_{t \leq u \leq rt} \frac{|B_u - B_t|}{\psi(t)}$$

Now taking the limit as $r \downarrow 1$, using Lemma 12.28 and the positivity of $\frac{\left|S_{\lfloor t \rfloor} - B_t\right|}{\psi(t)}$ we conclude $\lim_{t\to\infty} \frac{\left|S_{\lfloor t \rfloor} - B_t\right|}{\psi(t)} = 0$ a.s.

To get the next limit first we need a simple fact about deterministic sequences

Claim: If $\lim_{n\to\infty} a_n/n = 1$ then $\lim_{t\to\infty} \sup_{0\leq s \leq t} \left|a_{\lfloor s \rfloor} - s\right|/t = 0$.

Let $\epsilon > 0$ be arbitrary and pick $N > 0$ such that $1 - \epsilon < a_n/n < 1 + \epsilon$ for all $n \geq N$; we use this in the form $|a_n - n| < n\epsilon$. Now for every $t \geq N$ we use this bound to conclude

$$\sup_{0\leq s\leq t} \left|a_{\lfloor s \rfloor} - s\right| \leq \sup_{0 \leq s < N} \left|a_{\lfloor s \rfloor} - s\right| + \sup_{N \leq s \leq t} \left|a_{\lfloor s \rfloor} - \lfloor s \rfloor\right| + 1 \leq \sup_{0 \leq s < N} \left|a_{\lfloor s \rfloor} - s\right| + t\epsilon + 1$$

so dividing by t and taking the limit we get $\lim_{t\to\infty} \sup_{0\leq s\leq t} \left|a_{\lfloor s \rfloor} - s\right|/t \leq \epsilon$. Now let ϵ go to zero.

Now we define $\delta_t = \sup_{0\leq s\leq t} \left|\tau_{\lfloor s \rfloor} - s\right|$ and conclude from the previous claim that $\delta_t/t \xrightarrow{a.s.} 0$ as $t \to \infty$. Recall the definition of the modulus of continuity

$$w(f, t, h) = \sup_{\substack{0 \leq r,s \leq t \\ |r-s|<h}} |f(r) - f(s)|$$

and the fact that $\lim_{h\to\infty} w(f, t, h) = 0$ if and only if f is continuous (hence uniformly continuous) on $[0, t]$. With this notation in hand, let $\epsilon > 0$ and $h > 0$ be given and use a union bound and a rescaling of the Brownian motion by the factor $\sqrt{t}$ to bound

$$\begin{aligned}\mathbf{P}\{\frac{1}{\sqrt{t}} \sup_{0\leq s\leq t} \left|B_{\tau_{\lfloor s \rfloor}} - B_s\right| > \epsilon\} &\leq \mathbf{P}\{\delta_t \geq ht\} + \mathbf{P}\{w(B, t + ht, ht) > \sqrt{t}\epsilon\} \\ &= \mathbf{P}\{\delta_t \geq ht\} + \mathbf{P}\{w(B, 1 + h1, h) > \epsilon\}\end{aligned}$$

Since $\delta_t/t \xrightarrow{a.s.} 0$ we know that $\delta_t/t \xrightarrow{P} 0$ and taking the limit as $t \to \infty$ we get $\lim_{t\to\infty} \mathbf{P}\{\delta_t \geq ht\} = 0$. Then because Brownian motion is almost surely continuous hence almost surely uniformly continuous on every finite interval and we know that $w(B, T, h) \xrightarrow{P} 0$ as $h \to 0$ for every fixed $T > 0$ so we get $\lim_{h\to 0} \mathbf{P}\{w(B, 1+$

$h1, h) > \epsilon\} = 0$ and thus we conclude

$$\lim_{t \to \infty} \mathbf{P}\{\frac{1}{\sqrt{t}} \sup_{0 \le s \le t} \left| B_{\tau_{\lfloor s \rfloor}} - B_s \right| > \epsilon\} = 0$$

and the result is proven.

In the above proof we glossed over a measurability question that we backfill for completeness.

Claim: For fixed t and h, $w(f, t, h)$ is a measurable function of f on $C([0, \infty); \mathbb{R}) \cap \mathbb{R}^{[0,\infty)}$.

The basic point is that the supremum in the definition of the modulus of continuity can be restricted to the rationals without changing the definition. Let

$$w^{\mathbb{Q}}(f, t, h) = \sup_{\substack{0 \le r, q \le t; r, q \in \mathbb{Q} \\ |r-q|<h}} |f(r) - f(q)|$$

and we clearly have $w^{\mathbb{Q}}(f, t, h) \le w(f, t, h)$. In the other direction let $\epsilon > 0$ be given and pick x, y be such that $|f(x) - f(y)| > w(f, t, h) - \epsilon$. Now by density of rationals and continuity of f we can find rational numbers r, q such that $|r - q| < h$ and

$$|f(r) - f(q)| \ge |f(x) - f(y)| - |f(r) - f(x)| - |f(q) - f(y)| > w(f, t, h) - \epsilon/2$$

Since $\epsilon > 0$ was arbitrary we conclude that $w^{\mathbb{Q}}(f, t, h) = w(f, t, h)$. Now for any $v \ge 0$ we have

$$\{w^{\mathbb{Q}}(f, t, h) \le v\} = \cap_{\substack{0 \le r, q \le t; r, q \in \mathbb{Q} \\ |r-q|<h}} \{|f(r) - f(q)| \le v\}$$

and each $\{|f(r) - f(q)| \le v\}$ is easily seen to be measurable as it depends on evaluation of f at a finite number of points. ☐

The approximation result just proven can be turned into a weak convergence result if we put a little bit of work into defining the function spaces in which the convergence occurs. There are several choices one may make about how to do this. For the result we are to prove, we consider a random walk to be a piecewise constant function and therefore we look for convergence in a space of discontinuous functions.

We define the space of functions that have left limits and are continuous from the right (cadlag functions)

$$D[0, 1] = \{f : [0, 1] \to \mathbb{R} \mid \lim_{x \to a^+} f(x) = f(a) \text{ and } \lim_{x \to a^-} f(x) \text{ exists for all } x \in [0, 1]\}$$

and provide it with the supremum norm $\|f\|_\infty = \sup_{0 \le x \le 1} |f(x)|$. We take the σ-algebra on $D[0, 1]$ generated by the evaluations $\pi_t(f) = f(t)$ (i.e. we consider $D[0, 1] \cap \mathbb{R}^{[0,1]}$ as required in the definition of a stochastic process). We note that since each π_t is continuous in the sup norm, this σ-algebra is a sub-algebra of the Borel σ-algebra. It is in fact true that this σ-algebra is a proper subalgebra of the Borel σ-algebra so we have defined a setting in which we cannot apply our notions of convergence in distribution and therefore we have to be a bit barehanded about how we phrase and prove the desired result.

The key remaining technical lemma is the following one which can be regarded as a combination of Slutsky's Lemma and the Continuous Mapping Theorem tailor made for our scenario.

DEFINITION 12.30. Let $X : (\Omega, \mathcal{A}) \to D[0,1]$ be a process with paths in $D[0,1]$ and let $\phi : D[0,1] \to \mathbb{R}$ be a function. We say ϕ is almost surely continuous at X if

$$\sup_{\substack{A \in \mathcal{A} \\ X(A) \cap D_\phi = \emptyset}} \mathbf{P}\{A\} = 1$$

where

$$D_\phi = \{f \in D[0,1] \mid \phi \text{ is not continuous at } f\}$$

LEMMA 12.31. *Let $X^1, X^2, \dots$ and $Y, Y^1, Y^2, \dots$ be cadlag processes in $D[0,1]$ such that $Y^n \overset{d}{=} Y$ for all $n > 0$ and $\|X^n - Y^n\|_\infty \overset{P}{\to} 0$. Let $\phi : D[0,1] \to \mathbb{R}$ be measurable and almost surely continuous at Y, then $\phi(X^n) \overset{d}{\to} \phi(Y)$.*

PROOF. Let $T = [0,1] \cap \mathbb{Q}$ and note that $\mathbb{R}^T$ with the product σ-algebra is a Borel space (TODO: where do we prove this). Therefore using Lemma 8.40 we can construct a sequence processes $\overline{X}^n$ on T such that

(i) $\overline{X}^n \overset{d}{=} X^n$ for all $n > 0$
(ii) $(\overline{X}^n, Y) \overset{d}{=} (X^n, Y^n)$ for all $n > 0$

The first order of business is to verify that $\overline{X}^n$ can be extended to processes with paths in $D[0,1]$.

Claim: For every $n > 0$, $\overline{X}^n$ is almost surely bounded and has finitely many upcrossings on every finite interval.

The point is that these properties follow from the distribution of $\overline{X}^n$ and since they are true of X^n they hold for $\overline{X}^n$. Specifically as for almost sure boundedness

$$\mathbf{P}\{\cap_{m=1}^\infty \left\|\overline{X}^n\right\|_\infty > m\} = \lim_{m\to\infty} \mathbf{P}\{\left\|\overline{X}^n\right\|_\infty > m\} = \lim_{m\to\infty} \mathbf{P}\{\|X^n\|_\infty > m\} = \mathbf{P}\{\cap_{m=1}^\infty \|X^n\|_\infty > m\} = 0$$

TODO: Do the details on the upcrossings using the defintions from Doob Upcrossing.

Based on the previous claim, we see that we can define

$$\tilde{X}^n_t = \lim_{s\to t^+} \overline{X}^n_s$$

and since $\lim_{s\to t^-} \tilde{X}^n_t$ exists for every $t \in [0,1]$ we know that $\tilde{X}^n$ is a process with paths in $D[0,1]$ (measurability of $\tilde{X}^n_t$ follows from the fact that it is defined as a limit of measurable functions (Lemma 2.14)).

Claim: $(\tilde{X}^n, Y) \overset{d}{=} (X^n, Y^n)$ for all $n > 0$.

We know from construction that $\mathbf{P}\{(\tilde{X}^n, Y) \in A\} = \mathbf{P}\{(X^n, Y^n) \in A\}$ for all $A \in \mathcal{B}(\mathbb{R})^T$ so it suffices to show that $\mathcal{B}(\mathbb{R})^T$ generates $D[0,1] \cap \mathcal{B}(\mathbb{R})^{[0,1]}$ (Lemma 2.71). This follows from right continuity of members of $D[0,1]$ as for any $t \in [0,1]$ we have $\pi_t = \lim_{n\to\infty} \pi_{q_n}$ where q_n is an sequence of elements of T such that $q_n \downarrow t$.

Claim: $\|\cdot\|_\infty$ is measurable on $D[0,1]$ and subtraction $(f,g) \to f - g$ is measurable on $D[0,1] \times D[0,1] \to D[0,1]$.

By right continuity and density of $\mathbb{Q}$,

$$\{\|f\|_\infty \leq x\} = \{\sup_{t\in T} |f(t)|\} = \cap_{t\in T}\{f(t) \leq x\}$$

As for subtraction, again using right continuity and density of $\mathbb{Q}$,

$$\{f(t) - g(t) \leq x\} = \cap_{\substack{q \geq t \\ q \in \mathbb{Q}}} \{f(q) \leq x - q\} \times \{g(q) \leq q\}$$

From the two previous claims we know that $\left\|\tilde{X}^n - Y\right\|_\infty \overset{d}{=} \|X^n - Y^n\|_\infty$.

Claim: $\phi(\tilde{X}^n) \overset{P}{\to} \phi(Y)$.

First, since $\|X^n - Y^n\|_\infty \overset{P}{\to} 0$ we have $\|X^n - Y^n\|_\infty \overset{d}{\to} 0$ and since $\left\|\tilde{X}^n - Y\right\|_\infty \overset{d}{=} \|X^n - Y^n\|_\infty$ we conclude that $\left\|\tilde{X}^n - Y\right\|_\infty \overset{d}{\to} 0$; as the weak limit is a deterministic constant by Lemma 5.33 we get $\left\|\tilde{X}^n - Y\right\|_\infty \overset{P}{\to} 0$.

We know that $\left\|\tilde{X}^n - Y\right\|_\infty \overset{P}{\to} 0$ if and only if every subsequence has a further subsequence that converges almost surely (Lemma 5.10). Let N be a subsequence of $\left|\phi(\tilde{X}^n) - \phi(Y)\right|$ and select a further subsequence $N' \subset N$ such that $\left\|\tilde{X}^n - Y\right\|_\infty \overset{a.s.}{\to} 0$ along N'. By the almost sure continuity of ϕ at Y we conclude that $\left|\phi(\tilde{X}^n) - \phi(Y)\right| \overset{a.s.}{\to} 0$ along N' hence we conclude $\phi(\tilde{X}^n) \overset{P}{\to} \phi(Y)$. Since ϕ was assumed measurable we have $\phi(\tilde{X}^n) \overset{d}{=} \phi(X^n)$ and therefore we conclude $\phi(X^n) \overset{d}{\to} \phi(Y)$. □

Theorem 12.32 (Donsker's Invariance Principle). *Let $\xi_1, \xi_2, \dots$ be an i.i.d. sequence of random variable with mean 0 and variance 1 and define for all $t \in [0,1]$ and $n \in \mathbb{N}$,*

$$S_n = \sum_{k=1}^n \xi_k$$

$$X_t^n = \frac{1}{\sqrt{n}} \sum_{k=1}^{\lfloor nt \rfloor} \xi_k = \frac{1}{\sqrt{n}} S_{\lfloor nt \rfloor}$$

Let B be a Brownian motion on $[0,1]$ and let $\phi : D[0,1] \to \mathbb{R}$ be measurable and almost surely continuous at B, then $\phi(X^n) \overset{d}{\to} \phi(B)$.

Proof. Define $Y_t^n = \frac{1}{\sqrt{t}} B_{nt}$ and note that by scaling we have $Y^n \overset{d}{=} B$ for all $n \in \mathbb{N}$. Note that

$$\|X^n - Y^n\|_\infty = \frac{1}{\sqrt{n}} \sup_{0 \le t \le 1} \left|S_{\lfloor nt \rfloor} - B_{nt}\right| = \frac{1}{\sqrt{n}} \sup_{0 \le t \le n} \left|S_{\lfloor t \rfloor} - B_t\right|$$

and therefore by Theorem 12.29 we conclude $\|X^n - Y^n\|_\infty \overset{P}{\to} 0$. Now we apply the previous Lemma 12.31 to conclude $\phi(X^n) \overset{d}{\to} \phi(B)$ and we are done. □

Theorem 12.33 (Law of Iterated Logarithm). *Let B_t be a standard Brownian motion then*

$$\limsup_{t \to \infty} \frac{B_t}{\sqrt{2t \log \log t}} = 1 \ a.s.$$

Proof. The basic idea of the proof is to examine the behavior of Brownian paths sampled along the values of a geometric sequence q^n for some number $q > 1$. Because we need to interpolate between sampling points we must consider segments of the Brownian path between sampling points.

To get started pick a number $q \in \mathbb{Q}$ such that $q > 1$ and pick an $\epsilon > 0$ that we will later send to zero. To clean up the notation a bit define $\psi(t) = \sqrt{2t \log \log t}$ and let $A_n = \{\sup_{0 \leq t \leq q^n} B_t \geq (1+\epsilon)\psi(q^n)\}$. Using the distribution of the Brownian maximum process (Lemma 12.22), rescaling to a standard normal random variable and the Gaussian tail bounds from Lemma 7.23 we know that

$$\begin{aligned}
\mathbf{P}\{A_n\} &= \mathbf{P}\{|B_{q^n}| \geq (1+\epsilon)\psi(q^n)\} \\
&= \mathbf{P}\{\frac{|B_{q^n}|}{\sqrt{q^n}} \geq \frac{(1+\epsilon)\psi(q^n)}{\sqrt{q^n}}\} \\
&\leq \frac{\sqrt{q^n}}{(1+\epsilon)\psi(q^n)} e^{-(1+\epsilon)^2\psi^2(q^n)/2q^n} \\
&= \frac{1}{(1+\epsilon)\sqrt{2\log\log(q^n)}} e^{-(1+\epsilon)^2 \log\log(q^n)}
\end{aligned}$$

and there exists an N_q depending only on q such that the leading constant is less than 1 for $n \geq N_q$, so we have

$$\mathbf{P}\{A_n\} \leq \frac{1}{(n\log q)^{(1+\epsilon)^2}} \text{ for } n \geq N_q$$

which shows that $\sum_{n=1}^{\infty} \mathbf{P}\{A_n\} < \infty$. The Borel Cantelli Theorem implies that almost surely at most finitely many A_n occur. Thus almost surely there is an N_ω such that $|B_t| < (1+\epsilon)\psi(q^n)$ for all $n \geq N_\omega$ and all $0 \leq t \leq q^n$. We now have to provide a bound using $\psi(t)$ rather than $\psi(q^n)$. For any $t \geq 1$ pick $n \geq 1$ such that $q^{n-1} \leq t < q^n$, and use the fact that $\psi(t)/t = \sqrt{\frac{2\log\log t}{t}}$ is a decreasing function of t for large t (for example $t \geq e^e$ works since $\frac{d}{dt}\psi(t)/t = \frac{\frac{1}{\log t} - \log\log t}{t^2}$) to bound

$$\frac{B(t)}{\psi(t)} = \frac{B(t)}{\psi(q^n)} \frac{\psi(q^n)}{q^n} \frac{t}{\psi(t)} \frac{q^n}{t} \leq (1+\epsilon)q$$

for $t > q^{N_\omega} \wedge e^e$ and therefore $\limsup_{t\to\infty} \frac{B(t)}{\psi(t)} \leq (1+\epsilon)q$. Since $\epsilon > 0$ and $q > 1$ were arbitrary we conclude $\limsup_{t\to\infty} \frac{B(t)}{\psi(t)} \leq 1$.

Now for the other direction, again pick $q > 1$ and consider the events

$$D_n = \{B_{q^n} - B_{q^{n-1}} \geq \psi(q^n - q^{n-1})\}$$

We know that since $q \leq q^2 \leq \cdots$ so the D_n are independent events and $(B_{q^n} - B_{q^{n-1}})/\sqrt{q^n - q^{n-1}}$ is $N(0,1)$ so we can apply Lemma 7.23 to see that for any $x \geq x_0$ we have

$$\mathbf{P}\{(B_{q^n} - B_{q^{n-1}})/\sqrt{q^n - q^{n-1}} \geq x\} \geq \frac{x}{x^2+1} e^{-x^2/2} \geq \frac{x_0^2}{x_0^2+1} \frac{1}{x} e^{-x^2/2}$$

so if we let $c_1 = \frac{2\log\log q}{2\log\log q + 1}$ then

$$\begin{aligned}
\mathbf{P}\{D_n\} &= \mathbf{P}\{(B_{q^n} - B_{q^{n-1}})/\sqrt{q^n - q^{n-1}} \geq \psi(q^n - q^{n-1})/\sqrt{q^n - q^{n-1}}\} \\
&\geq c_1 \frac{e^{-\log\log(q^n - q^{n-1})}}{\sqrt{2\log\log(q^n - q^{n-1})}} \geq c_1 \frac{e^{-\log\log q^n}}{\sqrt{2\log\log q^n}} \geq \frac{c_2}{n\log n}
\end{aligned}$$

so by the integral test we see that $\sum_{n=1}^{\infty} \mathbf{P}\{D_n\} = \infty$. By Borel Cantelli we know that almost surely there exists N_1 such that $B_{q^n} \geq B_{q^{n-1}} + \psi(q^n - q^{n-1})$ for all $n \geq N_1$ (where N_1 depends on q and $\omega \in \Omega$). To turn this into a lower bound on B_{q^n} alone we use the fact $-B_t$ is also a Brownian motion so we know from the upper bound that we have already proven

$$\liminf_{t\to\infty} \frac{B_t}{\psi(t)} = -\limsup_{t\to\infty} \frac{-B_t}{\psi(t)} \geq -1 \text{ a.s.}$$

If we pick an arbitrary $\epsilon > 0$ then almost surely there exists N_2 such that for all $n \geq N_2$ $B_{q^n} \geq -(1+\epsilon)\psi(q^n)$. Therefore we have for all $n \geq N_1 \wedge N_2$,

$$\frac{B_{q^n}}{\psi(q^n)} \geq \frac{B_{q^{n-1}} + \psi(q^n - q^{n-1})}{\psi(q^n)} \geq \frac{-(1+\epsilon)\psi(q^{n-1}) + \psi(q^n - q^{n-1})}{\psi(q^n)}$$

Now we can provide lower bounds for $\psi(t)$ in the expressions above. Using the fact that $\psi(t)/\sqrt{t} = \sqrt{2\log\log(t)}$ is increasing we have

$$\frac{\psi(q^{n-1})}{\psi(q^n)} = \frac{\psi(q^{n-1})}{\sqrt{q^{n-1}}} \frac{\sqrt{q^n}}{\psi(q^n)} \frac{1}{\sqrt{q}} \leq \frac{1}{\sqrt{q}}$$

and using the fact that $\psi(t)/t$ is decreasing for large t we have

$$\frac{\psi(q^n - q^{n-1})}{\psi(q^n)} \geq \frac{q^n - q^{n-1}}{q^n} = 1 - \frac{1}{q}$$

for sufficiently large n so putting these facts together we get

$$\limsup_{t\to\infty} \frac{B_t}{\psi(t)} \geq \limsup_{n\to\infty} \frac{B_{q^n}}{\psi(q^n)} \geq \frac{-(1+\epsilon)}{\sqrt{q}} + 1 - \frac{1}{q} \text{ a.s.}$$

Now taking the intersection of countably many events of probability 1 over all $q \in \mathbb{Q}$ this bound exists almost surely for all rational numbers $q > 1$ so we may take the limit as $q \to \infty$ and conclude that $\limsup_{t\to\infty} \frac{B_t}{\psi(t)} \geq 1$. ☐

An additional scaling argument allows us to get a Law of Iterated Logarithm for the limit as $t \to 0$,

COROLLARY 12.34. *Let B_t be a standard Brownian motion then*

$$\limsup_{t\to 0} \frac{B_t}{\sqrt{2t\log\log(1/t)}} = 1$$

PROOF. We know that the rescaled process $X_t = tB_{1/t}$ for $t > 0$ is a standard Brownian motion. Therefore letting $h = 1/t$,

$$1 = \limsup_{h\to\infty} \frac{X_h}{\sqrt{2h\log\log(h)}} = \limsup_{t\to 0} \frac{X_{1/t}}{\sqrt{2/t\log\log(1/t)}} = \limsup_{t\to 0} \frac{B_t}{\sqrt{2t\log\log(1/t)}}$$

☐

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 12.35. *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* $[0, 1]$

$$S_n^*(t) = \frac{1}{\sqrt{n}} S(nt) \qquad \qquad \text{for } t \in [0, 1]$$

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 $\square$

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.

CHAPTER 13

Markov Processes

TODO: Thinking about Markov processes as dynamical/deterministic systems with (transduced) noise.

1. Markov Processes

The basic intuition of what a Markov process comprises is that it is a stochastic process X on a time scale T such that for every time $t \in T$ the future behavior of X_u for $u \geq t$ only depends on the past through the current value of X_t. Alas, in practice the types of problems that we concern ourselves with Markov process leads us to a definition of a significantly more complicated object. Rather than pummel the reader with the definition we take the approach of starting from simple intuition and building in the complexity by stages. Some readers may prefer to first jump to the end of this section to peer at the final defintion so that it can be kept in mind during the journey.

1.1. The Markov Property.

Definition 13.1. 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 13.2 (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 $\sigma(X_{t_0}, \dots, X_{t_{n-1}}, \mathcal{F}_{t_n}) \perp\!\!\!\perp_{X_{t_n}} X_{t_{n+1}}$. By Lemma 8.21 we conclude that $\mathcal{F}_{t_n} \perp\!\!\!\perp_{X_{t_0}, \dots, X_{t_n}} X_{t_{n+1}}$ for all $n \geq 0$; because $\mathcal{F}_{t_0} \subset \mathcal{F}_{t_n}$ we get $\mathcal{F}_{t_0} \perp\!\!\!\perp_{X_{t_0}, \dots, X_{t_n}} X_{t_{n+1}}$ for all $n \geq 0$. Another application of Lemma 8.21 shows that $\mathcal{F}_{t_0} \perp\!\!\!\perp_{X_{t_0}} \sigma(X_{t_1}, X_{t_2}, \dots)$.

Since the union of the σ-algebras $\sigma(X_{t_1}, X_{t_2}, \dots)$ 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 8.19). □

TODO: Merge previous lemma into this propostion. By Lemma 8.20 the conditional independence in the Markov property can be captured by statements about conditional probabilities; it useful to have alternative equivalent characterizations of the Markov property in terms of conditional expectations of more general functions.

PROPOSITION 13.3. *Let X be a process in $(S, \mathcal{S})$ with time scale T which is adapted to a filtration $\mathcal{F}_t$. Then the following are equivalent*

(i) *X has the Markov property*

(ii) *for all $t \in T$ and non-negative bounded $\sigma(\bigvee_{u \geq t} X_u)$-measurable random variables ξ we have*

$$\boldsymbol{E}[\xi \mid \mathcal{F}_t] = \boldsymbol{E}[\xi \mid X_t] \text{ a.s.}$$

(iii) *for all $s \leq t \in T$ and non-negative or bounded Borel measurable $f : S \to \mathbb{R}$ we have*

$$\boldsymbol{E}[f(X_t) \mid \mathcal{F}_s] = \boldsymbol{E}[f(X_t) \mid X_s] \text{ a.s.}$$

PROOF. To see that (i) implies (ii) suppose that X has the Markov property and let $A \in \sigma(\bigvee_{u \geq t} X_u)$. By Lemma 13.2 and Lemma 8.20 we have

$$\mathbf{P}\{A \mid X_t\} = \mathbf{P}\{A \mid \mathcal{F}_t, X_t\} = \mathbf{P}\{A \mid \mathcal{F}_t\} \text{ a.s.}$$

thus (ii) holds for indicator functions. By linearity of conditional expectation (ii) holds for $\sigma(\bigvee_{u \geq t} X_u)$-measurable simple functions. By approximation by simple functions and Monotone Convergence for conditional expectations (ii) holds for non-negative $\sigma(\bigvee_{u \geq t} X_u)$-measurable functions. By linearity of conditional expectations we get (ii) for bounded $\sigma(\bigvee_{u \geq t} X_u)$-measurable functions.

(ii) implies (iii) is immediate as any $f(X_t)$ if $\sigma(\bigvee_{u \geq t})$-measurable.

To see that (iii) implies (i) let $A \in \mathcal{S}$ and $s \leq t \in T$, then (iii) implies that

$$\mathbf{P}\{X_t \in A \mid X_s\} = \mathbf{P}\{X_t \in A \mid \mathcal{F}_s\} = \mathbf{P}\{X_t \in A \mid \mathcal{F}_s, X_s\} \text{ a.s.}$$

and Lemma 8.20 implies that $\mathcal{F}_s \perp\!\!\!\perp_{X_s} X_t$. □

1.2. Markov Transition Kernels. 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 8.34). 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 13.4. *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*

$$\boldsymbol{P}\{(X_{t_1}, \dots, X_{t_n}) \in A\} = \nu_{t_1} \otimes \mu_{t_1,t_2} \otimes \cdots \otimes \mu_{t_{n-1},t_n}(A)$$
$$\boldsymbol{P}\{(X_{t_1}, \dots, X_{t_n}) \in A \mid \mathcal{F}_{t_0}\}(\omega) = \mu_{t_0,t_1} \otimes \cdots \otimes \mu_{t_{n-1},t_n}(X_{t_0}(\omega), A)$$

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 distintegration

(Theorem 8.35) 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 13.2 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 13.5. Let μ and ν be probability kernels from S to S. Then we define the probability kernel $\mu\nu$ from S to S by

$$\mu\nu(s, A) = (\mu \otimes \nu)(s, S \times A) = \iint \mathbf{1}_{S\times A}(t, u)\, \nu(t, du)\mu(s, dt) = \int \nu(t, A)\mu(s, dt)$$

for all $s \in S$ and $A \in \mathcal{S}$.

Example 13.6. Let S be a finite set and view μ and ν as $S \times S$ matrices as in Example 8.25. Then $\mu\nu$ is just matrix multiplication:

$$\mu\nu(s, \{t\}) = \int \nu(u, \{t\})\mu(s, du) = \sum_{u\in S} \nu_{u,t}\mu_{s,u} = (\mu\nu)_{s,t}$$

Corollary 13.7 (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 8.34 that regular versions $\mu_{s,t}$ exist. By definition of $\mu_{s,t}\mu_{t,u}$, Lemma 13.4 and the uniqueness clause of Theorem 8.34

$$\begin{aligned}
\mu_{s,t}\mu_{t,u}(X_s, A) &= (\mu_{s,t} \otimes \mu_{t,u})(X_s, S \times A) \\
&= \mathbf{P}\{(X_t, X_u) \in S \times A \mid \mathcal{F}_s\} \\
&= \mathbf{P}\{X_u \in A \mid \mathcal{F}_s\} \\
&= \mathbf{P}\{X_u \in A \mid X_s\} \\
&= \mu_{s,u}(X_s, A) \text{ a.s.}
\end{aligned}$$

Therefore for each $A \in \mathcal{S}$,

$$\nu_s(\mu_{s,t}\mu_{t,u}(\cdot, A) \neq \mu_{s,u}(\cdot, A)) = \mathbf{P}\{\mu_{s,t}\mu_{t,u}(X_s, A) \neq \mu_{s,u}(X_s, A)\} = 0$$

Since S is Borel we can choose a common null set for all $A \in \mathcal{S}$ (when $S = [0, 1]$ just pick the union of null sets for A an interval with rational endpoints, show that this null set works for all intervals by continuity of measure and then use monotone classes; for general S just use the Borel isomorphism to reduce to the above case). ☐

1.3. Existence of Markov Processes. TODO: Example of process with the Markov property but for which the Chapman-Kolmogorov relations do not hold identically.

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 it 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 13.8. A *Markov transition kernel* on time scale T and state space $(S, \mathcal{S})$ is a probability kernel $\mu_{s,t} : S \times \mathcal{S} \to [0, 1]$ for each $s \leq t \in T$ such that

$$\mu_{s,t}\mu_{t,u} = \mu_{s,u} \text{ everywhere on } S \text{ for each } s \leq t \leq u$$

Definition 13.9. Let stochastic process X_t on a time scale $T \subset \mathbb{R}_+$ and state space $(S, \mathcal{S})$ such that X_t is adapted to a filtration $\mathcal{F}_t$. We say that X_t is a *Markov process* if there exists a Markov transition kernel $\mu_{s,t}$ such that for all $s \leq t \in T$

$$\mathbf{P}\{X_t \in \cdot \mid \mathcal{F}_s\} = \mu_{s,t}(X_s, \cdot) \text{ a.s.}$$

Note that we have not specified in the definition that a Markov process posesses the Markov property; showing that it does is not hard however.

Proposition 13.10. *A Markov process has the Markov property. Moreover for any non-negative or bounded function $f : S \to \mathbb{R}$ we have*

$$\boldsymbol{E}[f(X_t) \mid \mathcal{F}_s] = \boldsymbol{E}[f(X_t) \mid X_s] = \int f(u)\, \mu_{s,t}(X_s, du)$$

PROOF. Since we know $\mathbf{P}\{X_t \in A \mid \mathcal{F}_s\} = \mu_{s,t}(X_s, A)$ it follows that $\mathbf{P}\{X_t \in A \mid \mathcal{F}_s\}$ is X_s-measurable and therefore $\mathbf{P}\{X_t \in A \mid \mathcal{F}_s\} = \mathbf{P}\{X_t \in A \mid X_s\}$ a.s. The Markov property follows by Proposition 13.3.

Since $\mu_{s,t}(X_s, \cdot)$ is a regular version for $\mathbf{P}\{X_t \in \cdot \mid \mathcal{F}_s\}$ we apply Theorem 8.35 to see

$$\mathbf{E}\left[f(X_t) \mid \mathcal{F}_s\right] = \int f(u)\, \mu_{s,t}(X_s, du)$$

□

TODO: Note that in the discrete (or countable?) state space case we can in fact assume that Chapman-Kolmogorov are satisfied identically.

In lieu of general techinique for proving that a process is Markov from general principles, we give a result that shows that we can construct them from a set of transition kernels that obey the Chapman-Kolmogorov relations.

TODO: There are other ways of proving a process is Markov : the semigroup approach, the stochastic differential equation approach and the martingale problem approach. These are things we'll get to but not quite yet!

THEOREM 13.11. *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. This is an application of the Daniell-Kolmogorov Theorem. We first define the finite dimensional distributions and show that they form a projective family. For every $n \in \mathbb{N}$ and $0 \leq t_1 \leq \cdots \leq t_n$ we define

$$\begin{aligned}\nu_{t_1,\dots,t_n}(A) &= \nu\mu_{0,t_1} \otimes \mu_{t_1,t_2} \otimes \cdots \otimes \mu_{t_{n-1},t_n}(A) \\ &= \nu \otimes \mu_{0,t_1} \otimes \mu_{t_1,t_2} \otimes \cdots \otimes \mu_{t_{n-1},t_n}(S \times A)\end{aligned}$$

Let $A \in \mathcal{S}^{\otimes n-1}$ and let $1 \leq k \leq n$. Define

$$A_k = \{(x_1, \dots, x_n) \in S^n \mid (x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n) \in A\}$$

and calculate

$$\begin{aligned}&\nu_{t_1,\dots,t_n}(A_k) = (\nu\mu_{0,t_1} \otimes \mu_{t_1,t_2} \otimes \cdots \otimes \mu_{t_{n-1},t_n})(A_k) \\ &= \int \left[\int \left[\cdots \left[\int \mathbf{1}_{A_k}(s_1, \dots, s_n)\, \mu_{t_{n-1},t_n}(s_{n-1}, ds_n)\right] \cdots \right] \mu_{t_1,t_2}(s_1, ds_2)\right] \nu\mu_{0,t_1}(ds_1) \\ &= \int \left[\int \left[\cdots \left[\int \mathbf{1}_{A}(s_1, \dots, s_{k-1}, s_{k+1}, \dots, s_n)\, \mu_{t_{n-1},t_n}(s_{n-1}, ds_n)\right] \cdots \right] \mu_{t_1,t_2}(s_1, ds_2)\right] \nu\mu_{0,t_1}(ds_1)\end{aligned}$$

The point here is that the integral

$$\int \left[\cdots \left[\int \mathbf{1}_{A}(s_1, \dots, s_{k-1}, s_{k+1}, \dots, s_n)\, \mu_{t_{n-1},t_n}(s_{n-1}, ds_n)\right] \cdots \right] \mu_{t_k,t_{k+1}}(s_{k+1}, ds_{k+2})$$

is a function of $s_1, \dots, s_{k-1}, s_{k+1}$ only (i.e. it has no dependence on s_k). From the Chapman-Kolmogorov relation $\mu_{t_{k-1},t_k}\mu_{t_k,t_{k+1}} = \mu_{t_{k-1},t_{k+1}}$ we know that for any function of $f : S \to \mathbb{R}$ we have

$$\int \left[\int f(s_{k+1}) \, \mu_{t_k,t_{k+1}}(s_k, ds_{k+1}) \right] \mu_{t_{k-1},t_k}(s_{k-1}, ds_k) = \int f(s_{k+1}) \, \mu_{t_{k-1},t_{k+1}}(s_{k-1}, ds_{k+1})$$

which when applied to the integral above with $s_1, \dots, s_{k-2}$ fixed yields

$$\begin{aligned}
&\int \left[\cdots \left[\int \mathbf{1}_A(s_1, \dots, s_{k-1}, s_{k+1}, \dots, s_n) \, \mu_{t_{n-1},t_n}(s_{n-1}, ds_n) \right] \cdots \right] \mu_{t_{k-1},t_k}(s_{k-1}, ds_k) \\
&= \int \left[\cdots \left[\int \mathbf{1}_A(s_1, \dots, s_{k-1}, s_{k+1}, \dots, s_n) \, \mu_{t_{n-1},t_n}(s_{n-1}, ds_n) \right] \cdots \right] \mu_{t_{k-1},t_{k+1}}(s_{k-1}, ds_{k+1})
\end{aligned}$$

Now we can use this to conclude that

$$\begin{aligned}
\nu_{t_1,\dots,t_n}(A_k) &= (\nu\mu_{0,t_1} \otimes \mu_{t_1,t_2} \otimes \cdots \otimes \mu_{t_{n-1},t_n})(A_k) \\
&= (\nu\mu_{0,t_1} \otimes \mu_{t_1,t_2} \otimes \cdots \otimes \mu_{t_{k-1},t_{k+1}} \otimes \cdots \otimes \mu_{t_{n-1},t_n})(A) \\
&= \nu_{t_1,\dots,t_{k-1},t_{k+1},\dots,t_n}(A)
\end{aligned}$$

and we have show that the $\nu_{t_1,\dots,t_n}$ are a projective family. Now we can apply the Daniell-Kolmogorov Theorem 9.11 to conclude that there is an S valued process X on T such that

$$\mathcal{L}(X_{t_1}, \dots, X_{t_n}) = \nu_{t_1,\dots,t_n} = \nu\mu_{0,t_1} \otimes \mu_{t_1,t_2} \otimes \cdots \otimes \mu_{t_{n-1},t_n}$$

for all $n \in \mathbb{N}$ and $0 \leq t_1 \leq \cdots \leq t_n$. The case $n = 1$ and $t_1 = 0$ shows us that $\mathcal{L}(X_0) = \nu\mu_{0,0} = \nu$.

For every $t \in T$ define $\mathcal{F}_t = \sigma(X_s; s \leq t)$ to be filtration induced by X. We must show that X_t is a Markov process with transition kernels $\mu_{s,t}$ (the fact that the initial distribution is ν was already noted). Let $s \leq t$ be given and suppose that we have $s_1 \leq \cdots \leq s_n = s$. Pick $A \in \mathcal{S}^{\otimes n}$ and $B \in \mathcal{S}$ and calculate using the FDDs of X_t and the expectation rule (Lemma 3.7)

$$\begin{aligned}
&\mathbf{P}\{X_t \in B; (X_{s_1}, \dots, X_{s_n}) \in A\} = \mathbf{P}\{(X_{s_1}, \dots, X_{s_n}, X_t) \in A \times B\} = \nu_{s_1,\dots,s_n,t}(A \times B) \\
&= \int \left[\int \left[\cdots \left[\int \mathbf{1}_A(u_1, \dots, u_n)\mathbf{1}_B(u_{n+1}) \, \mu_{s,t}(u_n, du_{n+1}) \right] \cdots \right] \mu_{s_1,s_2}(u_1, du_2) \right] \nu\mu_{0,s_1}(du_1) \\
&= \int \left[\int \left[\cdots \left[\int \mathbf{1}_A(u_1, \dots, u_n)\mu_{s,t}(u_n, B) \, \mu_{s_{n-1},s_n}(u_{n-1}, du_n) \right] \cdots \right] \mu_{s_1,s_2}(u_1, du_2) \right] \nu\mu_{0,s_1}(du_1) \\
&= \mathbf{E}\left[\mu_{s,t}(X_s, B); (X_{s_1}, \dots, X_{s_n}) \in A\right]
\end{aligned}$$

Sets of the form $(X_{s_1}, \dots, X_{s_n}) \in A$ for $s_1 \leq \cdots \leq s_n = s$ are a π-system generating $\mathcal{F}_s$ and therefore by a monotone class argument (specifically Lemma 8.8) we may conclude that $\mathbf{E}\left[X_t \in \cdot \mid \mathcal{F}_s\right] = \mu_{s,t}(X_s, \cdot)$ a.s. $\square$

The previous theorem constructs a Markov process with an arbitrary initial distribution. As it turns out in many cases it is useful to consider the collection of Markov processes indexed by the initial distribution. Such a collection has a nice structure that results from the Markov property. The uncover the structure the first thing to do is to move all of the constructed Markov processes into the canonical picture so that we have a family of probability measures on S^T.

DEFINITION 13.12. 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 13.13. *The family $\boldsymbol{P}_x$ is a kernel from S to S^T. 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 $B \in \mathcal{S}^{\otimes n}$ and $\pi_t : S^T \to S$ is the evaluation map $\pi_t f = f(t)$. We can use Lemma 13.4 and the expectation rule Lemma 3.7 to compute for any ν,

$$\begin{aligned}
\mathbf{P}_\nu\{A\} &= \mathbf{P}_\nu\{(\pi_{t_1}, \dots, \pi_{t_n}) \in B\} \\
&= \mathbf{E}_\nu\left[\mathbf{P}\{(\pi_{t_1}, \dots, \pi_{t_n}) \in B \mid \mathcal{F}_0\}\right] \\
&= \mathbf{E}_\nu\left[\mu_{0,t_1} \otimes \dots \otimes \mu_{t_{n-1},t_n}(X_0, B)\right] \\
&= \int \mu_{0,t_1} \otimes \dots \otimes \mu_{t_{n-1},t_n}(x, B)\, \nu(dx)
\end{aligned}$$

In particular, for $\nu = \delta_x$ we get

$$\mathbf{P}_x\{A\} = \mu_{0,t_1} \otimes \dots \otimes \mu_{t_{n-1},t_n}(x, B)$$

which shows both that $\mathbf{P}_x\{A\}$ is a measurable function of x for fixed A (Lemma 8.29) 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 8.27 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 2.71). $\square$

TODO: This may not be the correct definition of a Markov process to settle on. We may want to select the picture of a Markov process as being a single stochastic process with a family of probability measures $\mathbf{P}_x$ for $x \in S$ such that under $\mathbf{P}_x$ the stochastic process is Markov (as above) starting at x. This definition assumes that we have a kernel property (so Lemma 13.13 proves such a kernel property holds in the "canonical" case). The work we have done to this point shows that a set of transition kernels gives rise to a Markov process with on the canonical space S^T. The interpretation as a family of measures without assuming the probability space is S^T is apparently useful (e.g. when we want to assume randomization variables exist for some construction). I still find the variety of interpretations of what a Markov process is to be very confusing. Perhaps we should define this latter concept as a Markov family and keep the current notion as a Markov process (I think Karatzas and Shreve do this). In the Karatzas and Shreve definition we wind up with an interesting new concept which is that the kernel $\mathbf{P}_x$ in a Markov family is only assumed to be *universally measurable* which is a looser condition than Borel measurability (the universal σ-algebra being the intersection of the completions of the Borel σ-algebra under all probability measures; hence being a superset of the Borel σ-algebra). This loosening seems to come up as important in the context of stochastic control. I am not at all clear on how important it is in the context of Markov processes as we are likely to develop it; it seems from Karatzas and

Shreve that this loosening comes up in Markov process theory when trying to find a right-continuous complete filtration with respect to which a Markov process (in particular Brownian motion) gives us a Markov family. So, we have shown that a by Kolmogorov existence we can construct a Markov family given a set of transition kernels however the filtration is not right continuous or complete and this construction results in a Borel measurable kernel $\mathbf{P}_x$. However if one tries to modify the construction to get a right-continuous complete filtration (usual conditions) then one has to give up Borel measurability in the kernel and make due with universal measurability. Perhaps it is worth having a definition of a Markov family and a "relaxed" or "complete" Markov family. What I don't have any intuition of is the circumstances under which we are forced to pass to impose the usual conditions and/or require measurability of $\mathbf{P}_x\{A\}$ for arbitrary $A \in \mathcal{A}$. The most obvious answer is that we can't consider mixed initial states $\mathbf{P}_\nu\{A\}$ to be defined (via $\mathbf{P}_\nu\{A\} = \int \mathbf{P}_x\{A\}\, \nu(dx)$) unless we have measurability of $\mathbf{P}_x\{A\}$ and therefore we don't even get measures $\mathbf{P}_\nu$ on all of $\mathcal{A}$ until we deal with the measurability issue (but we can always define $\mathbf{P}_\nu$ on $\mathcal{F}^X_\infty$). Note also that the concept of universal measurability comes up in the general theory of processes in which we have the Debut Theorem that states that every hitting time is universally measurable. It can also be shown that analytic sets are universally measurable.

Definition 13.14. A *time homogeneous Markov family* is a stochastic process X_t with a probability space $(\Omega, \mathcal{A})$, a time scale $T \subset \mathbb{R}_+$, a filtration $\mathcal{F}_t$, a ($\mathcal{F}_t$?) measureable $\theta_t : \Omega \to \Omega$, a state space $(S, \mathcal{S})$ and a family of probability measures $\mathbf{P}_x$ on Ω for $x \in S$ such that

(i) $\mathbf{P}_x$ is a (universally measurable?) kernel from S to Ω. TODO: Is this what we want? Blumenthal and Getoor say that $\mathbf{P}_x\{X_t \in A\}$ is $\mathcal{S}$ measurable for every $0 \leq t < \infty$ and $A \in \mathcal{S}$; so $\mathbf{P}_x$ is a kernel to $(\Omega, \mathcal{F}^X_\infty)$ but not necessarily to $(\Omega, \mathcal{A})$ or $(\Omega, \mathcal{F}_\infty)$.

(ii) $\mathbf{P}_x\{X_0 = x\} = 1$ for all $x \in S$ (note Blumenthal and Getoor do not assume this).

(iii) $\mathcal{F}_s \perp\!\!\!\perp_{X_s} X_t$ under $\mathbf{P}_x$ for all $s \leq t$ and $x \in S$ (i.e. for all $x \in S$, $A \in \mathcal{S}$ and $s \leq t$ we have $\mathbf{E}_x\left[X_t \in A \mid \mathcal{F}_s\right] = \mathbf{E}_x\left[X_t \in A \mid X_s\right]$ $\mathbf{P}_x$-a.s.) Alternatively do we just say:

$$\mathbf{E}\left[f(X_t \circ \theta_s) \mid \mathcal{F}_s\right] = \mathbf{E}_{X_s}\left[f(X_t)\right]$$

(iv) there exists a regular version $\mu^x_{s,t} : S \times \mathcal{S} \to [0,1]$ of $\mathbf{P}_x\{X_t \in \cdot \mid \mathcal{F}_s\}$ for each $s \leq t$ and $x \in S$ (is there any coherence requirement with respect to $x \in S$ here???).

Lemma 13.15. *Let $(S, \mathcal{S})$ be a measurable space and take a point $\Delta \notin S$, let $S^\Delta = S \cup \{\Delta\}$ and let $\mathcal{S}^\Delta$ be the σ-algebra on S^Δ generated by $\mathcal{S}$, then*

$$\mathcal{S}^\Delta = \mathcal{S} \cup \{A \cup \{\Delta\} \mid A \in \mathcal{S}\}$$

Proof. Since $S \in \mathcal{S}$ and $\{\Delta\} = S^\Delta \setminus S$ it follows that $\{\Delta\}$ is $\mathcal{S}^\Delta$-measurable and therefore the right hand side is included in $\mathcal{S}^\Delta$. It suffices to show that the right hand side is a σ-algebra. Clearly $\emptyset \in \mathcal{S} \subset \mathcal{S}^\Delta$ and for any $A \in \mathcal{S}$ we have both $S^\Delta \setminus A = (S \setminus A) \cup \{\Delta\}$ and $S^\Delta \setminus (A \cup \{\Delta\}) = S \setminus A$ and therefore the right hand side is closed under set complement. Given countable index sets I and J and sets $A_i \in \mathcal{S}$ and $B_j \in \mathcal{S}$ we have

$$\cup_{i \in I} A_i \cup \cup_{j \in J} (B_j \cup \{\Delta\}) = \begin{cases} (\cup_{i \in I} A_i \cup \cup_{j \in J} B_j) \cup \{\Delta\} & \text{if } J \neq \emptyset \\ \cup_{i \in I} A_i & \text{if } J = \emptyset \end{cases}$$

which shows that the right hand side is closed under countable unions. □

Adjoining a point to a measurable space as in the previous lemma will be referred to as *augmenting* the space with the point Δ.

Definition 13.16. A *Markov family* is a probability space $(\Omega, \mathcal{A})$ with a distinguished point ω_Δ, a time scale $T \subset [0, \infty]$ with $\infty \in T$, a filtration $\mathcal{F}_t$, a measurable $\theta_t : \Omega \to \Omega$ such that $\theta_\infty \equiv \omega_\Delta$, a state space $(S, \mathcal{S})$ augmented with a point Δ, an $\mathcal{F}$-adapted stochastic process X_t with time scale T and state space S^Δ and a family of probability measures $\mathbf{P}_x$ on $(\Omega, \mathcal{A})$ for $x \in S^\Delta$ such that

(i) $X_t(\omega_\Delta) \equiv \Delta$ for all $t \in T$
(ii) $X_t(\omega) = \Delta$ for some $t \in T$ and $\omega \in \Omega$ then $X_u(\omega) = \Delta$ for all $u \geq t$.
(iii) $X_\infty(\omega) = \Delta$ for all $\omega \in \Omega$.
(iv) $\mathbf{P}_\Delta\{X_0 = \Delta\} = 1$
(v) For all $t \in T \setminus \{\infty\}$, $A \in \mathcal{S}$, $\mathbf{P}_x\{X_t \in A\}$ is $\mathcal{S}$-measurable (i.e. $\mathbf{P}_x \circ X_t^{-1}$ is a kernel $S \times \mathcal{S} \to [0,1]$ for all $t \in T \setminus \{\infty\}$).
(vi) For all $s, t \in T$, $A \in \mathcal{S}^\Delta$ and $x \in S^\Delta$
$$\mathbf{P}_x\{X_t \circ \theta_s \in A \mid \mathcal{F}_s\} = \mathbf{P}_{X_s}\{X_t \in A\}$$

If in addition we have $X_t \circ \theta_s = X_{t+s}$ for all $t, s \in T$ then we say that the Markov family is *time-homogeneous.*

It is worth calling out some subtle points of the defintion. One of the more significant subtleties is the fact that we do not require that $\mathbf{P}_x : S^\Delta \times \mathcal{A} \to [0,1]$ is a kernel; while for fixed $x \in S^\Delta$ we know that $\mathbf{P}_x$ is a probability measure we do not assume that for fixed $A \in \mathcal{A}$ we have $\mathbf{P}_x\{A\}$ is $\mathcal{S}^\Delta$-measurable. This issue will be discussed in some detail later on. The other most notable issue is that our guiding intuition is that under the probability measure $\mathbf{P}_x$ X_t is a Markov process starting at $x \in S$, yet we haven't stated that clearly as part of the definition. At least the fact that X_t is a Markov process under $\mathbf{P}_x$ X_t can be proven.

Proposition 13.17. *Let $(\Omega, \mathcal{A}, \mathcal{F}_t, X_t, \theta_t, \boldsymbol{P}_x)$ be a Markov family then*

(i) *For all $t \in T$, $A \in \mathcal{S}^\Delta$, $\boldsymbol{P}_x\{X_t \in A\}$ is $\mathcal{S}^\Delta$-measurable (i.e. $\boldsymbol{P}_x \circ X_t^{-1}$ is a kernel $S^\Delta \times \mathcal{S}^\Delta \to [0,1]$ for all $t \in T$).*
(ii) *If X_t is time homogeneous then θ_t is $\mathcal{F}^X_\infty / \mathcal{F}^X_\infty$-measurable.*
(ii) *If X_t is time homogeneous then $\mu_t(x, A) = \boldsymbol{P}_x(X_t \in A)$ defines a Markov transition kernel and the pair X_t, μ_t is a Markov process under $\boldsymbol{P}_x$.*

Proof. To see (i) note that for $T = \infty$ we have for all $A \in \mathcal{S}^\Delta$ and $x \in S^\Delta$

$$\mathbf{P}_x\{X_\infty \in A\} = \mathbf{P}_x\{\Delta \in A\} = \begin{cases} 1 & \text{if } \Delta \in A \\ 0 & \text{if } \Delta \notin A \end{cases}$$

so is a constant function of x and therefore $\mathcal{S}^\Delta$-measurable. For $t \in T$ with $t \neq \infty$ and $A \in \mathcal{S}$ then as a function of S^Δ,

$$\mathbf{P}_x\{X_t \in A\} = \mathbf{1}_S(x)\mathbf{P}_x\{X_t \in A\} + \mathbf{1}_\Delta(x)\mathbf{P}_\Delta\{X_t \in A\}$$

which is $\mathcal{S}^\Delta$-measurable since $S, \{\Delta\} \in \mathcal{S}^\Delta$ and $\mathbf{P}_x\{X_t \in A\}$ is $\mathcal{S}$-measurable.

For $t \in T$ with $t \neq \infty$ and $B \in \mathcal{S}^\Delta \setminus \mathcal{S}$ by Lemma 13.15 we know that $B = A \cup \{\Delta\}$ for some $A \in \mathcal{S}$ and thus

$$\mathbf{P}_x\{X_t \in A \cup \{\Delta\}\} = \mathbf{P}_x\{X_t \in A\} + \mathbf{P}_x\{X_t = \Delta\} = \mathbf{P}_x\{X_t \in A\} + 1 - \mathbf{P}_x\{X_t \in S\}$$

is seen to be $\mathcal{S}^\Delta$-measurable.

To see (ii) note that by time homogeneity we have for $s, t \in T$ and $A \in \mathcal{S}^\Delta$, $\theta_s^{-1} X_t^{-1}(A) = X_{t+s}^{-1}(A)$ which shows that θ_s is $\mathcal{F}_t^X / \mathcal{F}_{t+s}^X$-measurable. In particular, (ii) follows.

To see (iii) we first show that $\mu_t(x, A) = \mathbf{P}_x\{X_t \in A\}$ is a Markov transition kernel. By (i) we have shown that it is a kernel from S^Δ to $\mathcal{P}(S^\Delta)$ so it remains to show the Chapman Kolmogorov relations. Let $t, s \in T$, $x \in S^\Delta$ and $A \in \mathcal{S}^\Delta$ then by the tower property of conditional expectation, property (vi) of the defintion of a Markov family and the Expectation Rule Lemma 3.7 (applied to $\mathbf{P}_X \circ X_s^{-1} = \mu_s(x, \cdot)$)

$$\begin{aligned}
\mu_{s+t}(x, A) &= \mathbf{P}_x\{X_{s+t} \in A\} \\
&= \mathbf{E}_x\left[\mathbf{P}_x\{X_{t+s} \in A \mid \mathcal{F}_s\}\right] \\
&= \mathbf{E}_x\left[\mathbf{P}_{X_s}\{X_t \in A\}\right] \\
&= \mathbf{E}_x\left[\mu_t(X_s, A)\right] \\
&= \int \mu_t(u, A) \mu_s(x, du) = \mu_s \mu_t(x, A)
\end{aligned}$$

showing the Chapman-Kolmogorov relations. Now property (vi) of the definition of a Markov family shows $\mathbf{P}_x\{X_{t+s} \in A \mid \mathcal{F}_s\} = \mu_t(X_s, A)$ and therefore X_t is a Markov process under $\mathbf{P}_x$ for every $x \in S^\Delta$. TODO: A non-homogeneous version of this proof and the proper definition of a non-homogeneous Markov family???? □

Note that Bass doesn't require that $\mathbf{P}_x$ is a kernel $S \to \mathcal{P}(\Omega)$ rather he only requires that $\mathbf{P}_x \circ X^{-1}$ is a kernel from S to $\mathcal{P}(S^T)$ (equivalently for each $t \in T$ and $A \in \mathcal{S}$ we have $\mathbf{P}_x\{X_t \in A\}$ is a measurable function of x or again equivalently $\mathbf{P}_x$ is a kernel on the natural filtration $\mathcal{F}_\infty^X$ (this is the same as Blumenthal and Getoor as I mention above). What I don't know if whether a universally measurable kernel $S \to \mathcal{P}(\Omega)$ is necessarily Borel measurable when restricted to $\mathcal{F}_\infty^X$; heck this isn't necessarily a well posed question since S is not assumed to be a topological space at this point. It is worth noting that Blumenthal and Getoor do show that every Markov family extends to a completed one with the state space given by the universal completion and they assume this completion is in place for much of the subsequent theory. That said, they don't modify the defintion of Markov family in doing so.

TODO: Question: Given a Markov family as above then given an arbitrary initial distribution ν on S we can define $\mathbf{P}_\nu$ by $\mathbf{P}_\nu\{A\} = \int \mathbf{P}_x\{A\}\, d\nu(x)$. Is X a Markov process with initial distribution ν under $\mathbf{P}_\nu$? Blumenthal and Getoor do this but alas universal measurability arises here as well (not when defining $\mathbf{P}_\nu\{A\}$ for $A \in \sigma(\bigvee_{t \in T} X_t)$ but only when trying to extend to $\mathcal{F}_\infty$)!

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 13.18. 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 13.19. *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 3.7) to see that

$$\int f(x+y)\,\mu(z, dx) = \int f(x)\,d\left[\mu(z) \circ t_y^{-1}\right] = \int f(x)\,\mu(y+z, dx)$$

$\square$

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 13.20. 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 13.21. *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 13.13 that $\mathbf{P}_x\{(X_{t_1}, \dots, X_{t_n}) \in A\} = \mu_{0,t_1} \otimes \dots \otimes \mu_{t_{n-1},t_n}(x, A)$, 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 13.19

$$\begin{aligned}
&\mathbf{P}_x\{(X_{t_1}, \ldots, X_{t_n}) \in A\} \\
&= \mu_{0,t_1} \otimes \ldots \otimes \mu_{t_{n-1},t_n}(x, A) \\
&= \int \int \mathbf{1}_A(x_1, x_2, \ldots, x_n) \mu_{t_1,t_2} \otimes \cdots \otimes \mu_{t_{n-1},t_n}(x_1, dx_2, \ldots, dx_n)\, \mu_{0,t_1}(x, dy) \\
&= \int \int \mathbf{1}_A(x_1 + x, x_2, \ldots, x_n) \mu_{t_1,t_2} \otimes \cdots \otimes \mu_{t_{n-1},t_n}(x_1, dx_2, \ldots, dx_n)\, \mu_{0,t_1}(0, dy) \\
&= \int \int \mathbf{1}_{A-x}(x_1, x_2, \ldots, x_n) \mu_{t_1,t_2} \otimes \cdots \otimes \mu_{t_{n-1},t_n}(x_1, dx_2, \ldots, dx_n)\, \mu_{0,t_1}(0, dy) \\
&= \mu_{0,t_1} \otimes \ldots \otimes \mu_{t_{n-1},t_n}(0, A - x) \\
&= \mathbf{P}_0\{(X_{t_1}, \ldots, X_{t_n}) \in A - x\}
\end{aligned}$$

Now we complete the result by a monotone class argument. We know that sets of the form $\{(X_{t_1}, \ldots, X_{t_n}) \in A\}$ are a generating π-system so by the π-λ Theorem (Theorem 2.27) 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 2.30) shows

$$\mathbf{P}_x\{\cup_n A_n\} = \lim_{n\to\infty} \mathbf{P}_x\{A_n\} = \lim_{n\to\infty} \mathbf{P}_0\{A_n - x\} = \mathbf{P}_0\{\cup_n A_n - x\}$$

$\square$

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 13.22. 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 13.23. *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 $\mathbf{P}\{X_t \mid \mathcal{F}_s\}$; 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 $A \in \mathcal{S}$,

$$\begin{aligned} \mathbf{P}\{X_t - X_s \in A \mid \mathcal{F}_s\} &= \int \mathbf{1}_A(x - X_s)\, \mu_{s,t}(X_s, dx) && \text{by Theorem 8.35} \\ &= \int \mathbf{1}_A(x)\, \mu_{s,t}(0, dx) && \text{by Lemma 13.19} \\ &= \mu_{s,t}(0, A) \end{aligned}$$

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 2.87 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 8.35 (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 8.20. 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. $\square$

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 13.24. A family of transition kernels $\mu_{s,t}$ $\mathbb{Z}_+$ or $\mathbb{R}_+$ is said to be *time homogeneous* if and only if there exist a family of kernels $\tilde{\mu}_t$ such that $\mu_{s,t}(x, B) = \tilde{\mu}_{t-s}(x, B)$ for all $x \in S$ and $B \in \mathcal{A}$. A Markov process X is said to be time homogeneous if it has a family of time homogeneous transition kernels.

Proposition 13.25. *If X is a time homogeneous Markov process then for all $s, t, u \in T$ and $B \in \mathcal{S}^T$ we have $\boldsymbol{E}[X_t \in B \mid \mathcal{F}_s] = \boldsymbol{E}[X_{t+u} \in B \mid \mathcal{F}_{s+u}]$.*

Proof. Immediate from the defintions,

$$\mathbf{E}\left[X_t \in B \mid \mathcal{F}_s\right] = \mu_{t-s}(X_s, B) = \mathbf{E}\left[X_{t+u} \in B \mid \mathcal{F}_{s+u}\right]$$

□

3. Strong Markov Property

In dealing with Markov processes we make a lot of use of constructions that involve the following

Definition 13.26. If T is equal to $\mathbb{Z}_+$ or $\mathbb{R}_+$, for each $t \in T$ we define the *shift operator* $\theta_t : S^T \to S^T$ by $\theta_t f(s) = f(s + t)$.

It is clear that for a fixed $t \in T$ the shift operator θ_t is measurable but we often need a stronger property the requires some more assumptions.

Lemma 13.27. *For any fixed $t \in T$ the shift operator $\theta_t : S^T \to S^T$ is measurable. If U is equal to S^∞, $C(T;S)$ or $D(T;S)$, then $\theta_t X$ defines a measurable function $\theta : U \cap S^T \times T \to U \cap S^T$.*

Proof. First let $t \in T$ be fixed pick $s \in T$ and $A \in \mathcal{S}$. Then $\theta_t^{-1}\{f(s) \in A\} = \{f(s+t) \in A\} \in \mathcal{S}^T$. Therefore since sets of the form $\{f(s) \in A\}$ generate $\mathcal{S}^T$, we see that θ_t is measurable by Lemma 2.12.

Now let U be as above. It is clear that the shift operator preserves the necessary continuity and limit properties and thus is well defined as a function $\theta : U \cap S^T \times T \to U \cap S^T$. To see measurability of θ, first note that the evaluation map $\pi : U \cap S^T \times T \to S$ given by $\pi(f,t) = f(t)$ is measurable (e.g. this follows by considering the process defined by the identity $U \cap S^T \to U \cap S^T$ and using Lemma 9.89 to see that it is jointly measurable). Now let $s \in T$ and $A \in \mathcal{S}$ as before and calculate

$$\begin{aligned}\{(f,t) \mid \theta_t f \in \pi_s^{-1} A\} &= \{(f,t) \mid \theta_t f(s) \in A\} = \{(f,t) \mid \theta_s f(t) \in A\} \\ &= (\theta_s, id)^{-1}\{(f,t) \mid f(t) \in A\} = (\theta_s, id)^{-1}\pi^{-1}A\end{aligned}$$

which is measurable by the joint measurability of π noted above and the measurability of θ_s for fixed $s \in T$. □

When considering Markov processes on the canonical space there is a very useful construction of time shifting optional times. Intuitively the construction is that given two optional times σ and τ one constructs the random time which is "the first time τ happens after σ happens". The following Lemma makes the construction precise and shows that under some assumption on the path space that the construction gives us a weak optional time.

Lemma 13.28. *Let S be a metric space and let σ and τ be weakly optional times on any of the canonical spaces S^∞, $C([0,\infty);S)$ or $D([0,\infty);S)$ provided with the canonical filtration $\mathcal{F}$. Then*

$$\gamma = \begin{cases} \sigma + \tau \circ \theta_\sigma & \text{when } \sigma < \infty \\ \infty & \text{when } \sigma = \infty \end{cases}$$

is also weakly $\mathcal{F}$-optional.

Proof. Let X be the canonical process (i.e. X_t is the evaluation function π_t).

First we claim that γ is measurable. This follows by noting that θ_σ is measurable by writing it as $\theta \circ (id, \sigma)$ and using by Lemma 13.27. Therefore γ is measurable by the measurability of θ_σ, σ and τ and application of Lemma 2.13 and Lemma 2.19.

Next we claim that if we pull back $\mathcal{F}_t$ by θ_σ then result should only depend on values X_s for $\sigma \leq s \leq \sigma + t$ hence should be $\mathcal{F}^+_{\sigma+t}$-measurable. We have to be a bit careful with this claim, because σ can be infinite in which case θ_σ isn't defined. To make the claim precise and to prove it pick $n \geq 0$ and note that by either discreteness or by continuity of sample paths together with Lemma 9.89 we know that X is $\mathcal{F}$-progressively measurable. By $\mathcal{F}^+$-optionality of $\sigma \wedge n$ and Lemma 9.90 we know that $X_{\sigma\wedge n+s} = X_s \circ \theta_{\sigma\wedge n}$ is $\mathcal{F}^+_{\sigma\wedge n+s}$-measurable for all $s \geq 0$. Now fix $t \geq 0$ then for $0 \leq s \leq t$, pick a measurable set $B \in \mathcal{S}$ and let $A = \{X_s \in B\}$; we note that $\theta^{-1}_{\sigma\wedge n} A = (X_s \circ \theta_{\sigma\wedge n})^{-1}(B) = X^{-1}_{\sigma\wedge n+s}(B) \in \mathcal{F}^+_{\sigma\wedge n+s} \subset \mathcal{F}^+_{\sigma\wedge n+t}$. Since

$\{A \mid \theta^{-1}_{\sigma\wedge n} A \in \mathcal{F}^+_{\sigma\wedge n+t}\}$ is a σ-algebra (Lemma 2.8) and sets of the form $\{X_s \in B\}$ for $0 \leq s \leq t$ generate $\mathcal{F}_t$, we know that $\theta^{-1}_{\sigma\wedge n}\mathcal{F}_t \subset \mathcal{F}^+_{\sigma\wedge n+t}$ for all $t \geq 0$ and $n \geq 0$.

Now fix $0 \leq t < \infty$, let $n = \lfloor t \rfloor + 1$ and note that

$$\begin{aligned}
\{\gamma < t\} &= \cup_{\substack{0<r<t \\ r\in\mathbb{Q}}} \{\sigma < r; \tau \circ \theta_\sigma < t - r\} \\
&= \cup_{\substack{0<r<t \\ r\in\mathbb{Q}}} \{\sigma \wedge n < r; \tau \circ \theta_{\sigma\wedge n} < t - r\}
\end{aligned}$$

Since τ is weakly $\mathcal{F}$-optional we know that $\{\tau < t - r\} \in \mathcal{F}_{t-r}$ hence $\theta^{-1}_{\sigma\wedge n}\{\tau < t-r\} \in \mathcal{F}^+_{\sigma\wedge n+t-r}$ and therefore using Lemma 9.69 applied to the stopped σ-algebra $\mathcal{F}^+_{\sigma\wedge n+t-r}$ we get

$$\{\sigma \wedge n < r; \tau \circ \theta_{\sigma\wedge n} < t - r\} = \{\sigma \wedge n + t - r < t\} \cap \theta^{-1}_{\sigma\wedge n}\{\tau < t - r\} \in \mathcal{F}_t$$

and therefore γ is weakly $\mathcal{F}$-optional. ∎

Note: Kallenberg's proof of the above Lemma is a little bit different and from what I can tell has a small error. He first proves the result for σ bounded, and then claims that $\gamma_n = \sigma \wedge n + \tau \circ \theta_{\sigma\wedge n} \uparrow \gamma$ enabling us to apply the result for the bounded case to γ_n and to conclude that $\gamma = \sup_n \gamma_n$ is weakly $\mathcal{F}$-optional via Lemma 9.72. The problem is that γ_n as defined is not increasing. To see a counter example let $S = \{H, T\}$ and consider the result for S^∞ (here time is $\mathbb{Z}_+$). Define

$$\tau = \min\{n \mid n \text{ is even and } X_n = H\}$$

It is easy to see that τ is a stopping time as

$$\{\tau = n\} = \begin{cases} \{X_0 = H\} & \text{for } n = 0 \\ \{X_n = H\} \cap \{X_{n-2} = T\} \cap \cdots \cap \{X_0 = T\} & \text{if } n \text{ is even and } n > 0 \\ \emptyset & \text{if } n \text{ is odd} \end{cases}$$

Now let σ be a suitably large deterministic time (say $\sigma = 2$) so that for $n \leq 2$ we have $\gamma_n = n + \tau \circ \theta_n$. Consider $\omega = (T, H, H, H, \dots) \in S^\infty$. Note that $\tau(\omega) = 2$ thus $\gamma_0(\omega) = 2$ but $\tau(\theta_1(\omega)) = 0$ and therefore $\gamma_1(\omega) = 1 < \gamma_0(\omega)$.

It is worth noting that even when we are not considering the canonical case many optional times of interest (in particular hitting times) are pull backs of optional times on the path space (i.e. are of the form $\tau \circ X$ where τ is an optional time defined on S^T). If we are given a pair of these optional times then we can apply the time shift construction of the optional times on the path space and pull back (i.e. forming $\sigma \circ X + \tau \circ \theta_{\sigma \circ X} \circ X$). The notation for the non-canonical case is a bit ugly so sometimes we will simply use the notation $\sigma + \tau \circ \theta_\sigma$ as a shorthand.

Theorem 13.29 (Strong Markov Property). *Let X be a time homogeneous Markov process on $\mathbb{Z}_+$ or $\mathbb{R}_+$ and let τ be an optional time with at most countably many values. Then for every measurable $A \subset S^T$,*

$$\boldsymbol{P}\{\theta_\tau X \in A \mid \mathcal{F}_\tau\}(\omega) = \boldsymbol{P}_{X_{\tau(\omega)}}\{A\} \text{ for almost all } \omega \text{ such that } \tau(\omega) < \infty$$

Proof. Before starting on the proof we first need to make some remarks about the well-definedness of the quantities in the result. Specifically we have not defined $\theta_\tau X$ nor $\mathbf{P}_{X_\tau}\{A\}$ when $\tau = \infty$ but neither have we assumed that τ is almost surely finite. The first point is that we can extend $\mathbf{P}_{X_\tau}\{A\}$ can be defined to be an arbitrary value on $\{\tau = \infty\}$ without affecting the values of $\mathbf{P}_{X_\tau}\{A\}$ on $\{\tau < \infty\}$ hence the assertion of the result. By locality of conditional expectation (Lemma 8.14) and the $\mathcal{F}_\tau$-measurability of τ (Lemma 9.27) we can define $\theta_\tau X$ arbitrarily on $\{\tau = \infty\}$ without affecting the values of $\mathbf{P}\{\theta_\tau X \in A \mid \mathcal{F}_\tau\}$ on $\{\tau < \infty\}$ hence the assertion of the result. Therefore the result makes sense assuming that such extensions have made and is independent of the extensions chosen.

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, $t_1 \leq \cdots \leq t_n$, $B \in \mathcal{S}^{\otimes n}$, $A = (\pi_{t_1}, \dots, \pi_{t_n})^{-1}(B)$ and calculate using Lemma 13.4, time homogeneity and the proof of Lemma 13.13

$$\begin{aligned}
\mathbf{P}\{\theta_t X \in A \mid \mathcal{F}_t\} &= \mathbf{P}\{((\theta_t X)_{t_1}, \dots, (\theta_t X)_{t_n}) \in B \mid \mathcal{F}_t\} \\
&= \mathbf{P}\{(X_{t+t_1}, \dots, X_{t+t_n}) \in B \mid \mathcal{F}_t\} \\
&= \mu_{t,t+t_1} \otimes \cdots \otimes \mu_{t+t_{n-1},t+t_n}(X_t, B) \\
&= \mu_{0,t_1} \otimes \cdots \otimes \mu_{t_{n-1},t_n}(X_t, B) \\
&= \mathbf{P}_{X_t}\{A\}
\end{aligned}$$

Now we know that sets of the form $(\pi_{t_1}, \dots, \pi_{t_n})^{-1}(B)$ are a generating π-system for the σ-algebra $\mathcal{S}^T$, the full result for deterministic times t follows from a monotone class argument. Specifically, 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 using Monotone Convergence and the result for deterministic times

$$\begin{aligned}
\mathbf{E}\left[\mathbf{1}_A(\theta_\tau X); B\right] &= \sum_t \mathbf{E}\left[\mathbf{1}_A(\theta_t X); \{\tau = t\} \cap B\right] \\
&= \sum_t \mathbf{E}\left[\mathbf{P}_{X_t}\{A\}; \{\tau = t\} \cap B\right] \\
&= \mathbf{E}\left[\mathbf{P}_{X_\tau}\{A\}; B\right]
\end{aligned}$$

so the result follows by the definition of conditional expectation.

An alternative argument that extends the case of deterministic times to countable optional times uses the localization of the stopped filtration Lemma 9.31 and the local property of conditional expectations Lemma 8.14. Let t be a value in the range of τ, combining these two results and using the result for deterministic times we know that on the set $\{\tau = t\}$ we have

$$\mathbf{P}\{\theta_\tau X \in A \mid \mathcal{F}_\tau\} = \mathbf{P}\{\theta_t X \in A \mid \mathcal{F}_t\} = \mathbf{P}_{X_t}\{A\} = \mathbf{P}_{X_\tau}\{A\} \text{ a.s.}$$

Let the set where the above inequality fails be called N_t. Since we have assumed the set of values of τ is countable, the union of the N_t is also a null set and the result holds off of this null set.

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 13.32 below where we make this implication explicit). Double check that we don't assume it in the proof above (note that Ethier and Kurtz do make the progressive assumption in their discussion of the strong markov property). $\square$

The key part of the above proof is the computation of finite dimensional distributions as a bridge to lift the simple Markov property $\mathbf{P}\{X_{t+s} \in A \mid \mathcal{F}_s\} = \mu_t(X_s, A) = \mathbf{P}_{X_s}\{X_t \in A\}$ on one dimensional distributions to the full σ-algebra $\mathcal{S}^T$. For an arbitrary optional time τ we have an analogous argument using finite dimensional distributions to show the strong Markov property for the one dimensional case is sufficient to prove the full strong Markov property for that optional time. This result will be used later in the text when we want to show that special classes of Markov processes have the strong Markov property. Note that in this case we are now dealing with arbitrary optional times τ and therefore we must assume the process X is progressive so that X_τ is $\mathcal{F}_\tau$-measurable (Lemma 9.90).

Proposition 13.30. *Let X be a progressive time homogeneous Markov process on $\mathbb{R}_+$ and let τ be a finite optional time. If for every $s, t \geq 0$ and $B \in \mathcal{S}$ we have*

$$\boldsymbol{P}\{X_{\tau+t+s} \in B \mid \mathcal{F}_{\tau+s}\} = \mu_t(X_{\tau+s}, B) \text{ a.s.}$$

then for every measurable $A \subset S^T$,

$$\boldsymbol{P}\{\theta_\tau X \in A \mid \mathcal{F}_\tau\} = \boldsymbol{P}_{X_\tau}\{A\} \text{ a.s.}$$

Proof. The crux of the proof is to show the result for finite dimensional distributions.

Claim 13.30.1. Let $n \in \mathbb{N}$ and $f_1, \ldots, f_n$ be bounded measurable functions from S to $\mathbb{R}$ and let $t_1 \leq \ldots \leq t_n$ then

$$\mathbf{E}\left[\prod_{i=1}^n f_i(X_{\tau+t_i}) \mid \mathcal{F}_\tau\right] = \int \prod_{i=1}^n f_i(u_i)\, \mu_{t_1} \otimes \cdots \otimes \mu_{t_n - t_{n-1}}(X_\tau, du_1, \ldots, du_n)$$

The proof is by induction on n. For $n = 1$ we have by hypothesis

$$\mathbf{P}\{X_{\tau+t} \in \cdot \mid \mathcal{F}_\tau\} = \mu_t(X_\tau, \cdot)$$

and therefore by Theorem 8.35 we have

$$\mathbf{E}\left[f(X_{\tau+t}) \mid \mathcal{F}_\tau\right] = \int f(u)\, \mu_t(X_\tau, du)$$

which is the claim for $n = 1$. For the induction step, assume the result for all $m \in \mathbb{N}$ such that $1 \leq m \leq n$ and let $f_1, \ldots, f_{n+1}$ and $t_1 \leq \cdots \leq t_{n+1}$ be given then we apply the tower and pullout properties of conditional expectation, the induction hypothesis for cases $n = 1$ and n and then the definition of kernel products to see

$$\begin{aligned}
&\mathbf{E}\left[\prod_{i=1}^{n+1} f_i(X_{\tau+t_i}) \mid \mathcal{F}_\tau\right] \\
&= \mathbf{E}\left[\mathbf{E}\left[\prod_{i=1}^{n+1} f_i(X_{\tau+t_i}) \mid \mathcal{F}_{\tau+t_n}\right] \mid \mathcal{F}_\tau\right] \\
&= \mathbf{E}\left[\prod_{i=1}^{n} f_i(X_{\tau+t_i}) \mathbf{E}\left[f_{n+1}(X_{\tau+t_{n+1}}) \mid \mathcal{F}_{\tau+t_n}\right] \mid \mathcal{F}_\tau\right] \\
&= \mathbf{E}\left[\prod_{i=1}^{n} f_i(X_{\tau+t_i}) \int f_{n+1}(u_{n+1})\, \mu_{t_{n+1}-t_n}(X_{\tau+t_n}, du_{n+1}) \mid \mathcal{F}_\tau\right] \\
&= \int \prod_{i=1}^{n} f_i(u_i) \left[\int f_{n+1}(u_{n+1})\, \mu_{t_{n+1}-t_n}(u_n, du_{n+1})\right] \mu_{t_1} \otimes \cdots \otimes \mu_{t_n - t_{n-1}}(X_\tau, du_1, \dots, du_n) \\
&= \int \prod_{i=1}^{n+1} f_i(u_i) \mu_{t_1} \otimes \cdots \otimes \mu_{t_{n+1}-t_n}(X_\tau, du_1, \dots, du_{n+1})
\end{aligned}$$

and the claim is proved.

From the claim and the proof of Lemma 13.13 we see that for arbitrary $B_1, \dots B_n \in \mathcal{S}$ and $t_1 \leq \cdots \leq t_n$ if we define $A = (\pi_{t_1}, \dots, \pi_{t_n})^{-1}(B_1 \times \cdots \times B_n)$ (where $\pi_t : S^T \to S$ is the evaluation map $\pi_t f = f(t)$) we have

$$\begin{aligned}
&\mathbf{P}\{\theta_\tau X \in A \mid \mathcal{F}_\tau\} \\
&= \mathbf{P}\{(X_{\tau+t_1}, \dots, X_{\tau+t_n}) \in B_1 \times \cdots \times B_n \mid \mathcal{F}_\tau\} \\
&= \mu_{t_1} \otimes \cdots \otimes \mu_{t_n - t_{n-1}}(X_\tau, B_1 \times \cdots \times B_n) \\
&= \mathbf{P}_{X_\tau}\{A\}
\end{aligned}$$

The set of such A is clearly a π-system and generates $\mathcal{S}^T$ (the latter being generated by the one dimensional $\pi_t^{-1}(B)$ in fact). By montone classes as in Theorem 13.29 we get the result for all $\mathcal{S}^T$. ☐

The previous result that shows how to establish the Strong Markov property for finite optional times is in fact sufficient to establish the Strong Markov property for all optional times by virtue of the following argument.

Proposition 13.31. *Let X be a progressive time homogeneous Markov process on $\mathbb{R}_+$ and suppose that for all finite optional times τ and all measurable sets $A \subset S^T$ we have*

$$\boldsymbol{P}\{\theta_\tau X \in A \mid \mathcal{F}_\tau\} = \boldsymbol{P}_{X_\tau}\{A\} \text{ a.s.}$$

then for all optional times τ and measurable sets $A \subset S^T$ we have

$$\boldsymbol{P}\{\theta_\tau X \in A \mid \mathcal{F}_\tau\} = \boldsymbol{P}_{X_\tau}\{A\} \text{ a.s. on } \tau < \infty$$

Proof. The fact that the terms in the conclusion of the result are in fact well defined see the discussion in Theorem 13.29. Given an arbitrary optional time τ, let $n \in \mathbb{N}$ and note that $\mathcal{F}_\tau = \mathcal{F}_{\tau \wedge n}$, $\theta_\tau X = \theta_{\tau \wedge n}$ and $\mathbf{P}_{X_\tau}\{A\} = \mathbf{P}_{X_{\tau \wedge n}}\{A\}$ on

$\{\tau \leq n\} = \{\tau = \tau \wedge n\}$ (see Proposition 9.32 for the first assertion). Observe that for $t \geq n$,

$$\{\tau \leq n\} \cap \{\tau \wedge n \leq t\} = \{\tau \leq n\} \in \mathcal{F}_n$$

and for $t < n$

$$\{\tau \leq n\} \cap \{\tau \wedge n \leq t\} = \{\tau \leq t\} \in \mathcal{F}_t \subset \mathcal{F}_n$$

Thus $\{\tau \leq n\} \in \mathcal{F}_\tau \cap \mathcal{F}_{\tau \wedge n} = \mathcal{F}_{\tau \wedge n}$ and we may apply localization of conditional expectations Lemma 8.14 to see that on $\{\tau \leq n\}$,

$$\mathbf{P}\{\theta_\tau X \in A \mid \mathcal{F}_\tau\} = \mathbf{P}\{\theta_{\tau \wedge n} X \in A \mid \mathcal{F}_{\tau \wedge n}\} = \mathbf{P}_{X_{\tau \wedge n}}\{A\} = \mathbf{P}_{X_\tau}\{A\} \text{ a.s.}$$

Now write $\{\tau < \infty\} = \cup_{n=1}^\infty \{\tau \leq n\}$ and the the union of a countable number of null sets. $\square$

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 13.23 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 13.32. *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 $X - X_0 \overset{d}{=} \theta_\tau X - X_\tau$

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 8.35 and space homogeneity of $\mathbf{P}_x$ (Lemma 13.21) 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 $\theta_\tau X - X_\tau \overset{d}{=} X - X_0$.

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 8.35 and Lemma 13.21 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. □

Definition 13.33. Let X be a time homogeneous Markov process with transition kernel μ_t we say an initial distribution ν is *invariant* if $\nu\mu_t = \nu$ for all $t \in T$, i.e. we have

$$\int \mu_t(x, A)\,\nu(dx) = \nu(A)$$

for all $t \in T$ and $A \in \mathcal{S}$.

Definition 13.34. Let X be a stochastic process with time scale T then we say X is *stationary* if $\theta_t X \overset{d}{=} X$ for all $t \in T$.

Lemma 13.35. *Let X be a time homogeneous Markov process with transition kernel μ and an invariant initial distribution ν, then X is stationary.*

Proof. Fix $t \in T$, $s_1 < \cdots < s_n$ and $A \in \mathcal{S}^{\otimes n}$, then using Lemma 13.4 and time homogeneity we compute

$$\begin{aligned}\mathbf{P}\{(X_{t+s_1}, \dots, X_{t+s_n}) \in A\} &= \nu_{t+s_1} \otimes \mu_{s_2-s_1} \otimes \cdots \otimes \mu_{s_n-s_{n-1}}(A) \\ &= \nu_{s_1} \otimes \mu_{s_2-s_1} \otimes \cdots \otimes \mu_{s_n-s_{n-1}}(A) = \mathbf{P}\{(X_{s_1}, \dots, X_{s_n}) \in A\}\end{aligned}$$

Since the finite dimensional distributions characterize the distribution of $\theta_t X$ (Lemma 9.6) it follows that X is stationary. □

4. Discrete Time Markov Chains

In this section we discuss the theory of Markov processes on a time scale $\mathbb{Z}_+$. This part of Markov process theory has many applications and we'll be able to construct lots of important examples that both illustrate and motivate the accompanying theory. Moreover much of the theory in discrete time illustrates concerns that are also present in more general cases but with fewer technical distractions.

One of our first concerns is to think about constructing examples of Markov processes. The obvious way to approach this is the way we have done it up until now: specify a transition kernel and an initial distribution. As it turns out, it can be surprisingly difficult to get a handle on the transition kernel of a concrete process and it is desirable to have alternative ways of constructing and characterizing Markov processes. In the discrete time case we can think of a Markov process as a deterministic system that is perturbed by noise (or alternatively a "transduced" noise sequence). We make this precise in the following theorem.

Theorem 13.36. *Let X be a process on time scale $\mathbb{Z}_+$ with a Borel state space S, then X is Markov if and only if there exist a measurable space $(T, \mathcal{T})$, measurable functions $f_1, f_2, \dots : S \times T \to S$ and i.i.d. random elements $\vartheta_1, \vartheta_2, \dots \perp\!\!\!\perp X_0$ such that $X_n = f_n(X_{n-1}, \vartheta_n)$ a.s. for all $n \in \mathbb{N}$. If X is Markov we may find such* *a representation with $T = [0,1]$ and ϑ_n i.i.d. $U(0,1)$ random variables. We may choose $f_1 = f_2 = \cdots$ if and only if X is time homogeneous.*

Proof. First assume that X has the hypothesized representation. Let $\mathcal{F}$ be the filtration generated by X. Let ν be the law of $\vartheta_1, \vartheta_2, \dots$. Pick a random element ϑ with law ν and for $A \in \mathcal{S}$ define $\mu_n(x, A) = \mathbf{P}\{f_n(x, \vartheta) \in A\}$. Note that it follows from a simple induction using $(\vartheta_1, \vartheta_2, \dots) \perp\!\!\!\perp X_0$ and the expression $X_n = f_n(X_{n-1}, \vartheta_n)$ that ϑ_n is independent of X_m for all $m = 0, \dots, n-1$. Therefore $\mathbf{P}\{\mathcal{F}_{n-1} \mid \vartheta_n \in \cdot\} = \mathbf{P}\{\vartheta \in \cdot\} = \nu$ and in particular has a regular version. Furthermore since X_{n-1} is $\mathcal{F}_{n-1}$-measurable we can apply Lemma 4.6 to compute for any $A \in \mathcal{S}$

$$\begin{aligned}\mathbf{P}\{\mathcal{F}_{n-1} \mid X_n \in A\} &= \mathbf{P}\{\mathcal{F}_{n-1} \mid f_n(X_{n-1}, \vartheta_n) \in A\} \\ &= \int \mathbf{1}_{f_n(X_{n-1}, s) \in A}\, \nu(ds) = \mathbf{P}\{f_n(X_{n-1}, \vartheta) \in A\} = \mu_n(X_{n-1}, A)\end{aligned}$$

which shows that X is Markov with transition kernel μ_n (recall in discrete time the Chapman Kolmogorov relations hold identically for free). Note also that $f_1 = f_2 = \dots$ if and only if $\mu_1 = \mu_2 = \dots$ which is to say that X is time homogeneous.

Now let X be Markov. Since S is Borel we may apply Lemma 8.31 to each transition kernel μ_n and construct a measurable function $f_n : S \times [0,1] \to S$ such that for a $U(0,1)$ random variable ϑ we know that $\mathbf{P}\{f_n(s, \vartheta) \in \cdot\} = \mu_n(s, \cdot)$. Let $\tilde{X}_0$ be a random element such that $\tilde{X}_0 \stackrel{d}{=} X_0$ (e.g. just take the identity on $(S, \mathcal{S})$ provided with the probability measure $\mathcal{L}(X_0)$). By extending the probability space of $\tilde{X}_0$ if necessary we can assume the existence of i.i.d. $U(0,1)$ random variables $\tilde{\vartheta}_1, \tilde{\vartheta}_2, \dots$. Recursively define $\tilde{X}_n = f_n(\tilde{X}_{n-1}, \tilde{\vartheta}_n)$ for $n \in \mathbb{N}$ and apply the first part of this theorem to conclude that $\tilde{X}$ is a Markov process with transition kernels μ_n and initial distribution $\mathcal{L}(\tilde{X}_0) = \mathcal{L}(X_0)$. We now apply Lemma 13.4 to conclude that the of $X \stackrel{f.d.d.}{=} \tilde{X}$ and thus $X \stackrel{d}{=} \tilde{X}$ by Lemma 9.6. Now since $[0,1]^\infty$ is a Borel space we may apply Lemma 8.40 to conclude there are random variables $\vartheta_1, \vartheta_2, \dots$ such that $(X, (\vartheta_1, \vartheta_2, \dots)) \stackrel{d}{=} (\tilde{X}, (\tilde{\vartheta}_1, \tilde{\vartheta}_2, \dots))$. By considering marginal distributions we conclude that $\vartheta_1, \vartheta_2, \dots$ are i.i.d. $U(0,1)$ and that $(\vartheta_1, \vartheta_2, \dots) \perp\!\!\!\perp X_0$. Also using $(X, (\vartheta_1, \vartheta_2, \dots)) \stackrel{d}{=} (\tilde{X}, (\tilde{\vartheta}_1, \tilde{\vartheta}_2, \dots))$, the measurability of the diagonal $\Delta \subset S \times S$ and the definition of $\tilde{X}$ we conclude that for each $n \in \mathbb{N}$

$$\begin{aligned}\mathbf{P}\{X_n = f_n(X_{n-1}, \vartheta_n)\} &= \mathbf{P}\{(X_n, f_n(X_{n-1}, \vartheta_n)) \in \Delta\} \\ &= \mathbf{P}\{(\tilde{X}_n, f_n(\tilde{X}_{n-1}, \tilde{\vartheta}_n)) \in \Delta\} = 1\end{aligned}$$

and we are done. ∎

The representation of a Markov process as in the preceeding theorem is refered to as a *random mapping representation* and we'll soon put it use in constructing examples of Markov processes.

We proceed to study the special subclass of time homogenous Markov processes with time scale $\mathbb{Z}_+$. A further important specialization occurs when the state space S is countable or finite. We establish some terminology while recording the definitions.

Definition 13.37. A time homogeneous Markov process X with time scale $\mathbb{Z}_+$, transition kernels μ_n and state space S is called a *discrete time Markov process*. Furthermore,

(i) If S countable then X is a *discrete time Markov chain*
(ii) If S is finite then X is a *finite discrete time Markov chain*
(iii) For each $y \in S$ we let $\tau_y^+ = \inf\{n \in \mathbb{N} \mid X_n = y\}$ and then recursively define the *return times*

$$\begin{aligned}\tau_y^0 &= 0 \\ \tau_y^{k+1} &= \tau_y^k + \tau_y^+ \circ \theta_{\tau_y^k} \text{ for } k \in \mathbb{Z}_+\end{aligned}$$

(iv) For each $y \in S$ we define the *occupation time* at $y \in S$ as

$$\kappa_y = \sup\{k \in \mathbb{Z}_+ \mid \tau_y^k < \infty\}$$

(v) For each $x, y \in S$ we define the *hitting probability*

$$r_{xy} = \mathbf{P}_x\{\tau_y^+ < \infty\} = \mathbf{P}_x\{\kappa_y > 0\}$$

(vi) For each $x, y \in S$ and we define the *transition probabilities*

$$p_{xy} = \mu_1(x, \{y\}) \qquad p_{xy}^n = \mu_n(x, \{y\}) \text{ for } n \in \mathbb{N}$$

For the case of a discrete time Markov chain, the transition probabilities p_{xy} characterize the transition kernels and recall from Example 8.30 it is convenient to interpret the p_{xy} as being the entries of a *transition matrix* we shall call p. Moreover from Example 8.30 and the Chapman Kolmogorov relations we have

$$\mu_n(x, \{y\}) = \mu_1^n(x, \{y\}) = p_{xy}^n$$

where in the last equality we are taking the $(x, y)^{th}$ entry of the n-fold product of the matrix p. This explains the use of the notation in the above definition of transition probabilites and also shows that for Markov chains the notation is consistent with transition matrix point of view. We emphasize that p_{xy}^n does not signify p_{xy} raised to the n^{th} power!

We have two initial goals in our study of Markov chains. The first is to develop a little macroscopic structure theory of Markov chains.

PROPOSITION 13.38. *Let X be a discrete time Markov process on state space S. Then for $y \in S$ we have*

$$\kappa_y = \sum_{n=1}^{\infty} \mathbf{1}_{X_n = y}$$

Moreover for all $x, y \in S$ and $n \in \mathbb{N}$,

$$\boldsymbol{P}_x\{\kappa_y \geq n\} = \boldsymbol{P}_x\{\tau_y^n < \infty\} = r_{xy} r_{yy}^{n-1}$$

If $r_{xy} = 0$ then $\kappa_y = 0$ P_x-almost surely, if $r_{xy} > 0$ and $r_{yy} = 1$ then $\boldsymbol{P}_x\{\kappa_y = \infty\} = r_{xy} > 0$, otherwise κ_y is integrable with expectation

$$\boldsymbol{E}_x\left[\kappa_y\right] = \frac{r_{xx}}{1 - r_{yy}} = \sum_{n=1}^{\infty} p_{xy}^n$$

PROOF. First to see that $\kappa_y = \sum_{n=1}^{\infty} \mathbf{1}_{X_n = y}$, simply note that both represent the number of times that X visits y.

To see that $\mathbf{P}_x\{\kappa_y \geq n\} = \mathbf{P}_x\{\tau_y^n < \infty\}$ simply note that equality holds at the level of events: $\{\kappa_y \geq n\}$ if and only if $\{\tau_y^n < \infty\}$. Since $\tau_y^{n+1} < \infty$ if and only if $\tau_y^n < \infty$ and $\theta_{\tau_y^n} \circ \tau_y^+ < \infty$ we can use the Strong Markov property to calculate

$$\begin{aligned}
\mathbf{P}_x\{\tau_y^{n+1} < \infty\} &= \mathbf{P}_x\{\tau_y^n < \infty; \theta_{\tau_y^n} \circ \tau_y^+ < \infty\} \\
&= \mathbf{E}_x\left[\tau_y^n < \infty; \mathbf{P}\{\theta_{\tau_y^n} \circ \tau_y^+ < \infty \mid \mathcal{F}_{\tau_y^n}\}\right] \\
&= \mathbf{E}_x\left[\tau_y^n < \infty; \mathbf{P}_y\{\tau_y^+ < \infty\}\right] = \mathbf{P}_x\{\tau_y^n < \infty\}\mathbf{P}_y\{\tau_y^+ < \infty\}
\end{aligned}$$

which we use in an induction argument to get $\mathbf{P}_x\{\tau_y^n < \infty\} = r_{xy} r_{yy}^{n-1}$.

Now we apply this fact along with Lemma 3.8 to see that

$$\mathbf{E}_x\left[\kappa_y\right] = \sum_{n=1}^{\infty} \mathbf{P}_x\{\kappa_y \geq n\} = r_{xy} \sum_{n=1}^{\infty} r_{yy}^{n-1} = \frac{r_{xy}}{1 - r_{yy}}$$

The rest of the statements in the proposition are trivial consequences of what we have proven. □

By virtue of this result we can see that for every $x \in S$ there is a dichotomy: either we have $r_{xx} = 1$ in which case $\kappa_x = \infty$ P_x-a.s. (almost surely X returns to x infinitely many times) or $0 \leq r_{xx} < 1$ in which case the number of times that X returns to x has finite expectation $\frac{r_{xx}}{1-r_{xx}}$. This attribute of states is worthy of a definition.

Definition 13.39. Let X be a discrete time Markov process on state space S, we say a state $x \in S$ is *recurrent* if and only if X returns to x infinitely many times P_x-a.s. We say $x \in S$ is *transient* if and only if X returns to x only finitely many times P_x-a.s.

The theory of Markov processes tends to be concerned with long term behavior of the process and therefore recurrent states are more important than transient states (just wait long enough and you'll never see a transient state again!) Being able to detect recurrent states is therefore a useful thing to be able to do. A simple and useful criterion can be found when there is an invariant distribution for X.

Proposition 13.40. *Let X be a discrete time Markov process with state space S and assume that an invariant distribution ν exists, then for every $x \in S$ if $\nu(x) > 0$ it follows that x is recurrent.*

Proof. Using the invariance of ν we get for every $n \in \mathbb{N}$

$$0 < \nu(x) = \int p_{xy}^n \, \nu(dy)$$

Therefore using the fact that $r_{yx} \leq 1$ for all $x, y \in S$, Proposition 13.38 and Tonelli's Theorem 2.88 we get

$$\frac{1}{1 - r_{xx}} \geq \int \frac{r_{yx}}{1 - r_{xx}} \, \nu(dy) = \int \sum_{n=1}^{\infty} p_{yx}^n \, \nu(dy) = \sum_{n=1}^{\infty} \int p_{yx}^n \, \nu(dy) = \infty$$

and thus it follows that $r_{xx} = 1$. □

Definition 13.41. Let p_{xy}^n be the transition probabilities of a discrete time Markov process on S. The *period* of a state $x \in S$ is

$$d_x = \gcd\{n \in \mathbb{N} \mid p_{xx}^n > 0\}$$

If $d_x = 1$ then we say that the state x is *aperiodic*.

PROPOSITION 13.42. *Let p^n_{xy} be the transition probabilities of a discrete time Markov process on S, if x has period d then there exists an $N > 0$ such that $p^{nd}_{xx} > 0$ for all $n \geq N$.*

PROOF. We need the following number theoretic fact:

LEMMA 13.43. *Let $A \subset \mathbb{Z}_+$ then there exists an integer m_A such that for all $m \geq m_S$ there exist constants $c_1, \dots, c_n \in \mathbb{Z}_+$ and $x_1, \dots, x_n \in A$ such that $m \gcd A = c_1x_1 + \cdots + c_nx_n$.*

PROOF. To prove the lemma we first recall that the greatest common divisor of a set is an integer linear combination of elements of the set.

Claim: For any subset $B \subset \mathbb{Z}_+$ there exist elements $x_1, \dots, x_n \in B$ and constants $c_1, \dots, c_n \in \mathbb{Z}$ such that $\gcd B = c_1x_1 + \cdots + c_nx_n$.

To see this, let g^*_B be smallest element in the set

$$C = \{c_1x_1 + \cdots + c_nx_n > 0 \mid n \in \mathbb{N}, c_1, \dots, c_n \in \mathbb{Z} \text{ and } x_1, \dots, x_n \in B\}$$

Note that g^*_B divides every $x \in B$; for if not then there is an x such that we can write $x = cg^*_B + r$ with $c \in \mathbb{Z}_+$ and $0 < r < g^*_S$ thus $r = x - g^*_B \in C$. Therefore it follows that $\gcd B$ divides g^*_B. On the other hand, since g^*_B is an integer linear combination of a finite number of elements of B it follows that $\gcd B$ divides g^*_B and therefore $\gcd B = g^*_B$.

Claim: For any set $B \subset \mathbb{Z}_+$ there is a finite subset $F \subset B$ such that $\gcd F = \gcd B$.

To see this consider the sequence $g_n = \gcd B \cap \{0, \dots, n\}$. Clearly, g_n is non-increasing and non-negative so there exists an $N > 0$ such that $g_n = g_N$ for all $n \geq N$. It is also clear that g_N divides every element of B since every element of B is in some $B \cap \{0, \dots, n\}$ and it follows by a similar argument that $\gcd S \leq g_N$. Thus $\gcd S = g_N$.

From the previous claim note that it suffices to prove the lemma for finite sets A. To prove the lemma for finite sets we proceed by induction on the cardinality of A.

The result is vacuous for singleton sets so let $A = \{a, b\}$ and let $g = \gcd A$. For every $m \in \mathbb{N}$ we can write $mg = ca + db$ for some $c, d \in \mathbb{Z}$. By replacing c and d by $c + kb$ and $d - ka$ for suitable $k \in \mathbb{Z}$ we may assume that $0 \leq c < b$ as well. Thus in this case, define $m_A = (ab - a - b)/g + 1$ and note that for any $m \geq m_A$ we have

$$mg = ca + db \geq (ab - a - b) + g > ab - a - b$$

with $0 \leq c < b$ which implies

$$(d + 1)b > ab - a - ca \geq 0$$

which in turn implies $d \geq 0$. Thus the result is proven for a two point set.

Now we do induction on the cardinality of A. Suppose the result is proven for all A with cardinality less than or equal to n. Let A be a finite subset of $\mathbb{Z}_+$ with $A = \{a_1, \dots, a_n\}$ and $\gcd A = g_A$. Let $a \in \mathbb{Z}_+ \setminus A$ and note the facts that $\gcd(A \cup \{a\}) = \gcd(\gcd A, a)$ and $\gcd(A \cup \{a\})$ divides $\gcd A$. Define $g = \gcd(A \cup \{a\})$ and

$$m_{A\cup\{a\}} = (m_{\{a,g_A\}}g + m_A g_A)/g$$

and pick any $m \geq m_{A \cup \{a\}}$: trivially we have $mg \geq m_{\{a,g_A\}} g + m_A g_A$. It follows from the fact that $g = \gcd(\gcd A, a)$ that g divides g_A and therefore there is a $\tilde{m} \in \mathbb{Z}_+$ such that $mg - m_A g_A = \tilde{m} g \geq m_{\{a,g_A\}} g$. By the definition of $m_{\{a,g_A\}}$ we know that there are integers $c, d \geq 0$ such that $mg - m_A g_A = ca + dg_A$. Therefore

$$mg = ca + (d + m_A) g_A = ca + \sum_{j=1}^n c_j a_j$$

with suitable $c_1, \ldots c_n \in \mathbb{Z}_+$ and the lemma is proved. $\square$

Now to prove the proposition, let $x \in S$, let $A = \{n \in \mathbb{N} \mid p^n_{xx} > 0\}$ assume that $\gcd A = d$. Applying the lemma we see that there is an $N > 0$ such that for all $n \geq N$, $nd = c_1 n_1 + \cdots + c_k n_k$ for suitable $k \in \mathbb{N}$, $n_1, \ldots, n_k \in A$ and $c_1, \ldots, c_k \in \mathbb{Z}_+$. On the other hand suppose $n, m \in A$ and note that by the Chapman Kolmogorov relations we have

$$\begin{aligned} p^{n+m}_{xx} &= \mu_{n+m}(x, \{x\}) = \mu_n \mu_m(x, \{x\}) = \int \mu_m(y, \{x\}) \, \mu_n(x, dy) \\ &\geq \int \mu_m(y, \{x\}) \mathbf{1}_{x=y} \, \mu_n(x, dy) = \mu_m(x, \{x\}) \mu_n(x, \{x\}) > 0 \end{aligned}$$

which shows that A is closed under addition. It follows that $nd \in A$ and the result is proven. $\square$

Definition 13.44. Let X be a discrete time Markov process with initial distribution ν and transition kernel μ. We say that X is *reversible* if for every non-negative measurable or integrable $f : S \times S \to \mathbb{R}$ we have

$$\begin{aligned} \int f(x,y) \, (\nu \otimes \mu)(dx, dy) &= \iint f(x,y) \, \mu(x, dy) \, \nu(dx) \\ &= \iint f(y,x) \, \mu(x, dy) \, \nu(dx) = \int f(y,x) \, (\nu \otimes \mu)(dx, dy) \end{aligned}$$

There are a couple of immediate consequences of reversibility that follow by looking at the finite dimensional distributions of a reversible X. The first very useful implication is that reversibility implies stationarity.

Proposition 13.45. *Let X be a reversible discrete time Markov process with initial distribution ν and transition kernel μ, then ν is invariant for X.*

Proof. Let $A \in \mathcal{S}$ then using Lemma 13.4 and reversibility

$$\begin{aligned} \mathbf{P}_\nu \circ X^{-1} = \nu\mu(A) = (\nu \otimes \mu)(S \times A) &= \iint \mathbf{1}_A(y) \, \mu(x, dy) \, \nu(dx) \\ &= \iint \mathbf{1}_A(x) \, \mu(x, dy) \, \nu(dx) = \int \mathbf{1}_A(x) \, \nu(dx) = \nu(A) \end{aligned}$$

$\square$

The next implication explains the origin of the term reversible. Prosaically one says that a reversible Markov process looks the same if run backwards.

Proposition 13.46. *Let X be a reversible discrete time Markov process then for all $n, k \geq 0$ and $A \in \mathcal{S}^{\otimes n}$*

$$\boldsymbol{P}\{(X_k, \ldots, X_{n+k}) \in A\} = \boldsymbol{P}\{(X_{n+k}, \ldots, X_k) \in A\}$$

PROOF. Because ν is invariant, it follows that X is a stationary process (Lemma 13.35) and therefore it suffices to prove the result for $k = 0$. In fact we prove a bit more; we show that

(14)

$$\int f(x_0, \ldots, x_n)\nu \otimes \mu^{\otimes n}(dx_0, \ldots, dx_n) = \int f(x_n, \ldots, x_0)\nu \otimes \mu^{\otimes n}(dx_0, \ldots, dx_n)$$

for all $n \in \mathbb{N}$ and all non-negative measurable functions $f : S^{n+1} \to [0, \infty)$. By Lemma 13.4 the current result follows from (4). The proof is by induction on n with the case $n = 1$ being part of the definition of reversibility.

Now supposing the result is true for $n - 1$, we use Lemma 13.4, Tonelli's Theorem and two applications of the induction hypothesis

$$\begin{aligned}
&\int f(s_0, \ldots, s_n)\, \mu(s_{n-1}, ds_n) \cdots \mu(s_0, ds_1)\nu(ds_0) \\
&= \int \left[\int f(s_0, \ldots, s_n)\, \mu(s_{n-1}, ds_n)\right] \mu(s_{n-2}, ds_{n-1}) \cdots \mu(s_0, ds_1)\nu(ds_0) \\
&= \int f(s_{n-1}, \ldots, s_0, s_n)\, \mu(s_0, ds_n)\mu(s_{n-2}, ds_{n-1}) \cdots \mu(s_0, ds_1)\nu(ds_0) \\
&= \int \left[\int f(s_{n-1}, \ldots, s_0, s_n)\, \mu(s_{n-2}, ds_{n-1}) \cdots \mu(s_0, ds_1)\right] \mu(s_0, ds_n)\nu(ds_0) \\
&= \int f(s_{n-1}, \ldots, s_n, s_0)\, \mu(s_{n-2}, ds_{n-1}) \cdots \mu(s_n, ds_1)\mu(s_0, ds_n)\nu(ds_0) \\
&= \int f(t_n, \ldots, t_1, t_0)\, \mu(t_{n-1}, dt_n) \cdots \mu(t_0, t_1)\nu(dt_0)
\end{aligned}$$

where in the last line we have defined new integration variables $t_0 = s_0$, $t_1 = s_n$ and $t_k = s_{k-1}$ for $2 \leq k \leq n$. ☐

We now make the transition to discussing discrete time Markov chains (that is to say we restrict ourselves to countable state spaces.

Recurrence is a somewhat contagious property; if you start with a recurrent state x and can reach a state y from x with positive probability then it will follow that y is recurrent. Intuitively this can be seen by making the following observations:

- If x is recurrent and I can reach y from x with positive probability then I must be able to reach x from y with positive probability; otherwise with positive probability x reaches y (returning to itself only a finite number of times on the way) and then never again returns to itself contradicting recurrence.
- One way for y to return to itself is to first travel to x, then return to itself some number of times, then to make the return trip from x to y; since x is recurrent with positive probability this may be done in infinitely many ways hence y is also recurrent.

These facts and a few more are captured less prosaically in the following lemma.

LEMMA 13.47. *Let X be a discrete time Markov chain with state space S, let $x \in S$ be recurrent and define $S_x = \{y \in S \mid r_{xy} > 0\}$. Then for all $y \in S_x$, it follows that y is recurrent and for every $y, z \in S_x$ we have $r_{yz} = 1$.*

PROOF. We first handle the case of showing that $r_{yx} = 1$. For this, we use a union bound, the Strong Markov property and the fact that $X_{\tau_y^+} = y$ on $\{\tau_y^+ < \infty\}$ to see

$$\begin{aligned} 0 &= \mathbf{P}_x\{\tau_x^+ = \infty\} \geq \mathbf{P}_x\{\tau_y^+ < \infty; \theta_{\tau_y^+} \circ \tau_x^+ = \infty\} \\ &= \mathbf{E}_x\left[\tau_y^+ < \infty; \mathbf{P}\{\theta_{\tau_y^+} \circ \tau_x^+ = \infty \mid \mathcal{F}_{\tau_y^+}\}\right] \\ &= \mathbf{P}_x\{\tau_y^+ < \infty; \mathbf{P}_y\{\tau_x^+ = \infty\}\} = \mathbf{P}_x\{\tau_y^+ < \infty\}\mathbf{P}_y\{\tau_x^+ = \infty\} = r_{xy}(1 - r_{yx}) \end{aligned}$$

which implies $r_{yx} = 1$ since we assumed $r_{xy} > 0$.

Now we turn to the task of showing that all $y \in S_x$ are recurrent. We know that $r_{xy} > 0$ and $r_{yx} > 0$ and therefore there exist $m, n \in \mathbb{N}$ such that $p_{xy}^n > 0$ and $p_{yx}^m > 0$. Thus, by Proposition 13.38 and two applications of the Chapman Kolmogorov relations and the recurrence of x we get

$$\begin{aligned} \mathbf{E}_y\left[\kappa_y\right] &= \sum_{j=1}^\infty p_{yy}^j \geq \sum_{j=1}^\infty p_{yy}^{j+m+n} = \sum_{j=1}^\infty \sum_{z\in S}\sum_{w\in S} p_{yz}^m p_{zw}^j p_{wy}^n \\ &\geq \sum_{j=1}^\infty p_{yx}^m p_{xx}^j p_{xy}^n = \infty \end{aligned}$$

which implies that y is recurrent. Knowing that y is recurrent and having already shown that $r_{yx} = 1 > 0$, we know that $x \in S_y$ and we can apply the first argument in the proof to conclude that $r_{xy} = 1$ as well.

Lastly let $y, z \in S_x$. We use the fact that one way for X to get from y to z is by passing through x first. Formally we use a union bound and the Strong Markov Property to see

$$\begin{aligned} r_{yz} &= \mathbf{P}_y\{\tau_z^+ < \infty\} \geq \mathbf{P}_y\{\tau_x^+ < \infty; \theta_{\tau_x^+} \circ \tau_z^+ < \infty\} \\ &= \mathbf{E}_y\left[\tau_x^+ < \infty; \mathbf{P}\{\theta_{\tau_x^+} \circ \tau_z^+ < \infty \mid \mathcal{F}_{\tau_x^+}\}\right] \\ &= \mathbf{P}_y\{\tau_x^+ < \infty\}\mathbf{P}_x\{\tau_z^+ < \infty\} = r_{yx}r_{xz} = 1 \end{aligned}$$

which shows us that $r_{yz} = 1$. □

DEFINITION 13.48. Let X be a discrete time Markov chain with state space S then we say that X is *irreducible* if $r_{xy} > 0$ for all $x, y \in S$. If X is not irreducible we say that X is *reducible*.

There are generalizations of the notion of irreducibility to the general discrete time Markov process case but they will be dealt with later; the countable state space case is historically the first to be handled and provides important motivation while avoid some subtle points. The first thing is to record some alternative characterizations of irreducibility; in the sequel we'll fell free to use these equivalences without explicit mention. They are all just slightly different ways of capturing the notion that a Markov chain is irreducible if it is possible for the chain to reach any part of state space regardless of the starting point.

PROPOSITION 13.49. *Let X be a discrete time Markov chain with state space S then X is irreducible if and only if for every $x, y \in S$ there exists $n \in \mathbb{N}$ such that $p_{xy}^n > 0$.*

PROOF. Suppose X is irreducible and let $x, y \in S$; it follows that $\mathbf{P}_x\{\tau_y^+ < \infty\} > 0$. Writing $\mathbf{P}_x\{\tau_y^+ < \infty\} = \cup_{n=1}^{\infty} \mathbf{P}_x\{\tau_y^+ = n\}$ we conclude there is an $n \in \mathbb{N}$ such that $\mathbf{P}_x\{\tau_y^+ = n\} > 0$. Now observe that by a union bound

$$0 < \mathbf{P}_x\{\tau_y^+ = n\} \leq \mathbf{P}_x\{X_n = y\} = p_{xy}^n$$

On the other hand suppose that $p_{xy}^n > 0$. Then we know that

$$\{X_n = y\} \subset \{\tau_y^+ \leq n\} \subset \{\tau_y^+ < \infty\}$$

and therefore $0 < p_{xy}^n \leq \mathbf{P}_x\{\tau_y^+ < \infty\}$. □

PROPOSITION 13.50. *Let X be an irreducible discrete time Markov chain, then*

(i) *Either every $x \in S$ is transient or every $x \in S$ is recurrent. Moreover $r_{xy} = 1$ for every $x, y \in S$.*
(ii) *Every $x \in S$ has the same period*
(iii) *If ν is an invariant distribution then $\nu(x) > 0$ for every $x \in S$.*

PROOF. Property (i) is an immediate consequence of Lemma 13.47 since for irreducible X we have $S = S_x$ for any $x \in S$.

To see (ii), let $x, y \in S$ and pick $m, n \in \mathbb{N}$ such that $p_{xy}^n > 0$ and $p_{yx}^m > 0$. Now by the Chapman Kolmogorov relations we see that for all $j \in \mathbb{Z}_+$

$$p_{yy}^{j+m+n} = \sum_{z \in S} \sum_{w \in S} p_{yz}^m p_{zw}^j p_{wy}^n \geq p_{yx}^m p_{xx}^j p_{xy}^n$$

If we choose $j = 0$ then we get inequality $p_{yy}^{m+n} \geq p_{yx}^m p_{xy}^n > 0$ which implies that d_y divides $m + n$. With this fact in hand, we see that for $j > 0$ for which $p_{xx}^j > 0$ it follows that $p_{yy}^{j+m+n} > 0$ and therefore d_y divides j as well. By definition of the period we then get $d_y \leq d_x$. The argument we just made is symmetric in x and y so the opposite inequality holds as well and we conclude that $d_x = d_y$.

To see (iii), suppose that ν is an invariant distribution and pick an $x \in S$ such that $\nu(x) > 0$. If we let $y \in S$ by irreducibility we find $n > 0$ such that $p_{xy}^n > 0$ and by invariance of ν we get

$$\nu(y) = \sum_{x \in S} p_{xy}^n \nu(x) \geq \nu(x) p_{xy}^n > 0$$

□

We now move to the theorem that gives us a useful criterion for the existence of an invariant distribution for a discrete time Markov chain and also shows that in a strong sense any initial distribution converges to that invariant distribution.

THEOREM 13.51. *Let X be an irreducible and aperiodic discrete time Markov chain with countable state space $(S, \mathcal{S})$. Then exactly one of the following holds*

(i) *There exists a unique invariant distribution ν for which $\nu(x) > 0$ for all $x \in S$ and moreover for every initial distribution μ we have*

$$\lim_{n \to \infty} \sup_{A \in \mathcal{S}^\infty} \left| \boldsymbol{P}_\mu \circ \theta_n^{-1}\{A\} - \boldsymbol{P}_\nu\{A\} \right| = 0 \tag{15}$$

(ii) *An invariant distribution does not exist and*

$$\lim_{n \to \infty} p_{xy}^n = 0 \textit{ for all } x, y \in S$$

The proof breaks down is a few different lemmas. The proof technique used here is referred to as a *coupling* argument; it will reappear with increasing levels of sophistication later in this book. The common thread in coupling arguments is the construction of a joint distribution on a product space (called a *coupling*) and its use to compare a process under study to one with simpler properties.

The first part of the coupling argument is the construction of the process on the product space. In this case a pair of independent Markov chains suffices but we need a few details of about such products of Markov chains to execute the coupling argument.

LEMMA 13.52. *Let X and Y be independent discrete time Markov chains with state space S and T and transition matrices p_{xy} and q_{xy} respectively. Then (X,Y) is an irreducible discrete Markov chain with state space $S \times T$ and transition matrix $r_{xz,yw} = p_{xy}q_{zw}$. If X and Y are both irreducible and aperiodic then (X,Y) is as well. If in addition invariant distributions exists for both X and Y then it follows that (X,Y) is recurrent.*

PROOF. The fact that (X,Y) is a discrete time Markov chain with transition matrix $p_{xy}q_{zw}$ is a special case of Exercise 50. If we assume that X is irreducible and aperiodic then for all $x,y \in S$ we know that there exists $n \in \mathbb{N}$ such that $p^n_{xy} > 0$ by irreducibility and furthermore by aperiodicity we know that $p^m_{yy} > 0$ for all by finitely many $m \in \mathbb{N}$ (Proposition 13.42) and therefore $p^{m+n}_{xy} \geq p^n_{xy}p^myy > 0$ for all by finitely many $m \in \mathbb{N}$. Applying the same argument to Y we see that for each $x,y \in S$ and $z,w \in T$ we have $r^n_{xz,yw} = p^n_{xy}q^n_{zw} > 0$ for all but finitely many $n \in \mathbb{N}$. Thus (X,Y) is irreducible and aperiodic.

If we assume that ν and μ are invariant distributions for X and Y respectively then it follows the fact that the transition kernel of (X,Y) is a product measure that the product measure $\nu \otimes \mu$ is invariant for (X,Y). Now apply Proposition 13.40 to see that (X,Y) has a recurrent state $(x,y) \in S \times T$ and Proposition 13.50 to see that (X,Y) is recurrent. ∎

We now apply the coupling to compare the behavior of a pair Markov chains with the same transition matrix but different initial distributions.

LEMMA 13.53. *Let X and Y be independent discrete time Markov chains both with state space S and transition matrix p_{xy} but with initial distributions ν and μ respectively. If (X,Y) is irreducible, aperiodic and recurrent then*

$$\lim_{n\to\infty} \sup_{A \in \mathcal{S}^\infty} \left| \boldsymbol{P}_\nu \circ \theta_n^{-1}\{A\} - \boldsymbol{P}_\mu \circ \theta_n^{-1}\{A\} \right| = 0$$

PROOF. By Lemma 13.52 we know that (X,Y) is a Markov chain with transition matrix $p_{xy}p_{zw}$. Let $\mathcal{F}$ be the induced filtration of (X,Y) and note that by the independence of X and Y each of X and Y is Markov with respect to $\mathcal{F}$. Consider the optional time $\tau = \min\{n \in \mathbb{N} \mid X_n = Y_n\}$ and note that by recurrence of (X,Y) we can apply Lemma 13.47 see that τ is almost surely finite (in fact for every $x \in S$, $\min\{n \in \mathbb{N} \mid X_n = Y_n = x\} < \infty$ almost surely). Let $A \in \mathcal{S}^\infty$, then since τ is countably valued and almost surely finite by the Strong Markov Property applied to X and Y,

$$\mathbf{P}\{\theta_\tau X \in A \mid \mathcal{F}_\tau\} = \mathbf{P}_{X_\tau}\{A\} = \mathbf{P}_{Y_\tau}\{A\} = \mathbf{P}\{\theta_\tau Y \in A \mid \mathcal{F}_\tau\}$$

From this and the $\mathcal{F}_\tau$-measurability of X^τ and τ it follows that $(X^\tau, \tau, \theta_\tau X) \overset{d}{=} (X^\tau, \tau, \theta_\tau Y)$. Define $\psi : S^\infty \times \mathbb{Z}_+ \times S^\infty \to S^\infty$ by

$$\psi(s,n,t)_m = \begin{cases} s_m & \text{if } m < n \\ t_{m-n} & \text{if } m \geq n \end{cases}$$

and note that for $A \in \mathcal{S}$,

$$\{\psi_m \in A\} = \cup_{n<m}\{s_m \in A\} \times \{n\} \times S^\infty \cup S^\infty \times \{n\} \times \cup_{n\geq m}\{s_{m-n} \in A\}$$

and therefore ψ is measurable. Define $\tilde{X} = \psi(X^\tau, \tau, \theta_\tau Y)$ so that

$$\tilde{X}_n = \begin{cases} X_n & \text{if } n < \tau \\ Y_n & \text{if } n \geq \tau \end{cases}$$

and also note that $X = \psi(X^\tau, \tau, \theta_\tau X)$. It follows from the Expectation Rule that $X \overset{d}{=} \tilde{X}$ and therefore for any $A \in \mathcal{S}^\infty$

$$\begin{aligned} |\mathbf{P}\{\theta_n X \in A\} - \mathbf{P}\{\theta_n Y \in A\}| &= \left|\mathbf{P}\{\theta_n \tilde{X} \in A\} - \mathbf{P}\{\theta_n Y \in A\}\right| \\ &= \left|\mathbf{P}\{\theta_n \tilde{X} \in A; \tau > n\} - \mathbf{P}\{\theta_n Y \in A; \tau > n\}\right| \leq 2\mathbf{P}\{\tau > n\} \end{aligned}$$

and therefore since τ is almost surely finite we have

$$\lim_{n\to\infty} \sup_{A \in \mathcal{S}^\infty} |\mathbf{P}\{\theta_n X \in A\} - \mathbf{P}\{\theta_n Y \in A\}| \leq 2 \lim_{n\to\infty} \mathbf{P}\{\tau > n\} = 0$$

$\square$

The proof of the existence of an invariant distribution also benefits from the coupling argument of the previous lemma.

Lemma 13.54. *Let X be an irreducible aperiodic Markov chain with state space S and transition matrix p_{xy} such that there exists $x_0, y_0 \in S$ for which $\limsup_{n\to\infty} p^n_{x_0 y_0} > 0$, then an invariant distribution for X exists.*

Proof. Take a subsequence N such that $\lim_{n\to\infty} p^n_{x_0 y_0}$ exists and is positive. By countability of S we can use a diagonal argument to pass to a further subsequence if necessary and assume that there are non-negative constants c_y for $y \in S$ with $c_{y_0} > 0$ such that $\lim_{n\to\infty} p^n_{x_0 y} = c_y$ along N for all $y \in S$. Note that by Fatou's Lemma

$$0 < \sum_{y\in S} c_y \leq \liminf_{n\to\infty} \sum_{y \in S} p^n_{x_0 y} = 1$$

Claim: $\lim_{n\to\infty} p^n_{xy} = c_y$ along N for all $x, y \in S$.

The proof of the claim uses the coupling argument. Pick an $x \in S$ and let Y be an Markov chain independent of X with transition matrix p_{xy} and initial distribution δ_x, then Y is also irreducible and aperiodic thus it follows from Lemma 13.52 that (X,Y) is an irreducible and aperiodic Markov chain with transition matrix $r_{xz,yw} = p_{xy}p_{zw}$. Suppose that (X,Y) is transient then it follows from Proposition 13.38 that

$$\sum_{n=1}^\infty (p^n_{x_0 y_0})^2 = \sum_{n=1}^\infty r^n_{x_0 x_0, y_0 y_0} < \infty$$

which would imply $\lim_{n\to\infty} p^n_{x_0 y_0} = 0$ which is a contradiction. Thus we know that (X, Y) is recurrent and we may apply Lemma 13.53 to conclude that $\lim_{n\to\infty}(p^n_{xy} - p^n_{x_0 y}) = 0$ for all $y \in S$ and therefore the claim follows.

Now note from the Chapman Kolomogorov relation that for each $x, y \in S$ and $n \in \mathbb{N}$

$$\sum_{z\in S} p^n_{xz} p_{zy} = p^{n+1}_{xy} = \sum_{z\in S} p_{xz} p^n_{zy}$$

Note that $p_{xz} p^n_{zy} \leq p_{xz}$ and $\sum_{z\in S} p_{xz} = 1$ and so we may use Dominated Convergence when taking limits in the second sum. In the first sum we can only use Fatou so we get

$$\sum_{z\in S} c_z p_{zy} \leq \lim_{n\to\infty} \sum_{z\in S} p^n_{xz} p_{zy} = \lim_{n\to\infty} \sum_{z\in S} p_{xz} p^n_{zy} = c_y \sum_{z\in S} p_{xz} = c_y$$

where all of the limits are taken along the subsequence N. Now suppose we have a strict inequality for some $y \in S$, then summing over y and using Tonelli's Theorem and the finiteness of $\sum_{z\in S} c_z$ we get

$$\sum_{z\in S} c_z = \sum_{y\in S}\sum_{z\in S} p^n_{xz} p_{zy} = \sum_{z\in S}\sum_{y\in S} p^n_{xz} p_{zy} < \sum_{z\in S} c_z$$

which is a contradiction. Thus we in fact have $\sum_{z\in S} c_z p_{zy} = c_y$ for all $y \in S$. We have observed that $\sum_{z\in S} c_z > 0$ and therefore we may define $\nu(x) = c_x / \sum_{z\in S} c_z$ to get an invariant distribution. $\square$

It remains to assemble the pieces into the proof of the theorem.

Proof. By Lemma 13.54 if X has no invariant distribution then $\limsup_{n\to\infty} p^n_{xy} = 0 \leq \liminf_{n\to\infty} p^n_{xy}$ for all $x, y \in S$; thus $\lim_{n\to\infty} p^n_{xy} = 0$ for all $x, y \in S$. Now suppose that an invariant distribution ν exists. Since X is irreducible we know that $\nu(x) > 0$ for all $x \in S$ by Proposition 13.50. Furthermore by the existence of ν and Lemma 13.52, if we let Y be an independent discrete time chain with transition matrix p_{xy} and initial distribution ν we know that (X, Y) is irreducible, aperiodic and recurrent. Thus we may apply Lemma 13.53 and the fact that $\mathbf{P}_\nu \circ \theta_n^{-1} = \nu$ to conclude that

$$\lim_{n\to\infty} \sup_{A\in\mathcal{S}^\infty} \left| \mathbf{P}_\mu \circ \theta_n^{-1}\{A\} - \mathbf{P}_\nu\{A\} \right| = 0$$

To see uniqueness of ν suppose that we have a second invariant distribution $\tilde{\nu}$ and note that by invariance of $\tilde{\nu}$ and the convergence property (13.51) $\sup_{A\in\mathcal{S}^\infty} |\mathbf{P}_{\tilde{\nu}}\{A\} - \mathbf{P}_\nu\{A\}| = 0$ which implies $\nu = \tilde{\nu}$. $\square$

Proposition 13.55. *Let X be a discrete time Markov chain with state space S and let $x, y \in S$ with y aperiodic then it follows that*

$$\lim_{n\to\infty} p^n_{xy} = \frac{\boldsymbol{P}_x\{\tau^+_y < \infty\}}{\boldsymbol{E}_y\left[\tau^+_y\right]}$$

Proof. Let's first consider the case in which $x = y$. Suppose that x is transient. In that case Proposition 13.38 implies $\sum_{n=1}^\infty p^n_{xx} = \mathbf{E}_x[\kappa_x] < \infty$ and thus $\lim_{n\to\infty} p^n_{xx} = 0$. Moreover when x is transient we know that $\mathbf{P}_x\{\tau^+_x = \infty\} = 1 - r_{xx} > 0$ and therefore $\mathbf{E}_x[\tau^+_x] = \infty$ and therefore the result holds in this case. So we now suppose that x is recurrent. Let $S_x = \{y \in S \mid r_{xy} > 0\}$ be the irreducible component containing x. We may restrict X to S_x and then by Proposition

13.47 it follows that the restriction is irreducible and recurrent and by Proposition 13.50 it follows that the restriction is aperiodic. Now we may apply Theorem 13.53 to conclude that $\lim p^n_{xx}$ exists.

Note that if we let $\xi_1 = \tau^+_x$ and $\xi_{n+1} = \tau^{n+1}_x - \tau^n_x$ for $n \in \mathbb{N}$ then by the Strong Markov property the ξ_n are an i.i.d. sequence with respect to $\mathbf{P}_x$. Moreover $\mathbf{E}_< [\tau^+_x] \infty$ (TODO: I don't believe I've shown this...)

TODO: Finish ☐

Definition 13.56. Let P be a finite discrete time Markov chain on S, we say a function $h : S \to \mathbb{R}$ is *harmonic* if for all $x \in S$, $\sum_{y \in S} P(x,y)h(y) = h(x)$.

Lemma 13.57. *Let P be an irreducible finite Markov chain on S and let $h : S \to \mathbb{R}$ be harmonic, then h is constant.*

Proof. Let M be the maximum value of h and let $x \in S$ be such that $h(x) = M$. Suppose there exists $y \in S$ such that $P(x,y) > 0$ and $h(y) < M$. It would then follow that

$$M = h(x) = \sum_{y \in S} h(x)P(x,y) < M \sum_{y \in S} P(x,y) = M$$

which is a contradiction. Thus we know that $h(y) = M$ for all $y \in S$ such that $P(x,y) > 0$. Now we do an induction. Suppose $h(y) = M$ for all $y \in S$ such that $P^{n-1}(x,y) > 0$ and suppose $z \in S$ is such that $P^n(x,z) > 0$. It follows from the expression of matrix multiplication $P^n(x,z) = \sum_{y \in S} P^{n-1}(x,y)P(y,z)$ that there exists a $y \in S$ such that $P^{n-1}(x,y) > 0$ and $P(y,z) > 0$. So by the induction hypothesis we know that $h(y) = M$ and by replaying the case of $n = 1$ with y we get that $h(z) = M$.

By irreducibility we know that for every $y \in S$, there exists $n \geq 0$ such that $P^n(x,y) > 0$ and thus we have $h(y) = M$ for every $y \in S$. ☐

Lemma 13.58. *Let P be an irreducible finite Markov chain, if the invariant distribution exists, then is unique.*

Proof. Let I denote the $\mathrm{card}(S) \times \mathrm{card}(S)$ identity matrix. By Lemma 13.57 we know that the matrix $P - I$ has a one dimensional null space given by the constant functions. Thus column rank of $P - I$ is $\mathrm{card}(S) - 1$ and the same is true for the row rank; thus there is a unique solution of $\pi(P - I) = 0$ that satisfies $\sum_{x \in S} \pi(x) = 1$. Note that this does not guarantee the existence of a invariant distribution as that requires that the entries of π be non-negative. ☐

Lemma 13.59. *If $\pi(x)P(x,y) = \pi(y)P(y,x)$ for all $x, y \in S$ then $\pi \cdot P = \pi$.*

Proof. This is a simple computation for each $y \in S$,

$$(\pi \cdot P)(y) = \sum_{x \in S} \pi(x)P(x,y) = \sum_{x \in S} \pi(y)P(y,x) = \pi(y) \sum_{x \in S} P(y,x) = \pi(y)$$

☐

The detail balance equation says "the probability of starting at x and making a transition to y is equal to the probability of starting at y and making a transition to x". To be more concise we may say that with starting distribution π, the probability of a trajectory $x \to y$ is the same as the probability of a trajectory $y \to x$. This is a type of symmetry that is sometime described as the equivalence running the chain forward and running the chain backward. By induction it is not hard to see that this symmetry extends to reversing trajectories of arbitrary finite length. We shall prove something more general by showing how to "reverse" a Markov chain that doesn't necessarily satisfy the detail balance equations.

DEFINITION 13.60. The *time reversal* of an irreducible Markov chain with transition matrix P and invariant distribution π is given by

$$\hat{P}(x,y) = \frac{\pi(y)P(y,x)}{\pi(x)}$$

LEMMA 13.61. *The time reversal is a stochastic matrix and π is invariant for $\hat{P}$. Moreover, for every $x_0,\dots,x_n \in S$, we have*

$$\boldsymbol{P}_\pi\{X_0 = x_0; \cdots; X_n = x_n\} = \boldsymbol{P}_\pi\{\hat{X}_0 = x_n; \cdots; \hat{X}_n = x_0\}$$

PROOF. By stationarity of π with respect to P for all $x \in S$,

$$\sum_{y\in S} \hat{P}(x,y) = \sum_{y\in S} \frac{\pi(y)P(y,x)}{pi(x)} \frac{1}{\pi(x)} \sum_{y\in S} \pi(y)P(y,x) = 1$$

To see π is invariant for $\hat{P}$, compute for all $y \in S$,

$$(\pi \cdot \hat{P})(y) = \sum_{x\in S} \pi(x)\hat{P}(x,y) = \sum_{x\in S} \pi(y)P(y,x) = \pi(y)$$

The last fact follows from an induction argument where the case $n = 1$ is the definition of the time reversal matrix $\hat{P}$. If we assume that the result holds for $n-1$ then

$$\begin{aligned}
\mathbf{P}_\pi\{X_0 = x_0; \cdots; X_n = x_n\} &= \pi(x_0)P(x_0,x_1)\cdots P(x_{n-1},x_n) \\
&= \hat{P}(x_1,x_0)\pi(x_1)P(x_1,x_2)\cdots P(x_{n-1},x_n) \\
&= \hat{P}(x_1,x_0)\pi(x_n)\hat{P}(x_n,x_{n-1})\cdots \hat{P}(x_2,x_1) \\
&= \mathbf{P}_\pi\{\hat{X}_0 = x_n; \cdots; \hat{X}_n = x_0\}
\end{aligned}$$

□

5. Poisson Process

The Poisson process is the standard example of a continuous time stochastic process that has discontinuous sample paths. It is a Markov process and is (almost) a martingale.

5.1. Poisson Random Variables.

DEFINITION 13.62. A Poisson distribution of rate r is the probability measure on $\mathbb{Z}_+$ given by $\mathbf{P}\{n\} = e^{-r}\frac{r^n}{n!}$. A random variable ξ whose law is a Poisson distribution is said to be a *Poisson random variable*.

PROPOSITION 13.63. *The Poisson distribution of rate r is a probability measure with mean r and variance r.*

Proof. To see that the Poisson distribution is a probability measure simply note that $\mathbf{P}\{\mathbb{Z}_+\} = \sum_{n=0}^{\infty} e^{-r}\frac{r^n}{n!} = 1$ by the power series for the exponential function. To calculate the mean

$$e^{-r}\sum_{n=0}^{\infty} n\frac{r^n}{n!} = e^{-r}\left(\sum_{n=0}^{\infty}(n+1)\frac{r^n}{n!} - e^r\right) = e^{-r}\frac{d}{dr}re^r - 1 = r$$

to calculate the variance we calculate the second moment in a similar way,

$$e^{-r}\sum_{n=0}^{\infty} n^2\frac{r^n}{n!} = e^{-r}\sum_{n=0}^{\infty}(n+2)(n+1)\frac{r^n}{n!} - 3r - 2 = e^{-r}\frac{d^2}{dr^2}r^2e^r - 3r - 2 = 2 + 4r + r^2 - 3r - 2 = r^2 + r$$

from which it follows that variance is $r^2 + r - r^2 = r$. $\square$

5.2. Exponential Random Variables. The standard construction of the Poisson process uses sums of a sequence of i.i.d. exponential random variables so it is therefore useful to discuss such random variables first. As explained below exponential random variables will figure prominently in subsequent theory of Markov processes as well so it will be a good investment of time to get familiar with them.

Definition 13.64. Given a parameter $\lambda > 0$, the probability measure on $\mathbb{R}_+$ given by $\mu(A) = \lambda \int_A e^{-\lambda x}\, dx$ is called the *exponential distribution with rate* λ. A random variable ξ whose law is an exponential distribution is said to be a *exponential random variable*.

The reader may have learned at some point that incandescent lightbulbs have peculiar property; the probability that such a light bulb will fail does not depend on the age of the light bulb. Expressed using our notation, if we let ξ be age of a light bulb when it fails we are saying that for all $t > s$ we have $\mathbf{P}\{\xi > t \mid \xi > s\} = \mathbf{P}\{\xi > t - s\}$ or equivalently

$$\mathbf{P}\{\xi > t\} = \mathbf{P}\{\xi > t; \xi > s\} = \mathbf{P}\{\xi > t \mid \xi > s\}\mathbf{P}\{\xi > s\} = \mathbf{P}\{\xi > t - s\}\mathbf{P}\{\xi > s\}$$

While the stated fact about light bulbs is only approximately true, it is a concrete illustration of a property that we call memorylessness. The reason that exponential random variables figure so prominently in subsequent theory is that they are precisely the random variables that have the property of being memoryless.

Proposition 13.65. *Let γ be an exponential random variable then for each $t, s \geq 0$ we have the functional equation*

$$\boldsymbol{P}\{\gamma > t + s\} = \boldsymbol{P}\{\gamma > t\}\boldsymbol{P}\{\gamma > s\} \tag{16}$$

Moreover if γ is a nonnegative random variable that is not almost surely equal to 0 *and satisfies* (16), *it follows that γ is exponential.*

Proof. The memorylessness property of exponential random variable is a trivial computation,

$$\mathbf{P}\{\gamma > t + s\} = e^{-\lambda(t+s)} = e^{-\lambda t}e^{-\lambda s} = \mathbf{P}\{\gamma > t\}\mathbf{P}\{\gamma > s\}$$

If we let $\mathbf{P}\{\gamma > 1\} = e^{-c}$ for some $c \in [0, \infty]$, then from the functional equation (16) we immediately see that for every $n \in \mathbb{N}$, $\mathbf{P}\{\gamma > n\} = \mathbf{P}\{\gamma > 1\}^n = e^{-cn}$ and then for all positive rationals $p/q \in \mathbb{Q}_+$ we have $\mathbf{P}\{\gamma > p/q\} = e^{-cp/q}$. Now since $\mathbf{P}\{\gamma > t\}$ is right continuous we can conclude that $\mathbf{P}\{\gamma > t\} = e^{-ct}$ for all $0 \leq t < \infty$. By our assumption that there exists some $t \geq 0$ such that $\mathbf{P}\{\gamma > t\} > 0$ it follows that $c < \infty$ and we have shown that γ is exponentially distributed. $\square$

PROPOSITION 13.66. *Let $\gamma_1, \dots, \gamma_n$ be a sequence of i.i.d. exponential random variables with rate λ then for all $t > s$ we have*

$$\boldsymbol{P}\{\gamma_1 + \cdots + \gamma_n > t; \gamma_1 > s\} = \boldsymbol{P}\{\gamma_1 + \cdots + \gamma_n > t - s\}\boldsymbol{P}\{\gamma_1 > s\}$$

PROOF. We proceed by induction. The initial case is just the memorylessness of a single exponential random variable. For $n \geq 2$ we compute using Fubini's theorem (specifically Lemma 4.6) and the non-negativity of exponential random variables

$$\begin{aligned}
&\mathbf{P}\{\gamma_1 + \cdots + \gamma_n > t; \gamma_1 > s\} \\
&= \mathbf{P}\{\gamma_1 + \cdots + \gamma_n > t; \gamma_1 > s; \gamma_n < t - s\} + \mathbf{P}\{\gamma_1 + \cdots + \gamma_n > t; \gamma_1 > s; \gamma_n \geq t - s\} \\
&= \mathbf{E}\left[\mathbf{P}\{\gamma_1 + \cdots + \gamma_{n-1} > t - u; \gamma_1 > s\} \mid_{u=\gamma_n}; \gamma_n < t - s\right] \\
&+ \mathbf{P}\{\gamma_1 + \cdots + \gamma_n > t; \gamma_1 > s; \gamma_n \geq t - s\} \\
&= \mathbf{E}\left[\mathbf{P}\{\gamma_1 + \cdots + \gamma_{n-1} > t - u - s\} \mid_{u=\gamma_n}; \gamma_n < t - s\right] \mathbf{P}\{\gamma_1 > s\} \\
&+ \mathbf{P}\{\gamma_n \geq t - s\}\mathbf{P}\{\gamma_1 > s\} \\
&= \left(\mathbf{P}\{\gamma_1 + \cdots + \gamma_n > t - s; \gamma_n < t - s\} + \mathbf{P}\{\gamma_n \geq t - s\}\right) \mathbf{P}\{\gamma_1 > s\} \\
&= \mathbf{P}\{\gamma_1 + \cdots + \gamma_n > t - s\}\mathbf{P}\{\gamma_1 > s\}
\end{aligned}$$

□

It is also worth having the density and cumulative distribution of a sum of i.i.d. exponential random variables handy

PROPOSITION 13.67. *Let $\gamma_1, \dots, \gamma_n$ be i.i.d. exponential random variables with rate λ, then the density of $\gamma_1 + \cdots + \gamma_n$ is $\lambda^n e^{-\lambda t} \frac{t^{n-1}}{(n-1)!}$ and*

$$\boldsymbol{P}\{\gamma_1 + \cdots + \gamma_n > t\} = e^{-\lambda t} \sum_{j=1}^{n-1} \frac{\lambda^k t^k}{k!}$$

PROOF. Straightforward induction calculation using the convolution formula. □

We are now in a position to show that Poisson processes exist.

THEOREM 13.68. *Let $\gamma_1, \gamma_2, \dots$ be i.i.d. exponential random variables with rate $\lambda > 0$. For each $n \in \mathbb{N}$ define $S_n = \gamma_1 + \ldots + \gamma_n$ and for each $t \in \mathbb{R}_+$ let*

$$N_t = \max\{n \in \mathbb{N} \mid S_n \leq t\}$$

where the maximum of the empty set is taken to be 0. Then N is a homogeneous Poisson process with rate λ.

PROOF. Since $\{N_t \geq m\} = \{S_m \leq t\}$ the measurability of N_t follows from the measurability of the γ_n and thus N is a stochastic process.

It remains to show that N has independent increments. Let $0 \leq s < t < \infty$ and consider the computation of $\mathbf{P}\{N_t - N_s \in \cdot \mid \mathcal{F}_s\}$. Let $\mathcal{F}$ be the filtration generated by N, let $\mathcal{G}_0 = \{\emptyset, \Omega\}$ and for each $n \in \mathbb{N}$ let $\mathcal{G}_n = \sigma(\gamma_1, \dots, \gamma_n)$. We wish to do this computation locally on events of the form $\{N_s = n\}$ for $n \in \mathbb{Z}_+$ by reducing to a conditional expectation with respect to $\mathcal{G}_n$.

Rather than appealing to the general Lemma 8.14 we use the following simple version.

Claim: Let $0 \leq s < \infty$, $n \in \mathbb{Z}_+$ and $A \in \mathcal{F}_s$. There exists a $B \in \mathcal{G}_n$ such that $A \cap \{N_s = n\} = B \cap \{N_s = n\}$.

Note first that the set $\mathcal{C}$ of all $A \in \mathcal{F}_s$ for which an appropriate $B \in \mathcal{G}_n$ exists is a σ-algebra. This is elementary since if $A \cap \{N_s = n\} = B \cap \{N_s = n\}$ then it follows that $A^c \cap \{N_s = n\} = B^c \cap \{N_s = n\}$ and morever if $A_m \cap \{N_s = n\} = B_m \cap \{N_s = n\}$ for all $m \in \mathbb{N}$ then

$$\begin{aligned}(\cup_{m=1}^\infty A_m) \cap \{N_s = n\} &= \cup_{m=1}^\infty (A_m \cap \{N_s = n\}) = \cup_{m=1}^\infty (B_m \cap \{N_s = n\}) \\ &= (\cup_{m=1}^\infty B_m) \cap \{N_s = n\}\end{aligned}$$

Since $\mathcal{C}$ is a σ-algebra is suffices to that that $\{N_u = m\} \in \mathcal{C}$ for all $0 \leq u \leq s$ and $m \in \mathbb{Z}_+$ since such sets generate $\mathcal{F}_s$. To see this note that $\{N_u = m\} \cap \{N_s = n\} = \emptyset$ for $m > n$, $\{N_u = m\} \cap \{N_s = n\} = \{S_m \leq u < S_{m+1}\} \cap \{N_s = n\}$ for $m < n$ and $\{N_u = n\} \cap \{N_s = n\} = \{S_n \leq u\} \cap \{N_s = n\}$.

We now use the claim to calculate $\mathbf{P}\{N_t - N_s \in \cdot \mid \mathcal{F}_s\}$. Let $A \in \mathcal{F}_s$ and for each $n \in \mathbb{Z}_+$ we pick $B_n \in \mathcal{G}_n$ such that $A \cap \{N_s = n\} = B_n \cap \{N_s = n\}$. We let $k \in \mathbb{Z}_+$ and use the definition of N_t, the independence of the γ, Lemma 4.6 and Proposition 13.66

$$\begin{aligned}
&\mathbf{P}\{N_t - N_s \leq k; A\} = \sum_{n=0}^\infty \mathbf{P}\{N_t - N_s \leq k; N_s = n; A\} = \sum_{n=0}^\infty \mathbf{P}\{N_t - N_s \leq k; N_s = n; B_n\} \\
&= \sum_{n=0}^\infty \mathbf{P}\{S_{n+k+1} > t; S_{n+1} > s; s \geq S_n; B_n\} \\
&= \sum_{n=0}^\infty \mathbf{E}\left[\mathbf{P}\{\gamma_{n+1} + \cdots + \gamma_{n+k+1} > t - u; \gamma_{n+1} > s - u\} \mid_{u=S_n}; s \geq S_n; B_n\right] \\
&= \sum_{n=0}^\infty \mathbf{E}\left[\mathbf{P}\{\gamma_{n+1} + \cdots + \gamma_{n+k+1} > t - s\}\mathbf{P}\{\gamma_{n+1} > s - u\} \mid_{u=S_n}; s \geq S_n; B_n\right] \\
&= \mathbf{P}\{\gamma_1 + \cdots + \gamma_{k+1} > t - s\} \sum_{n=0}^\infty \mathbf{P}\{S_{n+1} > s; s \geq S_n; B_n\} \\
&= \mathbf{P}\{\gamma_1 + \cdots + \gamma_{k+1} > t - s\} \sum_{n=0}^\infty \mathbf{P}\{N_s = n; B_n\} \\
&= \mathbf{P}\{\gamma_1 + \cdots + \gamma_{k+1} > t - s\} \sum_{n=0}^\infty \mathbf{P}\{N_s = n; A\} \\
&= \mathbf{P}\{\gamma_1 + \cdots + \gamma_{k+1} > t - s\}\mathbf{P}\{A\}
\end{aligned}$$

which shows that

$$\mathbf{P}\{N_t - N_s \leq k \mid \mathcal{F}_s\} = \mathbf{P}\{\gamma_1 + \cdots + \gamma_{k+1} > t - s\} = e^{-\lambda(t-s)} \sum_{j=0}^k \frac{\lambda^j (t-s)^j}{j!}$$

Since the conditional probability is a constant it follows that $N_t - N_s \perp\!\!\!\perp \mathcal{F}_s$ and moreover by taking expectations it follows that $N_t - N_s$ is Poisson distributed with rate $\lambda(t - s)$. $\square$

A homogeneous Poisson process provides us with another important example of a continuous time martingale.

PROPOSITION 13.69. *Let N be a homogeneous Poisson process with rate λ then $N_t - \lambda t$ is a cadlag martingale.*

PROOF. It is clear that $N_t - \lambda t$ is a cadlag process, moreover since N_t is Poisson distributed with rate λt it follows that $N_t - \lambda t$ is integrable and has mean zero. The martingale property follows from the independent increments property

$$\mathbf{E}\left[N_t \mid \mathcal{F}_s\right] = \mathbf{E}\left[N_t - N_s \mid \mathcal{F}_s\right] + N_s = \lambda(t-s) + N_s$$

□

6. Pure Jump-Type Markov Processes

In this section we discuss a simple subclass of time homogeneous Markov Processes on $\mathbb{R}_+$.

DEFINITION 13.70. A time homogenous Markov process on $\mathbb{R}_+$ with values in a metric (topological?) space $(S, \mathcal{B}(S))$ is said to be *pure jump-type* if almost surely its sample paths are piecewise constant with isolated jump discontinuities.

The first goal is to get a more constructive description of the class of pure jump-type Markov processes. The key idea in achieving that goal is to study the random time to the jumps of the process; in fact these random times are optional with respect to the right continuous filtration generated by the process.

DEFINITION 13.71. Let X be a pure jump-type Markov process then the *first jump time* is the random time

$$\tau_1 = \inf\{t \geq 0 \mid X_t \neq X_0\}$$

the n^{th} *jump time* is defined to be

$$\tau_n = \tau_{n-1} + \tau_1 \circ \theta_{\tau_{n-1}} = \inf\{t \geq \tau_{n-1} \mid X_t \neq X_{\tau_{n-1}}\} \text{ for } n > 1$$

and the 0^{th} *jump time* is $\tau_0 = 0$.

LEMMA 13.72. *Let X be a pure jump-type Markov process then τ_n is a weakly $\mathcal{F}$-optional time for all $n \geq 0$.*

PROOF. The case τ_0 is trivial as it is a deterministic time. For each $n \in \mathbb{N}$, define $\sigma_n = \min\{k/2^n \mid X_{k/2^n} \neq X_0\}$. Note that because of the right continuity of sample paths of X we have $\sigma_n \downarrow \tau_1$. Moreover we have

$$\{\sigma_n \leq t\} = \cup_{k=0}^{\lfloor 2^n t \rfloor} \{X_{k/2^n} \neq X_0\} \in \mathcal{F}_{\lfloor 2^n t \rfloor / 2^n} \subset \mathcal{F}_t$$

and therefore σ_n is $\mathcal{F}$-optional. Therefore by Lemma 9.72 we see that $\tau_1 = \lim_{n\to\infty} \sigma_n = \inf_n \sigma_n$ is weakly $\mathcal{F}$-optional.

The fact that τ_n is weakly optional follows by induction using Lemma 13.28 applied to the expression $\tau_n = \tau_{n-1} + \tau_1 \circ \theta_{\tau_{n-1}}$. □

The definition of the optional time τ_1 allows us to define an important property of elements of S.

DEFINITION 13.73. A state $x \in S$ is said to be *absorbing* if $\mathbf{P}\{X_t \equiv x\} = \mathbf{P}_x\{\tau_1 = \infty\} = 1$. If x is not absorbing we say it is *non-absorbing.*

By the Markov property we see that if a pure jump-type Markov process X reaches an absorbing state x it remains there indefinitely almost surely. If X is in a non-absorbing state one might ask whether there is a positive probability that it remains there forever (i.e. the state is "partially absorbing"). In fact in a non-absorbing state it is almost sure that a jump to a new state will occur (a kind of 0-1 law). This fact is a corollary of the following result that describes the distribution to the next jump from a non-absorbing state.

Lemma 13.74. *Let X be a pure jump-type Markov process and let $x \in S$ be nonabsorbing, then under $\boldsymbol{P}_x$ the optional time τ_1 is exponentially distributed and independent of $\theta_{\tau_1} X$.*

Proof. To see that τ_1 is exponentially distributed note that

$$\mathbf{P}_x\{\tau_1 > t + s\} = \mathbf{P}_x\{\tau_1 > s; \tau_1 \circ \theta_s > t\} = \mathbf{P}_x\{\tau_1 > s\}\mathbf{P}_x\{\tau_1 > t\}$$

By our assumption that x is nonabsorbing we know that $\tau_1 > 0$ with positive probability and therefore we can apply Proposition 13.65 to conclude that τ_1 is exponentially distributed.

Recall from Lemma 9.70 that when restricted to $D([0,\infty); S)$, one can think of τ_1 as being a composition of the process X with a measurable function on $S^{[0,\infty)}$ which we call $\tilde{\tau}_1$ (of course if X is the canonical process $\tau_1 = \tilde{\tau}_1$). Let B be a measurable set in $S^{[0,\infty)}$ and define the set

$$\tilde{B} = \{f \in D([0,\infty) \mid \theta_{\tilde{\tau}_1(f)} f \in B\}$$

By writing the indicator of $\tilde{B}$ as the composition

$$D([0,\infty); S) \overset{(id,\tilde{\tau}_1)}{\to} D([0,\infty); S) \times [0,\infty) \overset{\theta}{\to} D([0,\infty); S) \overset{\mathbf{1}_B}{\to} \mathbb{R}$$

we see that $\tilde{B}$ is also measurable (recall that θ as above is measurable by Lemma 13.27). It is also noted that we have the equality $\{X \in \tilde{B}\} = \{\theta_{\tau_1} X \in B\}$.

Let $\tau_1^t = \inf\{s \geq t \mid X_s \neq X_t\}$ and note that $\tau_1^t(X) = \tilde{\tau}_1(\theta_t X) + t$. From this we get

$$\left(\theta_{\tau_1^t} X\right)_s = X(\tilde{\tau}_1(\theta_t X) + t + s) = \left(\theta_{\tilde{\tau}_1(\theta_t X)} \theta_t X\right)_s$$

Now we can compute (in rather excruciating detail I might add) using the fact that $\tau_1^t = \tau_1$ on the set $\{\tau_1 > t\}$, the Markov Property of X, the definition of $\tilde{B}$ and the fact that $X_t = x$ on $\{\tau_1 > t\}$ to see

$$\begin{aligned}
\mathbf{P}\{\tau_1 > t; \theta_{\tau_1} X \in B\} &= \mathbf{P}\{\tau_1 > t; \theta_{\tau_1^t} X \in B\} \\
&= \mathbf{P}\{\tau_1 > t; \mathbf{E}\left[\theta_{\tilde{\tau}_1(\theta_t X)} \theta_t X \in B \mid \mathcal{F}_t\right]\} \\
&= \mathbf{P}\{\tau_1 > t; \mathbf{P}\{\theta_t X \in \tilde{B} \mid \mathcal{F}_t\}\} \\
&= \mathbf{P}\{\tau_1 > t; \mathbf{P}_{X_t}\{\tilde{B}\}\} \\
&= \mathbf{P}\{\tau_1 > t\}\mathbf{P}_x\{\tilde{B}\} \\
&= \mathbf{P}\{\tau_1 > t\}\mathbf{P}\{\theta_{\tau_1} X \in B\}
\end{aligned}$$

□

With the distribution of first jump time available we can now see that a the first jump time is either almost surely finite or almost surely infinite depending on whether the process starts in a non-absorbing or absorbing state.

COROLLARY 13.75. *Let $\boldsymbol{P}_x$ be a Markov family for pure jump-type Markov process and let τ_1 be the first jump time then*

$$\boldsymbol{P}_x\{\tau_1 < \infty\} = \begin{cases} 0 & \textit{when } x \textit{ is non-absorbing} \\ 1 & \textit{when } x \textit{ is absorbing} \end{cases}$$

PROOF. By Lemma 13.74 we know that for x non-absorbing $\mathbf{E}_x[\tau_1] < \infty$ which implies $\mathbf{P}_x\{\tau_1 < \infty\} < \infty$. $\square$

It should be noted that in the literature it is very uncommon to make the subtle distinction between the interpretation of τ_1 as either a random variable or a function on $D([0,\infty);\mathbb{R})$. On the one hand, authors may deal with the issue by glossing over the distinction and abusing notation through the use of τ_1 to denote both functions. On the other hand authors may try to define the problem away by restricting attention to the canonical case; this restriction later biting the reader when results proven in the canonical case are implicitly extended to the non-canonical case. At some point we will start to take the abuse of notation approach but we want to have some examples in which all of the fine distinctions are made so that the reader can refer back to them in times of confusion.

Based on the previous result we see that the distribution of the first jump of a pure jump type Markov process boils down to two independent distributions: the first being an exponential distribution that describes when a jump happens and the second being a general distribution that describes where the jump goes to. This observation can be used to give us a nice description of the entire process. Before providing the construction we settle on some terminology.

DEFINITION 13.76. Given a pure jump Markov process X with a first jump time τ_1 we define the *rate function* to be

$$c(x) = \begin{cases} 1/\mathbf{E}_x[\tau_1] & \text{if } x \text{ is non-absorbing} \\ 0 & \text{if } x \text{ is absorbing} \end{cases}$$

the *jump transition kernel* to be

$$\mu(x, B) = \begin{cases} \mathbf{P}_x\{\theta_{\tau_1} X \in B\} & \text{if } x \text{ is non-absorbing} \\ \delta_x(B) & \text{if } x \text{ is absorbing} \end{cases}$$

and the *rate kernel* to be $\alpha(x, B) = c(x)\mu(x, B)$.

Note that in the above definition we are thinking of the Markov process as the family of measures $\mathbf{P}_x$ on $S^{[0,\infty)}$ and interpreting τ_1 as a function from $S^{[0,\infty)}$ to $\mathbb{R}_+$.

Before proceeeding to our structure theory for pure jump type Markov processes we establish the basic measurability properties of the functions just defined.

LEMMA 13.77. *The rate function $c(x)$ is a measurable function on S and the jump transition kernel and rate kernel are both kernels from S to $S^{[0,\infty)}$. The rate kernel is a measurable function of the jump transition kernel.*

PROOF. We know that $\mathbf{P}_x$ is a kernel by Lemma 13.13 and therefore $\mathbf{E}_x[\tau_1]$ is a measurable function of x by Lemma 8.29. Lastly we see that

$$\{x \text{ is non-absorbing}\} = \{\mathbf{P}_x\{\tau_1 < \infty\} = 1\}$$

is measurable because $\mathbf{P}_x$ is a kernel; thus $c(x)$ is measurable.

The fact that $\mu(x, B)$ is a measurable function of x for fixed B follows from the fact $\mathbf{P}_x$ is a kernel. The fact that for fixed $\mu(x, B)$ is a probability measure for fixed x follows from measurability of the mapping taking X to $\theta_{\tau_1} X$ and Lemma 2.53.

To see that $\mu(x, B)$ is a measurable function of $\alpha(x, B)$ just observe that

$$\mu(x, B) = \begin{cases} \alpha(x, B)/\alpha(x, S) & \text{if } \alpha(x, S) \neq 0 \\ \delta_x(B) & \text{if } \alpha(x, S) = 0 \end{cases}$$

□

Extending these ideas further we will see that every pure jump-type Markov process decomposes into a discrete time Markov chain that describes the state transition of the jumps that occur and a sequence of independent exponential random variables that describe the time between jumps. This make intuitive sense given the last lemma and the Strong Markov property: our process begins by waiting for an exponentially distributed time then makes an independent jump to a new state; by the Strong Markov property the process starts afresh in the new state waits for another independent exponentially distributed time and makes another independent jump and so on. One subtlety arises because the heuristic argument just given ignores the fact that our process may jump into an absorbing state. The other subtlety is that the mean time to the next jump depends on the current state. If we normalize by the rate function of the current state then the means are all unity and we might be able "integrate" the waiting times into the single source of randomness that a sequence of i.i.d. exponential random variables would provide. Handling these problems and making things precise is the job of the next theorem.

THEOREM 13.78. *Let X be a pure jump Markov process with rate kernel $\alpha = c\mu$ and jump times $\tau_0, \tau_1, \dots$, then there is a Markov process Y on $\mathbb{Z}_+$ with transition kernel μ and a sequence of i.i.d. exponential random variables $\gamma_0, \gamma_1, \dots$ of rate 1 that are independent of Y such that for all $n \geq 1$*

$$\tau_n = \begin{cases} \sum_{k=0}^{n-1} \frac{\gamma_k}{c(Y_k)} & \text{when } c(Y_k) \neq 0 \text{ for all } k = 0, \dots, n-1 \\ \infty & \text{when } c(Y_k) = 0 \text{ for some } k = 0, \dots, n-1 \end{cases}$$

and

$$X_t = Y_n \text{ a.s. for } \tau_n \leq t < \tau_{n+1}$$

when $\tau_n < \infty$. If $\tau_n = \infty$ for some n then let $N = \max\{n \mid \tau_n < \infty\}$, then we have $Y_n = Y_{N-1} = X_{\tau_N}$ for all $n > N$.

PROOF. To simply notation, in the case in which $\tau_n = \infty$ for some n, let $X_\infty = X_{\tau_N}$ where N is defined in the statement of the Theorem (it is the position of X after its last jump). With that definition in hand we know that the result of the Theorem requires that we define $Y_n = X_{\tau_n}$. The work is in constructing the γ_n and validating the Markov property.

Our first real task is to understand the relationship between the condition $\{\tau_n < \infty\}$ and the condition $\{c(Y_{n-1}) \neq 0\}$ in order to make proper sense of the expression for τ_n.

Claim: $\tau_n < \infty$ almost surely when $c(Y_{n-1}) \neq 0$ and $\tau_{n-1} < \infty$ (i.e. $\mathbf{P}\{\tau_n < \infty; c(Y_{n-1}) \neq 0; \tau_{n-1} < \infty\} = \mathbf{P}\{c(Y_{n-1}) \neq 0; \tau_{n-1} < \infty\}$).

First note that for any $x \in S$, by definition $c(x) \neq 0$ implies that $\mathbf{E}_x[\tau_1] < \infty$ which certainly implies that $\mathbf{P}_x\{\tau_1 < \infty\} = 1$. Now for all $n \geq 1$ we can calculate using the tower property and pullout property of conditional expectations and the Strong Markov property

$$\begin{aligned}
&\mathbf{P}\{\tau_n < \infty; c(Y_{n-1}) \neq 0; \tau_{n-1} < \infty\} \\
&= \mathbf{E}\left[\mathbf{P}\{\tau_n < \infty; c(Y_{n-1}) \neq 0; \tau_{n-1} < \infty \mid \mathcal{F}_{\tau_{n-1}}\}\right] \\
&= \mathbf{E}\left[\mathbf{P}\{\tau_1(\theta_{\tau_{n-1}(X)} X) < \infty \mid \mathcal{F}_{\tau_{n-1}}\}; c(Y_{n-1}) \neq 0; \tau_{n-1} < \infty\right] \\
&= \mathbf{E}\left[\mathbf{P}_{Y_{n-1}}\{\tau_1 < \infty\}; c(Y_{n-1}) \neq 0; \tau_{n-1} < \infty\right] \\
&= \mathbf{P}\{c(Y_{n-1}) \neq 0; \tau_{n-1} < \infty\}
\end{aligned}$$

and the claim is proved.

Claim: $\{c(Y_{n-1}) = 0\} = \{\tau_n = \infty\}$ a.s.

What does this mean? I think $\mathbf{P}\{\{c(Y_{n-1}) = 0\} \triangle \{\tau_n = \infty\}\} = 0$. Calculate

$$\begin{aligned}
&\mathbf{P}\{c(Y_{n-1}) = 0; \tau_n < \infty\} \\
&= \mathbf{P}\{c(Y_{n-1}) = 0; \tau_{n-1} < \infty; \tau_1(\theta_{\tau_{n-1}(X)}(X)) < \infty\} \\
&= \mathbf{P}\{c(Y_{n-1}) = 0; \tau_{n-1} < \infty; \mathbf{P}\{\tau_1(\theta_{\tau_{n-1}(X)}(X)) < \infty \mid \mathcal{F}_{\tau_{n-1}}\}\} \\
&= \mathbf{P}\{c(Y_{n-1}) = 0; \tau_{n-1} < \infty; \mathbf{P}_{Y_{n-1}}\{\tau_1 < \infty\}\} = 0 \text{ by Corollary 13.75}
\end{aligned}$$

TODO: Here is the crux of where I get confused. Kallenberg says the following: let $\gamma'_1, \gamma'_2, \ldots$ be i.i.d. exponentially distributed of mean 1 and such that $\gamma'_n \perp\!\!\!\perp X$ which means we must be willing to break out of the canonical case (this is technically not true of Kallenberg's proof since he states that all randomization variables are assumed to exist in the canonical process setup). Define

$$\gamma_n = (\tau_n - \tau_{n-1}) c(Y_{n-1}) \mathbf{1}_{\tau_n < \infty} + \gamma'_n \mathbf{1}_{\tau_n = \infty}$$

and we claim that if $c(x) > 0$ then we have

$$\mathbf{P}_x\{\gamma_1 > t; Y_1 \in B\} = \mathbf{P}_x\{\tau_1 c(x) > t; Y_1 \in B\} = e^{-t}\mu(x, B)$$

and that if $c(x) = 0$ then

$$\mathbf{P}_x\{\gamma_1 > t; Y_1 \in B\} = \mathbf{P}_x\{\gamma'_1 > t; Y_1 \in B\} = e^{-t}\mu(x, B)$$

and this is where I get hung up on a subtlety. The measure $\mathbf{P}_x$ was defined to be on the path space but γ_1 is not defined on the path space but on an extension. Probably the right way to make sense of this is to consider a Markov family as in Definition 13.14 and then consider $\mathbf{P}_x$ in that context. Of course, we have not proven that Markov families exist (though I believe that is implicit in the proof of Markov processes and the proof of Daniell-Kolmogorov) nor have we proven that Markov families are preserved under extension (see Blumenthal and Getoor for an exercise that is sufficient for the case of extending by $[0, 1]$). In any case if we succeed in doing that then $\mathbf{P}_x$ is the probability measure on Ω under which X is a

Markov process with $X_0 = x$ almost surely and computation is a straightforward application of Lemma 13.74 and the independence of γ'_n and X. The thing that I am unsatisfied with in this context is the fact that the statement of the result does not involve or require Markov families. Also, what if X has a non-point mass initial distribution? Of course the other issue is that Kallenberg's formula for γ_n is wrong!!!! He writes

$$\gamma_n = (\tau_n - \tau_{n-1})c(Y_n)\mathbf{1}_{\tau_{n-1}<\infty} + \gamma'_n \mathbf{1}_{\tau_{n-1}=\infty}$$

$\square$

It is useful to turn this description of a pure jump-type Markov around and use it to construct a pure jump-type Markov process.

THEOREM 13.79. *Let $\alpha = c\mu$ be a kernel on S such that $\alpha(x, \{x\}) \equiv 0$, let Y be a Markov chain with transition kernel μ and let $\gamma_1, \gamma_2, \dots$ be i.i.d. exponential random variables of mean 1 such that $\gamma_1, \gamma_2, \dots \perp\!\!\!\perp Y$. Pick an arbitrary element $s_0 \in S$ and define $\tau_0 = 0$ and for $n \in \mathbb{N}$ we define*

$$\tau_n = \sum_{j=1}^{n} \frac{\gamma_j}{c(Y_{j-1})}$$

and

$$X_t = \begin{cases} Y_n \text{ for } \tau_n \leq t < \tau_{n+1} \\ s_0 \text{ for } t \geq \sup_n \tau_n \end{cases}$$

If $\lim_{n\to\infty} \tau_n = \infty$ a.s. for every initial distribution of Y then X is a pure jump-type Markov process with rate kernel α.

PROOF. We consider (Y, γ) as a Markov chain on the state space $S \times \mathbb{R}_+$. (TODO: Show that independence of Y and $\gamma_1, \gamma_2, \dots$ implies this is valid). Define τ_n and X as in the statement of the theorem, let $\mathcal{G}_n$ be the filtration generated by (Y, γ) and let $\mathcal{F}_t$ be the filtration generated by X.

We need leverage our knowledge that (Y, γ) is a Markov process to show that X has the Markov property. In order to do this we want to be able use information about conditional expectations with respect to $\mathcal{G}$ in order to compute conditional expectations with respect to $\mathcal{F}$; thus we first clarify the relationship between the two filtrations. The trick is that in general a given X_t can be equal to any Y_n and therefore to restrict the set of possible Y_n we must restrict the number of jumps that occur before t. In other words we must restrict the possible values of some τ_n; in this way the random variables $\gamma_1, \gamma_2, \dots$ enter the picture.

Claim: Let $t \geq 0$ and $n \in \mathbb{N}$ be fixed, then $\mathcal{G}_n \vee \{\tau_{n+1} > t\}$ and $\mathcal{F}_t$ agree on $\{\tau_n \leq t < \tau_{n+1}\}$. Furthermore $\{\tau_n \leq t < \tau_{n+1}\}$ is $\mathcal{G}_n \vee \{\tau_{n+1} > t\} \cap \mathcal{F}_t$-measurable.

The $\mathcal{G}_n \vee \{\tau_{n+1} > t\}$ of $\{\tau_n \leq t < \tau_{n+1}$ is immediate as τ_n is a function of $\gamma_1, \dots, \gamma_n$ and $Y_0, \dots, Y_{n-1}$ and therefore is $\mathcal{G}_n$ measurable. To see $\mathcal{F}_t$-measurability first note that, by construction, τ_n is the n^{th} jumping time of X (TODO: What about the fact that we have probability zero event that $Y_m = Y_{m+1}$? This seems like a real issue since X cannot detect τ_n unless the value of X changes there. What we do know is that if X sees n jumps then at least n of the timers γ have gone off; maybe this is enough...) TODO: Note that I believe Blumenthal and Getoor handle this issue as part of their construction which they redefine the probability space by removing the set of probability zero where $Y_m = Y_{m+1}$. This seems to indicate that either Kallenberg needs the flexibility to much with the probability space in a similar way.

Note that even here we have Y assumed to be a Markov family and we are constructing X as a Markov family.

TODO: Finish and properly understand Kallenberg's proof here.

This is how Blumenthal and Getoor do this. Not clear that they allow for the existence of absorbing states but they certainly do allow for explosion by appending the cemetary state.

Let $T = S \times [0, \infty)$, $\mathcal{T} = \mathcal{S} \otimes \mathcal{B}([0, \infty))$, $\Omega = T^\infty$ and $\mathcal{A} = \mathcal{T}^\infty$. We write a generic element $\omega \in \Omega$ as $\omega = ((x_0, t_0), (x_1, t_1), \dots)$. Define $Z_n : \Omega \to T$ as $Z_n(\omega) = (x_n, t_n)$ to be the n^{th} coordinate projection and let $Y_n : \Omega \to S$ and $\tau_n : \Omega \to [0, \infty)$ be defined by $Y_n(\omega) = x_n$ and $\tau_n(\omega) = t_n$ respectively. By the definition of the product σ-algebra we know that each of Z_n, Y_n and τ_n is measurable.

Claim 13.79.1. For each $A \in \mathcal{T}$ let

$$\tilde{\mu}((x, t), A) = \int \mathbf{1}_A(y, s) \mathbf{1}_{[t, \infty)}(s)\, \mu(x, dy) c(x) e^{-c(x)(s-t)} ds$$

then $\tilde{\mu}$ is a probability kernel and moreover $\tilde{\mu}$ is translation invariant in t.

By the monotone class argument of Lemma 8.27 we can reduce to considering set $A = B \times C$ and since

$$\tilde{\mu}((x, t), B \times C) = \mu(x, B) \int_t^\infty \mathbf{1}_C(s) c(x) e^{-c(x)(s-t)} ds$$

and μ is assumed a kernel we must only show $\int_t^\infty \mathbf{1}_C(s) c(x) e^{-c(x)(s-t)} ds$ is a probability kernel. The fact that it is a probability measure is elementary calculus and the fact that $c(x) > 0$,

$$\int_t^\infty c(x) e^{-c(x)(s-t)} ds = e^{c(x)t} \int_t^\infty c(x) e^{-c(x)s} ds = e^{c(x)t} e^{-c(x)t} = 1$$

The fact that it is a kernel follows from Lemma 8.29. To see that $\tilde{\mu}$ is translation invariant we consider $B \times C \in \mathcal{S} \times \mathcal{B}([0, \infty))$ and calculate as above and using a change of integration variable

$$\begin{aligned}
\tilde{\mu}((x, t), B \times C) &= \mu(x, B) \int_t^\infty \mathbf{1}_C(s) c(x) e^{-c(x)(s-t)} ds \\
&= \mu(x, B) \int_{t+v}^\infty \mathbf{1}_C(s - v) c(x) e^{-c(x)(s-v-t)} ds \\
&= \mu(x, B) \int_{t+v}^\infty \mathbf{1}_{C+v}(s) c(x) e^{-c(x)(s-(t+v))} ds \\
&= \tilde{\mu}((x, t + v), B \times C + (0, v))
\end{aligned}$$

By Lemma 2.71 probability measures are uniquely determined by values on a generating π-system it follows that $\tilde{\mu}((x, t), A) = \tilde{\mu}((x, t + v), A + (0, v))$ for all $A \in \mathcal{S} \times \mathcal{B}([0, \infty))$.

Now can apply the Ionescu Tulcea Theorem ??? to see that for every probability measure ν on T there exists a probability measure $\mathbf{P}_\nu$ on Ω such that Z_n is a discrete time Markov process with initial distribution, the transition kernel $\tilde{\mu}$ and the natural filtration $\mathcal{G}_n = \sigma(Z_0, Z_1, \dots, Z_n)$.

Observe that for all $\alpha > 0$ we have from the tower property of conditional expectation and Theorem 8.35 the following

$$\begin{aligned}
\mathbf{E}_\nu\left[e^{-\alpha(\tau_{n+1}-\tau_n)}\right] &= \mathbf{E}_\nu\left[\mathbf{E}_\nu\left[e^{-\alpha(\tau_{n+1}-\tau_n)} \mid \mathcal{G}_n\right]\right] \\
&= \mathbf{E}_\nu\left[\int e^{-\alpha(s-\tau_n)}\,\mathbf{1}_{[\tau_n,\infty)}(s)\mu(Y_n,dy)c(Y_n)e^{-c(Y_n)(s-\tau_n)}ds\right] \\
&= \mathbf{E}_\nu\left[e^{(\alpha+c(Y_n))\tau_n}c(Y_n)\int_{\tau_n}^\infty e^{-(\alpha+c(Y_n))s}\,ds\right] \\
&= \mathbf{E}_\nu\left[e^{(\alpha+c(Y_n))\tau_n}c(Y_n)\frac{e^{-(\alpha+c(Y_n))\tau_n}}{\alpha+c(Y_n)}\right] = \mathbf{E}_\nu\left[\frac{c(Y_n)}{\alpha+c(Y_n)}\right]
\end{aligned}$$

By the above computation and the Monotone Convergence Theorem

$$\begin{aligned}
0 = \lim_{\alpha\to\infty}\mathbf{E}_\nu\left[\frac{c(Y_n)}{\alpha+c(Y_n)}\right] &= \lim_{\alpha\to\infty}\mathbf{E}_\nu\left[e^{-\alpha(\tau_{n+1}-\tau_n)}\right] \\
\geq \lim_{\alpha\to\infty}\mathbf{E}_\nu\left[e^{-\alpha(\tau_{n+1}-\tau_n)};\tau_{n+1}\leq\tau_n\right] &\geq \mathbf{P}_\nu\{\tau_{n+1}\leq\tau_n\}
\end{aligned}$$

Thus we have $\tau_{n+1} > \tau_n$ ν-a.s. for every initial measure ν. A similar but simpler computation shows

$$\begin{aligned}
\mathbf{P}_\nu\{Y_{n+1}\neq Y_n\} = \mathbf{E}_\nu\left[\mathbf{P}_\nu\{Y_{n+1}\neq Y_n \mid \mathcal{G}_n\}\right] &= \mathbf{E}_\nu\left[\int \mathbf{1}_{S\setminus\{Y_n\}}(y)\,\mu(Y_n,dy)\right] \\
&= \mathbf{E}_\nu\left[1-\mu(Y_n,\{Y_n\})\right] = 1
\end{aligned}$$

Lastly if we let $\mathbf{P}_x$ be the shorthand for $\mathbf{P}_{\delta_{(x,0)}}$ then $\mathbf{P}_x\{\tau_0 = 0\} = 1$ for all $x \in S$. Thus let

$$\begin{aligned}
\tilde{\Omega} &= \cap_{n=0}^\infty\{\tau_{n+1}>\tau_n\}\cap\cap_{n=0}^\infty\{Y_{n+1}>Y_n\}\cap\{\tau_0=0\} \\
&= \{((x_0,t_0),(x_1,t_1),\dots) \mid t_0=0, x_n\neq x_{n+1} \text{ and } t_n<t_{n+1} \text{ for all } n\in\mathbb{Z}_+\}
\end{aligned}$$

and it follows that $\tilde{\Omega} \in \mathcal{A}$ and $\mathbf{P}_x\{\tilde{\Omega}\} = 1$ for all $x \in S$.

TODO: Add the definitions of all of the elements of the Markov family. Now let

$$\tilde{\Omega} = \cap_{n=0}^\infty\{Y_n\neq Y_{n+1}\}\cap\cap_{n=0}^\infty\{\tau_n<\tau_{n+1}\}\cap\{\tau_0=0\}$$

From the above discussion we know that $\tilde{\Omega}$ is $\mathcal{A}$-measurable and $\mathbf{P}_x\{\tilde{\Omega}\} = 1$. Thus we may define the restrict the probability measures, the *sigma*-algebra $\mathcal{A}$ and the filtration $\tilde{\mathcal{G}}_n$ to $\tilde{\Omega}$ and Z_n remains a Markov process which considered with the probability space $(\tilde{\Omega}, \tilde{\mathcal{A}})$ and the filtration $\tilde{\mathcal{G}}_n$.

Claim 13.79.2. $\tilde{\mathcal{G}}_n$ is the σ-algebra generated by the coordinate projections Z_j for $j = 0, \dots, n$.

TODO.

At this point we redefine Ω as $\tilde{\Omega}$ with a measurable point ω_Δ adjoined. For each $x \in S$, we extend the measure $\mathbf{P}_x$ by defining $\mathbf{P}_x\{Delta\} = 0$. We define $\mathbf{P}_A\Delta = \delta_\Delta(A)$. Define $Y_n(\omega_\Delta) = \Delta$ and $\tau_n(\omega_\Delta) = \infty$ for all $n \in \mathbb{Z}_+$.

For $\omega \in \tilde{\Omega}$, let $\zeta(\omega) = \lim_{n\to\infty} \tau_n(\omega)$ which as a limit of measurable function is measurable (Lemma 2.14) ; define $\zeta(\omega_\Delta) = 0$. For $\omega \in \tilde{\Omega}$, define

$$X_t(\omega) = \begin{cases} Y_n(\omega) & \text{for } \tau_n(\omega) \leq t < \tau_{n+1}(\omega) \\ \Delta & \text{for } t \geq \zeta(\omega) \end{cases}$$

and $X_\infty(\omega) = \Delta$. Define $X_t(\omega_\Delta) = \Delta$ for all $0 \leq t \leq \infty$.

X_t is seen to be measurable by writing for each $A \in \mathcal{S}^\Delta$,

$$X_t^{-1}(A) = (\{\Delta \in A\} \cap \{t \geq \zeta\}) \cup \cup_{n=0}^\infty (\{Y_n \in A\} \cap \{\tau_n \leq t < \tau_{n+1}\})$$

By definition and the fact that $\tau_n < \tau_{n+1}$ everywhere on Ω we see that X_t is a step function on the interval $[0, \zeta)$ (in particular it is cadlag when S is given the discrete topology).

Now introduce the following notation, let $\mu_\alpha(x, A) = \frac{c(x)}{\alpha + c(x)} \mu(x, A)$ and then recursively define the famility $\mu_\alpha^0(x, A) = \delta_x(A)$ and for $n \in \mathbb{N}$,

$$\mu_\alpha^n(x, A) = \mu_\alpha \mu_\alpha^{n-1}(x, A) = \int \mu_\alpha(y, A)\, \mu_\alpha^{n-1}(x, dy)$$

(note that $\mu_\alpha^1(x, A) = \int \mu_\alpha(y, A)\, \delta_x(dy) = \mu_\alpha(x, A)$ so the chosen notation is consistent). In a similar way let

$$\tilde{\mu}^n((x, t), A) = \int \tilde{\mu}((y, s), A)\, \tilde{\mu}((x, t), dy, ds)$$

and more generally for every positive measurable or integrable $f : S \times [0, \infty) \to \mathbb{R}$ we have

$$\int f(u, v), \tilde{\mu}^n((x, t), du, dv) = \iint f(u, v) \tilde{\mu}((y, s), du, dv)\, \tilde{\mu}((x, t), dy, ds)$$

TODO: Move this to conditioning chapter or an exercise.

Claim 13.79.3. If μ and ν are translation invariant kernels then so is $\mu \cdot \nu$. Moreover if f is measurable and either non-negative or integrable then $\int f(s)\, \mu(x, ds) = \int f(s - y)\, \mu(x + y, ds)$.

We prove the second claim first. Note that by translation invariance

$$\int \mathbf{1}_A(y)\, \mu(x, dy) = \mu(x, A) = \mu(x + u, A + u) = \int \mathbf{1}_{A+u}(y)\, \mu(x + u, dy)$$
$$= \int \mathbf{1}_A(y - u)\, \mu(x + u, dy)$$

so the result holds for indicator functions. Now apply the standard machinery. The first claim follows from the second,

$$\mu \cdot \nu(x, A) = \int \nu(y, A)\, \mu(x, dy) = \int \nu(y - u, A)\, \mu(x + u, dy) = \int \nu(y, A + y)\, \mu(x + u, dy)$$
$$= \mu \cdot \nu(x + u, A + u)$$

Claim 13.79.4. $\int_0^\infty e^{-\lambda(s-t)} \tilde{\mu}^n((x, t), A, ds) = \mu_\lambda^n(x, A)$.

By translation invariance we know that $\int_0^\infty e^{-\lambda(s-t)} \tilde{\mu}^n((x, t), A, ds) = \int_0^\infty e^{-\lambda s} \tilde{\mu}^n((x, 0), A, ds)$ so it suffices to use the right hand side. For $n = 0$ and $n = 1$ we compute directly,

$$\int_0^\infty e^{-\lambda s} \tilde{\mu}^0((x, 0), A, ds) = \int_0^\infty e^{-\lambda s} \mathbf{1}_A(y) \delta_{(x,0)}(dy, ds) = \mathbf{1}_A(x) = \mu_\alpha^0(x, A)$$

and

$$\int_0^\infty e^{-\lambda s} \tilde{\mu}((x,0), A, ds) = \mu(x,A)c(x) \int_0^\infty e^{-s(\lambda + c(x)} \, ds = \mu(x,A) \frac{c(x)}{\lambda + c(x)} = \mu_\lambda(x,A)$$

For general n we use an induction argument,

$$\begin{aligned}
\int_0^\infty e^{-\lambda s} \tilde{\mu}^{n+1}((x,0), A, ds) &= \iint e^{-\lambda s} \tilde{\mu}^n((u,v), A, ds) \tilde{\mu}((x,0), du, dv) \\
&= \iint e^{-\lambda v} \mu_\lambda^n(u, A) \tilde{\mu}((x,0), du, dv) \\
&= \int \mu_\lambda^n(u,A) c(x) \left[\int_0^\infty e^{-v(\lambda + c(x))} dv \right] \mu(x, du) \\
&= \int \mu_\lambda^n(u,A) \frac{c(x)}{\lambda + c(x)} \mu(x, du) \\
&= \int \mu_\lambda^n(u,A) \mu_\lambda(x, du) = \mu_\lambda^{n+1}(x, A)
\end{aligned}$$

Claim 13.79.5. Let $g : S \to [0,\infty)$ be measurable and $\lambda \geq 0$ then for all $m \geq n$,

$$\mathbf{E}_x \left[g(Y_m)(e^{-\lambda \tau_m} - e^{-\lambda \tau_{m+1}}) \mid \mathcal{G}_n \right] = \lambda e^{-\lambda \tau_n} \int \frac{g(y)}{\lambda + c(y)} \, \mu_\lambda^{m-n}(Y_n, dy)$$

By the Markov property we know that $\mathbf{P}\{(Y_m, \tau_m) \in A \mid \mathcal{G}_n\} = \tilde{\mu}^{m-n}((Y_m, \tau_m), A)$ and therefore by Theorem 8.35 $\mathbf{E}\left[f(Y_m, \tau_m, Y_n, \tau_n) \mid \mathcal{G}_n \right] = \int f(y, s, Y_n, \tau_n) \, \tilde{\mu}^{m-n}((Y_n, \tau_n), dy, ds)$. We compute using two applications of this observation, the tower and pullout properties of conditional expectation and the previous claim to see

$$\begin{aligned}
&\mathbf{E}_x \left[g(Y_m)(e^{-\lambda \tau_m} - e^{-\lambda \tau_{m+1}}) \mid \mathcal{G}_n \right] = \mathbf{E}_x \left[g(Y_m) e^{-\lambda \tau_m} \mathbf{E}_x \left[1 - e^{-\lambda(\tau_{m+1} - \tau_m)} \mid \mathcal{G}_m \right] \mid \mathcal{G}_n \right] \\
&= \mathbf{E}_x \left[g(Y_m) e^{-\lambda \tau_m} \left(1 - \int_{\tau_m}^\infty e^{-\lambda(s - \tau_m)} c(Y_m) e^{-c(Y_m)(s - \tau_m)} \, ds \right) \mid \mathcal{G}_n \right] \\
&= \mathbf{E}_x \left[g(Y_m) e^{-\lambda \tau_m} \left(1 - c(Y_m) e^{\tau_m(\lambda + c(Y_m))} \int_{\tau_m}^\infty e^{-s(\lambda + c(Y_m))} \, ds \right) \mid \mathcal{G}_n \right] \\
&= \mathbf{E}_x \left[g(Y_m) e^{-\lambda \tau_m} \left(1 - c(Y_m) e^{\tau_m(\lambda + c(Y_m))} \frac{e^{-\tau_m(\lambda + c(Y_m))}}{\lambda + c(Y_m)} \right) \mid \mathcal{G}_n \right] \\
&= \mathbf{E}_x \left[g(Y_m) e^{-\lambda \tau_m} \frac{\lambda}{\lambda + c(Y_m)} \mid \mathcal{G}_n \right] \\
&= \iint g(y) e^{-\lambda s} \frac{\lambda}{\lambda + c(y)} \, \tilde{\mu}^{m-n}((Y_n, \tau_n), dy, ds) \\
&= \lambda e^{-\lambda \tau_m} \iint e^{-\lambda(s - \tau_m)} \frac{g(y)}{\lambda + c(y)} \, \tilde{\mu}^{m-n}((Y_n, \tau_n), dy, ds) \\
&= \lambda e^{-\lambda \tau_m} \int \frac{g(y)}{\lambda + c(y)} \, \mu_\lambda^{m-n}(Y_n, dy)
\end{aligned}$$

We also need the following similar computation

CLAIM 13.79.6. Let $g : S \to [0, \infty)$ be measurable and $\lambda \geq 0$ then for all $n \geq 0$,

$$\mathbf{E}_x \left[e^{-\lambda \tau_{n+1}} g(Y_{n+1}); \tau_{n+1} > t \mid \mathcal{G}_n \right] = e^{c(Y_n)\tau_n - (\lambda + c(Y_n))(t \vee \tau_m)} \int g(y) \mu_\alpha(Y_n, dy)$$

This is just an application of the Disintegration Theorem as above

$$\begin{aligned}
&\mathbf{E}_x \left[e^{-\lambda \tau_{n+1}} g(Y_{n+1}); \tau_{n+1} > t \mid \mathcal{G}_n \right] \\
&= \iint e^{-\lambda s} g(y) \mathbf{1}_{(t,\infty)}(s) \mathbf{1}_{[\tau_n,\infty)}(s) c(Y_n) e^{-c(Y_n)(s-\tau_n)} \, ds \, \mu(Y_n, dy) \\
&= \int g(y) c(Y_n) e^{c(Y_n)\tau_n} \left[\int_{t \vee \tau_n}^\infty e^{-(\lambda + c(Y_n))s} \, ds \right] \mu(Y_n, dy) \\
&= \int g(y) e^{c(Y_n)\tau_n} \frac{c(Y_n)}{\lambda + c(Y_n)} e^{-(\lambda + c(Y_n)) t \vee \tau_n} \mu(Y_n, dy) \\
&= e^{c(Y_n)\tau_n - (\lambda + c(Y_n)) t \vee \tau_n} \int g(y) \, \mu_\lambda(Y_n, dy)
\end{aligned}$$

CLAIM 13.79.7. Let $f : S \to [0, \infty)$ be Borel measurable and extend to S^Δ by defining $f(\Delta) = 0$ then if we define

$$R_\lambda f(x) = \int_0^\infty e^{-\lambda s} T_s f(x) \, ds = \int_0^\infty e^{-\lambda s} \mathbf{E}_x \left[f(X_s) \right] \, ds = \mathbf{E}_x \left[\int_0^\infty e^{-\lambda s} f(X_s) \, ds \right]$$

then we have

$$R_\lambda f(x) = \sum_{n=0}^\infty \int \frac{f(y)}{\lambda + c(y)} \, \mu_\lambda^n(x, dy) = \frac{f(x)}{\lambda + c(x)} + \int R_\lambda f(y) \mu_\lambda(x, dy)$$

Using $f(\Delta) = 0$ and the fact that $X_t = \Delta$ on $[\zeta, \infty]$ we have

$$\begin{aligned}
\int_0^\infty e^{-\lambda s} f(X_s) \, ds = \int_0^\zeta e^{-\lambda s} f(X_s) \, ds &= \sum_{n=0}^\infty \int_{\tau_n}^{\tau_{n+1}} e^{-\lambda s} f(Y_n) \, ds \\
&= \sum_{n=0}^\infty f(Y_n) \frac{e^{-\lambda \tau_n} - e^{-\lambda \tau_{n+1}}}{\lambda}
\end{aligned}$$

So taking expectations using two applications of this fact, Lemma 2.44, the Claim 13.79.5 and the definition of μ^n_λ we get

$$\begin{aligned}
R_\lambda f(x) &= \mathbf{E}_x\left[\sum_{n=0}^\infty f(Y_n)\frac{e^{-\lambda\tau_n} - e^{-\lambda\tau_{n+1}}}{\lambda}\right] \\
&= \sum_{n=0}^\infty \int \frac{f(y)}{\lambda + c(y)}\, \mu^n_\lambda(x, dy) \\
&= \frac{f(x)}{\lambda + c(x)} + \sum_{n=1}^\infty \int \frac{f(y)}{\lambda + c(y)}\, \mu^n_\lambda(x, dy) \\
&= \frac{f(x)}{\lambda + c(x)} + \sum_{n=1}^\infty \int \left[\int \frac{f(y)}{\lambda + c(y)}\, \mu^{n-1}_\lambda(w, dy)\right] \mu_\lambda(x, dw) \\
&= \frac{f(x)}{\lambda + c(x)} + \int R_\lambda f(y)\, \mu_\lambda(x, dy)
\end{aligned}$$

CLAIM 13.79.8. Given $A \in \mathcal{F}^0_t$ then for all $n \in \mathbb{Z}_+$ there exists $A_n \in \mathcal{G}_n$ such that

$$A \cap \{\tau_n \leq t < \tau_{n+1}\} = A_n\{t < \tau_{n+1}\}$$

Let $\mathcal{C}$ be the set of sets in $\mathcal{A}$ satisfying the criteria of the claim; we show that $\mathcal{C}$ is a σ-algebra and that each X_s for $s \leq t$ is $\mathcal{C}$-measurable. First we show that $\mathcal{C}$ is a σ-algebra. Let $A \in \mathcal{C}$ and $n \in \mathbb{Z}_+$. Pick $A_n \in \mathcal{G}_n$ such that $A \cap \{\tau_n \leq t < \tau_{n+1}\} = A_n\{t < \tau_{n+1}\}$. Let $B_n = A^c_n \cap \{\tau_n \geq t\} \in \mathcal{G}_n$. If $\omega \in A^c \cap \{\tau_n \leq t < \tau_{n+1}\}$ then it must be that

$$\omega \notin B_n \cap \{t < \tau_{n+1}\} = A^c_n \cap \{\tau_n \geq t\} \cap \{t < \tau_{n+1}\} = A^c_n \cap \{t < \tau_{n+1}\}$$

since otherwise we would have $\omega \in A$. Since $\omega \in \{\tau_n \geq t\} \cap \{t < \tau_{n+1}\} = \{\tau_n \leq t < \tau_{n+1}\}$ it follows that $A^c \cap \{\tau_n \leq t < \tau_{n+1}\} \subset B_n \cap \{t < \tau_{n+1}\}$. On the other hand, if $\omega \in B_n \cap \{t < \tau_{n+1}\} = A^c_n \cap \{\tau_n \leq t < \tau_{n+1}\}$ then $\omega \notin A$ since otherwise we would get that $\omega \in A \cap \{\tau_n \leq t < \tau_{n+1}\} = A_n \cap \{t < \tau_{n+1} \subset A_n$ which contradicts $\omega \in B_n = \{t \geq \tau_n\} \setminus A_n$. Therefore we conclude $A^c \cap \{\tau_n \leq t < \tau_{n+1}\} = B_n \cap \{t < \tau_{n+1}\}$ and therefore $A^c \in \mathcal{C}$. Now let $A_1, A_2, \ldots \in \mathcal{C}$ and $n \in \mathbb{Z}_+$ and $A_{j,n} \in \mathcal{G}_n$ be chosen so that $A_j \cap \{\tau_n \leq t < \tau_{n+1}\} = A_{j,n} \cap \{t \geq \tau_n\}$. Now we have

$$\begin{aligned}
\left(\cap_{j=1}^\infty A_j\right) \cap \{\tau_n \leq t < \tau_{n+1}\} &= \cap_{j=1}^\infty \left(A_j \cap \{\tau_n \leq t < \tau_{n+1}\}\right) = \cap_{j=1}^\infty \left(A_{j,n} \cap \{t < \tau_{n+1}\}\right) \\
&= \left(\cap_{j=1}^\infty A_{j,n}\right) \cap \{t < \tau_{n+1}\}
\end{aligned}$$

and since $\cap_{j=1}^\infty A_{j,n} \in \mathcal{G}_n$ we see that $\cap_{j=1}^\infty A_j \in \mathcal{C}$. Clearly $\emptyset \in \mathcal{C}$ and it follows that $\mathcal{C}$ is a σ-algebra. To see that X_s is $\mathcal{C}$-measurable for $s \leq t$, let $A \in \mathcal{S}$ and note that

$$\begin{aligned}
\{X_s \in A\} \cap \{\tau_n \leq t < \tau_{n+1}\} &= \cup_{j=0}^n \left(\{X_s \in A\} \cap \{\tau_j \leq s < \tau_{j+1}\}\right) \cap \{\tau_n \leq t < \tau_{n+1}\} \\
&= \cup_{j=0}^n \left(\{Y_j \in A\} \cap \{\tau_j \leq s < \tau_{j+1}\}\right) \cap \{\tau_n \leq t < \tau_{n+1}\} \\
&= \left(\{Y_n \in A; \tau_n \leq s\} \cup \cup_{j=0}^{n-1}\{Y_j \in A; \tau_j \leq s < \tau_{j+1}\}\right) \cap \{\tau_n \leq t\} \cap \{\leq t < \tau_{n+1}\}
\end{aligned}$$

which shows $\{X_s \in A\} \in \mathcal{C}$ hence $\mathcal{F}^0_t \subset \mathcal{C}$ and the claim is proven.

The following claim is essentially the Laplace Transform version of the Markov property that we seek to prove. In terms of financial ideas, suppose we have a fixed interest rate λ. The next claim says the that the expected net present value of the cash flow $f(X_s)$ starting at t is obtained by taking the expected net present value starting at $s = 0$ assuming that we start at X_t and then discounting that amount by $e^{-t\lambda}$.

CLAIM 13.79.9. Let $f : S \to [0, \infty)$, $t \geq 0$, $\lambda \geq 0$ and $x \in S$. If we extend f to S^Δ by defining $f(\Delta) = 0$ then it follows that

$$\mathbf{E}_x \left[\int_t^\infty e^{-\lambda u} f(X_u)\, du \mid \mathcal{F}_t^0 \right] = e^{-\lambda t} R_\lambda f(X_t)$$

Let $A \in \mathcal{F}_t^0$. Since $f(\Delta) = 0$ and $X_t = \Delta$ for $t \geq \zeta$ it follows that

$$\mathbf{E}_x \left[\int_t^\infty e^{-\lambda u} f(X_u)\, du; A \cap \{t \geq \zeta\} \right] = 0$$

and moreover

$$\begin{aligned}
\mathbf{E}_x \left[e^{-\lambda t} R_\lambda f(X_t); A \cap \{t \geq \zeta\} \right] &= \mathbf{E}_x \left[e^{-\lambda t} R_\lambda f(\Delta); A \cap \{t \geq \zeta\} \right] \\
&= \mathbf{E}_x \left[e^{-\lambda t} \int_0^\infty e^{-\lambda s} \mathbf{E}_\Delta \left[f(X_s) \right] ds; A \cap \{t \geq \zeta\} \right] \\
&= \mathbf{E}_x \left[e^{-\lambda t} \int_0^\infty e^{-\lambda s} f(\Delta)\, ds; A \cap \{t \geq \zeta\} \right] = 0
\end{aligned}$$

Therefore we may assume that $A \subset \{t < \zeta\}$ and in particular by linearity we may assume that there exists $n \in \mathbb{Z}_+$ such that $A \subset \{\tau_n \leq t < \tau_{n+1}\}$. By the previous claim, we may choose an $A_n \in \mathcal{G}_n$ such that $A_n \subset \{\tau_n \leq t\}$ and $A = A_n \cap \{t < \tau_{n+1}\}$. Therefore we can decompose the integral

$$\begin{aligned}
&\mathbf{E}_x \left[\int_t^\infty e^{-\lambda u} f(X_u)\, du; A_n \cap \{t < \tau_{n+1}\} \right] \\
&= \mathbf{E}_x \left[\int_t^{\tau_{n+1}} e^{-\lambda u} f(X_u)\, du; A_n \cap \{t < \tau_{n+1}\} \right] + \\
&\quad \sum_{k=n+1}^\infty \mathbf{E}_x \left[\int_{\tau_k}^{\tau_{k+1}} e^{-\lambda u} f(X_u)\, du; A_n \cap \{t < \tau_{n+1}\} \right]
\end{aligned}$$

We compute in each of the two cases on the right hand side. For the first term using Claim 13.79.6 and recalling that $A_n \subset \{\tau_n \leq t\}$,

$$\begin{aligned}
&\mathbf{E}_x \left[\int_t^{\tau_{n+1}} e^{-\lambda u} f(X_u)\, du; A_n \cap \{t < \tau_{n+1}\} \right] \\
&= \mathbf{E}_x \left[f(Y_n) \lambda^{-1} (e^{-\lambda t} - e^{-\lambda \tau_{n+1}}); A_n \cap \{t < \tau_{n+1}\} \right] \\
&= \lambda^{-1} \mathbf{E}_x \left[f(Y_n); A_n; \mathbf{E}_x \left[(e^{-\lambda t} - e^{-\lambda \tau_{n+1}}); t < \tau_{n+1} \mid \mathcal{G}_n \right] \right] \\
&= \lambda^{-1} \mathbf{E}_x \left[f(Y_n); A_n; e^{-c(Y_n)(t-\tau_n)} e^{-\lambda t} (\mu_0(Y_n, S) - \mu_\lambda(Y_n, S)) \right] \\
&= \lambda^{-1} \mathbf{E}_x \left[f(Y_n); A_n; e^{-c(Y_n)(t-\tau_n)} e^{-\lambda t} (1 - \frac{c(Y_n)}{\lambda + c(Y_n)}) \right] \\
&= e^{-\lambda t} \mathbf{E}_x \left[A_n; e^{-c(Y_n)(t-\tau_n)} \frac{f(Y_n)}{\lambda + c(Y_n)} \right]
\end{aligned}$$

For the summands of the second term for $k \geq n+1$ we use two applications of Claim 13.79.5

$$
\begin{aligned}
&\mathbf{E}_x \left[\int_{\tau_k}^{\tau_{k+1}} e^{-\lambda u} f(X_u)\, du; A_n \cap \{t < \tau_{n+1}\} \right] \\
&= \mathbf{E}_x \left[f(Y_k) \lambda^{-1} (e^{-\lambda \tau_k} - e^{-\lambda \tau_{k+1}}); A_n \cap \{t < \tau_{n+1}\} \right] \\
&= \lambda^{-1} \mathbf{E}_x \left[\mathbf{E}_x \left[f(Y_k)(e^{-\lambda \tau_k} - e^{-\lambda \tau_{k+1}}) \mid \mathcal{G}_{n+1} \right]; A_n \cap \{t < \tau_{n+1}\} \right] \\
&= \mathbf{E}_x \left[e^{-\lambda \tau_{n+1}} \int \frac{f(y)}{\lambda + c(y)}\, \mu_\lambda^{k-n-1}(Y_{n+1}, dy); A_n \cap \{t < \tau_{n+1}\} \right] \\
&= \mathbf{E}_x \left[A_n; \mathbf{E}_x \left[e^{-\lambda \tau_{n+1}} \int \frac{f(y)}{\lambda + c(y)}\, \mu_\lambda^{k-n-1}(Y_{n+1}, dy); \cap \{t < \tau_{n+1}\} \mid \mathcal{G}_n \right] \right] \\
&= \mathbf{E}_x \left[A_n; e^{-c(Y_n)(t-\tau_n)} e^{-\lambda t} \int \left[\int \frac{f(y)}{\lambda + c(y)}\, \mu_\lambda^{k-n-1}(u, dy) \right] \mu_\lambda(Y_n, du) \right] \\
&= e^{-\lambda t} \mathbf{E}_x \left[A_n; e^{-c(Y_n)(t-\tau_n)} \int \frac{f(y)}{\lambda + c(y)}\, \mu_\lambda^{k-n}(Y_n, dy) \right]
\end{aligned}
$$

Putting this all together we get using Claim 13.79.7, Claim 13.79.6 and the fact that $A_n \subset \{\tau_n \leq t\}$

$$
\begin{aligned}
&\mathbf{E}_x \left[\int_t^\infty e^{-\lambda u} f(X_u)\, du; A_n \cap \{t < \tau_{n+1}\} \right] \\
&= e^{-\lambda t} \mathbf{E}_x \left[A_n; e^{-c(Y_n)(t-\tau_n)} \frac{f(Y_n)}{\lambda + c(Y_n)} \right] + \\
&\qquad \sum_{k=n+1}^\infty e^{-\lambda t} \mathbf{E}_x \left[A_n; e^{-c(Y_n)(t-\tau_n)} \int \frac{f(y)}{\lambda + c(y)}\, \mu_\lambda^{k-n}(Y_n, dy) \right] \\
&= e^{-\lambda t} \mathbf{E}_x \left[A_n; e^{-c(Y_n)(t-\tau_n)}; \sum_{k=0}^\infty \int \frac{f(y)}{\lambda + c(y)}\, \mu_\lambda^{k}(Y_n, dy) \right] \\
&= e^{-\lambda t} \mathbf{E}_x \left[A_n; R_\lambda f(Y_n); e^{-c(Y_n)(t-\tau_n)} \right] \\
&= e^{-\lambda t} \mathbf{E}_x \left[A_n; R_\lambda f(Y_n); \mathbf{P}_x\{t < \tau_{n+1} \mid \mathcal{G}_n\} \right] \\
&= \mathbf{E}_x \left[e^{-\lambda t} R_\lambda f(Y_n); A_n \cap \{t < \tau_{n+1}\} \right] \\
&= \mathbf{E}_x \left[e^{-\lambda t} R_\lambda f(X_t); A_n \cap \{t < \tau_{n+1}\} \right]
\end{aligned}
$$

and the claim follows.

Now we show the Markov property for X_t.

CLAIM 13.79.10. Suppose $f : S \to [0, \infty)$ is a bounded measurable function and $\varphi : [0, \infty) \to \mathbb{R}$ be a continuous function such that $\lim_{t\to\infty} \varphi(t) = 0$ then

$$
\mathbf{E}_x \left[\int_t^\infty \varphi(u) f(X_u)\, du \mid \mathcal{F}_t^0 \right] = \mathbf{E}_{X_t} \left[\int_0^\infty \varphi(u+t) R_\lambda f(X_u)\, du \right]
$$

Then for any $\varphi(t) = \sum_{j=1}^n a_j e^{-\lambda_j t}$ with $\lambda_j \geq 0$ and $-\infty < a_j < \infty$ use Claim 13.79.9 and linearity of expecations to see that $\mathbf{E}_x \left[\int_t^\infty \varphi(u) f(X_u)\, du \mid \mathcal{F}_t^0 \right] = \mathbf{E}_{X_t} \left[\int_0^\infty \varphi(u+t) R_\lambda f(X_u)\, du \right]$. Now let $\varphi : [0, \infty) \to \mathbb{R}$ be a continuous function such that $\lim_{t\to\infty} \varphi(t) = 0$ then we may extend φ to a continuous function $\varphi : [0,\infty] \to \mathbb{R}$ by defining $\varphi(\infty) = 0$. By the Stone Weierstrass Theorem 1.44 we may therefore approximate φ uniformly by finite linear combinations of exponentials. TODO: Show that the uniform approximation allows us to extend to such φ; to be honest I don't see this yet given that the domain of integration is infinite...

Now for any $s \geq 0$ we may choose continuous compactly supported φ_n such that φ_n approximate the delta function at $t+s$ e.g.

$$\varphi_n(u) = \begin{cases} n(u-t-s) & \text{for } t+s \leq u \leq t+s+\frac{1}{n} \\ n(t+s-u)+2 & \text{for } t+s+\frac{1}{n} \leq u \leq t+s+\frac{2}{n} \\ 0 & \text{otherwise} \end{cases}$$

By the right continuity of X_t for almost all $\omega \in \Omega$ there is an N_ω such that $\varphi_n(u) f(X_u) = f(X_{t+s})\varphi_n(u)$ and $\varphi_n(u+t) f(X_u) = \varphi_n(u+t) f(X_s)$ for all $n \geq N_\omega$. In particular $\int_t^\infty \varphi_n(u) f(X_u)\, du \stackrel{a.s.}{\to} f(X_{t+s})$ and $\int_0^\infty \varphi_n(u+t) f(X_u)\, du \stackrel{a.s.}{\to} f(X_s)$ and by bounded convergence we get

$$\begin{aligned} \mathbf{E}_x \left[f(X_{t+s}) \mid \mathcal{F}_t^0 \right] &= \lim_{n\to\infty} \mathbf{E}_x \left[\int_t^\infty \varphi_n(u) f(X_u)\, du \mid \mathcal{F}_t^0 \right] \\ &= \lim_{n\to\infty} \mathbf{E}_{X_t} \left[\int_0^\infty \varphi_n(u+t) f(X_u)\, du \right] \\ &= \mathbf{E}_{X_t} \left[f(X_s) \right] \end{aligned}$$

Note that to extend the above construction to account for absorbing states we need to change the definition of our probability space. Instead of constructing a probability measure on the set of sequences $((x_0,t_0),(x_1,t_1),\dots) \in S \times [0,\infty)$ as our starting point we need to account for the fact that we may have only finitely many jumps. This is done by allowing t_n to be infinite with the provisions that when t_n is infinite we must have $x_n = \Delta$ and that t_m must be infinite for all $m \geq n$. With this definition we should be able to relax the restriction that $c(x) > 0$. We should look to see if we can prove that $\mathbf{P}\{\tau_n = \infty, Y_n = \Delta\} = \mathbf{P}\{\tau_n = \infty\}$ and $\mathbf{P}\{\tau_n = \infty, \tau_{n+1} = \infty\} = \mathbf{P}\{\tau_n = \infty\}$ after applying Ionescu-Tulcea. □

TODO: Kolmogorov Backward equation.

Let X be a pure jump-type Markov process on state space S, τ be the first jump time and let $\sigma = \tau \wedge t$ for some $0 \leq t < \infty$. Now by the Strong Markov Property, the independence of X_τ and τ and the disintegration Lemma 4.6

$$\begin{aligned} T_t f(x) &= \mathbf{E}_x \left[f(X_t) \right] = \mathbf{E}_x \left[f((\theta_\sigma X)_{t-\sigma}) \right] = \mathbf{E}_x \left[\mathbf{E} \left[f((\theta_\sigma X)_{t-\sigma}) \mid \mathcal{F}_\sigma \right] \right] \\ &= \mathbf{E}_x \left[\mathbf{E}_{X_\sigma} \left[f(X_{t-\sigma}) \right] \right] = \mathbf{E}_x \left[T_{t-\sigma} f(X_\sigma) \right] \\ &= \mathbf{E}_x \left[T_{t-\sigma} f(X_\sigma); \tau > t \right] + \mathbf{E}_x \left[T_{t-\sigma} f(X_\sigma); \tau \leq t \right] \\ &= \mathbf{E}_x \left[T_0 f(X_0); \tau > t \right] + \mathbf{E}_x \left[T_{t-\tau} f(X_\tau); \tau \leq t \right] \\ &= f(x) \mathbf{P}_x\{\tau > t\} + \int_0^t \int c(x) e^{-sc(x)} T_{t-s} f(y)\, \mu(x,dy) ds \\ &= e^{-tc(x)} \left(f(x) + \int_0^t \int c(x) e^{-sc(x)} T_s f(y)\, \mu(x,dy) ds \right) \end{aligned}$$

TODO: There is some measurability subtlety in applying disintegration; we need to know that $T_t f(x)$ is jointly measurable in (t, x) and then apply it to (τ, X_τ) and use the fact that X_τ is progressively measurable by right continuity and thus X_τ is measurable by Lemma 9.90. TODO: How do the measurability considerations in the definition of Markov family enter into the picture here.

CLAIM 13.79.11. Let $(\Omega, \mathcal{A}, \mathcal{F}_t, X_t, \theta_t, P_x)$ be a Markov family with X jointly measurable, let f be a bounded or non-negative measurable function and define $T_t f(x) = \mathbf{E}_x [f(X_t)]$ then $T_t f(x)$ is a jointly measurable function $[0, \infty) \times S \to \mathbb{R}$.

To see the claim we just use the kernel property of P_x and write $T_t f(x) = \mathbf{E}_x [f(X_t)] = \int X(t, \omega) \, P_x(d\omega)$ and apply Exercise 22. TODO: Since X is $\mathcal{B}([0, \infty)) \otimes \mathcal{F}^X_\infty$ measurable I don't think the universal measurability issues enter into the discussion here.

CHAPTER 14

Stochastic Integration

1. Local Martingales

Definition 14.1. Let M_t be an $\mathcal{F}$-adapted process, we say M is a *local martingale* if there exists a sequence of optional times τ_n such that $\tau_n \uparrow \infty$ a.s. and $M^{\tau_n} - M_0$ is an $\mathcal{F}$-martingale for all n. We say that τ_n is a *localizing sequence* for M.

It is useful to note that a local martingale can be localized to martingales with nice properties. In the general case we can always assume that we localize to a uniformly integrable martingale.

Lemma 14.2. *Let M be a local martingale with a localizing sequence $\tau_n \uparrow \infty$, then $\tau_n \wedge n$ is a localizing sequence such that $M^{\tau_n \wedge n}$ is a uniformly integrable martingale for each $n \in \mathbb{N}$.*

Proof. It is clear that $\tau_n \wedge n$ is a sequence of optional times such that $\tau_n \wedge n \uparrow \infty$. Moreover, since $(M - M_0)^{\tau_n}$ is a cadlag martingale, n is a bounded optional time and $(M-M_0)^{\tau_n \wedge n}_t = (M-M_0)_{\tau_n \wedge n \wedge t} = ((M-M_0)^{\tau}_n)^n_t$ the Optional Sampling Theorem 9.82 tells us that $(M-M_0)^{\tau_n \wedge n}$ is closable hence uniformly integrable. $\square$

Even better is the fact is that continuous local martingales can always be localized to bounded martingales. This is one of the general facts that makes the theory of continuous local martingales easier than the general case.

Lemma 14.3. *Let M be a continuous local martingale and for each $n \in \mathbb{Z}_+$ let $\tau_n = \inf\{t \geq 0 \mid |M_t| \geq n\}$ then τ_n is a localizing sequence for M.*

Proof. By continuity and $\mathcal{F}$-adaptedness of M and the fact that $[t, \infty)$ is closed we know that τ_n is an optional time (Lemma 9.70). It is clear that $|M_t| \geq n$ implies $|M_t| \geq n-1$ and therefore τ_n is an increasing sequence. By continuity of M we know that M is bounded on bounded intervals and therefore $\tau_n \uparrow \infty$ a.s.

It remains to show that $(M - M_0)^{\tau_n}$ is a martingale for every $n \in \mathbb{Z}_+$. Let σ_m be a localizing sequence for M. From Optional Sampling we know that $(M - M_0)^{\tau_n \wedge \sigma_m} = ((M-M_0)^{\sigma_m})^{\tau_n}$ is a martingale for every $m, n \in \mathbb{Z}_+$. Furthermore for fixed n and every $m \in \mathbb{Z}_+$ since $\sigma_m \uparrow \infty$ a.s. we know that $(M-M_0)^{\tau_n \wedge \sigma_m}_t \overset{a.s.}{\to} (M - M_0)^{\tau_n}_t$. Morever $\left|(M - M_0)^{\tau_n \wedge \sigma_m}_t\right| = |M_{\tau_n \wedge \sigma_m \wedge t} - M_0| \leq |M_0| + n$. Since M_0 is integrable (TODO: Do we really know this with the Kallenberg definition of a local martingale?; if not what replaces it do we define $\tau_n = \inf\{t \geq 0 \mid |M_t - M_0| \geq n\}$?) so that by Dominated Convergence we get $(M - M_0)^{\tau_n \wedge \sigma_m}_t \overset{L^1}{\to} (M - M_0)^{\tau_n}_t$ as well. Using both forms of convergence and the martingale property of $M^{\tau_n \wedge \sigma_m}$, for each

$s < t$ we get the equality

$$\begin{aligned}\mathbf{E}\left[(M - M_0)_t^{\tau_n} \mid \mathcal{F}_s\right] &= \lim_{m\to\infty} \mathbf{E}\left[(M - M_0)_t^{\tau_n \wedge \sigma_m} \mid \mathcal{F}_s\right] \\ &= \lim_{m\to\infty} (M - M_0)_s^{\tau_n \wedge \sigma_m} = (M - M_0)_s^{\tau_n} \text{ a.s.}\end{aligned}$$

which shows that M^{τ_n} is a martingale. □

It will occasionally be important to know when we may conclude a local martingale is actually a martingale. The simplest case is that of a bounded local martingale (not necessarily continuous).

LEMMA 14.4. *Let M be a bounded local martingale then it follows that M is a uniformly integrable martingale.*

PROOF. Let τ_n be a localizing sequence for M so that $M_{t\wedge\tau_n} - M_0$ is a martingale so that $\mathbf{E}\left[M_{t\wedge\tau_n} - M_0 \mid \mathcal{F}_s\right] = M_{s\wedge\tau_n} - M_0$ for every $s < t$. Now boundedness of M implies boundedness of $M_{t\wedge\tau_n} - M_0$ and therefore we may apply Dominated Convergence for conditional expectations (Lemma 8.11) and the fact that $\tau_n \uparrow \infty$ a.s. to conclude that

$$\mathbf{E}\left[M_t - M_0 \mid \mathcal{F}_s\right] = \lim_{n\to\infty} \mathbf{E}\left[M_{t\wedge\tau_n} - M_0 \mid \mathcal{F}_s\right] = \lim_{n\to\infty} M_{s\wedge\tau_n} - M_0 = M_s - M_0$$

almost surely. Thus M is a martingale. Since M is bounded, in fact it is uniformly integrable by Example 5.50 □

There is property that characterizes when a local martingale is in fact a martingale. We give the definition; the concept will reappear in a subsequent chapter when discussing the Doob Decomposition for continuous time submartingales.

DEFINITION 14.5. An $\mathcal{F}$-adapted process X is said to be of *class D* if and only if the set $\{X_\tau \mathbf{1}_{\tau<\infty} \mid \tau \text{ is an } \mathcal{F}\text{-optional time}\}$ is uniformly integrable. X is said to be of *class DL* (or *locally of class D*) if for every $0 \leq C < \infty$ the set $\{X_\tau \mid \tau \text{ is an } \mathcal{F}\text{-optional time with } \tau \leq C\}$ is uniformly integrable.

TODO: Maybe a better exercise since I'm not sure we'll ever use this result.

PROPOSITION 14.6. *A local martingale is a martingale if and only if is of class DL.*

PROOF. The fact that a martingale is of class DL follows immediately from the Optional Sampling Theorem 9.82 and Corollary 9.52 for by Optional Sampling for any $0 \leq C < \infty$ and $\mathcal{F}$-optional time τ we have $X_\tau = \mathbf{E}\left[X_C \mid \mathcal{F}_\tau\right]$.

Now suppose that a local martingale X is of class DL. Since X is a local martingale we know that it is adapted. Apply the class DL property to constant optional times we see that X_t is integrable for every $0 \leq t < \infty$. Let τ_n be a localizing sequence for X so that for all $0 \leq s < t < \infty$ and $n \in \mathbb{N}$ we have $\mathbf{E}\left[X_{\tau_n\wedge t} \mid \mathcal{F}_s\right] = X_{\tau_n \wedge s}$ a.s. Since $\tau_n \uparrow \infty$ a.s. we know that $X_{\tau_n\wedge s} \overset{a.s.}{\to} X_s$ and $X_{\tau_n\wedge t} \overset{a.s.}{\to} X_t$. Since X is of class DL we know that $X_{\tau_n \wedge t}$ is uniformly integrable we conclude know that $\mathbf{E}\left[X_{\tau_n\wedge t} \mid \mathcal{F}_s\right] \overset{a.s.}{\to} \mathbf{E}\left[X_t \mid \mathcal{F}_s\right]$ and therefore stringing the equalities together we see that X is a martingale. □

LEMMA 14.7. *Let $\mathcal{F}$ be a right continuous filtration, M be a cadlag $\mathcal{F}$-local martingale with localizing sequence τ_n and let σ_n be an arbitrary sequence of bounded optional times such that $\sigma_n \uparrow \infty$, then $\tau_n \wedge \sigma_n$ is a localizing sequence for M. In particular the space of cadlag $\mathcal{F}$-local martingales is a linear space.*

Proof. First we claim that every local martingale M has localizing sequence of bounded optional times. This follows from picking an arbitrary localizing sequence τ_n and then noting that $\tau_n \wedge n$ is also a localizing sequence as $\tau_n \wedge n \uparrow \infty$ a.s. and $M^{\tau_n \wedge n} - M_0 = (M^{\tau_n})^n - M_0$ is a martingale from Optional Sampling (Theorem 9.82) since M^{τ_n} is a cadlag martingale, $\mathcal{F}$ is right continuous and n is a bounded optional time.

Given τ_n and σ_n as in the hypothesis and by our first claim we assume that each τ_n and σ_n is bounded. It is clear that $\tau_n \wedge \sigma_n$ is a sequence of optional times such that $\tau_n \wedge \sigma_n \uparrow \infty$ and again applying Optional Sampling we see that $M^{\tau_n \wedge \sigma_n} - M_0 = (M^{\tau_n})^{\sigma_n} - M_0$ is a martingale.

Lastly if we are given M and N local martingales, take τ_n and σ_n to be bounded localizing sequences for M and N respectively and by the previous claim, we know that $\tau_n \wedge \sigma_n$ is a joint localizing sequence for M and N. Therefore $(aM + bN)^{\tau_n \wedge \sigma_n} - aM_0 - bN_0$ is a martingale for all $n \geq 0$. □

Lemma 14.8. *Let τ_n be a sequence of optional times such that $\tau_n \uparrow \infty$ a.s. and let M be an $\mathcal{F}$-adapted process. Then M is a local martingale if an only if M^{τ_n} is for all $n \geq 0.9$*

Proof. TODO: □

Lemma 14.9. *Let M be a continuous local martingale with locally bounded variation then $M = M_0$ a.s.*

Proof. We first reduce to the case in which M is a martingale with locally bounded variation and $M_0 = 0$. Let τ_n be a localizing sequence for M then if we can show that $M_{\tau_n \wedge t} - M_0 = 0$ a.s. for all $n \geq 0$ and $t \geq 0$ then as $\tau_n \to \infty$ we can conclude that $M_t = M_0$ a.s. for all $t \geq 0$.

Next note that since M is locally of bounded variation we have optional times τ_n such that $\tau_n \uparrow \infty$ such that M^{τ_n} is of bounded variation. This implies that M is of bounded variation on every interval $[0, t]$. Therefore we can define the total variation process $V_t = TV_0^t(M)$. Since M is continuous, V_t is continuous (Lemma 2.121) and by defintion of total variation it is clear that V_t is $\mathcal{F}$-adapted. Now define $\sigma_n = \inf\{t \geq 0 \mid V_t = n\}$; we know by continuity of V_t that σ_n is an optional time (Lemma 9.70) and that $M_{\sigma_n \wedge t}$ is a continuous martingale. Since M is of locally finite variation we know that $\sigma_n \uparrow \infty$ and as before if we can show that $M_{\sigma_n \wedge t} = 0$ a.s. for all $n \geq 0$ and $t \geq 0$ then it will follow that $M_t = 0$ for all $t \geq 0$.

Now we have reduced to the case in which M is a continuous martingale with $M_0 = 0$ and bounded variation. So fix $t > 0$ and define the partition $t_{n,k} = kt/n$ for all $n > 0$ and $k = 0, 1, \ldots, n$. If we define

$$\zeta_n = \sum_{k=1}^{n} \left(M_{t_{n,k}} - M_{t_{n,k-1}}\right)^2 \leq V_t \max_{1 \leq k \leq n} \left|M_{t_{n,k}} - M_{t_{n,k-1}}\right|$$

then using the continuity of M we know that M is uniformly continuous on $[0, t]$ and therefore we have $\lim_{n\to\infty} \zeta_n = 0$ a.s. Moreover we have

$$\zeta_n \leq \sum_{k=1}^{n} \sum_{j=1}^{n} \left|M_{t_{n,k}} - M_{t_{n,k-1}}\right| \left|M_{t_{n,j}} - M_{t_{n,j-1}}\right| = V_t^2$$

Since V_t is bounded we can apply Dominated Convergence, the martingale property of M_t and the fact that $M_0 = 0$ to conclude

$$0 = \lim_{n \to \infty} \mathbf{E}\left[\zeta_n\right] = \sum_{k=1}^{n} \mathbf{E}\left[M^2_{t_{n,k}}\right] - \mathbf{E}\left[M^2_{t_{n,k-1}}\right] = \mathbf{E}\left[M_t^2\right]$$

and from this we conclude that $M_t = 0$ a.s. Taking the union of a countable number of sets of probability zero we see that almost surely $M_q = 0$ for all $q \in \mathbb{Q}_+$. Since M_t is continuous we conclude that almost surely $M_t = 0$ for all $t \in \mathbb{R}_+$. ☐

2. Stieltjes Integrals

There are a few simple facts about Stieltjes integrals that we want to describe in the stochastic setting as they will play a part in the general theory of stochastic integration. First we record the formula for the restriction of a Lebesgue-Stieltjes measure to an interval.

Lemma 14.10. *Let F be a right continuous function of bounded variation on $[a, b]$, let $[c, d] \subset [a, b]$. If we let μ_F denote the signed Lebesgue-Stieltjes measure associated with F and we let*

$$F^{[c,d]}(s) = F((s \wedge d) \vee c) = \begin{cases} F(c) & \textit{if } s < c \\ F(s) & \textit{if } c \leq s \leq d \\ F(d) & \textit{if } d < s \end{cases}$$

then $F^{[c,d]}$ is right continuous of bounded variation on $[a, b]$ and $\mu_F \mid_{[c,d]} = \mu_{F^{[c,d]}}$.

Proof. First suppose that F is non-decreasing and right continuous. It is elementary that $F^{[c,d]}$ is also non-decreasing and right continuous. For any half open interval $(x, y] \subset [a, b]$ we have

$$\begin{aligned} \mu_F \mid_{[c,d]} ((x, y]) &= \mu_F([c, d] \cap (x, y]) = \mu_F((d \wedge x) \vee c, (d \wedge y) \vee c]) \\ &= F((d \wedge y) \vee c) - F((d \wedge x) \vee c) = F^{[c,d]}(y) - F^{[c,d]}(x) = \mu_{F^{[c,d]}}((x, y]) \end{aligned}$$

and as we know that $\mu_F \mid_{[c,d]}$ is locally finite, by Lemma 2.112 we get $\mu_F \mid_{[c,d]} = \mu_{F^{[c,d]}}$.

In the case of F is right continuous of bounded variation, then if we write $F = F_+ - F_-$ as a difference of right continuous non-decreasing functions then it is also true $F^{[c,d]} = F_+^{[c,d]} - F_-^{[c,d]}$ and clearly each $F_\pm^{[c,d]}$ is non-descreasing which show us that $F^{[c,d]}$ is of bounded variation. Moreover, using the result for non-descreasing functions

$$\mu_F \mid_{[c,d]} = \mu_{F_+} \mid_{[c,d]} - \mu_{F_-} \mid_{[c,d]} = \mu_{F_+^{[c,d]}} - \mu_{F_-^{[c,d]}} = \mu_{F^{[c,d]}}$$

and we are done. ☐

The simplest type of stochastic integral arises for a process that has right continuous paths with locally finite variation. In this case, we can just apply the ordinary theory of Lebesgue-Stieltjes integrals pointwise to the process.

Definition 14.11. Let F be an cadlag adapted process and locally finite variation and let V be a jointly measurable process then we define a new process $\int V_s \, dF_s$

by

$$\left(\int V_s\,dF_s\right)_t(\omega) = \int_0^t V_s(\omega)\,dF(\omega)_s \qquad \text{for all } t \geq 0 \text{ and } \omega \in \Omega$$

We usually write $\left(\int V_s\,dF_s\right)_t = \int_0^t V_s\,dF_s$.

The fact that the integral defined as above is actually a process requires verification. In addition we show that when V is progressive then the resulting process is adapted.

Lemma 14.12. *If F is a cadlag process of locally finite variation (not necessarily adapted) and V is a jointly measurable process then $\int_0^t V_s\,dF_s$ is a cadlag process of locally finite variation. If in addition F is $\mathcal{F}$-adapted and V is $\mathcal{F}$ - progressively measurable then $\int_0^t V_s\,dF_s$ is $\mathcal{F}$-adapted.*

Proof. If we denote by μ_F the signed Lebesgue-Stieltjes measure constructed from F and let $\cup_{j=1}^n (a_j, b_j]$ be a disjoint union of intervals, then we have by finite additivity $\mu_F(\cup_{j=1}^n (a_j, b_j]) = \sum_{j=1}^n (F(b_j) - F(a_j))$ which measurable by the measurability of F. As the set of disjoint unions of half open intervals is a ring (Example 2.83) and therefore a π-system that generates the Borel σ-algebra we know μ_F is a kernel by monotone classes (specifically Lemma 8.27). If V is jointly measurable then the same is true of $\mathbf{1}_{[0,t]}V$ for every $t \geq 0$ and therefore $\int_0^t V_s\,dF_s$ is measurable by Lemma 8.29. The fact that $\int_0^t V_s\,dF_s$ is cadlag and has locally finite variation follow pointwise from Corollary 2.126.

Note also that for any $t \geq 0$ we have by Lemma 2.57 and Lemma 14.10

$$\int_0^t V_s\,dF_s = \int_0^\infty \mathbf{1}_{[0,t]} V_s\,dF_s = \int_0^\infty V_s^t\,dF\mid_{[0,t]}(s) = \int_0^\infty V_s^t\,dF_s^t$$

where $F^t(s) = F(t \wedge s)$ and $V_s^t = V_{t \wedge s}$. If we assume that F is adapted it follows that F_s^t is $\mathcal{F}_t$ measurable for all $s \geq 0$ and by the argument above we see that μ_{F^t} is an $\mathcal{F}_t$-measurable kernel. If V is progressive then by writing $V^t(\omega, s) = V\mid_{\Omega \times [0,t]} (\omega, s \wedge t)$ which shows that V^t is $\mathcal{F}_t \otimes \mathcal{B}([0,\infty))$-measurable. Now applying Lemma 8.29 we get $\mathcal{F}_t$-measurability of $\int_0^t V_s\,dF_s$. □

Because of the previous result we make the following definition for the space of integrands that we'll initially concern ourselves with.

Definition 14.13. If F is a cadlag process of locally finite variation then let $L(F)$ be the space of progressive processes V that are pointwise integrable with respect to F.

Because we use stochastic Stieltjes integrals in defining general stochastic integrals we record the following simple facts. Both of these facts have analogues for general stochastic integrals as well.

Lemma 14.14. *Let F be a cadlag process of locally finite variation, let $V \in L(F)$ and let U be a progressive process. $U \in L(\int V\,dF)$ if and only if $UV \in L(F)$ and moreover*

$$\int_0^t U_s V_s\,dF_s = \int_0^t U_s\,d\int V_s\,dF_s$$

PROOF. Initially assume that U and V are both positive. Note that by definition of the Lebesgue-Stieltjes measure we have pointwise for any finite interval $(a, b]$,

$$\mu_{\int V_s\, dF_s}((a,b]) = \int_0^b V_s\, dF_s - \int_0^a V_s\, dF_s = \int_0^\infty \mathbf{1}_{(a,b]} V_s\, dF_s$$

and therefore we have $\mu_{\int V_s\, dF_s} = V \cdot \mu_F$ (i.e. V is a μ_F-density of $\mu_{\int V_s\, dF_s}$); the result now follows from Lemma 2.57. The rest of the result follows from writing $U = U_+ - U_-$ and $V = V_+ - V_-$ and using linearity. $\square$

We also want to record the behavior of a stochastic Stieltjes integral under stopping.

LEMMA 14.15. *Let F be a cadlag process of locally finite variation, let $V \in L(F)$ and let τ be an optional time then*

$$\int_0^{t\wedge\tau} V_s\, dF_s = \int_0^t \mathbf{1}_{[0,\tau]} V_s\, dF_s = \int_0^t V_s\, F_s^\tau$$

PROOF. This follows immediately by writing $\int_0^\infty \mathbf{1}_{[0,t]}\mathbf{1}_{[0,\tau]} V_s\, dF_s$ and pointwise using the fact that $\mu_F \mid_{[0,\tau]} = \mu_{F^\tau}$ (Lemma 14.10). $\square$

3. Stochastic Integrals

The process of defining stochastic integrals follows the standard path of defining integrals for a subclass of integrands for which the definition and existence of the associated integral is easy to see. Then one uses approximations to extend the class of integrands. We begin by defining that initial subclass of integrands and define integrals of them with respect to an arbitrary martingale.

DEFINITION 14.16. Let $\tau_1 \leq \tau_2 \leq \cdots$ be optional times, let $\xi_1, \xi_2, \ldots$ be bounded random variables and assume ξ_k is $\mathcal{F}_{\tau_k}$-measurable. Then we say that

$$V_t = \sum_{k=1}^\infty \xi_k \mathbf{1}_{\tau_k > t}$$

is a *predictable step process*. Given a predictable step process and a process M we define the *elementary stochastic integral*

$$\int_0^t V\, dM = \sum_{k=1}^\infty \xi_k \left(M_t - M_{\tau_k \wedge t}\right)$$

In case $\tau_n = \tau_{n+1} = \cdots$ and $\xi_n = \xi_{n+1} = \cdots$ we say that V is a *finite predictable step process.*

Note that in the definition of a stochastic integral for a predictable step process there is no need to consider convergence questions since for each $t \geq 0$ the sum that defines the integral has only finitely many non-zero terms.

TODO: The definition of the elementary stochastic integral isn't quite justified as we haven't shown that it only depends on V and not a particular representation of $V_t = \sum_{k=1}^\infty \xi_k \mathbf{1}_{\tau_k > t}$. To show this it seems like it would be helpful to have a canonical representation for a predictable step process. At some point we also may need the fact that the space of such processes (at least the finite linear combinations) is a vector space or algebra (as per Rogers and Williams).

If one defines the vector space spanned by $\xi \mathbf{1}_{(\sigma,\tau]}$ then there is a standard (but not unique) form $\sum_{j=1}^n \xi_j \mathbf{1}_{(\sigma_j,\tau_j]}$ where σ_j and τ_j are optional times satisfying $\sigma_1 \leq \tau_1 \leq \sigma_2 \leq \tau_2 \leq \cdots \leq \sigma_n \leq \tau_n$. To see this we first need a simple preliminary fact. If σ and τ are optional times and ξ is either $\mathcal{F}_\sigma$-measurable then $\xi \mathbf{1}_{\sigma < \tau}$ is $\mathcal{F}_{\sigma\wedge\tau}$-measurable. This follows from noting that for all $t \in \mathbb{R}$,

$$\{\xi \mathbf{1}_{\sigma<\tau} \leq t\} = \begin{cases} \{\sigma \geq \tau\} \cup (\{\xi \leq t\} \cap \{\sigma < \tau\}) & \text{if } t \geq 0 \\ \{\xi \leq t\} \cap \{\sigma < \tau\} & \text{if } t < 0 \end{cases}$$

and since $\{\sigma \geq \tau\}$ is $\mathcal{F}_{\sigma\wedge\tau}$-measurable it suffices to show that $\{\xi \leq t\} \cap \{\sigma < \tau\} \in \mathcal{F}_{\sigma\wedge\tau}$ for all $t \in \mathbb{R}$. Thus pick $s \in \mathbb{R}$ and using the $\mathcal{F}_\sigma$-measurability of ξ and the $\mathcal{F}_{\sigma\wedge\tau}$-measurability of $\{\sigma < \tau\}$ we get

$$\{\xi \leq t\} \cap \{\sigma < \tau\} \cap \{\sigma \wedge \tau \leq s\} = \{\xi \leq t\} \cap \{\sigma \leq s\} \cap \{\sigma < \tau\} \cap \{\sigma \wedge \tau \leq s\} \in \mathcal{F}_s$$

Now considering the decomposition of the intersection of two half open intervals in $\mathbb{R}$ into 3 disjoint parts we see

$$\begin{aligned}
&\xi_1 \mathbf{1}_{(\sigma_1,\tau_1]} + \xi_2 \mathbf{1}_{(\sigma_2,\tau_2]} = \\
&(\xi_1 \mathbf{1}_{\sigma_1<\sigma_2} + \xi_2 \mathbf{1}_{\sigma_2<\sigma_1}) \mathbf{1}_{(\sigma_1\wedge\sigma_2,(\sigma_1\vee\sigma_2)\wedge\tau_1\wedge\tau_2]} + \\
&(\xi_1 + \xi_2) \mathbf{1}_{(\sigma_1\vee\sigma_2,\tau_1\wedge\tau_2\vee\sigma_1\vee\sigma_2]} + \\
&(\xi_1 \mathbf{1}_{\tau_1>\tau_2} + \xi_2 \mathbf{1}_{\tau_2>\tau_1}) \mathbf{1}_{(\sigma_1\vee\sigma_2\vee(\tau_1\wedge\tau_2),\tau_1\vee\tau_2]}
\end{aligned}$$

By our claim above get that $\xi_1 \mathbf{1}_{\sigma_1<\sigma_2} + \xi_2 \mathbf{1}_{\sigma_2<\sigma_1}$ is $\mathcal{F}_{\sigma\wedge\tau}$-measurable. By $\mathcal{F}_{\sigma_1}$-measurability of ξ_1 and $\mathcal{F}_{\sigma_2}$-measurability of ξ_2 we get $\mathcal{F}_{\sigma_1\vee\sigma_2}$-measurability of $\xi_1 + \xi_2$. Lastly we know also that $\{\tau_1 > \tau_2\}$ and $\{\tau_2 > \tau_1\}$ are $\mathcal{F}_{\tau_1\wedge\tau_2}$-measurable so $\xi_1 \mathbf{1}_{\tau_1>\tau_2} + \xi_2 \mathbf{1}_{\tau_2>\tau_1}$ is $\mathcal{F}_{\sigma_1\vee\sigma_2\vee(\tau_1\wedge\tau_2)}$-measurable. Moreover it is clear that we have the inequalities

$$\sigma_1 \wedge \sigma_2 \leq (\sigma_1 \vee \sigma_2) \wedge \tau_1 \wedge \tau_2 \leq \sigma_1 \vee \sigma_2 \leq \tau_1 \wedge \tau_2 \vee \sigma_1 \vee \sigma_2 \leq \sigma_1 \vee \sigma_2 \vee (\tau_1 \wedge \tau_2) \leq \tau_1 \vee \tau_2$$

and therefore the result is shown.

The representation for a predictable step process we have given in the definition is occasionally not the most convenient one. Given $V_t = \sum_{k=1}^\infty \xi_k \mathbf{1}_{\tau_k > t}$ if we define $\eta_n = \sum_{k=1}^n \xi_k$ and therefore

$$\begin{aligned}
V_t &= \sum_{k=1}^\infty \xi_k \mathbf{1}_{t>\tau_k} = \sum_{k=1}^\infty \xi_k \sum_{j=k}^\infty \mathbf{1}_{(\tau_j,\tau_{j+1}]}(t) \\
&= \sum_{j=1}^\infty \sum_{k=1}^j \xi_k \mathbf{1}_{(\tau_j,\tau_{j+1}]}(t) = \sum_{j=1}^\infty \eta_j \mathbf{1}_{(\tau_j,\tau_{j+1}]}(t)
\end{aligned}$$

and

$$\begin{aligned}
\int_0^t V \, dM &= \sum_{k=1}^\infty \xi_k \left(M_t - M_{\tau_k\wedge t}\right) = \sum_{k=1}^\infty \xi_k \sum_{j=k}^\infty \left(M_{\tau_{j+1}\wedge t} - M_{\tau_j\wedge t}\right) \\
&= \sum_{j=1}^\infty \sum_{k=1}^j \xi_k \left(M_{\tau_{j+1}\wedge t} - M_{\tau_j\wedge t}\right) = \sum_{j=1}^\infty \eta_j \left(M_{\tau_{j+1}\wedge t} - M_{\tau_j\wedge t}\right)
\end{aligned}$$

In what follows we will feel free to switch between these representations without comment.

The first order of business is to establish conditions under which an elementary stochastic integral is a martingale. To do this we need the following characterization of the martingale property.

LEMMA 14.17. *Let M_t be an integrable adapted process on an index set T. Then M is a martingale if and only if $\boldsymbol{E}[M_\sigma] = \boldsymbol{E}[M_\tau]$ for all T-valued optional times σ and τ that take at most two values.*

PROOF. Restricting M_t to the union of the ranges of τ and σ we can apply Lemma 9.37 to conclude $\mathbf{E}[M_\sigma] = M_0 = \mathbf{E}[M_\tau]$. In the other direction, let $s, t \in T$ with $s < t$. Let $A \in \mathcal{F}_s$ and define $\sigma = s\mathbf{1}_{A^c} + t\mathbf{1}_A$ and note that σ is an optional time. Now, applying our hypothesis to the optional time σ and the deterministic optional time s, we get $\mathbf{E}[M_t; A] = \mathbf{E}[M_\sigma] - \mathbf{E}[M_s; A^c] = \mathbf{E}[M_s] - \mathbf{E}[M_s; A^c] = \mathbf{E}[M_s; A]$ which shows $\mathbf{E}[M_t \mid \mathcal{F}_s] = M_s$ a.s. □

LEMMA 14.18. *Suppose $\mathcal{F}$ is a filtration, $\tau_1 \leq \tau_2 \leq \cdots \leq \tau_n$ are bounded $\mathcal{F}$-optional times, M_t is a martingale and either*

(i) *each τ_k is countably valued*
(ii) *$\mathcal{F}$ and M are right continuous*

Then if

$$V_t = \sum_{k=1}^n \xi_k \boldsymbol{1}_{\tau_k > t}$$

is a finite predictable step process then $\int_0^t V\,dM$ is a martingale. If we assume that M is a local martingale then $\int_0^t V\,dM$ is a local martingale.

PROOF. By definition of elementary stochastic integral and linearity, it suffices to show that $N_t = \xi\left(M_t - M_{\tau\wedge t}\right)$ is a martingale whenever either τ is a countably valued optional time or $\mathcal{F}$ and M are right continuous and ξ is a bounded $\mathcal{F}_\tau$-measurable random variable. In the first case, by restricting M_t to the range of τ we can apply the Optional Sampling Theorem 9.38 to the bounded optional time $\tau \wedge t$ to conclude that $M_{\tau\wedge t}$ is integrable and in the second case we can apply the continuous time Optional Sampling Theorem 9.82 to conclude that $M_{\tau\wedge t}$ is integrable. This together with the integrability of M_t and boundedness of ξ shows that N_t is integrable. If we note that $N_t = \xi\mathbf{1}_{\tau\leq t}\left(M_t - M_{\tau\wedge t}\right)$ then because $\xi\mathbf{1}_{\tau\leq t}$ and M_t are $\mathcal{F}_t$-measurable and $M_{\tau\wedge t}$ is $\mathcal{F}_{\tau\wedge t}$-measurable (hence $\mathcal{F}_t$-measurable) we see that N_t is adapted. Lastly let σ be an countably valued optional time then by the $\mathcal{F}_\tau$-measurability of ξ we have and either the Optional Sampling Theorem 9.38 or the Optional Sampling Theorem 9.82 we get

$$\mathbf{E}[N_\sigma \mid \mathcal{F}_\tau] = \xi\mathbf{E}[M_\sigma - M_{\tau\wedge\sigma} \mid \mathcal{F}_\tau] = \xi\left(M_{\tau\wedge\sigma} - M_{\tau\wedge\sigma}\right) = 0$$

and by the tower property of conditional expectations we get $\mathbf{E}[N_\sigma] = 0$. Now by Lemma 14.17 we see that N_t is a martingale.

Now let us assume that M is a local martingale. To see that $\int_0^t V\,dM$ is a local martingale let σ_n be a localizing sequence and note that

$$\left(\int_0^t V\,dM\right)^{\sigma_n} = \sum_{k=1}^n \xi_k\left(M_{\sigma_n\wedge t} - M_{\sigma_n\wedge\tau_k\wedge t}\right) = \int_0^t V\,dM^{\sigma_n}$$

is a martingale with localizing sequence σ_n by the first part of the Lemma. □

LEMMA 14.19. *Suppose $\mathcal{F}$ is a filtration, $\tau_1 \leq \tau_2 \leq \cdots \leq \cdots$ are bounded $\mathcal{F}$-optional times with $\tau_n \uparrow \infty$, M_t is an L^2 martingale with M_0, V_t is a predictable step process with $|V_t| \leq 1$ and either*

(i) *each τ_k is countably valued*
(ii) *$\mathcal{F}$ and M are right continuous*

then $\int_0^t V\,dM$ is an L^2-martingale and $\boldsymbol{E}\left[\left(\int_0^t V\,dM\right)^2\right] \leq \boldsymbol{E}\left[M_t^2\right]$.

PROOF. We let $V_t = \sum_{k=1}^\infty \eta_k \mathbf{1}_{(\tau_k,\tau_{k+1}]}$. We start by taking an arbitrary $n > 0$ and defining $V_t^n = \sum_{k=1}^n \eta_k \mathbf{1}_{(\tau_k,\tau_{k+1}]}$ so that V^n is a finite predictable step process. By Lemma 14.18 shows that $\int_0^t V^n\,dM$ is a martingale. The L^2 bound for V^n follows from Optional Sampling (Theorem 9.38 or Theorem 9.82 depending on which hypothesis we choose). The critical point is that for any $1 \leq k < j \leq n$ we have for each cross term term of the stochastic integral

$$\begin{aligned}
&\mathbf{E}\left[\eta_j\eta_k\left(M_{\tau_{j+1}\wedge t} - M_{\tau_j\wedge t}\right)\left(M_{\tau_{k+1}\wedge t} - M_{\tau_k\wedge t}\right)\right] \\
&= \mathbf{E}\left[\eta_j\eta_k\left(M_{\tau_{j+1}\wedge t} - M_{\tau_j}\right)\left(M_{\tau_{k+1}} - M_{\tau_k}\right); t > \tau_j\right] \\
&= \mathbf{E}\left[\eta_j\eta_k\mathbf{E}\left[M_{\tau_{j+1}\wedge t} - M_{\tau_j} \mid \mathcal{F}_{\tau_j}\right]\left(M_{\tau_{k+1}} - M_{\tau_k}\right); t > \tau_j\right] = 0
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{E}\left[\left(M_{\tau_{k+1}\wedge t} - M_{\tau_k\wedge t}\right)^2\right] &= \mathbf{E}\left[M^2_{\tau_{k+1}\wedge t}\right] - 2\mathbf{E}\left[M_{\tau_{k+1}\wedge t}M_{\tau_k\wedge t}\right] + \mathbf{E}\left[M^2_{\tau_k\wedge t}\right] \\
&= \mathbf{E}\left[M^2_{\tau_{k+1}\wedge t}\right] - 2\mathbf{E}\left[\mathbf{E}\left[M_{\tau_{k+1}\wedge t} \mid \mathcal{F}_{\tau_k\wedge t}\right]M_{\tau_k\wedge t}\right] + \mathbf{E}\left[M^2_{\tau_k\wedge t}\right] \\
&= \mathbf{E}\left[M^2_{\tau_{k+1}\wedge t}\right] - \mathbf{E}\left[M^2_{\tau_k\wedge t}\right]
\end{aligned}$$

Using the above facts, the fact that $M_0 = 0$ and the bound on V

$$\begin{aligned}
\mathbf{E}\left(\int_0^t V^n\,dM\right)^2 &= \mathbf{E}\sum_{k=1}^n \eta_k^2\left(M_{\tau_{k+1}\wedge t} - M_{\tau_k\wedge t}\right)^2 \\
&\leq \mathbf{E}\sum_{k=1}^n \left(M_{\tau_{k+1}\wedge t} - M_{\tau_k\wedge t}\right)^2 \\
&= \sum_{k=1}^n \mathbf{E}\left[M^2_{\tau_{k+1}\wedge t}\right] - \mathbf{E}\left[M^2_{\tau_k\wedge t}\right] \\
&\leq \sum_{k=1}^\infty \mathbf{E}\left[M^2_{\tau_{k+1}\wedge t}\right] - \mathbf{E}\left[M^2_{\tau_k\wedge t}\right] \\
&= \lim_{n\to\infty} \mathbf{E}\left[M^2_{\tau_n\wedge t}\right] = \mathbf{E}\left[M_t^2\right]
\end{aligned}$$

Now in the general case, we get the L^2 bound by Fatou's Lemma Theorem 2.45

$$\mathbf{E}\left(\int_0^t V\,dM\right)^2 = \liminf_{n\to\infty} \mathbf{E}\left(\int_0^t V^n\,dM\right)^2 \leq \mathbf{E}\left[M_t^2\right]$$

In addition, the L^2 bound shows that the family $\int_0^t V\,dM, \int_0^t V^1\,dM, \int_0^t V^2\,dM, \dots$ is uniformly integrable (Lemma 5.51) and therefore for every $t \geq 0$ and the martingale property of $\int_0^t V^n\,dM$ we get for $u < t$,

$$\begin{aligned}
\mathbf{E}\left[\int_0^t V\,dM \mid \mathcal{F}_u\right] &= \mathbf{E}\left[\lim_{n\to\infty}\int_0^t V^n\,dM \mid \mathcal{F}_u\right] \\
&= \lim_{n\to\infty}\mathbf{E}\left[\int_0^t V^n\,dM \mid \mathcal{F}_u\right] \\
&= \lim_{n\to\infty}\int_0^u V^n\,dM = \int_0^u V\,dM
\end{aligned}$$

(to exchange the limits with the conditional expectation, use the fact that for each $A \in \mathcal{F}_u$ we can see that $\mathbf{1}_A \int_0^t V^n\,dM$ is uniformly integrable then use Theorem 5.58) showing $\int_0^t V\,dM$ is an L^2-martingale. □

Definition 14.20. Two processes X and Y on a time scale T are said to be *versions* of one another if $\mathbf{P}\{X_t = Y_t\} = 1$ for every fixed $t \in T$. One also says that Y is a modification of X (and vice versa). Two processes X and Y on a time scale T are said to be *indistinguishable* if $\mathbf{P}\{X_t = Y_t \text{ for all } t \in T\} = 1$.

While it is trivial that two indistinguishable processes are versions of one another it is also simple that for continuous processes on time scales that are separable the two notions are equivalent. The following is the case that is most important for us.

Proposition 14.21. *Let X and Y be cadlag processes on a time scale $\mathbb{R}_+$. Then X and Y are versions of one another if and only if they are indistinguishable.*

Proof. Let A_X be the event that X has cadlag sample paths and similarly with Y. Let $B = \cap_{\substack{q\geq 0\\ q\in\mathbb{Q}}}\{X_q = Y_q\}$. If X and Y are versions then by taking a countable intersection of almost sure events we have $\mathbf{P}\{A_X \cap A_Y \cap B\} = 1$. Morever on the event $A_X \cap A_Y \cap B$ by right continuity we have for all $t \geq 0$, $X_t = \lim_{q\downarrow t} X_q = \lim_{q\downarrow t} Y_q = Y_t$ and therefore $A_X \cap A_Y \cap B = A_X \cap A_Y \cap \{X_t = Y_t \text{ for all } t \geq 0\}$ and it follows that $\mathbf{P}\{X_t = Y_t \text{ for all } t \geq 0\} \geq \mathbf{P}\{A_X \cap A_Y \cap \{X_t = Y_t \text{ for all } t \geq 0\}\} = 1$. □

When constructing spaces of processes is often the case that we'll need to identify indistinguishable processes, thus we make explicit note of the following simple fact.

Proposition 14.22. *Indistinguishability of processes is an equivalence relation.*

Proof. Reflexivity and symmetry are immediate from the defintion, and transitivity follows from the fact that $\{X_t = Z_t \text{ for all } t \in T\} \supset \{X_t = Y_t \text{ for all } t \in T\} \cap \{Y_t = Z_t \text{ for all } t \in T\}$ and $\mathbf{P}\{\{X_t = Y_t \text{ for all } t \in T\} \cap \{Y_t = Z_t \text{ for all } t \in T\}\} = 1$. □

Definition 14.23. Fix a probability space $(\Omega, \mathcal{A}, P)$ and suppose $\mathcal{F}$ is a right continuous and complete filtration. Let $\mathcal{M}^2$ be the space of L^2 bounded continuous $\mathcal{F}$-martingales such that $M_0 = 0$ a.s. up to indistinguishability. That is to say that for all $M \in \mathcal{M}^2$ there exists $C \geq 0$ such that for all $0 \leq t < \infty$ we have $\|M_t\|_2 \leq C$. For $M, N \in \mathcal{M}^2$, define $\langle M, N\rangle = \langle M_\infty, N_\infty\rangle = \mathbf{E}[M_\infty N_\infty]$.

Lemma 14.24. *The space $\mathcal{M}^2$ is a Hilbert space.*

Proof. The fact that $\mathcal{M}^2$ is a vector space follows immediately from linearity of conditional expectation, the linearity of the space $C([0,\infty);\mathbb{R})$ and the triangle inequality of the L^2 norm on $C([0,\infty);\mathbb{R})$.

To see that we have an inner product on $\mathcal{M}^2$, first observe that if $\langle M, M\rangle = \|M_\infty\|_2^2 = 0$ then $M_\infty = 0$ a.s. hence since M is closable it follows that $M_t = \mathbf{E}\left[M_\infty \mid \mathcal{F}_t\right] = 0$ a.s. for all $0 \leq t < \infty$. Since M is continuous it follows that $M = 0$ a.s. Symmetry of $\langle \cdot, \cdot \rangle$ follows immediately from symmetry of the L^2 inner product on $C([0,\infty);\mathbb{R})$. Supposing M and N are both L^2 bounded continuous martingales we know they are closable hence $M_t = \mathbf{E}\left[M_\infty \mid \mathcal{F}_t\right]$ and similarly for N. It then follows from linearity of conditional expectation that $(aM + bN)_\infty = aM_\infty + bN_\infty$ for any $a, b \in \mathbb{R}$. From this fact we see that for all $M, N, R \in \mathcal{M}^2$ and $a, b \in \mathbb{R}$ we have $\langle aM + bN, R\rangle = \langle aM_\infty + bN_\infty, R_\infty\rangle = a\langle M, R\rangle + b\langle N, R\rangle$.

We now show that $\mathcal{M}^2$ is complete. Suppose $M^1, M^2, \dots$ is Cauchy in $\mathcal{M}^2$, then $M^1_\infty, M^2_\infty, \dots$ is Cauchy in L^2 and therefore has a limit ξ in L^2 that is $\mathcal{F}_\infty$-measurable. Define $M_t = \mathbf{E}\left[\xi \mid \mathcal{F}_t\right]$ so we know that M_t is a martingale and $M_\infty = \xi$ a.s. TODO: Do we know at this point that M is L^2 bounded? Now by the Doob L^2 inequality Lemma 9.78 applied to the closed martingale M_t on $[0,\infty]$ we have $\left\|\sup_{0\leq s\leq\infty}(M^n_s - M_s)\right\|_2 \leq 2\left\|M^n_\infty - M_\infty\right\|_2$. From this we get that $\lim_{n\to\infty}\left\|\sup_{0\leq s\leq\infty}(M^n_s - M_s)\right\|_2 = 0$ hence $\sup_{0\leq s\leq\infty}(M^n_s - M_s) \xrightarrow{P} 0$ (Lemma 5.6) and therefore $\sup_{0\leq s\leq\infty}(M^n_s - M_s) \xrightarrow{a.s.} 0$ along a subsequence (Lemma 5.10) which shows that M is has almost surely continuous sample paths (Lemma 1.39). ☐

4. Quadratic Variation

The crux of the problem in defining stochastic integrals is the fact that sample paths of continuous martingales almost surely have infinite total variation and therefore Lebesgue-Stieltjes integrals cannot be defined.

Lemma 14.25. *Let M and N be continuous local martingales and let τ be an optional time, then $M^\tau(N - N^\tau)$ is a local martingale.*

Proof. First let us assume that N is a martingale, τ is an optional time and η is an $\mathcal{F}_\tau$-measurable bounded random variable. We claim that $\eta(N_t - N^\tau_t)$ is a martingale. By the Optional Sampling Theorem 9.82 we know that $\tau \wedge t$ is a bounded optional time hence $N_{\tau\wedge t}$ is integrable and therefore by boundedness of η we know that $\eta(N_t - N_{\tau\wedge t})$ is integrable. To see adaptedness, note that $\eta(N_t - N_{\tau\wedge t}) = \eta\mathbf{1}_{\tau\leq t}(N_t - N_{\tau\wedge t})$ so that by $\mathcal{F}_\tau$-measurability of η we also have $\mathcal{F}_t$-measurability of $\eta\mathbf{1}_{\tau\leq t}$. Furthermore, $N_{\tau\wedge t}$ is $\mathcal{F}_{\tau\wedge t}$-measurable and since $\tau\wedge t \leq t$ we see that it is also $\mathcal{F}_t$-measurable. To see that $\eta(N_t - N_{\tau\wedge t})$ is a martingale we let σ be any bounded optional time and then note by Optional Sampling, the tower and pullout properties of conditional expectation

$$\mathbf{E}\left[\eta(N_\sigma - N_{\tau\wedge\sigma})\right] = \mathbf{E}\left[\eta\mathbf{E}\left[(N_\sigma - N_{\tau\wedge\sigma}) \mid \mathcal{F}_\tau\right]\right] = \mathbf{E}\left[\eta(N_{\tau\wedge\sigma} - N_{\tau\wedge\sigma})\right] = 0$$

which independent of σ hence we can apply Lemma 14.17 to conclude that $\eta(N - N^\tau)$ is a martingale.

TODO: Finish ☐

TODO: We used continuity of the local martingale M to reduce ourselves to the case of bounded martingales which was used to obtain integrability. Do general local martingales localize to L^2 bounded martingales or something else that would allow us to get integrability?

THEOREM 14.26 (Quadratic Covariation). *Let M and N be continuous local martingales, there exists an almost surely unique continuous process $[M, N]$ of locally finite variation such that $[M, N]_0 = 0$ and $MN - [M, N]$ is a local martingale. The pairing $[M, N]$ is bilinear and symmetric and for every optional time τ satisfies*

$$[M, N]^\tau = [M^\tau, N^\tau] = [M^\tau, N] \text{ a.s.}$$

If we define $[M] = [M, M]$ then it is the case that $[M]$ is almost surely non-decreasing. If M is a bounded martingale then $M^2 - [M]$ is an L^2 bounded martingale. The process $[M, N]$ is called the quadratic covariation *of M and N and the process $[M]$ is called the* quadratic variation *of M.*

PROOF. We first consider the case when $M = N$ and we first assume that M is a bounded martingale such that $M_0 = 0$ and $|M_t| \leq C$ for some deterministic constant $C > 0$. To motivate the construction recall the basic fact that for a function f of bounded variation we have the Lebesgue-Stieltjes integral $f^2 = 2 \int f \, df$. We suspect that in a stochastic setting such an identity won't quite work (because the Stieltjes integral doesn't work). That suspicion is correct and what does turn out to be true is that once we have defined a stochastic integral, $M^2 - 2\int M \, dM = [M]$. Of course our plan is to use the quadratic variation to define stochastic integrals so this reasoning is getting pretty circular here; nonetheless if we suspend belief for moment and define something that *looks like* it could be $\int M \, dM$ then we might get the right definition for quadratic variation. Motivated by these observations, our first step is to come up with an approximation of M by predictable step processes so we can create an approximation of $\int M \, dM$. For each $n > 0$ define the sequence of optional times $\tau_0^n, \tau_1^n, \ldots$ by $\tau_0^n = 0$ and

$$\begin{aligned}
\tau_k^n &= \inf\{t > \tau_{k-1}^n \mid \left| M_t - M_{\tau_{k-1}^n} \right| = 2^{-n}\} \text{ for } k > 0 \\
&= \tau_{k-1}^n + \tau_1^n \circ \theta_{\tau_{k-1}^n}
\end{aligned}$$

CLAIM 14.26.1. $\lim_{k\to\infty} \tau_k^n = \infty$ almost surely for every $n \in \mathbb{N}$.

This follows from the continuity of M_t. Indeed if $\lim_{k\to\infty} \tau_k^n = T < \infty$ then there exists a δ such that $|M_t - M_T| < 2^{-n-1}$ for all $T - \delta < t \leq T$. Because $\lim_{k\to\infty} \tau_k^n = T$ there exists a K such that $T - \delta < \tau_K^n < \tau_{K+1}^n \leq T$ and thus by the triangle inequality we have $\left| M_{\tau_{K+1}^n} - M_{\tau_K^n} \right| < 2^{-n}$. On the other hand, by the definition of τ_{K+1}^n we know that $\left| M_{\tau_{K+1}^n} - M_{\tau_K^n} \right| = 2^{-n}$ which is a contradiction.

CLAIM 14.26.2. Either $\tau_k^n = \infty$ or $M_{\tau_k^n} = j/2^n$ for some random $j \in \mathbb{Z}$.

This is a simple induction on k for each n using continuity of M_t, the fact that $M_0 = 0$ and $\tau_0^n = 0$.

CLAIM 14.26.3. Suppose $M_t = j/2^n$ for some $n \in \mathbb{N}$ and $j \in \mathbb{Z}$ and let $K = \max\{k \mid \tau_k^n \leq t\}$, then $M_{\tau_K^n} = j/2^n$.

The claim is trivially true if $\tau_K^n = t$. If this is not true by the previous claim we have $i \in \mathbb{Z}$ such that $M_{\tau_k^n} = i/2^n$. By definition of K, we have $\tau_K^n < t < \tau_{K+1}^n$

and by definition of τ^n_{K+1}, continuity of M_t and the intermediate value theorem we know that $\left|M_t - M_{\tau^n_K}\right| = |i-j|\, 2^{-n} < 2^{-n}$. Thus $i = j$ and the claim is verified.

Claim 14.26.4. For every $n \geq 0$ and $k \geq 0$ there exists a random integer $l \geq 0$ such that $\tau^n_k = \tau^{n+1}_l$.

We proceed by induction on k with the case $k = 0$ being true because $\tau^n_0 = 0$ for all $n \geq 0$. Having assumed $M_0 = 0$ we see that we can pick $0 \leq j < \infty$ such that $M_{\tau^n_k} = j/2^n$. (TODO: what to say when $\tau^n_k = \infty$; not clear that we can assert some $\tau^{n+1}_l = \infty$ since we may be oscillating with small enough amplitude?) Let i be the largest index such that $\tau^{n+1}_i \leq \tau^n_k$. Since $M_{\tau^n_k} = j/2^n = 2j/2^{n+1}$ we can apply the previous claim to see that $M_{\tau^{n+1}_i} = M_{\tau^n_k}$. By the intermediate value theorem we know that $M_{\tau^n_{k-1}} < M_{\tau^{n+1}_i} \leq M_{\tau^n_k}$. Because $\left|M_{\tau^{n+1}_i} - M_{\tau^n_{k-1}}\right| = \left|M_{\tau^n_k} - M_{\tau^n_{k-1}}\right| = 2^{-n}$ by definition of τ^n_k we know that $\tau^{n+1}_i \geq \tau^n_k$ and therefore $\tau^{n+1}_i = \tau^n_k$.

Define

$$V^n_t = \sum_{k=0}^{\infty} M_{\tau^n_k} \mathbf{1}_{(\tau^n_k, \tau^n_{k+1}]}(t)$$

where despite the fact that we have written an infinite sum we don't have to worry about convergence since for any fixed t the sum is finite. Clearly, each V^n is a bounded predictable step process and it is also clear that V^n is an approximation of M (though we won't yet belabor the exact sense in which this is true). Pick $t \geq 0$ and let K be the random index such that $\tau^n_K < t \leq \tau^n_{K+1}$ then we can compute using high school algebra and the fact that $M_{\tau^n_0} = M_0 = 0$

$$\begin{aligned}
2\int_0^t V^n\, dM &= 2\sum_{k=0}^{\infty} M_{\tau^n_k}\left(M_{t\wedge\tau^n_{k+1}} - M_{t\wedge\tau^n_k}\right) \\
&= 2M_{\tau^n_K} M_t - 2M^2_{\tau^n_K} + 2\sum_{k=0}^{K-1} M_{\tau^n_k} M_{\tau^n_{k+1}} - 2\sum_{k=0}^{K-1} M^2_{\tau^n_k} \\
&= 2M_{\tau^n_K} M_t - M^2_{\tau^n_K} + 2\sum_{k=0}^{K-1} M_{\tau^n_k} M_{\tau^n_{k+1}} - \sum_{k=0}^{K-1} M^2_{\tau^n_k} - \sum_{k=0}^{K-1} M^2_{\tau^n_{k+1}} \\
&= 2M_{\tau^n_K} M_t - M^2_{\tau^n_K} - \sum_{k=0}^{K-1}\left(M_{\tau^n_{k+1}} - M_{\tau^n_k}\right)^2 \\
&= M^2_t - \left(M_t - M_{\tau^n_K}\right)^2 - \sum_{k=0}^{K-1}\left(M_{t\wedge\tau^n_{k+1}} - M_{t\wedge\tau^n_k}\right)^2 \\
&= M^2_t - \sum_{k=0}^{\infty}\left(M_{t\wedge\tau^n_{k+1}} - M_{t\wedge\tau^n_k}\right)^2
\end{aligned}$$

So if we define

$$Q^n_t = \sum_{k=0}^{\infty}\left(M_{t\wedge\tau^n_{k+1}} - M_{t\wedge\tau^n_k}\right)^2$$

we have the identity

$$M_t^2 = 2\int_0^t V^n \, dM + Q_t^n$$

Since V^n is a bounded predictable step process and M is an L^2 continuous martingale, we know that $\int V^n \, dM$ is a continuous L^2 martingale (Lemma 14.19). Furthermore by construction we have $\sup_{0 \leq t < \infty} |V_t^n - M_t| < 2^{-n}$ and therefore $\sup_{0 \leq t < \infty} |V_t^n - V_t^m| < 2^{-n+1}$ for all $n \leq m$.

$$\begin{aligned}
\left\| \int V^n \, dM - \int V^m \, dM \right\|_2 &= \left\| \int (V^n - V^m) \, dM \right\|_2 \\
&= \left\| \lim_{t \to \infty} \int_0^t (V^n - V^m) \, dM \right\|_2 \\
&\leq \varliminf_{t \to \infty} \left\| \int_0^t (V^n - V^m) \, dM \right\|_2 && \text{by Fatou's Lemma} \\
&\leq \varliminf_{t \to \infty} 2^{-n+1} \left\| M_t \right\|_2 && \text{by Lemma 14.19} \\
&= 2^{-n+1} \left\| M_\infty \right\|_2 = 2^{-n+1} \left\| M \right\|_2 && \text{since } M_t \xrightarrow{L^2} M_\infty
\end{aligned}$$

which shows that $\int V^n \, dM_s$ is a Cauchy sequence in $\mathcal{M}^2$. (TODO: I am almost certain that we know $\int (V^n - V^m) \, dM$ is a bounded L^2 martingale so that in fact we have $\int_0^t (V^n - V^m) \, dM \xrightarrow{L^2} \int_0^\infty (V^n - V^m) \, dM$ and we don't need Fatou above, we have equality). By completeness of $\mathcal{M}^2$ (Lemma 14.24) there is $N \in \mathcal{M}^2$ such that $\int V_s^n \, dM_s$ converges to N. Define $[M] = M^2 - 2N$ and use the Doob L^2 inequality $\sup_{0 \leq t \leq \infty} \left| N_t - \int_0^t V^n \, dM \right| \leq 2 \left\| N - \int V^n \, dM \right\| \to 0$ to get

$$\begin{aligned}
\sup_{0 \leq t < \infty} |Q_t^n - [M]_t| &= \sup_{0 \leq t < \infty} \left| Q_t^n - M_t^2 + 2N_t \right| \\
&= 2 \sup_{0 \leq t < \infty} \left| N_t - \int_0^t V^n \, dM \right| \xrightarrow{P} 0
\end{aligned}$$

Therefore $\sup_{0 \leq t < \infty} |Q_t^n - [M]_t| \xrightarrow{a.s.} 0$ along a subsequence (Lemma 5.10). Define the random set $T = \{\tau_k^n \mid n, k \in \mathbb{N}\}$. We have shown above that for any two elements $s < t \in T$ for sufficiently large n such that $s = \tau_k^n$ and $t = \tau_j^n$ for appropriate $k, j \in \mathbb{Z}_+$ (where k and j depend on n of course). From the definition of Q_t^n it follows that $Q_s^n \leq Q_t^n$ for all such n; thus $[M]$ is almost surely non-decreasing on T. By continuity of $[M]$ we can extend this to conclude that almost surely $[M]$ is non-decreasing on the closure $\overline{T}$. To see that $[M]$ is non-decreasing everywhere, we know that $\mathbb{R}_+ \setminus \overline{T}$ is a countable union of open intervals so it suffices to show that $[M]$ is constant on any open interval $(a, b) \subset \mathbb{R}_+ \setminus \overline{T}$. If $[M]$ is not constant on (a, b) then we can find suitable s, t such that $a < s < t < b$ and $X_s = k/2^n$ and $X_t = (k+1)/2^n$ or $X_t = (k-1)/2^n$ for some $k, n \in \mathbb{Z}$. Pick the largest i such that $\tau_i^n \leq s$. As $(a, b) \cap \overline{T} = \emptyset$ we know that $\tau_i^n < s$. By our previous claim we know that $X_{\tau_i^n} = X_s$ and therefore $\left| X_t - X_{\tau_i^n} \right| = |X_t - X_s| = 2^{-n}$ which implies $\tau_i^n < s < \tau_{i+1}^n \leq t$ which contradicts $(a, b) \cap \overline{T} = \emptyset$.

Now we need to extend the definition of the quadratic variation to unbounded martingales M. Let $\tau_n = \inf\{t \geq 0 \mid |M_t| = n\}$ which is an optional time by continuity of sample paths of M and Lemma 9.70. By what we have proven, we know that $[M^{\tau_n}]$ exists and is almost surely non-decreasing. TODO: Finish the result by extending to the unbounded case.

Having defined the quadratic variation $[M]$, we now extend it to the quadratic covariation $[M, N]$ for general local martingales M and N. First we establish the uniqueness. Note that if we are given processes of locally bounded variation Q and R such that $Q_0 = R_0 = 0$ and $MN - Q$ and $MN - R$ are local martingales, then $Q - R = (MN - R) - (MN - Q)$ is a local martingale of locally bounded variation and Lemma 14.9 implies that $Q - R = Q_0 - R_0 = 0$ a.s. From the uniqueness we immediately see that $[M, N]$ is bilinear and symmetric.

Now we reduce the definition of $[M, N]$ to the case $[M]$ with $M_0 = 0$ by a pair of reductions.

Claim 14.26.5. $[M - M_0, N - N_0] = [M, N]$ a.s.

Simply note that

$$MN - (M - M_0)(N - N_0) = M_0 N_0 + M_0 N + N_0 M$$

is a local martingale and therefore $(M - M_0)(N - N_0) - [M, N] = -(MN - (M - M_0)(N - N_0)) + MN - [M, N]$ is a local martingale.

Claim 14.26.6. $[M, N] = \frac{1}{4}([M + N] - [M - N])$

Note that

$$4MN - [M + N] + [M - N] = ((M + N)^2 - [M + N]) - ((M - N)^2 - [M - N])$$

is a local martingale.

Lastly we prove the behavior of localization under optional times.

Claim 14.26.7. Let τ be an optional time, then $[M, N]^\tau = [M^\tau, N^\tau] = [M^\tau, N]$ a.s.

For the first reduction, suppose that τ is an optional time then we know that

$$(MN - [M, N])^\tau = M^\tau N^\tau - [M, N]^\tau$$

which is a local martingale by Lemma 14.8 and moreover $[M, N]^\tau$ is of locally finite variation therefore we see $[M^\tau, N^\tau] = [M, N]^\tau$ a.s. We also know that $M^\tau(N^\tau - N)$ is a local martingale (TODO:this is supposed to follow for martingales from optional sampling (doesn't that actually show $M_\tau(N - N^\tau)$ is a martingale) then given a localizing sequence τ_n for M and σ_n for N we know that $\tau_n \wedge \sigma_n$ is a localizing sequence for both M and N) and therefore

$$M^\tau N - [M, N]^\tau = M^\tau(N - N^\tau) + M^\tau N^\tau - [M, N]^\tau$$

is a local martingale which shows that $[M, N] = [M^\tau, N]$ a.s. ∎

The behavior of continuous local martingales is intimately connected with the behavior of its quadratic variation process. This is actually a very useful thing because, as we will see later on, the quadratic variation is generally a simpler object (in particular the quadratic variation of a stochastic integral will be an random Stieltjes integral rather than a full blown stochastic integral). The source of the close connection is the local martingale property of $M^2 - [M]$. The connection is much more direct in the case that $M^2 - [M]$ is a regular martingale since then we conclude $\mathbf{E}\left[M_t^2\right] = \mathbf{E}\left[[M]_t\right]$. For this reason it is useful to have some conditions that guarantee that $M^2 - [M]$ is a martingale. In the proof of Theorem 14.26 we have already seen that this is true when M is a bounded continuous martingale (in fact in this case $M^2 - [M]$ is an L^2 bounded martingale). Without too much more work we show that $M^2 - [M]$ is a martingale whenever M is L^2 bounded.

PROPOSITION 14.27. *A continuous local martingale M is an L^2 bounded martingale if and only if $\boldsymbol{E}\left[M_0^2\right] < \infty$ and $\boldsymbol{E}\left[[M]_\infty\right] < \infty$. In this case it is also true that $M^2 - [M]$ is a uniformly integrable martingale.*

PROOF. First we reduce the first equivalence to the case in which $M_0 = 0$. Suppose we have the equivalence in this case. If M is an L^2 bounded martingale then it follows that M_0 is L^2 and by the triangle inequality $M_t - M_0$ is an L^2 bounded martingale starting at zero thus we know $\mathbf{E}\left[[M]_\infty\right] = \mathbf{E}\left[[M - M_0]_\infty\right] < \infty$. On the other hand if M is a continuous local martingale with $\mathbf{E}\left[M_0^2\right] < \infty$ and $\mathbf{E}\left[[M]_\infty\right] < \infty$ then $M - M_0$ is a continuous local martingale starting at zero with $\mathbf{E}\left[[M - M_0]_\infty\right] = \mathbf{E}\left[[M]_\infty\right] < \infty$ thus $M - M_0$ is an L^2 bounded martingale; it follows that M is an L^2 bounded martingale.

Next we note a simple fact that we'll use a few times during the remaining proof. Suppose that τ_n is a localizing sequence for M such that M^{τ_n} is a bounded martingale. Then from Theorem 14.26 we have $(M^{\tau_n})^2 - [M^{\tau_n}] = (M^2 - [M])^{\tau_n}$ is an L^2 bounded martingale; in particular τ_n is a localizing sequence for $M^2 - [M]$ as well.

Suppose that M is an L^2 bounded martingale. By the L^2 Martingale Convergence Theorem 9.81 we know that there is a square integrable random variable M_∞ such that $M_t \overset{a.s.}{\to} M_\infty$, $M_t \overset{L^2}{\to} M_\infty$ and $M_t = \mathbf{E}\left[M_\infty \mid \mathcal{F}_t\right]$ for all $0 \leq t < \infty$. Letting τ_n be a localizing sequence for M with M^{τ_n} a bounded martingale. Then by the above comment we know that $\mathbf{E}\left[M^2_{\tau_n \wedge t}\right] = \mathbf{E}\left[[M]_{\tau_n \wedge t}\right]$ for all $0 \leq t < \infty$ and $n \in \mathbb{N}$. By Doob's L^2 inequality we have $\mathbf{E}\left[\sup_{0 \leq t \leq \infty} M_t^2\right] \leq 4\mathbf{E}\left[M_\infty^2\right]$ and therefore by two applications of Dominated Convergence

$$\lim_{t \to \infty} \lim_{n \to \infty} \mathbf{E}\left[M^2_{\tau_n \wedge t}\right] = \mathbf{E}\left[M^2_\infty\right]$$

By two applications of Monotone Convergence and the above

$$\mathbf{E}\left[[M]_\infty\right] = \lim_{t \to \infty} \lim_{\tau_n \to \infty} \mathbf{E}\left[[M]_{\tau_n \wedge t}\right] = \lim_{t \to \infty} \lim_{\tau_n \to \infty} \mathbf{E}\left[M^2_{\tau_n \wedge t}\right] = \mathbf{E}\left[M^2_\infty\right] < \infty$$

Conversely suppose that M is a continuous local martingale starting at zero such that $\mathbf{E}\left[[M]_\infty\right] < \infty$. Again, let τ_n localize M to a bounded martingale so that we have $\mathbf{E}\left[M^2_{\tau_n \wedge t}\right] = \mathbf{E}\left[[M]_{\tau_n \wedge t}\right]$ for all $0 \leq t < \infty$ and $n \in \mathbb{N}$. Now we apply Fatou's Lemma and the fact that $[M]$ is increasing to see that

$$\mathbf{E}\left[M_t^2\right] \leq \lim_{n \to \infty} \mathbf{E}\left[M^2_{\tau_n \wedge t}\right] = \lim_{n \to \infty} \mathbf{E}\left[[M]_{\tau_n \wedge t}\right] \leq \mathbf{E}\left[[M]_\infty\right] < \infty$$

and thus M is L^2 bounded. The same reasoning without taking the limits shows that for a fixed $0 \leq t < \infty$ and all $n \in \mathbb{N}$

$$\mathbf{E}\left[M^2_{\tau_n \wedge t}\right] = \mathbf{E}\left[[M]_{\tau_n \wedge t}\right] \leq \mathbf{E}\left[[M]_\infty\right] < \infty$$

so by Lemma 5.51 we know that $M_{\tau_n \wedge t}$ is uniformly integrable. Thus from $\tau_n \uparrow \infty$ a.s. we conclude that for every $s < t$ we have

$$\mathbf{E}\left[M_t \mid \mathcal{F}_s\right] = \lim_{n \to \infty} \mathbf{E}\left[M_{\tau_n \wedge t} \mid \mathcal{F}_s\right] = \lim_{n \to \infty} M_{\tau_n \wedge s} = M_s \text{ a.s.}$$

and we see that M is a martingale.

Now suppose that M is an L^2 bounded martingale. Again we have M_∞ with $M_t \stackrel{L^2}{\to} M_\infty$ and by Doob's L^2 inequality 9.78 we have $\mathbf{E}\left[\sup_{0\leq t\leq\infty} M_t^2\right] \leq 4\mathbf{E}\left[M_\infty^2\right] < \infty$. Moreover by the first part of the proposition we know that $\mathbf{E}\left[[M]_\infty\right] < \infty$. From the fact that $[M]$ is increasing we know that for all $0 \leq t < \infty$ we have $\left|M_t^2 - [M]_t\right| \leq \sup_{0\leq u\leq\infty} M_u^2 + [M]_\infty$ where we have noted the right hand side is integrable. Thus by Example 5.50 we know that $M^2 - [M]$ is uniformly integrable.

To show the martingale property, let τ_n be a localizing sequence for $M^2 - [M]$ and let $0 \leq s < t < \infty$ be fixed. We know that

$$\mathbf{E}\left[M^2_{\tau_n\wedge t} - [M]_{\tau_n\wedge t} \mid \mathcal{F}_s\right] = M^2_{\tau_n\wedge s} - [M]_{\tau_n\wedge s} \text{ a.s.}$$

As above we know that $\left|M^2_{\tau_n\wedge t} - [M]_{\tau_n\wedge t}\right| \leq \sup_{0\leq u\leq\infty} M_u^2 + [M]_\infty$ and therefore $M^2_{\tau_n\wedge t} - [M]_{\tau_n\wedge t}$ is uniformly integrable. Therefore

$$\begin{aligned}\mathbf{E}\left[M_t^2 - [M]_t \mid \mathcal{F}_s\right] &= \lim_{n\to\infty} \mathbf{E}\left[M^2_{\tau_n\wedge t} - [M]_{\tau_n\wedge t} \mid \mathcal{F}_s\right] \\ &= \lim_{n\to\infty} M^2_{\tau_n\wedge s} - [M]_{\tau_n\wedge s} = M_s^2 - [M]_s \text{ a.s.}\end{aligned}$$

□

We can leverage the above facts to prove a type of continuity property of quadratic variation that is the root of what will become our dominated convergence theorem for stochastic integrals.

LEMMA 14.28. *Let M_n be a sequence of continuous local martingales, then $M_n^* \stackrel{P}{\to} 0$ if and only if $[M_n]_\infty \stackrel{P}{\to} 0$.*

PROOF. First we assume that $M_n^* \stackrel{P}{\to} 0$. Let $\epsilon > 0$ be given and define $\tau_n = \inf\{t \geq 0 \mid (M_n)_t > \epsilon\}$ which is an optional time because of the continuity of M_n and the right continuity of $\mathcal{F}$ (Lemma 9.70 and Lemma 9.69). Moreover, we know that $M_n^{\tau_n}$ is a bounded continuous martingale and therefore $(M_n^2 - [M_n])^{\tau_n} = (M_n^{\tau_n})^2 - [M_n^{\tau_n}]$ is a martingale starting at zero which shows that for all $t \geq 0$,

$$\mathbf{E}\left[[M_n^{\tau_n}]_t\right] = \mathbf{E}\left[(M_n^{\tau_n})_t^2\right] \leq \epsilon^2$$

Now we can use a Markov bound to see that

$$\begin{aligned}\mathbf{P}\{[M_n]_\infty > \epsilon\} &\leq \mathbf{P}\{[M_n]_\infty > \epsilon; \tau_n < \infty\} + \mathbf{P}\{[M_n]_\infty > \epsilon; \tau_n = \infty\} \\ &\leq \mathbf{P}\{\tau_n < \infty\} + \mathbf{P}\{[M_n]_{\tau_n} > \epsilon\} \\ &\leq \mathbf{P}\{M_n^* > \epsilon\} + \epsilon^{-1}\mathbf{E}\left[[M_n]_{\tau_n}\right] \\ &\leq \mathbf{P}\{M_n^* > \epsilon\} + \epsilon\end{aligned}$$

To see that this shows convergence in probability, first note that by our assumption that $M_n^* \stackrel{P}{\to} 0$ we have $\lim_{n\to\infty} \mathbf{P}\{[M_n]_\infty > \epsilon\} \leq \epsilon$. But now note that the left hand limit is a decreasing function of ϵ and therefore

$$\lim_{n\to\infty} \mathbf{P}\{[M_n]_\infty > \epsilon\} \leq \lim_{\delta\to 0^+} \lim_{n\to\infty} \mathbf{P}\{[M_n]_\infty > \delta\} \leq \lim_{\delta\to 0^+} \delta = 0$$

thus as $\epsilon > 0$ was arbitrary we have shown $[M_n]_\infty \stackrel{P}{\to} 0$.

Now we assume that $[M_n]_\infty \stackrel{P}{\to} 0$. As before let $\epsilon > 0$ be given and this time define $\tau_n = \inf\{t \geq 0 \mid [M_n]_t > \epsilon^2\}$ which is an $\mathcal{F}$-optional time by $\mathcal{F}$-adaptedness of $[M_n]$, continuity of $[M_n]$ and the right continuity of $\mathcal{F}$.

Now we can apply Proposition 14.27 to the local martingale $M_n^{\tau_n}$ for which, by definition of τ_n, we have $[M_n^{\tau_n}]_\infty = [M_n]_{\tau_n} \leq \epsilon^2$ and therefore $\mathbf{E}\left[[M_n^{\tau_n}]_\infty\right] < \infty$. We conclude that $M_n^{\tau_n}$ is an L^2-bounded martingale and therefore $(M_n^{\tau_n})^2 - [M_n^{\tau_n}]$ is a uniformly integrable martingale starting at zero. We are now in a position to mimic the first part of the proof. By the martingale property and the defintion of τ_n we have for all $0 \leq t \leq \infty$,

$$\mathbf{E}\left[(M_n^{\tau_n})_t^2\right] = \mathbf{E}\left[[M_n^{\tau_n}]_t\right] = \mathbf{E}\left[[M_n]_{\tau_n \wedge t}\right] \leq \epsilon^2$$

and by a Markov bound and Doob's L^2 inequality applied to the L^2 bounded martingale $M_n^{\tau_n}$ we get,

$$\begin{aligned}
\mathbf{P}\{M_n^* \geq \epsilon\} &\leq \mathbf{P}\{M_n^* \geq \epsilon; \tau_n < \infty\} + \mathbf{P}\{M_n^* \geq \epsilon; \tau_n = \infty\} \\
&\leq \mathbf{P}\{\tau_n < \infty\} + \mathbf{P}\{(M_n^{\tau_n})^* \geq \epsilon\} \\
&\leq \mathbf{P}\{\tau_n < \infty\} + \epsilon^{-1}\mathbf{E}\left[(M_n^{\tau_n})^*\right] \\
&\leq \mathbf{P}\{[M_n]_\infty > \epsilon^2\} + 4\epsilon^{-1}\mathbf{E}\left[(M_n^{\tau_n})_\infty^2\right] \\
&\leq \mathbf{P}\{[M_n]_\infty > \epsilon^2\} + 4\epsilon
\end{aligned}$$

and as before take the limit as $n \to \infty$ and then as $\epsilon \to 0$ to see that $M_n^* \xrightarrow{P} 0$. ☐

Because the covariation process $[M, N]$ is of finite variation we can define a pointwise Lebesgue-Stieltjes integral $\int f(\omega, s)\, d[M, N]_s$ for any progressive process $f(\omega, t)$ (TODO: is jointly measurable enough? If we assume progressive then I guess we get a local martingale out of this). Note that there is the potential for ambiguity in interpreting an integral with respect to a process of finite variation when the integrand is a step process as we could also consider using the definition as an elementary stochastic integral. It does turn out that these two possible definitions agree but we'll defer addressing the question and instead we will always explicitly denote the integration variable when considering a pointwise Stieltjes integral as in the expression $\int_0^t U_s\, dM_s$. TODO: Validate that the elementary stochastic integral defined above is consistent with the definition of the pointwise Stieltjes integral; it is worth understand the point at which we need this fact as Rogers and Williams indicate that they don't require it for quite some time. Actually the consistency when integrands are step processes is trivial to see. The fact that Rogers and Williams defer is the deeper fact that once one has defined a stochastic integral for not necessarily continuous local martingales one has the possibility for a stochastic integral with an integrator of finite variation. This integral can be shown to agree with the pointwise Stieltjes integral.

Before we begin we record the following simple fact about Lebesgue-Stieltjes integrals.

TODO: Remove as we moved this into a separate section.

Lemma 14.29. *Let F be a function of finite variation and for each $t \geq 0$ define $F^t(s) = F(t \wedge s)$, then for all measurable g we have $\int_0^t g\, dF = \int g\, dF^t$.*

Proof. First suppose that F is a non-decreasing right continuous function, and consider the Stieltjes measure μ_t defined by F^t (see 2.112). For any interval

$[a, b]$ we have

$$\mu_t([a, b]) = F^t(b) - F^t(a) = F(b \wedge t) - F(a \wedge t) = \int_0^\infty \mathbf{1}_{[a,b]} \mathbf{1}_{[0,t]} \, dF$$

and therefore μ_t obtained by applying the density function $\mathbf{1}_{[0,t]}$ to the Stieltjes measure for F. Now by Lemma 2.57 we see that $\int g \, dF^t = \int g \mathbf{1}_{[0,t]} \, dF = \int_0^t g \, dF$. To finish the result, write a function of finite variation as a difference of two montone functions. □

LEMMA 14.30. *Let M and N be continuous local martingales and let U and V be finite predictable step processes with deterministic jump times, then*

$$[\int U \, dM, \int V \, dN] = \int U_s V_s \, d[M, N]_s \text{ a.s.}$$

PROOF. We know that each of $\int U \, dM$ and $\int U \, dM$ is a continuous local martingale by Lemma 14.19. In addition each of the expressions in the results is invariant under centering thus we may assume $M_0 = N_0 = 0$. Furthermore for any optional time τ we have by Theorem 14.26

$$[\int U \, dM, \int V \, dN]^\tau = [\left(\int U \, dM\right)^\tau, \left(\int V \, dN\right)^\tau] = [\int U \, dM^\tau, \int V \, dN^\tau]$$

and by Lemma 14.15

$$\left(\int U_s V_s \, d[M, N]_s\right)^\tau = \int U_s V_s \, d[M, N]_s^\tau$$

so if we choose a common localizing sequence $\tau_n \uparrow \infty$ it suffices prove the result for M^{τ_n}, N^{τ_n} and $[M, N]^{\tau_n}$. Thus, we may assume that M, N and $[M, N]$ is each bounded. Thus each of M, N and $MN - [M, N]$ is a bounded martingale hence each is closable and we may in fact assume each is a bounded martingale on $[0, \infty]$.

Now we first assume that $V = 1$ and let $U = \sum_{k=1}^n \eta_k \mathbf{1}_{(t_{k-1}, t_k]}$. By appending an extra term with $\eta_n = 0$ we may assume that $t_n = \infty$. Now we compute using the definitions and the martingale property of M, N and $MN - [M, N]$ to see

$$\begin{aligned}
\mathbf{E}\left[N_\infty \int_0^\infty U \, dM\right] &= \mathbf{E}\left[\sum_{k=1}^n \eta_k \left(M_{t_k} - M_{t_{k-1}}\right) \sum_{k=1}^n \left(N_{t_k} - N_{t_{k-1}}\right)\right] \\
&= \mathbf{E}\left[\sum_{k=1}^n \eta_k \left(M_{t_k} N_{t_k} - M_{t_{k-1}} N_{t_{k-1}}\right)\right] \\
&= \mathbf{E}\left[\sum_{k=1}^n \eta_k \left([M, N]_{t_k} - [M, N]_{t_{k-1}}\right)\right] \\
&= \mathbf{E}\left[\int_0^\infty U_s \, d[M, N]_s\right]
\end{aligned}$$

For an arbitrary optional time τ we can also apply this argument to M^τ and N^τ to see that

$$\begin{aligned}
\mathbf{E}\left[N_\tau \int_0^\tau U \, dM\right] &= \mathbf{E}\left[N_\infty^\tau \int_0^\infty U \, dM^\tau\right] \\
&= \mathbf{E}\left[\int_0^\infty U_s \, d[M^\tau, N^\tau]_s\right] = \mathbf{E}\left[\int_0^\tau U_s \, d[M, N]_s\right]
\end{aligned}$$

From Lemma 14.17 we see that $N_t \int_0^t U\, dM - \int_0^t U_s\, d[M,N]_s$ is a martingale and therefore $[\int U\, dM, N] = \int_0^t U_s\, d[M,N]_s$ a.s. by uniqueness of the quadratic covariation.

Now we finish by assuming a general $V = \sum_{k=1}^n \xi_k \mathbf{1}_{(t_{k-1},t_k]}$. Note that we can assume by redefining ξ_k and η_k appropriately that both U and V are defined with respect to the same sequence of deterministic jump times $0 = t_0 < t_1 < \cdots < t_n$ so in particular $UV = \sum_{k=1}^n \eta_k \xi_k \mathbf{1}_{(t_{k-1},t_k]}$. We can compute directly twice using the special case just proven

$$\begin{aligned}
[\int U\, dM, \int V\, dN]_t &= \int_0^t U_s\, d[M, \int V\, dN]_s \\
&= \sum_{k=1}^n \eta_k \left([M, \int V\, dN]_{t_k \wedge t} - [M, \int V\, dN]_{t_{k-1} \wedge t} \right) \\
&= \sum_{k=1}^n \eta_k \left(\int_0^{t_k \wedge t} V_u\, d[M,N]_u - \int_0^{t_{k-1} \wedge t} V_u\, d[M,N]_u \right) \\
&= \sum_{k=1}^n \eta_k \sum_{j=0}^n \xi_j \left([M,N]_{t_j \wedge t_k \wedge t} - [M,N]_{t_{j-1} \wedge t_k \wedge t} - [M,N]_{t_j \wedge t_{k-1} \wedge t} + [M,N]_{t_{j-1} \wedge t_{k-1} \wedge t} \right) \\
&= \sum_{k=1}^n \eta_k \xi_k \left([M,N]_{t_k \wedge t} - [M,N]_{t_{k-1} \wedge t} \right) \\
&= \int_0^t U_s V_s\, d[M,N]_s
\end{aligned}$$

and the full result is proven. $\square$

We have the following bounds on ruin probabilities as a corollary of Optional Sampling for continuous martingales.

LEMMA 14.31. *Let M be a continuous martingale with $M_0 = 0$ and such that $\boldsymbol{P}\{M^* > 0\} > 0$. If we define $\tau_x = \inf\{t > 0 \mid M_t = x\}$ then for every $a < 0 < b$ we have*

$$\boldsymbol{P}\{\tau_a < \tau_b \mid M^* > 0\} \leq \frac{b}{b-a} \leq \boldsymbol{P}\{\tau_a \leq \tau_b \mid M^* > 0\}$$

PROOF. We know that τ_a and τ_b are optional by continuity of M and Lemma 9.70. Define $\tau = \tau_a \wedge \tau_b$ which we know is optional as well. For every $t \geq 0$, by Optional Sampling we know that $\mathbf{E}[M_{\tau \wedge t}] = M_0 = 0$. Clearly $\lim_{t \to \infty} M_{\tau \wedge t} = M_\tau$ and by the definition of τ we know that $|M_{\tau \wedge t}| \leq -a \vee b < \infty$ and therefore we can apply Dominated Convergence to conclude that $\mathbf{E}[M_\tau] = 0$. Now we can establish bounds using two simple facts. First by continuity of M, we know that $\tau_a = \tau_b$ if and only if $\tau_a = \tau_b = \tau = \infty$. Secondly $\tau_a \neq \tau_b$ implies $M^* > 0$. With these observations in hand,

$$\begin{aligned}
0 &= \mathbf{E}\left[M_\tau; \tau_a < \tau_b\right] + \mathbf{E}\left[M_\tau; \tau_b < \tau_a\right] + \mathbf{E}\left[M_\infty; \tau_a = \tau_b = \infty\right] \\
&\leq a\mathbf{P}\{\tau_a < \tau_b\} + b\mathbf{P}\{\tau_b < \tau_a\} + b\mathbf{P}\{M^* > 0; \tau_a = \tau_b = \infty\} \\
&= a\mathbf{P}\{\tau_a < \tau_b\} + b\mathbf{P}\{M^* > 0; \tau_b \leq \tau_a\} \\
&= a\mathbf{P}\{\tau_a < \tau_b\} + b\mathbf{P}\{M^* > 0\} - b\mathbf{P}\{M^* > 0; \tau_a < \tau_b\} \\
&= b\mathbf{P}\{M^* > 0\} - (b-a)\mathbf{P}\{M^* > 0; \tau_a < \tau_b\}
\end{aligned}$$

which gives the first inequality. The second inequality is demonstrated in the same way but using a lower bound for M_∞ on $\tau_a = \tau_b = \infty$,

$$\begin{aligned}
0 &\geq a\mathbf{P}\{\tau_a < \tau_b\} + b\mathbf{P}\{\tau_b < \tau_a\} + a\mathbf{P}\{M^* > 0; \tau_a = \tau_b = \infty\} \\
&= a\mathbf{P}\{M^* > 0; \tau_a \leq \tau_b\} + b\mathbf{P}\{M^* > 0; \tau_b < \tau_a\} \\
&= a\mathbf{P}\{M^* > 0; \tau_a \leq \tau_b\} + b\mathbf{P}\{M^* > 0\} - b\mathbf{P}\{M^* > 0; \tau_a \leq \tau_b\} \\
&= b\mathbf{P}\{M^* > 0\} - (b-a)\mathbf{P}\{M^* > 0; \tau_a \leq \tau_b\}
\end{aligned}$$

□

THEOREM 14.32 (Burkholder-Davis-Gundy Inequalities). *For every $p > 0$ there exist a constant $0 < c_p < \infty$ such that for every continuous local martingale M with $M_0 = 0$ we have*

$$c_p^{-1} \boldsymbol{E}\left[[M]_\infty^{p/2}\right] \leq \boldsymbol{E}\left[(M^*)^p\right] \leq c_p \boldsymbol{E}\left[[M]_\infty^{p/2}\right]$$

PROOF. TODO: Perform reduction to the bounded martingale case via localization and optional sampling (Kallenberg indicates that we may also assume $[M]$ is bounded).

The following argument is quite elementary in each of its steps but is not entirely obvious so we spell it out in great detail. To derive the inequalities for expectations we'll use Lemma 3.8 and therefore we proceed by creating tail bounds for the random variables in question. We first work on the right hand inequality of the result. Let $r > 0$ be fixed and define $\tau = \inf\{t \geq 0 \mid M_t^2 = r\}$ (which is an optional time by sample path continuity and Lemma 9.70), $\tilde{M} = M - M^\tau$ and $N = \tilde{M}^2 - [\tilde{M}]$. Pick any $0 < c < 1$ (we'll later refine the required bounds on c) and write

$$\begin{aligned}
\mathbf{P}\{(M^*)^2 \geq 4r\} &= \mathbf{P}\{(M^*)^2 \geq 4r; [M]_\infty \geq cr\} + \mathbf{P}\{(M^*)^2 \geq 4r; [M]_\infty < cr\} \\
&\leq \mathbf{P}\{[M]_\infty \geq cr\} + \mathbf{P}\{(M^*)^2 \geq 4r; [M]_\infty < cr\}
\end{aligned}$$

we get

$$\mathbf{P}\{(M^*)^2 \geq 4r\} - \mathbf{P}\{[M]_\infty \geq cr\} \leq \mathbf{P}\{(M^*)^2 \geq 4r; [M]_\infty < cr\}$$

CLAIM 14.32.1. $\{(M^*)^2 \geq 4r; [M]_\infty < cr\} \subset \{N > -cr; \sup_t N_t > r - cr\}$

From Theorem 14.26 we compute

$$[\tilde{M}] = [M - M^\tau, M - M^\tau] = [M, M] - [M^\tau, M] - [M^\tau, M] + [M^\tau, M^\tau] = [M] - [M]^\tau$$

Since $[M]^\tau$ is non-negative it follows that $[\tilde{M}] \leq [M]$ and therefore $[M]_\infty < cr$ implies $[\tilde{M}]_\infty < cr$. Trivially $\tilde{M}^2 \geq 0$ so we know that $[M]_\infty < cr$ implies

$$N = \tilde{M}^2 - [\tilde{M}] > -cr$$

On $\{(M^*)^2 \geq 4r\}$ we know that $\tau < \infty$ and therefore $|M_\tau| = \sqrt{r}$ and for any $\epsilon > 0$ we can find a random $t \geq \tau$ such that $|M_t| \geq 2\sqrt{r} - \epsilon$, thus $|M_t - M_\tau| > \sqrt{r} - \epsilon$ which implies $\sup_t \tilde{M}_t^2 \geq r$. On $\{(M^*)^2 \geq 4r; [M]_\infty < cr\}$ we therefore get

$$\sup_t N_t = \sup_t (\tilde{M}_t^2 - [\tilde{M}]_t) \geq \sup_t \tilde{M}_t^2 - [\tilde{M}]_\infty \geq r - cr$$

and the claim follows.

From Claim 14.32.1 we apply a union bound and get

$$\begin{aligned}
\mathbf{P}\{(M^*)^2 \geq 4r\} - \mathbf{P}\{[M]_\infty \geq cr\} &\leq \mathbf{P}\{(M^*)^2 \geq 4r; [M]_\infty < cr\} \\
&\leq \mathbf{P}\{N > -cr; \sup_t N_t > r - cr\}
\end{aligned}$$

To get a further bound we note that N is a martingale with $N_0 = 0$ and introduce the hitting times τ_{-cr} and τ_{r-cr} for N. We can apply the upper bound in the Gambler's Ruin Lemma 14.31 with $-cr < 0 < r - cr$ and use the fact that $\{N > -cr; \sup_t N_t > r - cr\} \subset \{\tau_{-cr} > \tau_{r-cr}; N^* > 0\}$ to conclude that

$$\begin{aligned}
\mathbf{P}\{N > -cr; \sup_t N_t > r - cr\} &\leq \mathbf{P}\{\tau_{-cr} > \tau_{r-cr}; N^* > 0\} \\
&= \mathbf{P}\{N^* > 0\} - \mathbf{P}\{\tau_{-cr} \leq \tau_{r-cr}; N^* > 0\} \\
&\leq (1 - \frac{r - cr}{r - cr + cr})\mathbf{P}\{N^* > 0\} = c\mathbf{P}\{N^* > 0\}
\end{aligned}$$

It is clear from the nonnegativity of $[\tilde{M}]$ and the definition of N that $N^* > 0$ implies $\tilde{M}^* = (M - M^\tau)^* > 0$ which implies $\tau < \infty$ and therefore $(M^*)^2 > r$. Summarizing all of the bounds we get

$$\begin{aligned}
\mathbf{P}\{(M^*)^2 \geq 4r\} - \mathbf{P}\{[M]_\infty \geq cr\} &\leq \mathbf{P}\{(M^*)^2 \geq 4r; [M]_\infty < cr\} \\
&\leq \mathbf{P}\{N > -cr; \sup_t N_t > r - cr\} \\
&\leq c\mathbf{P}\{N^* > 0\} \leq c\mathbf{P}\{(M^*)^2 > r\}
\end{aligned}$$

Now we multiply by $\frac{p}{2} r^{p/2-1}$ and integrate to get

$$\begin{aligned}
&\frac{p}{2}\int_0^\infty r^{p/2-1}\mathbf{P}\{(M^*)^2 \geq 4r\}\, dr - \frac{p}{2}\int_0^\infty r^{p/2-1}\mathbf{P}\{[M]_\infty \geq cr\}\, dr \\
&\leq \frac{cp}{2}\int_0^\infty r^{p/2-1}\mathbf{P}\{(M^*)^2 > r\}\, dr
\end{aligned}$$

which yields upon changing integration variables ($2\sqrt{r}$, cr and $\sqrt{r}$ respectively)

$$\begin{aligned}
&\frac{p}{2^p}\int_0^\infty r^{p-1}\mathbf{P}\{M^* \geq r\}\, dr - c^{-p/2}\frac{p}{2}\int_0^\infty r^{p/2-1}\mathbf{P}\{[M]_\infty \geq r\}\, dr \\
&\leq cp\int_0^\infty r^{p-1}\mathbf{P}\{M^* > r\}\, dr
\end{aligned}$$

and applying Lemma 3.8

$$2^{-p}\mathbf{E}\left[(M^*)^p\right] - c^{-p/2}\mathbf{E}\left[|[M]_\infty|^{p/2}\right] \leq c\mathbf{E}\left[(M^*)^p\right]$$

Thus we get the right hand inequality for $c_p = c^{-p/2}/(2^{-p} - c)$ which is a positive constant for any $0 < c < 2^{-p}$.

The proof of the left hand inequality follows the same pattern but this time we define the optional time $\tau = \inf\{t \geq 0 \mid [M_t] = r\}$ and as before $\tilde{M} = M - M^\tau$ and $N = \tilde{M}^2 - [\tilde{M}]$. We let $r > 0$ be arbitrary, assuming that $0 < c < 1/4$. We give the entire computation at once and them make some comments about the details of the justification:

$$\begin{aligned}
\mathbf{P}\{[M]_\infty \geq 2r\} - \mathbf{P}\{(M^*)^2 \geq cr\} &\leq \mathbf{P}\{[M]_\infty \geq 2r; (M^*)^2 < cr\} \\
&\leq \mathbf{P}\{N < 4cr; \inf_t N_t < 4cr - r\} \\
&\leq 4c\mathbf{P}\{N^* > 0\} \\
&\leq 4c\mathbf{P}\{[M]_\infty \geq r\}
\end{aligned}$$

The first inequality follows as before by a simple union bound. To see the second inequality, note first that on $\{(M^*)^2 < cr\}$ by non-negativity of $[\tilde{M}]$ we have

$$N \leq \tilde{M}^2 \leq (|M| + |M^\tau|)^2 \leq (2M^*)^2 < 4cr$$

and also on $\{[M]_\infty \geq 2r\}$ we have $\tau < \infty$ and

$$[\tilde{M}]_\infty = [M]_\infty - [M]_\tau \geq 2r - r = r$$

To see the third inequality we again apply Gambler's Ruin Lemma 14.31 to N this time on $4cr - r < 0 < 4cr$ noting that $\mathbf{P}\{N < 4cr; \inf_t N_t < 4cr - r\} \leq \mathbf{P}\{\tau_{4cr-r} < \tau_{4c}; N^* > 0\} \leq 4c\mathbf{P}\{N^* > 0\}$. The final inequality again follows from noting that $\tau < \infty$ on $N^* > 0$ and therefore because $[M]$ is non-decreasing we have $[M]_\infty \geq [M]_\tau = r$.

Again we multiply by $(p/2)r^{p/2-1}$ and integrate to get

$$\begin{aligned}
&\frac{p}{2}\int_0^\infty r^{p/2-1}\mathbf{P}\{[M]_\infty \geq 2r\}\,dr - \frac{p}{2}\int_0^\infty r^{p/2-1}\mathbf{P}\{(M^*)^2 \geq cr\}\,dr \\
&\leq 4c\frac{p}{2}\int_0^\infty r^{p/2-1}\mathbf{P}\{[M]_\infty \geq r\}\,dr
\end{aligned}$$

which upon changing variables and applying Lemma 3.8

$$2^{p/2}\mathbf{E}\left[|[M]_\infty|^{p/2}\right] - c^{-p/2}\mathbf{E}\left[|M^*|^p\right] \leq 4c\mathbf{E}\left[|[M]_\infty|^{p/2}\right]$$

which yields the left hand inequality with $c_p = c^{-p/2}/(2^{-p/2} - 4c)$ which is positive for any $0 < c < 2^{-p/2-2}$. ∎

Later on we will need the following extension of the BDG inequalities to the vector valued case.

Corollary 14.33. *Let $M = (M^{(1)}, \ldots, M^{(d)})$ be a continuous vector valued local martingale such that $M_0 = 0$ and let*

$$\|M\|^*_t = \sup_{0 \leq s \leq t} \|M_s\|$$

$$[M]_t = \sum_{i=1}^d [M^{(i)}]_t$$

then for any $p > 0$ there exists a constant d_p such that

$$d_p^{-1} \boldsymbol{E}\left[[M]_\infty^{p/2} \right] \leq \boldsymbol{E}\left[(\|M\|^*)^p \right] \leq d_p \boldsymbol{E}\left[[M]_\infty^{p/2} \right]$$

Proof. Using definitions we have

$$(\|M\|^*)^p = \sup_{0 \leq s < \infty} \left((M^{(1)})^2 + \cdots + (M^{(d)})^2 \right)^{p/2}$$

From Proposition C.4 we have

$$\left((M_t^{(1)})^2 + \cdots + (M_t^{(d)})^2 \right)^{p/2} \leq \left(\left| M_t^{(1)} \right| + \cdots + \left| M_t^{(d)} \right| \right)^p \leq d^{p-1} \left(\left| M_t^{(1)} \right|^p + \cdots + \left| M_t^{(d)} \right|^p \right)$$

and therefore taking suprema and expectations and the scalar BDG inequality

$$\begin{aligned}
\mathbf{E}\left[(\|M\|^*)^p \right] &\leq d^{p-1} \left(\mathbf{E}\left[((M^{(1)})^*)^p \right] + \cdots + \mathbf{E}\left[((M^{(d)})^*)^p \right] \right) \\
&\leq d^{p-1} c_p \left(\mathbf{E}\left[[M^{(1)}]_\infty^{p/2} \right] + \cdots + \mathbf{E}\left[[M^{(d)}]_\infty^{p/2} \right] \right) \\
&\leq d^{p-1} c_p \mathbf{E}\left[[M]_\infty^{p/2} \right]
\end{aligned}$$

To get the reverse inequality first note

$$\left((M_t^{(1)})^2 + \cdots + (M_t^{(d)})^2 \right)^{p/2} \geq d^{-1} \left(\left| M_t^{(1)} \right|^p + \cdots + \left| M_t^{(d)} \right|^p \right)$$

and therefore taking suprema (using the general fact that $\sup_t(f_1(t)) + \cdots + \sup_t(f_d(t)) \leq d \max_{1 \leq i \leq d} \sup_t f_i(t) \leq d \sup_t(f_1(t) + \cdots + f_d(t))$) and expectations and using the scalar BDG inequality and the Power Mean Inequality

$$\begin{aligned}
\mathbf{E}\left[(\|M\|^*)^p \right] &\geq d^{-2} \left(\mathbf{E}\left[((M^{(1)})^*)^p \right] + \cdots + \mathbf{E}\left[((M_t^{(d)})^*)^p \right] \right) \\
&\geq d^{-2} c_p^{-1} \left(\mathbf{E}\left[[M^{(1)}]_\infty^{p/2} \right] + \cdots + \mathbf{E}\left[[M^{(d)}]_\infty^{p/2} \right] \right) \\
&\geq d^{p/2-2} c_p^{-1} \mathbf{E}\left[[M]_\infty^{p/2} \right]
\end{aligned}$$

□

In the following Lemma we remind the reader of the notation $\int g \; |dF|$ to denote integration with respect to the Lebesgue-Stieltjes measure determined by the total variation function of F.

Lemma 14.34. *Let M and N be continuous local martingales, then almost surely for every $t \geq 0$,*

$$|[M,N]_t| \leq \int_0^t |d[M,N]|_s \leq [M]_t^{1/2} [N]_t^{1/2} \tag{17}$$

Furthermore almost surely for any jointly measurable processes U and V we have

$$\int_0^t |U_s V_s| \; |d[M,N]|_s \leq \left(\int_0^t U_s^2 \, d[M]_s \right)^{1/2} \left(\int_0^t V_s^2 \, d[N]_s \right)^{1/2}$$

(TODO: Confirm that almost sure this holds for all U, V not that for each pair U, V this holds a.s.)

Proof. First we can use positivity and bilinearity of quadratic covariation to see that for a fixed $t \geq 0$ and $\lambda \in \mathbb{R}$ we have

$$0 \leq [M + \lambda N]_t = [M]_t + 2\lambda [M,N]_t + \lambda^2 [N]_t \text{ a.s.}$$

It follows that $\mathbf{P}\{\cap_{\lambda \in \mathbb{Q}} \{0 \leq [M]_t + 2\lambda[M,N]_t + \lambda^2 [N]_t\}\} = 1$ and by continuity of the quadratic polynomial we get that for fixed $t \geq 0$, almost surely $0 \leq [M]_t + 2\lambda[M,N]_t + \lambda^2[N]_t$ for all $\lambda \in \mathbb{R}$. Taking the discriminant of the quadratic polynomial and using the fact that it must be non-negative we see that for every $t \geq 0$ we have $[M,N]_t^2 \leq [M]_t [N]_t$ almost surely. Again, taking the intersection of a countable number of almost sure events we see that almost surely we have $[M,N]_q^2 \leq [M]_q [N]_q$ for all $q \in \mathbb{Q}$ with $q \geq 0$ and by continuity of the quadratic variation this implies that almost surely $[M,N]_t^2 \leq [M]_t [N]_t$ for all $t \geq 0$.

Now fix an $s \geq 0$ and consider the processes $M - M^s$ and $N - N^s$. Replaying our continuity argument once more we see that almost surely the inequality just proven will hold almost surely over all the processes $M - M^s$, $N - N^s$ and all $t \geq 0$. Using this fact and Theorem 14.26 we conclude that almost surely for all $s \geq 0$ and $s < t$ we have

$$|[M,N]_t - [M,N]_s| = |[M - M^s, N - N^s]_t| \leq ([M]_t - [M]_s)^{1/2} ([N]_t - [N]_s)^{1/2}$$

Suppose we are given a partition $s = t_0 < \cdots < t_n = t$ and use the triangle inequality, the Cauchy-Schwartz inequality for sequences and the above inequality gives us

$$\begin{aligned}
|[M,N]_t - [M,N]_s| &\leq \sum_{j=1}^n \left| [M,N]_{t_j} - [M,N]_{t_{j-1}} \right| \\
&\leq \sum_{j=1}^n \left([M]_{t_j} - [M]_{t_{j-1}} \right)^{1/2} \left([N]_{t_j} - [N]_{t_{j-1}} \right)^{1/2} \\
&\leq \left(\sum_{j=1}^n [M]_{t_j} - [M]_{t_{j-1}} \right)^{1/2} \left(\sum_{j=1}^n [N]_{t_j} - [N]_{t_{j-1}} \right)^{1/2} \\
&= ([M]_t - [M]_s)^{1/2} ([N]_t - [N]_s)^{1/2}
\end{aligned}$$

Again, note that this holds almost sure simultaneously for all $0 \leq s < t$, all $n \geq 0$ and all partitions $s = t_0 < \cdots < t_n = t$. We may then take the supremum over all partitions to get

$$|[M,N]_t - [M,N]_s| \leq \int_s^t |d[M,N]_s| \leq ([M]_t - [M]_s)^{1/2} ([N]_t - [N]_s)^{1/2}$$

and substituting $s = 0$ we get (17).

Before proceeding further it is helpful to name all of the random Lebesgue-Stieltjes measures floating around: let $\mu = d[M]$, $\nu = d[N]$ and $\rho = |d[M,N]|$.

Note that we have shown that almost surely for every closed interval $I \subset \mathbb{R}$ we have $\rho(I)^2 \leq \mu(I)\nu(I)$. By continuity of $[M]$, $[N]$ and $[M, N]$ the measures above have no atoms and therefore this inequality also holds for open intervals. Now if we let G be an arbitrary open set then we can write it as a disjoint union of open intervals (Lemma 1.16) $G = \cup_{n=1}^\infty I_n$. Then by countable additivity and Cauchy-Schwartz for sequences

$$\begin{aligned}\rho(G) &= \sum_{n=1}^\infty \rho(I_n) \leq \sum_{n=1}^\infty \mu(I_n)^{1/2}\nu(I_n)^{1/2} \\ &\leq \left(\sum_{n=1}^\infty \mu(I_n)\right)^{1/2}\left(\sum_{n=1}^\infty \nu(I_n)\right)^{1/2} = \mu(G)^{1/2}\nu(G)^{1/2}\end{aligned}$$

TODO: Extend to general Borel sets by monotone classes: I think we needed boundedness of the measures here.

Now let $f = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$ and $g = \sum_{i=1}^n b_i \mathbf{1}_{A_i}$ be positive simple functions. Then once again applying Cauchy-Schwartz for sequences we get

$$\begin{aligned}\int f(s)g(s)\ |d[M,N]_s| &= \sum_{i=1}^n a_i b_i \rho(A_i) \\ &\leq \sum_{i=1}^n a_i b_i \mu(A_i)^{1/2}\nu(A_i)^{1/2} \\ &\leq \left(\sum_{i=1}^n a_i^2 \mu(A_i)\right)^{1/2}\left(\sum_{i=1}^n b_i^2 \nu(A_i)\right)^{1/2} \\ &= \left(\int f^2(s)\, d[M]_s\right)^{1/2}\left(\int g^2(s)\, d[N]_s\right)^{1/2}\end{aligned}$$

For general positive measurable functions f and g we take positive simple approximations $f_n \uparrow f$ and $g_n \uparrow g$ and we get by Monotone Convergence

$$\begin{aligned}\int f(s)g(s)\ |d[M,N]_s| &= \lim_{n\to\infty} \int f_n(s)g_n(s)\ |d[M,N]_s| \\ &\leq \lim_{n\to\infty}\left(\int f_n^2(s)\, d[M]_s\right)^{1/2} \lim_{n\to\infty}\left(\int g_n^2(s)\, d[N]_s\right)^{1/2} \\ &= \left(\int f^2(s)\, d[M]_s\right)^{1/2}\left(\int g^2(s)\, d[N]_s\right)^{1/2}\end{aligned}$$

noting that this holds almost surely for all f and g positive and measurable.

TODO: Finish, is there anything subtle about applying to the processes? $\square$

Definition 14.35. Given a continuous local martingale M we let $L(M)$ denote the set of processes that are progressively measurable and for which $\int_0^t V_s^2 \, d[M]_s < \infty$ almost surely for all $t \geq 0$.

The space $L(M)$ gives the integrands for the extension of the stochastic integral with respect to the integrator M.

Theorem 14.36. *Let M be a continuous local martingale and $V \in L(M)$, there exists an almost surely unique continuous local martingale $\int V\, dM$ starting at zero* *and for which for every continuous local martingale N almost surely $[\int V\,dM, N]_t = \int_0^t V_s\,d[M,N]_s$ for all $t \geq 0$.*

Proof. First we show uniqueness as we shall use it during the existence argument. Suppose that M' and M'' are continuous local martingales starting at zero for which for every continuous local martingale $[M', N] = [M'', N] = \int V_s\,d[M,N]_s$ almost surely. By linearity of quadratic covariation, this tell us that for all N we have $[M' - M'', N] = 0$ almost surely. In particular this will be true if we pick $N = M' - M''$ so we know that $[M' - M''] = 0$ almost surely. By definition of the quadratic variation this implies that $(M' - M'')^2$ is a continuous local martingale starting at zero. Picking a localizing sequence τ_n and using the martingale property we see that $\mathbf{E}\left[(M'_{t\wedge\tau_n} - M''_{t\wedge\tau_n})^2\right] = 0$ which shows us that $(M'_{t\wedge\tau_n} - M''_{t\wedge\tau_n})^2$ almost surely. Taking the limit as $n \to \infty$ we get that $(M'_t - M''_t)^2 = 0$ almost surely for each $t \geq 0$ hence simultaneously for all $t \in \mathbb{Q}_+$ and then by continuity for all $t \geq 0$ almost surely.

We first assume that $\int_0^\infty V_s^2\,d[M]_s < \infty$ almost surely and we use the notation $\|V\|_M^2 = \int_0^\infty V_s^2\,d[M]_s$ to denote the corresponding value. Then if $N \in \mathcal{M}^2$ we have

$$
\begin{aligned}
\left|\mathbf{E}\left[\int_0^\infty V_s\,d[M,N]_s\right]\right| &\leq \mathbf{E}\left[\left|\int_0^\infty V_s\,d[M,N]_s\right|\right] \\
&\leq \mathbf{E}\left[\int_0^\infty |V_s|\;|d[M,N]_s|\right] && \text{by Lemma 2.122} \\
&\leq \mathbf{E}\left[\left(\int_0^\infty V_s^2\,d[M]_s\right)^{1/2}\left(\int_0^\infty d[N]_s\right)^{1/2}\right] && \text{by Lemma 14.34} \\
&= \mathbf{E}\left[\left(\int_0^\infty V_s^2\,d[M]_s\right)^{1/2}[N]_\infty^{1/2}\right] \\
&\leq \mathbf{E}\left[\int_0^\infty V_s^2\,d[M]_s\right]^{1/2}\mathbf{E}\left[[N]_\infty\right]^{1/2} && \text{by Cauchy Schwartz} \\
&= \|V\|_M\,\mathbf{E}\left[N_\infty^2\right]^{1/2} = \|V\|_M\,\|N\|
\end{aligned}
$$

which shows that $N \mapsto \mathbf{E}\left[\int_0^\infty V_s\,d[M,N]_s\right]$ is a continuous linear functional on $\mathcal{M}^2$. Thus since $\mathcal{M}^2$ is a Hilbert space with inner product given by $\langle M, N\rangle = \mathbf{E}\left[M_\infty N_\infty\right]$ (Lemma 14.24) we know that there exists an L^2-bounded martingale $\int V\,dM \in \mathcal{M}^2$ such that $\mathbf{E}\left[\int_0^\infty V_s\,d[M,N]_s\right] = \mathbf{E}\left[N_\infty \cdot \int_0^\infty V\,dM\right]$ for all $N \in \mathcal{M}^2$ (we emphasize that the use of the integral sign in the name $\int V\,dM$ we give to this martingale is only meant to be suggestive and the reader should not get confused trying to figure out that this element can be constructed by some kind of generalized sum; at this point it is no more and no less than the element of the Hilbert space corresponding to the linear functional we've defined).

Since V is progressive we know that $\int V_s\,d[M,N]_s$ is $\mathcal{F}$-adapted (Lemma 14.12) and we have just shown that it is integrable. Now let τ be an arbitrary optional time and apply the above construction to N^τ (TODO: Remind why $N^\tau \in \mathcal{M}^2$). in the following computation

$$\begin{aligned}
\mathbf{E}\left[\int_0^\tau V_s\, d[M,N]_s\right] &= \mathbf{E}\left[\int_0^\infty V_s\, d[M,N]_s^\tau\right] && \text{Lemma 14.15}\\
&= \mathbf{E}\left[\int_0^\infty V_s\, d[M,N^\tau]_s\right] && \text{Lemma 14.26}\\
&= \mathbf{E}\left[N_\infty^\tau \cdot \int_0^\infty V\, dM\right] && \text{definition of } \int V\, dM\\
&= \mathbf{E}\left[N_\tau \cdot \int_0^\infty V\, dM\right] &&\\
&= \mathbf{E}\left[N_\tau \mathbf{E}\left[\int_0^\infty V\, dM \mid \mathcal{F}_\tau\right]\right] && \text{Tower Property}\\
&= \mathbf{E}\left[N_\tau \int_0^\tau V\, dM\right] && \text{Optional Sampling}
\end{aligned}$$

We apply Lemma 14.17 to conclude that $N_t \int_0^t V\, dM - \int_0^t V_s\, d[M,N]_s$ is a martingale. By the continuity of $[M,N]$ we know that $\int_0^t V_s\, d[M,N]_s$ is continuous and has locally finite variation (Corollary 2.126); thus uniqueness and the defining property of quadratic covariation implies $\int V_s\, d[M,N]_s = [N, \int V\, dM]$ almost surely.

The next step is to extend the defining property of the integral to arbitrary continuous local martingales. For this we take a localizing sequence τ_n such that N^{τ_n} is bounded (hence in $\mathcal{M}^2$). Let A be the event that $\tau_n \uparrow \infty$ and for each n, let A_n be the event that $[N^{\tau_n}, \int V\, dM] = \int V_s\, d[M,N]_s$. For all $\omega \in A \cap (\cap_{n=1}^\infty A_n)$ and $t \geq 0$ we have

$$\begin{aligned}
[N, \int V\, dM]_t(\omega) &= \lim_{n\to\infty} [N, \int V\, dM]_t^{\tau_n}(\omega)\\
&= \lim_{n\to\infty} [N^{\tau_n}, \int V\, dM]_t(\omega)\\
&= \lim_{n\to\infty} \int_0^t V_s(\omega)\, d[M, N^{\tau_n}]_s(\omega)\\
&= \lim_{n\to\infty} \int_0^{t\wedge\tau_n} V_s(\omega)\, d[M,N]_s(\omega)\\
&= \int_0^t V_s(\omega)\, d[M,N]_s(\omega)
\end{aligned}$$

and as $\mathbf{P}\{A \cap (\cap_{n=1}^\infty A_n)\} = 1$ we have $[N, \int V\, dM] = \int V_s\, d[M,N]_s$ almost surely.

Lastly we must remove the assumption that $\int_0^\infty V_s^2\, d[M]_s < \infty$. We know that $\int_0^t V_s^2\, d[M]_s$ is a continuous process (Lemma 2.126) and therefore for every $n > 0$ we can define an optional time $\tau_n = \inf\{t \geq 0 \mid \int_0^t V_s^2\, d[M]_s = n\}$. We have

$$\int_0^\infty V_s^2\, d[M^{\tau_n}]_s = \int_0^{\tau_n} V_s^2\, d[M]_s = n < \infty$$

and by our assumption that $\int_0^t V_s^2 \, d[M]_s < \infty$ for all $t \geq 0$ we know that $\tau_n \uparrow \infty$. We apply the existing construction to define $\int V \, dM^{\tau_n}$ and it satisfies

$$[N, \int V \, dM^{\tau_n}]_t = \int_0^t V_s \, d[M,N]_s^{\tau_n} = \int_0^{t \wedge \tau_n} V_s \, d[M,N]_s$$

for every continuous local martingale N. Moreover for $m < n$, from the above fact and Lemma 14.26 we have

$$[N, \left(\int V \, dM^{\tau_n} \right)^{\tau_m}]_t = [N, \int V \, dM^{\tau_n}]_{t \wedge \tau_m} = \int_0^{t \wedge \tau_m} V_s \, d[M,N]_s$$

for all continuous local martingales N which by uniqueness of the stochastic integral shows $\left(\int V \, dM^{\tau_n} \right)^{\tau_m} = \int V \, dM^{\tau_m}$ so that $\int V \, dM^{\tau_m}$ and $\int V \, dM^{\tau_n}$ agree on the interval $[0, \tau_m]$. Therefore we can define $\int_0^t V \, dM$ as the limit of $\int_0^t V \, dM^{\tau_n}$ for any $\tau_n \geq t$. The fact that this defines an adapted process follows from writing $\int_0^t V \, dM = \sum_{n=1}^\infty \mathbf{1}_{\{\tau_{n-1} \leq t < \tau_n\}} \int_0^t V \, dM^{\tau_n}$ together with the facts that τ_n is optional and $\int V \, dM^{\tau_n}$ is adapted. Continuity at $t \geq 0$ follows by picking $\tau_n > t$ and noting that $\int_0^t V \, dM = \int_0^t V \, dM^{\tau_n}$ and continuity of $\int_0^t V \, dM^{\tau_n}$ at t. By Lemma 14.8 we know that $\int V \, dM$ is a continuous local martingale. Lastly by construction, for all $n \geq 0$ and each continuous local martingale there is a set A_n with $\mathbf{P}\{A_n\} = 1$ such that

$$\begin{aligned} [N, \int V \, dM]_t &= [N, \int V \, dM]_t^{\tau_n} = [N, \left(\int V \, dM \right)^{\tau_n}]_t = [N, \int V \, dM^{\tau_n}]_t \\ &= \int_0^t V_s \, d[M^{\tau_n}, N]_s = \int_0^{t \wedge \tau_n} V_s \, d[M,N]_s = \int_0^t V_s \, d[M,N]_s \end{aligned}$$

for all $0 \leq t \leq \tau_n$ on A_n. Thus taking the intersection of A_n we see that $[N, \int V \, dM] = \int V_s \, d[M,N]_s$ almost surely. $\square$

With this Theorem proven we know that the following defintion makes sense.

Definition 14.37. Given a continuous local martingale M and a progressive process V such that $\int_0^t V_s^2 \, d[M]_s < \infty$ for all $t \geq 0$, the *stochastic integral* $\int V \, dM$ is the almost surely unique continuous local martingale for which $[\int V \, dM, N]_t = \int_0^t V_s \, d[M,N]_s$ for all $t \geq 0$ almost surely for every continuous local martingale N.

Here we collect a few of the most elementary facts about the stochastic integral. In particular we call attention to the Ito Isometry which doesn't figure as prominently in our presentation as in others but is nonetheless very useful. We shall have more to say about this later.

Lemma 14.38. *Let M be a continuous local martingales. If $U, V \in L(M)$ such that $U_t = V_t$ for all $t \geq 0$ almost surely then $\int U \, dM = \int V \, dM$. The stochastic integral is bilinear in both the integrand and integrator. TODO: Be very precise about assumptions here! E.g. is $V \in L(aM + bN)$ equivalent to $V \in L(M)$ and $V \in L(N)$? Clearly the latter is at least as strong as the former. If M is a continuous local martingale and $V \in L(M)$ then we have $[\int V \, dM]_t = \int_0^t V_s^2 \, d[M]_s$ for all $t \geq 0$ almost surely. In particular, if M is a continuous martingale then we have the* Ito Isometry

$$\boldsymbol{E}\left[\left(\int_0^t V \, dM \right)^2 \right] = \int_0^t V_s^2 \, d[M]_s \text{ for all } t \geq 0$$

If M is a continuous martingale, τ is an optional time and almost surely $V_s = 0$ for all $s \leq \tau$ (i.e. $V_s \mathbf{1}_{s \leq \tau} = 0$ almost surely) then we have

$$\int_0^t V_s dM_s = 0 \textit{ almost surely on } \{t \leq \tau\}$$

PROOF. With the assumption that $U = V$ almost surely we see that for all continuous local martingales N we have

$$[\int U\, dM, N]_t = \int_0^t U_s\, d[M, N]_s = \int_0^t V_s\, d[M, N]_s = [\int V\, dM, N]_t$$

for all $t \geq 0$ almost surely. By the uniqueness property of the stochastic integral we have $\int U\, dM = \int V\, dM$.

Bilinearity boils down to a couple of simple computations using bilinearity of the Lebesgue-Stieltjes integral and the quadratic covariation

$$\begin{aligned}
[\int (aV + bU)\, dM, N]_t &= \int_0^t (aV_s + bW_s)\, d[M, N]_s \\
&= a \int_0^t V_s\, d[M, N]_s + b \int_0^t W_s\, d[M, N]_s \\
&= a[\int V\, dM, N]_t + b[\int W\, dM, N]_t \\
&= [a \int V\, dM + b \int W\, dM, N]_t
\end{aligned}$$

and

$$\begin{aligned}
[\int V\, d[aM + bN], R]_t &= \int_0^t V_s\, d[aM + bN, R]_s \\
&= a \int_0^t V_s\, d[M, R]_s + b \int_0^t V_s\, d[N, R]_s \\
&= a[\int V\, dM, R] + b[\int V\, dN, R]_t \\
&= [a \int V\, dM + b \int V\, dN, R]_t
\end{aligned}$$

Now apply the uniqueness criteria for stochastic integrals.

Using the defining property of the stochastic integral twice and Lemma 14.14 once we see

$$\begin{aligned}
[\int V\, dM]_t &= \int_0^t V_s\, d[M, \int V\, dM]_s = \int_0^t V_s\, d \int_0^s V_u\, d[M](u) \\
&= \int_0^t V^2(s)\, d[M]_s
\end{aligned}$$

In the special case that M is a martingale we know that $\left(\int V\, dM\right)^2 - [\int V\, dM]$ is a martingale starting at zero and thus taking expectations we get

$$\mathbf{E}\left[\left(\int_0^t V\, dM\right)^2\right] = \mathbf{E}\left[[\int V\, dM]_t\right] = \mathbf{E}\left[\int_0^t V^2(s)\, d[M]_s\right]$$

To see the last property note that since M is a martingale and τ is $\mathcal{F}$-optional $\left(\int_0^{t\wedge\tau} V\,dM\right)^2 - [\int V\,dM]_{t\wedge\tau}$ is a martingale starting at 0. Let $t \geq 0$ be fixed and consider

$$\begin{aligned}
\mathbf{E}\left[\left(\int_0^{t\wedge\tau} V_s\,dM_s\right)^2\right] &= \mathbf{E}\left[[\int V_s\,dM_s]_{t\wedge\tau}\right] = \mathbf{E}\left[\int_0^{t\wedge\tau} V_s^2\,d[M]_s]\right] \\
&= \mathbf{E}\left[\int_0^{\infty} \mathbf{1}_{s\leq t\wedge\tau} V_s^2\,d[M]_s]\right] \\
&\leq \mathbf{E}\left[\int_0^{\infty} \mathbf{1}_{s\leq \tau} V_s^2\,d[M]_s]\right] = 0
\end{aligned}$$

from which it follows that $\int_0^{t\wedge\tau} V_s\,dM_s$ almost surely. Taking a countable intersection of almost sure events we conclude that $\int_0^{t\wedge\tau} V_s\,dM_s$ for all $t \in \mathbb{Q}\cap\mathbb{R}_+$ almost surely and by continuity of $\int V_s\,dM_s$ we get $\int_0^{t\wedge\tau} V_s\,dM_s$ for all $0 \leq t < \infty$ almost surely. $\square$

It is common for the details of defining the stochastic integral to unfold a bit differently than our presentation. The alternative presentation begins just as we have defines the stochastic integral for predictable step process integrands but then notes the property of the Ito Isometry holds for such integrands. The basic idea is to show that predictable step processes are dense in an L^2 space and the use the Ito isometry to extend the definition of the stochastic integral by a completion argument. There is a subtlety to deal with. We note that the isometry holds for every *fixed* $t \geq 0$ and thus is a family of isometries between an L^2 space of integrands (predictable step processes on $[0, t]$) and an L^2 space of random variables; it is not a single isometry between a spaces of processes. There are two ways to proceed. In the first case (Steele, Peres and Morters, others) one stays with the *one t at a time* approach and shows that step processes are dense in the progressive processes in $L^2(\Omega \times [0, t])$ and then extends the stochastic integral pointwise in $t \geq 0$. An extra step is necessary at this point to show that one may find a version of the resulting stochastic integral process that is indeed a continuous martingale. In the second case (e.g. Karatzas and Shreve), one defines a norm on the space of L^2 continuous martingales (different from the Hilbert space structure we have used) and shows that the Ito isometries can be assembled into a single isometry between the space of integrands and this space of martingales; again one extends by completion. What about Rogers and Williams; they use the Ito isometry approach but I think the details are slightly different.

The basic continuity property of the stochastic integral is

LEMMA 14.39. *Let M_n be a sequence of continuous local martingales and let $V_n \in L(M_n)$, then $\left(\int V_n\,dM_n\right)^* \xrightarrow{P} 0$ if and only if $\int_0^\infty V_n^2(s)\,d[M_n](s) \xrightarrow{P} 0$.*

PROOF. Lemma 14.28 says that $\left(\int V_n\,dM_n\right)^* \xrightarrow{P} 0$ if and only if $[\int V_n\,dM_n]_\infty \xrightarrow{P} 0$ but Lemma 14.38 tells us that $[\int V_n\,dM_n]_\infty = \int_0^\infty V_n^2(s)\,d[M_n](s)$. $\square$

Before proceeding further we extend the class of integrators in what initially seems like a very ad-hoc manner. Indeed this extension follows the historical path of the development of stochastic integration which broadened the scope of definitions in exactly these ways. The reader is encouraged not to spend too much time trying to find the method in the madness as later we will prove a theorem that shows that the only continuous stochastic processes that make sense as integrators are the ones we define here.

DEFINITION 14.40. A *continuous semimartingale* X is a cadlag adapted process in $\mathbb{R}$ such that there is a continuous local martingale M and a continuous, adapted process of locally finite variation A with $A_0 = 0$ such that $X = M + A$. A cadlag adapted process $X = (X_1, \dots, X_d)$ in $\mathbb{R}^d$ is said to be a continuous semimartingale if and only each X_i is. Given a continuous semimartingale $X = M + A$ we let

$$L(X) = \{V \mid V^2 \in L([M]) \text{ and } V \in L(A)\}$$

that is to say $L(X) = L(M) \cap L(A)$ and for any $V \in L(X)$ we define $\int V\,dX = \int V\,dM + \int V_s\,dA_s$.

Note that the decomposition $X = M + A$ is almost surely unique as if $M + A = \tilde{M} + \tilde{A}$ then $M - \tilde{M} = \tilde{A} - A$ is a continuous local martingale of locally finite variation and is therefore 0 almost surely by Lemma 14.9. As such, we refer to this as the *canonical decomposition*.

We want to develop the primary properties of the stochastic integral with a continuous semimartingale integrator. Note that by the definition $\int V\,dX = \int V\,dM + \int V\,dA$ we can see that a stochastic integral with respect to a continuous semimartingale integrator is itself a continuous semimartingale. Thus we can consider a stochastic integral as an integrator and the first result is a generalization of the "chain rule" proven in Lemma 14.14.

LEMMA 14.41. *Let X be a continuous semimartingale and let $V \in L(X)$ then $U \in L(\int V\,dX)$ if and only if $UV \in L(X)$ and $\int U\,d\int V\,dX = \int UV\,dX$ a.s.*

PROOF. For X an adapted process of locally finite variation, this is proven in Lemma 14.14. Now suppose that $X = M$ is a continuous local martingale. In this case from the proof of Lemma 14.39 and Lemma 14.14 we have

$$\int_0^t U_s^2\,d[\int V\,dM]_s = \int_0^t U_s^2 d \int_0^s V_u^2 d[M]_u = \int_0^t U_s^2 V_s^2\,d[M]_s$$

which shows us that $U \in L(\int V\,dM)$ if and only if $UV \in L(M)$. Moreover, for any continuous local martingale N, we have

$$\begin{aligned}
[\int U\,d\int V\,dM, N]_t &= \int_0^t U_s\,d[\int V\,dM, N]_s = \int_0^t U_s\,d\int_0^s V_u\,d[M,N]_u \\
&= \int_0^t U_s V_s\,d[M,N]_s = [\int UV\,dM, N]_t
\end{aligned}$$

almost surely. Thus by the defining property of stochastic integrals with a continuous local martingale integrator we know that $\int U\,d\int V\,dM = \int UV\,dM$.

Lastly let X be a continuous semimartingale and let $X = M + A$ be the canonical decompostion of X. Since the canonical decomposition of $\int V\,dX$ is $\int V\,dM + \int V_s\,dA_s$ we have

$$L(\int V\,dX) = L(\int V\,dM) \cap L(\int V_s\,dA_s)$$

hence combining results for Stieltjes integrals and continuous local martingale we have $U \in L(\int V\,dX)$ if and only if $UV \in L(M)$ and $UV \in L(A)$ (i.e. $UV \in L(X)$).

Furthermore,

$$\begin{aligned}\int_0^t U\,d\int V\,dX &= \int_0^t U\,d\int V\,dM + \int_0^t U\,d\int V_s\,dA_s \\ &= \int_0^t UV\,dM + \int_0^t U_s V_s\,dA_s = \int_0^t UV\,dX\end{aligned}$$

and the result is proven. □

The other useful result is the behavior of stochastic integrals under stopping (a generalizaiton of Lemma 14.15).

LEMMA 14.42. *Let X be a continuous semimartingale, $V \in L(X)$ and τ an optional time then*

$$\left(\int V\,dX\right)^\tau = \int V\,dX^\tau = \int \mathbf{1}_{[0,\tau]} V\,dX$$

PROOF. The result is proven for Stieltjes integrals in Lemma 14.15, so consider next the case in which $X = M$ is a continuous local martingale. Suppose N is another continuous local martingale and compute

$$\begin{aligned}[\left(\int V\,dM\right)^\tau, N]_t &= [\int V\,dM, N]_t^\tau = \int_0^{t\wedge\tau} V_s\,d[M,N]_s = \int_0^t V_s\,d[M,N]_s^\tau \\ &= \int_0^t V_s\,d[M^\tau, N]_s = [\int V\,M^\tau, N]_t\end{aligned}$$

and similarly

$$[\left(\int V\,dM\right)^\tau, N]_t = [\int V\,dM, N]_t^\tau = \int_0^t \mathbf{1}_{[0,\tau]} V_s\,d[M,N]_s = [\int \mathbf{1}_{[0,\tau]} V\,dM, N]_t$$

and we appeal to the defining property of stochastic integrals with a continuous local martingale integrator.

For a general continuous semimartingale X, let $X = M + A$ be the canonical decomposition and then the fact that M^τ is a continuous local martingale and A^τ has locally finite variation to conclude that the canonical decomposition of X^τ is $M^\tau + A^\tau$ and use the results for the continuous local martingale case and the Stieltjes integral case to see

$$\left(\int V\,dX\right)^\tau = \left(\int V\,dM\right)^\tau + \left(\int V_s\,dA_s\right)^\tau = \int V\,dM^\tau + \int V_s\,dA_s^\tau = \int V\,dX^\tau$$

The second equality is equally trivial. □

The following Lemma will be a useful for exchanging limits and stochastic integrals and represents the fundamental continuity property of stochastic integrals.

LEMMA 14.43. *Let X be a continuous semimartingale and let $U, V, V_1, V_2, \ldots \in L(X)$ with $|V_n| \leq U$ and $V_n \overset{a.s.}{\to} V$ (TODO: Make precise what this means) then $\sup_{0\leq s\leq t} \left|\int_0^s V_n\,dX - \int_0^s V\,dX\right| \overset{P}{\to} 0$ for all $t \geq 0$.*

PROOF. Write $X = M + A$ so that $U^2 \in L([M])$ and $U \in L(A)$. By ordinary Dominated Convergence applied pointwise in Ω we know that almost surely

$\int_0^t V_n(u)\,dA(u) \to \int_0^t V(u)\,dA(u)$ and $\int_0^t V_n^2(u)\,d[M](u) \to \int_0^t V^2(u)\,d[M](u)$ for every $t \geq 0$. Because $|V_n| \leq U$ we have

$$\left| \int_0^t V_n(u)\,dA(u) \right| \leq \int_0^t |V_n(u)|\ d\,|A|\,(u) \leq \int_0^t U(u)\,d\,|A|\,(u)$$

and the uniform continuity of $\int_0^t U(u)\,d\,|A|\,(u)$ on every bounded interval we know that the family $\int_0^t V_n(u)\,dA(u)$ is uniformly equicontinuous on every bounded interval. Therefore the pointwise convergence $\int_0^t V_n(u)\,dA(u) \to \int_0^t V(u)\,dA(u)$ can be extended to uniform convergence on bounded intervals $\sup_{0 \leq s \leq t} \left| \int_0^s V_n(u)\,dA(u) - \int_0^s V(u)\,dA(u) \right| \xrightarrow{a.s.} 0$ and so it follows that $\sup_{0 \leq s \leq t} \left| \int_0^s V_n(u)\,dA(u) - \int_0^s V(u)\,dA(u) \right| \xrightarrow{P} 0$.

From $\int_0^t V_n^2(u)\,d[M](u) \xrightarrow{a.s.} \int_0^t V^2(u)\,d[M](u)$ we get $\int_0^\infty V_n^2(u)\,d[M^t](u) \xrightarrow{a.s.} \int_0^\infty V^2(u)\,d[M^t](u)$ (Lemma 14.15). By Lemma 14.39 the latter convergence statement implies $\left(\int V_n\,dM^t - \int V\,dM^t\right)^* \xrightarrow{P} 0$ and the Lemma follows since $\left(\int V_n\,dM^t - \int V\,dM^t\right)^* = \left(\int V_n\,dM - \int V\,dM\right)_t^*$ (TODO: Is this obvious from earlier or do we need to reference the general stopping property of stochastic integral) from Lemma 14.42. $\square$

Recall that in the proof of Theorem 14.26 we motivated the construction of the quadratic variation $[M]$ by pointing out that in the case of a bounded martingale starting at zero what we were doing was defining $[M] = M^2 - \int M\,dM$; the stochastic integral had not been defined at that point so the comment served the pedagogical purpose of motivating the formulae but wasn't mathematically justified. Now that we have defined the stochastic integral are in a position to state and prove a proper Theorem.

THEOREM 14.44 (Integration by parts). *Let X and Y be continuous semimartingales then*

$$XY = X_0 Y_0 + \int X\,dY + \int Y\,dX + [X, Y]$$

PROOF. First let us assume $X = Y$ (we will later use polarization to extend to the general case). Furthermore, let us assume that $X = M$ where $M \in \mathcal{M}^2$ is bounded and starts at zero. Recall that from the proof of Theorem 14.26, if we define for $n \geq 0$,

$$\begin{aligned}
\tau_k^n &= \inf\{t > \tau_{k-1}^n \mid \left| M_t - M_{\tau_{k-1}^n} \right| = 2^{-n}\} \text{ for } k > 0 \\
V_t^n &= \sum_{k=0}^\infty M_{\tau_k^n} \mathbf{1}_{(\tau_k^n, \tau_{k+1}^n]}(t) \\
Q_t^n &= \sum_{k=0}^\infty \left(M_{t \wedge \tau_{k+1}^n} - M_{t \wedge \tau_k^n} \right)^2
\end{aligned}$$

then we have the identity

$$M_t^2 = 2 \int_0^t V^n\,dM + Q_t^n$$

and the convergence results that $V^n \xrightarrow{a.s.} M$ and $\sup_{0 \leq t < \infty} |Q_t^n - [M]_t| \xrightarrow{P} 0$. While in the proof of Theorem 14.26 we weren't in a position to discuss the convergence of $\int_0^t V^n \, dM$ we now note that in addition we have $|V_t^n| \leq \sup_{0 \leq s \leq t} |M_s| < \infty$ so we can apply Lemma 14.43 to conclude that

$$\sup_{0 \leq s \leq t} \left| \int_0^s V^n \, dM - \int_0^t M \, dM \right| \xrightarrow{P} 0$$

for all $t \geq 0$. So we have $Q_t^n \overset{a.s.}{\to} [M]_t$ and $\int_0^s V^n \, dM \overset{a.s.}{\to} \int_0^t M \, dM$ along a common subsequence and therefore $M_t^2 = 2 \int_0^t M \, dM + [M]_t$ almost surely. For an arbitrary continuous local martingale M we take a localizing sequence τ_n such that each M^{τ_n} is bounded (Lemma 14.3) then using the result for bounded M, Lemma 14.42 and Theorem 14.26 we have for each $t \geq 0$, almost surely

$$\begin{aligned} M_t^2 &= \lim_{n \to \infty} M_{t \wedge \tau_n}^2 = \lim_{n \to \infty} 2 \int_0^t M^{\tau_n} \, dM^{\tau_n} + [M^{\tau_n}]_t \\ &= \lim_{n \to \infty} \left(2 \int_0^{t \wedge \tau_n} M \, dM + [M]_{t \wedge \tau_n} \right) = 2 \int_0^t M \, dM + [M]_t \end{aligned}$$

Note that by Tonelli's Theorem we know that for any measurable space S, any σ-finite measure μ and any positive measurable function $f : S \times S \to \mathbb{R}_+$ we have

$$\begin{aligned} \iint f(x,y) \, d\mu(x) \otimes d\mu(y) &= \int \left[\int f(x,y) \, d\mu(y) \right] \, d\mu(x) \\ &= \int \left[\int f(y,x) \, d\mu(x) \right] \, d\mu(y) = \iint f(y,x) \, d\mu(x) \otimes d\mu(y) \end{aligned}$$

so in particular the product measure is invariant under reflection along the diagonal. Using this fact, for $X = A$ with A of locally finite variation and $A_0 = 0$, we have by definition $[A] = 0$ and

$$A_t^2 = \int_0^t \int_0^t dA(u) \otimes dA(v) = 2 \int_0^t \left[\int_0^u dA(v) \right] \, dA(u) \; = 2 \int_0^t A(u) \, dA(u)$$

so the result holds for Stieltjes integrals.

Now assume that $X = M + A$ is a continuous semimartingale with $X_0 = 0$. Using the results for the continuous local martingale case and the Stieltjes integral case we have

$$\begin{aligned} X^2 = M^2 + A^2 + 2MA &= 2 \int M \, dM + 2 \int A_s \, dA_s + [M] + 2MA \\ &= 2 \int X \, dX - 2 \int A \, dM - 2 \int M_s \, dA_s + [X] + 2MA \end{aligned}$$

so the result will follow if we can show that $MA = \int A \, dM + \int M_s \, dA_s$ almost surely. For this we can proceed by defining approximations. Fix a $t \geq 0$ and for each $n > 0$ define processes $A_s^n = A_{(k-1)t/n}$ and $M_s^n = M_{tk/n}$ for $s \in (t(k-1)/n, tk/n]$. Note A^n is a predictable step process by construction and that

$$\begin{aligned} &\int_0^t A^n \, dM + \int_0^t M_s^n \, dA_s \\ &= \sum_{k=1}^n A_{t(k-1)/n} \left(M_{tk/n} - M_{t(k-1)/n} \right) + \sum_{k=1}^n M_{kt/n} \left(A_{tk/n} - A_{t(k-1)/n} \right) \\ &= A_t M_t \end{aligned}$$

for every $n > 0$. We have $A^n \stackrel{a.s.}{\to} A$ by continuity of A and therefore $\sup_{0 \leq s \leq t} \left| \int_0^s A^n \, dM - \int A \, dM \right| \stackrel{P}{\to} 0$ by Lemma 14.39 (TODO: we need domination!) and $M^n \stackrel{a.s.}{\to} M$ and therefore $\int_0^t M^n_s \, dA_s \to \int_0^t M_s \, dA_s$ by Dominated Convergence applied pointwise (TODO: We need domination!).

Now we remove the assumption $X_0 = 0$. Applying the result proven to $X - X_0$, we have

$$X^2 = (X - X_0)^2 + 2X_0 X - X_0^2 = 2 \int (X - X_0) \, d(X - X_0) + [X - X_0] + 2X_0 X - X_0^2$$
$$= 2 \int X \, dX - 2X_0(X - X_0) + [X] + 2X_0 X - X_0^2 = X_0^2 + 2 \int X \, dX + [X]$$

Lastly, we perform the polarization to extend to general X and Y, using bilinearity of the stochastic integral and bilinearity and symmetry of the quadratic covariation,

$$\begin{aligned} XY &= \frac{1}{4} \left((X+Y)^2 - (X-Y)^2 \right) \\ &= \frac{1}{4} \big((X_0 + Y_0)^2 + 2 \int (X+Y) \, d(X+Y) + [X+Y] \\ &- (X_0 - Y_0)^2 - 2 \int (X-Y) d(X-Y) - [X-Y] \big) \\ &= X_0 Y_0 + \int X \, dY + \int Y \, dX + [X, Y] \end{aligned}$$

□

TODO: Figure out where to put this...

It is also useful to have a short hand to discuss stochastic integrals with a vector value integrator and a matrix valued integrand.

Definition 14.45. Let $X^{(i)}$ be an semimartingale for each $1 \leq i \leq n$ and let V^{ij} for $1 \leq i \leq m$ and $1 \leq j \leq n$ such that $V^{ij} \in L(X^j)$, then we define $\int V_s \, dX_s$ to be the semimartingale in $\mathbb{R}^m$ such that

$$\left(\int_0^t V_s \, dX_s \right)^{(i)} = \sum_{j=1}^n \int_0^t V^{ij}_s \, dX^{(j)})_s \text{ for } 1 \leq i \leq m$$

We have mentioned that our introduction of the concept of semimartingales was not well motivated. Though there is a much deeper justification for the relevance of the concept to be provided later, note that the integration by parts formula gives us an inkling that the concept is robust. Even if we started with just local martingales, multiplying them together would not result in a local martingale but only a semimartingale. The integration by parts formula shows us that the space of semimartingales forms an algebra. Even more is true however. The following Theorem shows that the class of continuous semimartingales is closed under composition sufficiently smooth functions and provides a means of computing many stochastic integrals. It is probably the most important theorem in stochastic calculus.

Theorem 14.46 (Ito's Lemma). *Let X be a continuous semimartingale and let $f \in C^2(\mathbb{R}^d)$ then almost surely*

$$f(X) = f(X_0) + \int f'(X) \, dX + \frac{1}{2} \int f''(X)(s) \, d[X](s)$$

PROOF. Let $\mathcal{C}$ be set of all functions for which the result holds. First we show that $\mathcal{C}$ contains all polynomials and then extend to smooth functions via an approximation argument. It is trivial that it is true for $f = c$ a constant and for $f(x) = x$ the result is simply the fact that $\int_0^t dX = X - X_0$. To see that $\mathcal{C}$ contains all polynomials, if suffices to show that $\mathcal{C}$ is an algebra. Suppose that $f, g \in \mathcal{C}$, using integration by parts Theorem 14.44, the Chain Rule Lemma 14.41 and the defining property of stochastic integrals, we get almost surely

$$\begin{aligned}
&f(X)g(X) - f(X_0)g(X_0) = \int f(X)\, dg(X) + \int g(X)\, df(X) + [f(X), g(X)] \\
&= \int f(X)\, d \int g'(X)\, dX + \frac{1}{2} \int f(X)\, d \int g''(X)(s)\, d[X](s) \\
&+ \int g(X)\, d \int f'(X)\, dX + \frac{1}{2} \int g(X)\, d \int f''(X)(s)\, d[X](s) \\
&+ [\int f'(X)\, dX + \frac{1}{2} \int f''(X)(s)\, d[X](s), \int g'(X)\, dX + \frac{1}{2} \int g''(X)(s)\, d[X](s)] \\
&= \int f(X)g'(X)\, dX + \frac{1}{2} \int f(X)g''(X)(s)\, d[X](s) + \int g(X)f'(X)\, dX \\
&+ \frac{1}{2} \int g(X)f''(X)(s)\, d[X](s) + [\int f'(X)\, dX, \int g'(X)\, dX] \\
&= \int (fg)'(X)\, dX + \frac{1}{2} \int f(X)g''(X)(s)\, d[X](s) + \frac{1}{2} \int g(X)f''(X)(s)\, d[X](s) \\
&+ \int f'(X)g'(X)(s)\, d[X](s) \\
&= \int (fg)'(X)\, dX + \frac{1}{2} \int (fg)''(X)(s)\, d[X](s)
\end{aligned}$$

Now suppose that we have $f \in C^2(\mathbb{R})$. Let $t \geq 0$ be fixed and by the Weierstrass Approximation Theorem (Corollary 1.46) (TODO: We actually need approximation in $C((-\infty, \infty); \mathbb{R})$; i.e. uniform approximation on compact sets) find a polynomials $q_n(x)$ such that q_n uniformly approximates $f''(x)$ on every interval $[-c, c]$. Taking two antiderivatives of each $q_n(x)$ we get polynomials $p_n(x)$ such that

$$\lim_{n \to \infty} \sup_{-c \leq x \leq c} |f(x) - p_n(x)| \vee |f'(x) - p_n'(x)| \vee |f''(x) - p_n''(x)| = 0$$

for every $t \geq 0$. In particular, $p_n(X_t(\omega)) \to f(X_t(\omega))$ for every $t \geq 0$ and $\omega \in \Omega$. TODO: Finish $\square$

It can be notationally convenient to have a complex variables version of Ito's Lemma. We say that Z is a complex semimartingale if $Z = X + iY$ where each of X and Y is a real semimartingale.

COROLLARY 14.47. *Let Z be a complex semimartingale and let f be an entire function then*

$$f(Z_t) = f(Z_0) + \int_0^t f'(Z)\, dZ + \frac{1}{2} \int_0^t f''(Z)\, d[Z]$$

PROOF. Write $Z = X + iY$ where X and Y are each real valued semimartingales. If we write $f = g + ih$ then we know that g and h are analytic but in particular are in $C^2(\mathbb{R}^2)$ so we may apply Ito's Lemma to them. Unfolding our notation and using the Cauchy-Riemann equations $\frac{\partial g}{\partial x} = \frac{\partial h}{\partial y}$ and $\frac{\partial g}{\partial y} = -\frac{\partial h}{\partial x}$ we get

$$\begin{aligned}
&\int_0^t f'(Z)\,dZ = \int_0^t (\frac{\partial g}{\partial x}(X,Y) + i\frac{\partial h}{\partial x}(X,Y))\,d(X+iY) \\
&= \int_0^t \frac{\partial g}{\partial x}(X,Y)\,dX - \int_0^t \frac{\partial h}{\partial x}(X,Y)\,dY + i\int_0^t \frac{\partial h}{\partial x}(X,Y)\,dX + i\int_0^t \frac{\partial g}{\partial x}(X,Y)\,dY \\
&= \int_0^t \frac{\partial g}{\partial x}(X,Y)\,dX + \int_0^t \frac{\partial g}{\partial y}(X,Y)\,dY + i\int_0^t \frac{\partial h}{\partial x}(X,Y)\,dX + i\int_0^t \frac{\partial h}{\partial y}(X,Y)\,dY
\end{aligned}$$

Similarly recall that two application of Cauchy-Riemann implies $\frac{\partial^2 g}{\partial x^2} = \frac{\partial^2 h}{\partial x \partial y} = -\frac{\partial^2 g}{\partial x^2}$ and similarly with h to get

$$\begin{aligned}
&\int_0^t f''(Z)\,d[Z] = \int_0^t (\frac{\partial^2 g}{\partial x^2}(X,Y) + i\frac{\partial^2 h}{\partial x^2}(X,Y))\,d[X+iY] \\
&= \int_0^t \frac{\partial^2 g}{\partial x^2}(X,Y)\,d[X] - 2\int_0^t \frac{\partial^2 h}{\partial x^2}(X,Y)\,d[X,Y] - \int_0^t \frac{\partial^2 g}{\partial x^2}(X,Y)\,d[Y] \\
&+ i\int_0^t \frac{\partial^2 h}{\partial x^2}(X,Y)\,d[X] + 2i\int_0^t \frac{\partial^2 g}{\partial x^2}(X,Y)\,d[X,Y] - i\int_0^t \frac{\partial^2 h}{\partial x^2}(X,Y)\,d[Y] \\
&= \int_0^t \frac{\partial^2 g}{\partial x^2}(X,Y)\,d[X] + 2\int_0^t \frac{\partial^2 g}{\partial x \partial y}(X,Y)\,d[X,Y] + \int_0^t \frac{\partial^2 g}{\partial y^2}(X,Y)\,d[Y] \\
&+ i\int_0^t \frac{\partial^2 h}{\partial x^2}(X,Y)\,d[X] + 2i\int_0^t \frac{\partial^2 h}{\partial x \partial y}(X,Y)\,d[X,Y] + i\int_0^t \frac{\partial^2 h}{\partial y^2}(X,Y)\,d[Y]
\end{aligned}$$

Applying Ito's Lemma to each of g and h separately and using the above formulae we conclude that

$$f(Z) = g(X,Y) + ih(X,Y) = f(Z_0) + \int_0^t f'(Z)\,dZ + \frac{1}{2}\int_0^t f''(Z)\,d[Z]$$

□

The following lemma provides an intuitively appealing interpretation of the quadratic covariation that ties it together with the traditional notion of variation in measure theory. Note that the convergence of the approximation is in probability and not almost sure convergence. By making more assumptions about the underlying partitions one can prove an almost sure approximation (or simply pass to an appropriate subsequence).

LEMMA 14.48. *Let X and Y be continuous semimartingales, let $t \geq 0$ be fixed and suppose that we have a sequence of partitions $0 = t_{n,0} < t_{n,1} < \cdots < t_{n,k_n} = t$ such that $\lim_{n\to\infty} \max_{1\leq k\leq k_n}(t_{n,k} - t_{n,k-1}) = 0$, then*

$$\sum_{k=1}^{k_n} (X_{n,k} - X_{n,k-1})(Y_{n,k} - Y_{n,k-1}) \xrightarrow{P} [X,Y]_t$$

PROOF. Using $[X,Y] = [X - X_0, Y - Y_0]$ it is immediate that we may assume $X_0 = Y_0 = 0$. Given the partition $0 = t_{n,0} < t_{n,1} < \cdots < t_{n,k_n} = t$

we define predictable step processes $X^n_s = \sum_{k=1}^{k_n} X_{t_{k-1}} \mathbf{1}_{(t_{k-1},t_k]}(s)$ and $Y^n_s = \sum_{k=1}^{k_n} Y_{t_{k-1}} \mathbf{1}_{(t_{k-1},t_k]}(s)$. By a little algebra using the fact that the integrals $\int X^n \, dY$ and $\int Y^n \, dX$ are given by Riemann sums we see

$$\begin{aligned}
&\sum_{k=1}^{k_n} (X_{n,k} - X_{n,k-1})(Y_{n,k} - Y_{n,k-1}) \\
&= \sum_{k=1}^{k_n} X_{n,k}(Y_{n,k} - Y_{n,k-1}) - \int_0^t X^n \, dY \\
&= \sum_{k=1}^{k_n} (X_{n,k}Y_{n,k} - X_{n,k-1}Y_{n,k-1}) - \int_0^t X^n \, dY - \int_0^t Y^n \, dX \\
&= X_t Y_t - \int_0^t X^n \, dY - \int_0^t Y^n \, dX
\end{aligned}$$

By continuity of X and Y we see that $X^n \overset{a.s.}{\to} X$ and $X^n_t \leq X^*_t < \infty$. Since X is continuous, X^* is also continuous hence $X^* \in L(Y)$, therefore we may apply Lemma 14.43 to conclude that $\int_0^t X^n \, dY \overset{P}{\to} \int_0^t X \, dY$. In exactly the same way we see that $\int_0^t Y^n \, dX \overset{P}{\to} \int_0^t Y \, dX$. Now we can apply integration by parts Lemma 14.44 to conclude that

$$\sum_{k=1}^{k_n} (X_{n,k} - X_{n,k-1})(Y_{n,k} - Y_{n,k-1}) \overset{P}{\to} X_t Y_t - \int_0^t X \, dY - \int_0^t Y \, dX = [X, Y]_t$$

□

5. Approximation By Step Processes

We defined the stochastic integral in an elegant but somewhat abstract way as the representative of a linear functional on a Hilbert space. The uniqueness property of the integral showed us that this definition was consistent with intuitively clear defintion of the stochastic integral for step process integrands as Riemann sums. The uniqueness property of the stochastic integral has shown itself to be a very useful technical tool but is lacking somewhat in intuitive appeal. We repair this deficiency by showing that the continuity properties of the stochastic integral also characterize the extension from step process integrands. To see this requires that we understand the approximation by step processes in the spaces $L(M)$. We note that these approximation results also lead to an alternative path to defining the stochastic integral in the first place.

LEMMA 14.49. *Let X be a continuous semimartingale with canonical decomposition $X = M + A$ and let $V \in L(X)$. Then there exists processes $V_1, V_2, \ldots \in \mathcal{E}$ such that almost surely* $\lim_{n\to\infty} \int_0^t (V_n - V)^2(s) \, d[M](s) = 0$ *and* $\lim_{n\to\infty} \sup_{0 \leq s \leq t} \left| \int_0^s (V_n - V)(u) \, dA(u) \right| = 0$ *for all $t \geq 0$.*

PROOF. TODO: A bunch of stuff

Now suppose that A is a strictly increasing, continuous and adapted process with $A_0 = 0$. If one thinks for a moment about the case in which $A_t = t$ then it is more or less clear how to approximate any integrable function f by a continuous one: just define $f^h(t) = \frac{1}{h}\int_{t-h}^t f(s)\,ds$ for $h > 0$ and note that by the Fundamental Theorem of Calculus for almost all t we have $\lim_{h\to 0^+} f^h(t) = f(t)$. If we treat a general Stieltjes integral then we just have to use the fact that every Lebesgue-Stieltjes measure is of the form $\lambda \circ G^{-1}$. Specifically, from the proof of Lemma 2.112 recall that if F is nondecreasing and right continuous then the Lebesgue-Stieltjes measure associated with F is given by $\lambda \circ G^{-1}$ where $G(t) = \sup\{s \mid F(s) < t\}$. Let us apply this to our process A pointwise by defining the process, $T_t = \sup\{s \geq 0 \mid A_s < t\}$ for $t \geq 0$. TODO: Show T is a process. Because we have assumed that A is strictly increasing, T is strictly increasing and is an actual inverse satisfying $T(A(t)) = A(T(t)) = t$. We can now define the approximation for $h > 0$ and $t > 0$,

$$V_t^h = \frac{1}{h}\int_{T((A_t-h)\vee 0)}^t V(s)\,dA(s) = \frac{1}{h}\int_{(A_t-h)\vee 0}^{A_t} V(T(s))\,ds$$

where we have used the change of variables Lemma 2.55 and the fact that $T((A_t - h)\vee 0) \leq T(s) \leq t$ if and only if $(A_t - h)\vee 0 \leq s \leq A(t)$. TODO: What about $t = 0$? Having expressed the definition of V_t^h in terms of an ordinary Lebesgue integral, we can apply the Fundamental Theorem of Calculus to see that

$$\lim_{h\to 0} V^h(T(t)) = \lim_{h\to 0}\frac{1}{h}\int_{(t-h)\vee 0}^t V(T(s))\,ds = V(T(t))$$

for almost all $0 \leq t \leq A_1$. Now we can apply the Dominated Convergence Theorem to conclude

$$\lim_{h\to 0}\int_0^1 \left|V_s^h - V_s\right|\,dA_s = \lim_{h\to 0}\int_0^{A_1}\left|V^h(T(s)) - V(T(s))\right|\,ds = 0$$

$\square$

LEMMA 14.50. *Let V be a bounded $\mathcal{F}$-adapted process then there exist $V^n \in \mathcal{E}$ such that*

$$\sup_{0\leq T<\infty}\lim_{n\to\infty}\boldsymbol{E}\left[\int_0^T \left|V_s - V_s^n\right|^2\,ds\right] = 0$$

PROOF. First fix a $T \geq 0$ and we will approximate on the interval $[0, T]$. It is also notationally convenient to set $V_t = 0$ for all $t < 0$ in what follows. Set up the following family of approximations; for every $s \geq 0$ and $n \in \mathbb{N}$ define

$$V_t^{(n,s)}(\omega) = \sum_{j=0}^{\lceil 2^n T\rceil} V_{j/2^n+s}(\omega)\mathbf{1}_{(j/2^n+s,(j+1)/2^n+s]}(t)\mathbf{1}_{[0,T]}(t)$$

Note that $V^{(n,s)} \in \mathcal{E}$ and moreover it is jointly measurable in (s, t, ω). Also note that $V_t^{(n,s)} = V_t^{(n,s+1/2^n)}$ for all $s \geq 0$ and all $t \geq 0$.

CLAIM 14.50.1. Let $f \in L^2([0,T])$ then $\lim_{h\downarrow 0}\int_0^T (f(s) - f((s-h)\vee 0))^2\,ds = 0$.

By Lemma 8.6 we can find bounded continuous f_n such that $f_n \xrightarrow{L^2} f$. By the triangle inequality, continuity of f_n, Dominated Convergence and the translation invariance of Lebesgue measure we get for every n

$$\begin{aligned}
&\lim_{h\downarrow 0}\left(\int_0^T (f(s)-f((s-h)\vee 0))^2\,ds\right)^{1/2} \\
&\leq \left(\int_0^T (f(s)-f_n(s))^2\,ds\right)^{1/2} + \\
&\lim_{h\downarrow 0}\left(\int_0^T (f_n(s)-f_n((s-h)\vee 0))^2\,ds\right)^{1/2} + \\
&\lim_{h\downarrow 0}\left(\int_0^T (f_n((s-h)\vee 0)-f((s-h)\vee 0))^2\,ds\right)^{1/2} \\
&= \left(\int_0^T (f(s)-f_n(s))^2\,ds\right)^{1/2} + \lim_{h\downarrow 0}\left(\int_0^{T-h} (f_n(s)-f(s))^2\,ds\right)^{1/2} \\
&\leq 2\left\|f-f_n\right\|_2
\end{aligned}$$

so we now take the limit as $n\to\infty$.

It is a simple matter to extend this result to a bounded adapted process V. In this case we know that $\int_0^T (V_s - V_{(s-h)\vee 0})^2\,ds$ is bounded and therefore we conclude from Dominated Convergence and the result on $L^2([0,T])$ that

$$\lim_{h\downarrow 0}\mathbf{E}\left[\int_0^T (V_s - V_{(s-h)\vee 0})^2\,ds\right] = \mathbf{E}\left[\lim_{h\downarrow 0}\int_0^T (V_s - V_{(s-h)\vee 0})^2\,ds\right] = 0$$

Claim 14.50.2. $\lim_{n\to\infty}\mathbf{E}\left[\int_0^T\int_0^1 (V_t^{(n,s)} - V_t)^2\,ds\,dt\right] = 0$.

First off, from $V_t^{(n,s)} = V_t^{(n,s+1/2^n)}$, the definition of $V_t^{(n,s)}$ and a change of integration variable we write

$$\int_0^1 (V_t^{(n,s)} - V_t)^2\,ds = 2^n\int_0^{2^{-n}} (V_t^{(n,s)} - V_t)^2\,ds = 2^n\int_{t-2^{-n}}^t (V_s - V_t)^2\,ds = 2^n\int_0^{2^{-n}} (V_t - V_{t-h})^2\,dh$$

Now using this fact and Tonelli's Theorem

$$\mathbf{E}\left[\int_0^T\int_0^1 (V_t^{(n,s)} - V_t)^2\,ds\,dt\right] = 2^n\int_0^{2^{-n}}\mathbf{E}\left[\int_0^T (V_t - V_{t-h})^2\,dt\right]\,dh$$

By the previous claim for any $\epsilon > 0$ we can find $N > 0$ such that $\mathbf{E}\left[\int_0^T (V_t - V_{t-h})^2\,dt\right] < \epsilon$ for all $0\leq h\leq 2^{-N}$ and therefore for all $0\leq h\leq 2^{-n}$ for any $n\geq N$. Thus for any $n\geq N$ we have $\mathbf{E}\left[\int_0^T\int_0^1 (V_t^{(n,s)} - V_t)^2\,ds\,dt\right] < \epsilon$ and the claim is shown by letting $\epsilon\to 0$.

TODO: Make sure we deal with the boundary at 0 consistently (we're not at the moment).

Viewing $\mathbf{E}\left[\int_0^n (V_t^n - V_t)^2\,dt\right]$ as a random variable on the probability space $([0,1],\mathcal{B}([0,1]),\lambda)$ and applying Tonelli's Theorem to previous claim, conclude $\mathbf{E}\left[\int_0^T (V_t^{(n,s)} - V_t)^2\,dt\right] \xrightarrow{L^1} 0$ which implies $\mathbf{E}\left[\int_0^T (V_t^{(n,s)} - V_t)^2\,dt\right] \overset{a.s.}{\to} 0$ along some subsequence $N\subset\mathbb{N}$

(Lemma 5.6 and Lemma 5.10). Pick any $s \in [0,1]$ where the subsequence converges.

To finish the proof, for each $n \in \mathbb{N}$ we apply the result for fixed $T = n$ and find an element $V^n \in \mathcal{E}$ such that $\mathbf{E}\left[\int_0^n (V^n_t - V_t)^2\,dt\right] < 1/n$. Then given $T > 0$ and any $\epsilon > 0$ it holds for any $n > \epsilon^{-1} \vee T$ that

$$\mathbf{E}\left[\int_0^T (V^n_t - V_t)^2\,dt\right] < \mathbf{E}\left[\int_0^n (V^n_t - V_t)^2\,dt\right] < 1/n < \epsilon$$

and the result is proven. □

Lemma 14.51. *Let A be a non-decreasing, continuous and $\mathcal{F}$-adapted process such that $A_0 = 0$ and $\boldsymbol{E}[A_t] < \infty$ for all $t \geq 0$. Let σ and τ be bounded $\mathcal{F}$-optional times such that $\sigma \leq \tau$ and ξ be an $\mathcal{F}_\sigma$-measurable bounded random variable. Then there exist $V^n \in \mathcal{E}$ such that*

$$\sup_{0 \leq T < \infty} \lim_{n \to \infty} \boldsymbol{E}\left[\int_0^T \left|\xi \boldsymbol{1}_{(\sigma,\tau]}(s) - V^n(s)\right|^2\,dA(s)\right] = 0$$

Proof. Take the standard discrete approximation of optional times $\tau_n = \frac{1}{2^n}\lfloor 2^n\tau + 1\rfloor$ and $\sigma_n = \frac{1}{2^n}\lfloor 2^n\sigma + 1\rfloor$ (Lemma 9.71) so that $\tau_n \downarrow \tau$ and $\sigma_n \downarrow \sigma$. Note that $s \in (\sigma_n, \tau_n]$ if and only if there exists a k such that $\tau_n \geq k/2^n$, $\sigma_n \leq (k-1)/2^n$ and $(k-1)/2^n < s \leq k/2^n$. As $\tau_n = k/2^n$ when $(k-1)/2^n \leq \tau < k/2^n$ and likewise for σ_n we see that $\tau_n \geq k/2^n$ is equivalent to $\tau_n \geq (k-1)/2^n$ and $\sigma_n \leq (k-1)/2^n$ is equivalent to $\sigma < (k-1)/2^n$. From these facts and the boundedness of τ we see that

$$\mathbf{1}_{(\sigma_n,\tau_n]}(s) = \sum_{k=1}^N \mathbf{1}_{\{\sigma < (k-1)/2^n \leq \tau\}} \mathbf{1}_{((k-1)/2^n, k/2^n]}(s)$$

for some large N. Now we define

$$V^n = \xi \mathbf{1}_{(\sigma_n,\tau_n]}(s) = \sum_{k=1}^N \xi \mathbf{1}_{\{\sigma < (k-1)/2^n \leq \tau\}} \mathbf{1}_{((k-1)/2^n, k/2^n]}(s)$$

and claim that $V^n \in \mathcal{E}$.

Lastly we note that because $\sigma < \sigma_n \leq \tau < \tau_n$ and ξ is bounded (say $|\xi| \leq K$) we get

$$\begin{aligned}
\mathbf{E}\left[\int_0^T \left|\xi \mathbf{1}_{(\sigma,\tau]}(s) - V^n(s)\right|^2\,dA(s)\right] &= \mathbf{E}\left[\xi^2 \int_0^T (\mathbf{1}_{(\sigma,\tau]}(s) - \mathbf{1}_{(\sigma_n,\tau_n]}(s))^2\,dA(s)\right] \\
&= \mathbf{E}\left[\xi^2 (A_{\tau_n} - A_\tau)\right] + \mathbf{E}\left[\xi^2 (A_{\sigma_n} - A_\sigma)\right] \\
&\leq K^2 \mathbf{E}\left[(A_{\tau_n} - A_\tau)\right] + \mathbf{E}\left[(A_{\sigma_n} - A_\sigma)\right]
\end{aligned}$$

If we let C be a bound for τ, it follows that τ_n is bounded by $C+1$ for all n and by the non-decreasingness of A we have $|A_{\tau_n} - A_\tau| \leq 2A_{C+1}$ and similarly with σ, therefore by Dominated Convergence we get $\lim_{n\to\infty} \mathbf{E}\left[\int_0^T \left|\xi \mathbf{1}_{(\sigma,\tau]}(s) - V^n(s)\right|^2\,dA(s)\right] = 0$.

TODO: If we need the sup over T then we have that argument elsewhere; check if we really use it. □

LEMMA 14.52. *Let A be a non-decreasing, continuous and $\mathcal{F}$-adapted process with $A_0 = 0$ and $\boldsymbol{E}[A_t] < \infty$ for all $t \geq 0$. Let V be an $\mathcal{F}$-progressively measurable process such that*

$$\boldsymbol{E}\left[\int_0^t V_s^2 \, dA_s\right] < \infty$$

for every $t \geq 0$, then then there exist $V^n \in \mathcal{E}$ such that

$$\sup_{0 \leq T < \infty} \lim_{m \to \infty} \boldsymbol{E}\left[\int_0^T |V(s) - V^n(s)|^2 \, dA(s)\right] = 0$$

PROOF. Pick a $T \geq 0$ fixed and assume that $V_t = 0$ for all $t > T$ and that $V_t(\omega) \leq C$ for all $t \geq 0$ and $\omega \in \Omega$. Now we want to use the fact that a Lebesgue-Stieltjes integral can be reduced to an ordinary Lebesgue integral via change of variables: this will allow us to use Lemma 14.50. To make dealing with the change of variables a bit easier, consider $A_s + s$ which is a strictly increasing function; in this case we have genuine inverse T_s that is increasing. Moreover since $A_{T_s} + T_s = s$ and $A_s \geq 0$ we have $T_s \leq s$ and from the increasingness of T_s we have $\{T_s \leq t\} = \{s \leq A_t + t\} \in \mathcal{F}_t$; so in particular, each T_s is a bounded $\mathcal{F}$-optional time. Now define the process $W_s = V_{T_s}$ and the filtration $\mathcal{G}_s = \mathcal{F}_{T_s}$ and note that by $\mathcal{F}$-progressive measurability of V and Lemma 9.90 we know that W_s is $\mathcal{G}$-adapted. Also we compute

$$\mathbf{E}\left[\int_0^\infty W_s^2 \, ds\right] = \mathbf{E}\left[\int_0^\infty \mathbf{1}_{T_s \leq T}(s) V_{T_s}^2 \, ds\right] = \mathbf{E}\left[\int_0^{A_T + T} V_{T_s}^2 \, ds\right] \leq C(\mathbf{E}[A_T] + T) < \infty$$

so that in particular $\lim_{R \to \infty} \mathbf{E}\left[\int_R^\infty W_s^2 \, ds\right] = 0$. By our boundedness assumption and Lemma 14.50 we know that we can approximate W by $\mathcal{G}$-predictable step processes with deterministic jump times. Thus if we let $\epsilon > 0$ then we can find $R > 0$ such that $\mathbf{E}\left[\int_R^\infty W_s^2 \, ds\right] < \epsilon/2$ and $W_s^\epsilon = \xi_0 \mathbf{1}_{\{0\}}(s) + \sum_{j=1}^n \xi_j \mathbf{1}_{(s_{j-1}, s_j]}(s)$ such that $\mathbf{E}\left[\int_0^R |W_s - W_s^\epsilon|^2 \, ds\right] < \epsilon/2$ and by defining $W_s^\epsilon = 0$ for $s > R$ we have

$$\mathbf{E}\left[\int_0^\infty |W_s - W_s^\epsilon|^2 \, ds\right] = \mathbf{E}\left[\int_0^R |W_s - W_s^\epsilon|^2 \, ds\right] + \mathbf{E}\left[\int_R^\infty W_s^2 \, ds\right] < \epsilon$$

Now we undo our change of variables to see what type of approximation we have of V. Let

$$\begin{aligned} V_s^\epsilon = W_{A_s + s}^\epsilon &= \xi_0 \mathbf{1}_{\{0\}}(A_s + s) + \sum_{j=1}^n \xi_j \mathbf{1}_{(s_{j-1}, s_j]}(A_s + s) \\ &= \xi_0 \mathbf{1}_{\{0\}}(s) + \sum_{j=1}^n \xi_j \mathbf{1}_{(T_{s_{j-1}}, T_{s_j}]}(s) \end{aligned}$$

we claim that V^ϵ is $\mathcal{F}$-adapted. This follows from the fact that ξ_j is $\mathcal{F}_{s_{j-1}}$-measurable and for any $u > 0$ and $j \geq 1$,

$$\{\xi_j \mathbf{1}_{(T_{s_{j-1}}, T_{s_j}]}(s) \leq u\} = \{\xi_j \leq u\} \cap \{T_{s_{j-1}} < s\} \cap \{s \leq T_{s_j}\} \in \mathcal{F}_s$$

TODO: Why is $\{s \leq T_{s_j}\} \in \mathcal{F}_s$? Moreover, by the construction of Stieltjes integral we have

$$\mathbf{E}\left[\int_0^T |V_s - V_s^\epsilon|^2 \, dA_s\right] \leq \mathbf{E}\left[\int_0^\infty |V_s - V_s^\epsilon|^2 \, d(A_s + s)\right]$$
$$= \mathbf{E}\left[\int_0^\infty |W_s - W_s^\epsilon|^2 \, ds\right] < \epsilon$$

We are not quite done as V^ϵ has random jump times. However, we can apply Lemma 14.51 to find $V^{(m,n)} \in \mathcal{E}$ such that $\lim_{m\to\infty} \mathbf{E}\left[\int_0^T \left|V_s^{1/n} - V_s^{(m.n)}\right|^2 \, dA_s\right] = 0$ and then we find a $V^{(m_n,n)}$ such that $\lim_{n\to\infty} \mathbf{E}\left[\int_0^T \left|V_s - V_s^{(m_n.n)}\right|^2 \, dA_s\right] = 0$.

Now we remove the assumption that V is bounded. For a general V_s such that $\mathbf{E}\left[\int_0^T V_s^2 \, dA_s\right] < \infty$, let $V_s^n = V_s \mathbf{1}_{|V_s|\leq n}$ where by the Dominated Convergence Theorem we know that $\mathbf{E}\left[\int_0^T |V_s - V_s^n|^2 \, dA_s\right] = 0$. Since each V_s^n is bounded we can find a sequence $V_s^{(n,m)}$ such that $\lim_{m\to\infty} \mathbf{E}\left[\int_0^T \left|V_s^n - V_s^{(n,m)}\right|^2 \, dA_s\right] = 0$ and now an array argument shows we get a subsequence $V^{(n,m_n)}$ such that $\lim_{n\to\infty} \mathbf{E}\left[\int_0^T \left|V_s^{(n,m_n)} - V_s\right|^2 \, dA_s\right] = 0$.

Lastly it remains to remove the assumption that we are dealing with a fixed $T \geq 0$. By what we have proven thus far, if V is such that $\mathbf{E}\left[\int_0^t V_s^2 \, dA_s\right] < \infty$ for all $t \geq 0$, then for each $m > 0$ we have a sequence $V^{(n,m)} \in \mathcal{E}$ such that $\lim_{n\to\infty} \mathbf{E}\left[\int_0^m \left|V_s - V_s^{(n,m)}\right|^2 \, dA_s\right] = 0$, so in particular there is n_m such that $\mathbf{E}\left[\int_0^m \left|V_s - V_s^{(n_m,m)}\right|^2 \, dA_s\right] < \frac{1}{m}$. If we let $V_s^m = V_s^{(n_m,m)}$ then for every $T > 0$,

$$\lim_{m\to\infty} \mathbf{E}\left[\int_0^T \left|V_s - V_s^{(n_m,m)}\right|^2 \, dA_s\right] \leq \lim_{m\to\infty} \mathbf{E}\left[\int_0^m \left|V_s - V_s^{(n_m,m)}\right|^2 \, dA_s\right] = 0$$

and thus $\sup_{0\leq T<\infty} \lim_{m\to\infty} \mathbf{E}\left[\int_0^T \left|V_s - V_s^{(n_m,m)}\right|^2 \, dA_s\right] = 0$ and we are finally done. ☐

6. Brownian Motion and Continuous Martingales

The theme of this section is the centrality of Brownian motion in the universe of continuous martingales. We demonstrate through several different constructions that all continuous martingales can be derived from (or transformed into) a suitable Brownian motion. We start with a slightly different problem. Suppose we are given a Brownian motion, we ask whether we can identify a class of continuous martingales that can be constructed from it.

Definition 14.53. Let $B_t = (B_t^1, \ldots, B_t^d)$ be a d-dimensional Brownian motion and let $\mathcal{F}_t$ be completion of the filtration generated by B, we say that a cadlag (local) martingale that is adapted to $\mathcal{F}$ is a *Brownian (local) martingale*.

Note that we have not assumed that a Brownian martingale is continuous but only cadlag. Our goals is to show that all Brownian martingales may be constructed as stochastic integrals of suitable progressively measurable integrands. One corollary of this fact is that Brownian martingales are in fact continuous (even though the definition only assumes that they are cadlag). We start out working with L^2 continuous martingales so that we may leverage Hilbert space structures to assist in the analysis. First a basic decomposition result.

OOPS! Here we are using the covariation of not necessarily continuous martingales which we haven't defined! Better go back to Kallenberg to understand his proof that doesn't use this idea. Note that the result goes though with additional assumption of continuity but the results are actually strong enough that continuity of Brownian martingales is part of the conclusion.

LEMMA 14.54. *Let B be a one-dimensional Brownian motion and let M be a bounded L^2 Brownian martingale then there is $V \in L(B)$ and a bounded L^2 martingale Z such that $M = \int V\,dB + Z$ and $[Z, \int U\,dB] = 0$ for all $U \in L(B)$. Moreover such a decomposition is unique up to indistinguishability.*

PROOF. We first show the uniqueness part of the claim. Suppose that we have $M = \int V\,dB + Z = \int \tilde{V}\,dB + \tilde{Z}$. It then follows by linearity of the stochastic integral that $Z - \tilde{Z} = \int(\tilde{V} - V)\,dB$ is an L^2 bounded continuous martingale and therefore $[Z - \tilde{Z}] = [Z, \int(\tilde{V} - V)\,dB] - [\tilde{Z}, \int(\tilde{V} - V)\,dB] = 0$. Therefore Z and $\tilde{Z}$ are indistinguishable and it follows that $\int V\,dB$ and $\int \tilde{V}\,dB$ are indistinguishable which implies V and $\tilde{V}$ are indistinguishable by the Ito Isometry.

Now we reduce to demonstrating the decomposition for the stopped process M^t. To that end, suppose that we have $M^t = \int V\,dB + Z$, then clearly for $s < t$ we have

$$M^s = (M^t)^s = \left(\int V\,dB + Z\right)^s = \int \mathbf{1}_{[0,s]} V\,dB + Z^s$$

so we can define V and Z by extending from each interval $[0, t]$. It is clear that $M = \int V\,dB + Z$ and moreover for every $0 \leq t < \infty$ we have $M^t = \int \mathbf{1}_{[0,t]} V\,dB + Z^t$ with $[Z^t, \int U\,dB] = 0$ for all $U \in L(B)$. From this for every $U \in L(B)$, we have $[Z, \int U\,dB]^n = [Z^n, \int U\,dB] = 0$ for every $n \in \mathbb{N}$ and therefore it follows that $[Z, \int U\,dB] = 0$.

So we now fix $t > 0$ and suppose that $M_t = M_s$ for all $s \geq t$. We consider M_t as an element of $L^2(\Omega, \mathcal{F}_t)$.

Claim: The subspace of elements of the form $\int_0^t V\,dB$ is closed.

This follows from the Ito isometry as if $\int V^n\,dB$ is a convergent sequence in $L^2(\Omega, \mathcal{F}_t)$ then by the Ito Isometry we know that V^n is Cauchy in $L^2(\Omega \times [0, t])$ hence converges to a progressive process $V \in L^2(\Omega \times [0, t])$ (note it follows from Lemma 9.87 that the limit of progressive processes is progressive). Again by the Ito Isometry, it follows that $\int V^n\,dB \xrightarrow{L^2} \int V\,dB$.

From the claim we can write $M_t = \int_0^t V\,dB + Z_t$ where $\mathbf{E}\left[Z_t \int_0^t U\,dB\right] = 0$ for all progressive $U \in L^2(\Omega \times [0, t])$. Now let Z_s be a cadlag version of the martingale $\mathbf{E}\left[Z_t \mid \mathcal{F}_s\right]$ (noting that $Z_s = Z_t$ for all $s \geq t$) and by Jensen's inequality for conditional expectations

$$\mathbf{E}\left[Z_s^2\right] \leq \mathbf{E}\left[\mathbf{E}\left[Z_t^2 \mid \mathcal{F}_s\right]\right] = \mathbf{E}\left[Z_t^2\right] < \infty$$

which shows that Z_t is L^2-bounded. ☐

LEMMA 14.55. *Let M_t be a real continuous local martingale such that $M_0 = 0$ then $Z_t = e^{iM_t + \frac{1}{2}[M]_t}$ is a complex local martingale satisfying $Z_t = 1 + i\int_0^t Z dM$.*

PROOF. Apply Ito's Lemma Corollary 14.47 to the complex semimartingale $X_t = iM_t + \frac{1}{2}[M]_t$ and the entire function $f(z) = e^z$ to see that

$$\begin{aligned} Z_t &= 1 + \int_0^t Z\, dX + \frac{1}{2}\int_0^t Z_s\, d[X]_s \\ &= 1 + i\int_0^t Z\, dM - \frac{1}{2}\int_0^t Z_s\, d[M]_s + \frac{1}{2}\int_0^t Z_s\, d[M]_s = 1 + i\int_0^t Z\, dM \end{aligned}$$

The fact that Z_t is a complex local martingale follows from the fact that it is a stochastic integral. ☐

LEMMA 14.56. *Let $B_t = (B_t^1, \ldots, B_t^d)$ be a d-dimensional Brownian motion and let ξ be a B-measurable random variable with $\boldsymbol{E}[\xi] = 0$ and $\boldsymbol{E}\left[\xi^2\right] < \infty$. There exists $\boldsymbol{P}\times\lambda$ almost everywhere unique processes $V^1, \ldots, V^d$ such that $\boldsymbol{E}\left[\int_0^\infty (V^j(s))^2 ds\right] < \infty$ for each $j = 1, \ldots, d$ and $\xi = \sum_{j=1}^d \int_0^\infty V^j\, dB^j$ almost surely.*

PROOF. Let H be the subspace of $L^2(\Omega, \mathcal{A}, \mathbf{P})$ such that $\mathbf{E}[\xi] = 0$. Note that if $\xi_1, \xi_2, \ldots$ is a sequence in H and $\xi \in L^2$ such that $\xi_n \xrightarrow{L^2} \xi$ then by Jensen's Inequality, $\mathbf{E}[\xi]^2 = \lim_{j\to\infty} \mathbf{E}[\xi - \xi_j]^2 \leq \lim_{j\to\infty} \mathbf{E}\left[(\xi - \xi_j)^2\right] = 0$ and therefore H is closed hence a Hilbert space. Now let K be the subspace of elements of the form $\sum_{j=1}^d \int_0^\infty V^j\, dB^j$ where V^j are progressive processes with $\mathbf{E}\left[\int_0^\infty (V^j(s))^2\, ds\right] < \infty$.

Claim: $K \subset H$ is a closed subspace

First focus on a single B^j. Note that by the Ito Isometry we have

$$\mathbf{E}\left[\left(\int_0^t V^j\, dB^j\right)^2\right] = \mathbf{E}\left[\int_0^t (V^j(s))^2\, ds\right] \leq \mathbf{E}\left[\int_0^\infty (V^j(s))^2\, ds\right] < \infty$$

and therefore each $\int V^j\, dB^j$ is L^2-bounded and $\int_0^\infty V^j\, dB^j$ is defined and in L^2 (hence in H). Moreover we have the limit of the Ito Isometry $\mathbf{E}\left[\left(\int_0^\infty V^j\, dB^j\right)^2\right] = \mathbf{E}\left[\int_0^\infty (V^j(s))^2\, ds\right]$. Thus if $\int V^{n,j}\, dB^j$ converges in L^2 then it is Cauchy which implies $V^{n,j}$ is Cauchy and thus $V^{n,j}$ converges to some V^j by completeness of L^2. Another application of the Ito Isometry shows that $\int_0^\infty V^{n,j}\, dB^j \xrightarrow{L^2} \int_0^\infty V^j\, dB^j$ hence the space of $\int_0^\infty V^j\, dB^j$ is a closed subspace of H for each $j = 1, \ldots, d$. Lastly note that for $i \neq j$ we have $\mathbf{E}\left[\int_0^\infty V^i\, dB^i \int_0^\infty V^j\, dB^j\right] = \mathbf{E}\left[[\int V^i\, dB^i \int V^j\, dB^j]_\infty\right] = \mathbf{E}\left[\int_0^\infty V_s^i V_s^j\, d[B^i, B^j]_s\right] = 0$ since $[B^i, B^j] = 0$. Thus the space of $\sum_{j=1}^d \int_0^\infty V^j\, dB^j$ is the orthogonal sum of closed subspaces and is therefore closed.

The uniqueness claim of the Lemma also follows from the argument above since we have shown that K is an orthogonal sum of subspaces each of which is isometric to $L^2(\Omega \times \mathbb{R}_+, \mathcal{A} \otimes \mathcal{B}(\mathbb{R}_+), P \times \lambda)$.

Now for the existence portion of the argument, for any $\xi \in H$ we can write $\xi = \eta + \sum_{j=1}^d \int_0^\infty V^j\, dB^j$ with η orthogonal to K. It suffices to show that if η is B-measurable then $\eta = 0$; so let $\eta \in H \ominus K$. Suppose that $u^1, \ldots, u^d$ are deterministic functions in $L^2(\mathbb{R})$, $M_t = \sum_{j=1}^d \int_0^t u^j\, dB^j$ and $Z_t = e^{iM_t + \frac{1}{2}[M]_t}$. By Lemma 14.55

and Lemma 14.41 we know that

$$Z_t - 1 = \int_0^t Z\, dM = \sum_{j=1}^d \int_0^t Z\, d\int u^j\, dB^j = \sum_{j=1}^d \int_0^t Z u^j\, dB^j$$

Moreover we have $[M]_t = \sum_{j=1}^d \int_0^t (u^j(s))^2 ds$ is deterministic and therefore

$$\mathbf{E}\left[|Z_t|^2\right] = \mathbf{E}\left[e^{[M]_t}\right] = e^{\sum_{j=1}^d \int_0^t (u^j(s))^2 ds} \leq e^{\sum_{j=1}^d \left\|u^j\right\|_2} < \infty$$

so Z_t is L^2 bounded. Therefore $Z_\infty - 1 \in K$ and from $\eta \in H \ominus K$ we have

$$0 = \mathbf{E}\left[\eta(Z_\infty - 1)\right] = \mathbf{E}\left[\eta Z_\infty\right] = e^{\frac{1}{2}\sum_{j=1}^d \int_0^\infty (u^j(s))^2 ds} \mathbf{E}\left[\eta e^{i\sum_{j=1}^d \int_0^\infty u^j\, dB^j}\right]$$

and therefore $\mathbf{E}\left[\eta e^{i\sum_{j=1}^d \int_0^\infty u^j\, dB^j}\right] = 0$.

This expression looks quite a bit like a charactersitic function; we proceed to pick some strategic u^j so that it really becomes an honest one. Fix an arbitrary $n \in \mathbb{N}$ and let $(t_1, \dots, t_n) \in \mathbb{R}_+^n$ and $\theta^1, \dots, \theta^n \in \mathbb{R}^d$ be given. Define the step functions $u^j = \sum_{k=1}^n \theta_j^k \mathbf{1}_{[0,t_k]}$ for $j = 1, \dots, d$. Then $\sum_{j=1}^d \int_0^\infty u^j\, dB^j = \sum_{j=1}^d \sum_{k=1}^n \theta_j^k B_{t_k}^j = \sum_{k=1}^n \langle \theta^k, B_{t_k}\rangle$ and therefore we get

$$\mathbf{E}\left[\eta e^{\sum_{k=1}^n \langle \theta^k, B_{t_k}\rangle}\right] = 0$$

Writing $\eta = \eta_+ - \eta_-$ with $\eta_\pm \geq 0$ we note that by Lemma 2.57 and Lemma 2.55

$$\begin{aligned}\mathbf{E}\left[\eta_\pm e^{\sum_{k=1}^n \langle \theta^k, B_{t_k}\rangle}\right] &= \int e^{\sum_{k=1}^n \langle \theta^k, B_{t_k}\rangle}\, d(\eta_\pm \cdot \mathbf{P}) \\ &= \int e^{\sum_{k=1}^n \langle \theta^k, x_k\rangle}\, d\eta_\pm \cdot \mathbf{P} \circ (B_{t_1}, \dots, B_{t_n})^{-1}(x_1, \dots, x_n)\end{aligned}$$

is the Fourier transform of the measure $\eta_\pm \cdot \mathbf{P} \circ (B_{t_1}, \dots, B_{t_n})^{-1}$ on $\mathbb{R}^{nd}$. By uniqueness of the Fourier transform or measures we conclude that $\mathbf{E}\left[\eta; (B_{t_1}, \dots B_{t_n}) \in A\right] = 0$ for all $A \in \mathcal{B}(\mathbb{R}^{nd})$.

Since it is trivial that $\{(B_{t_1}, \dots B_{t_n}) \in A\} \cap \{(B_{s_1}, \dots B_{s_m}) \in C\} = \{(B_{t_1}, \dots B_{t_n}, B_{s_1}, \dots B_{s_m}) \in A \times C\}$ we see that sets of the form $\{(B_{t_1}, \dots B_{t_n}) \in A$ are a π-system and we know they generate $\vee_{t\geq 0}\sigma(B_t)$. Moreover if we let $\mathcal{C} = \{A \in \mathcal{A} \mid \mathbf{E}\left[\eta; A\right] = 0\}$ we have $A \subset C$ then $\mathbf{E}\left[\eta; C \setminus A\right] = \mathbf{E}\left[\eta; C\right] - \mathbf{E}\left[\eta; A\right] = 0$ and if $A_1 \subset A_2 \subset \cdots$ then by Dominated Convergence, $\mathbf{E}\left[\eta; \cup_{j=1}^\infty A_j\right] = \lim_{j\to\infty} \mathbf{E}\left[\eta; A_j\right] = 0$ which shows $\mathcal{C}$ is a λ-system. Now by the π-λ Theorem 2.27 we see that $\mathbf{E}\left[\eta; A\right] = 0$ for all $A \in \vee_{t\geq 0}\sigma(B_t)$ which shows that $\mathbf{E}\left[\eta \mid \vee_{t\geq 0}\sigma(B_t)\right] = 0$. If we also assume that η is B-measurable then we know that $\eta = \mathbf{E}\left[\eta \mid \vee_{t\geq 0}\sigma(B_t)\right]$ and we are done. □

Theorem 14.57 (Martingale Representation Theorem). *Let $B = (B_1, \dots, B_d)$ be a d-dimensional Brownian motion and let $\mathcal{F}_t$ be the complete filtration generated by B. Let M be a cadlag local $\mathcal{F}$-martingale. Then M is continuous and moreover there exists $\mathbf{P} \times \lambda$-almost everywhere unique progressive processes $V^1, \dots, V^d$ such that*

$$M = M_0 + \sum_{j=1}^d \int V^j\, dB^j$$

Proof. By applying the result to $M - M_0$ it is clear that we may assume that $M_0 = 0$. We first show that M is continuous. By Lemma 14.2 we may pick a localizing sequence τ_n such that M^{τ_n} is a uniformly integrable cadlag martingale. If we can show that every M^{τ_n} is almost surely continuous then it follows that M is almost surely continuous. To be precise, if we let $A = \{\tau_n \uparrow \infty\} \cap \cap_{n=1}^\infty \{M^{\tau_n} \text{ is continuous}\}$ then for all $\omega \in A$ and for every $t \geq 0$ there exists an N such that $\tau_n(\omega) \geq t+1$ for all $n \geq N$ and therefore $M_s(\omega) = M_s^{\tau_n}(\omega)$ for all $0 \vee t-1 < s < t+1$ and therefore M is continuous at t.

Thus we may assume that M is a uniformly integrable martingale starting at zero. By the Martingale Convergence Theorem 9.80 we have $\mathcal{F}_\infty$-measurable M_∞ such that $M \stackrel{L^1}{\to} M_\infty$. Since L^2 is dense in L^1 we may find $\xi^n \in L^2(\Omega, \mathcal{F}_\infty)$ such that $\xi^n \stackrel{L^1}{\to} M_\infty$. By $\mathcal{F}_\infty$-measurability of ξ^n and Lemma 14.56 we know that the martingale $M_t^n = \mathbf{E}[\xi^n \mid \mathcal{F}_t]$ is almost surely continuous. Denote $\Delta M_t = M_t - \lim_{s \uparrow t} M_s$ to be jump process associated with M and note that by the Doob Maximal Inequality (Lemma 9.77) we have for each $\epsilon > 0$ and $n \in \mathbb{N}$

$$\mathbf{P}\{\sup_{0 \leq t < \infty} |\Delta M_t| > 2\epsilon\} \leq \mathbf{P}\{\sup_{0 \leq t < \infty} |M_t^n - M_t| > \epsilon\} \leq \epsilon^{-1} \mathbf{E}[|\xi^n - M_\infty|]$$

and taking the limit as $n \to \infty$ we see that $\mathbf{P}\{\sup_{0 \leq t < \infty} |\Delta M_t| > 2\epsilon\} = 0$ for every $\epsilon > 0$ and thus $\mathbf{P}\{\sup_{0 \leq t < \infty} |\Delta M_t| \neq 0\} \leq \cup_{n=1}^\infty \mathbf{P}\{\sup_{0 \leq t < \infty} |\Delta M_t| > 1/n\} = 0$ which shows us that M_t is almost surely continous.

By the above argument we may now assume that M is a continuous local $\mathcal{F}$-martingale. By Lemma 14.3 we may assume that we have a localizing sequence τ_n such each M^{τ_n} is bounded (in particular L^2 bounded). Therefore $M_\infty^{\tau_n}$ exists and is in L^2. Since $M_\infty^{\tau_n}$ is $\mathcal{F}_\infty$-measurable and $\mathbf{E}[M_\infty^{\tau_n}] = 0$ we may apply Lemma 14.56 to conclude there are $V^{j,n} \in L(B^1)$ such that $M_\infty^{\tau_n} = \sum_{j=1}^d \int_0^\infty V^{j,n}\, dB^j$. Therefore from the fact that $M_t^{\tau_n}$ is a closable martingale we have $M_t^{\tau_n} = \mathbf{E}[M_\infty^{\tau_n} \mid \mathcal{F}_t] = \sum_{j=1}^d \int_0^t V^{j,n}\, dB^j$.

For $m < n$ we have $\tau_m \leq \tau_n$ and therefore using Lemma 14.42

$$\sum_{j=1}^d \int V^{j,m}\, dB^j = M^{\tau_m} = (M^{\tau_n})^{\tau_m} = \sum_{j=1}^d \int \mathbf{1}_{[0,\tau_m]} V^{j,n}\, dB^j$$

and by the almost sure uniqueness of the $V^{j,m}$ we conclude that $V^{j,n} \mid_{[0,\tau_m]} = V^{j,m}$. Therefore there exists V^j such that $V^j \mid_{[0,\tau_n]} = V^{j,n}$ and for any $t \geq 0$ using Lemma 14.42

$$\begin{aligned} M_t &= \lim_{n \to \infty} M_t^{\tau_n} = \lim_{n \to \infty} \sum_{j=1}^d \int_0^t V^{j,n}\, dB^j \\ &= \lim_{n \to \infty} \sum_{j=1}^d \int_0^t \mathbf{1}_{[0,\tau_n]} V^j\, dB^j \\ &= \lim_{n \to \infty} \sum_{j=1}^d \int_0^{\tau_n \wedge t} V^j\, dB^j = \sum_{j=1}^d \int_0^t V^j\, dB^j \end{aligned}$$

almost surely. TODO: Show $V^j \in L(B^1)$ and show a.s. uniqueness of V^j. $\square$

Another result that is important is the Levy's characterization of Brownian motion in terms of its covariance structure.

THEOREM 14.58 (Levy's Theorem). *Let $B_t = (B^1_t, \dots, B^d_t)$ be a process in $\mathbb{R}^d$ such that $B_0 = 0$, then B is an $\mathcal{F}$-Brownian motion if and only if B is a continuous local $\mathcal{F}$-martingale with $[B^i, B^j]_t = \delta_{ij} t$.*

PROOF. Suppose that B is a continuous local $\mathcal{F}$-martingale with $[B^i, B^j]_t = \delta_{ij} t$ and $B_0 = 0$. We need to show that for each $s < t$, $B_t - B_s$ is independent of $\mathcal{F}_s$ and Gaussian with covariance matrix $t - s$ times the identity. Fix $S < T$ and define the filtration $\tilde{\mathcal{F}}_t = \mathcal{F}_{t+S}$, $\tilde{B}^i_t = B_{t+S} - B_S$ for each $i = 1, \dots, d$ and $\tilde{B} = (\tilde{B}^1, \dots, \tilde{B}^d)$. Clearly, $\tilde{B}$ is a continuous local $\tilde{\mathcal{F}}$-martingale and moreover note that $[\tilde{B}^i, \tilde{B}^j]_t = [B^i, B^j]_{t+S} - [B^i, B^j]_S = t\delta_{ij}$. Let $u = (u_1, \dots, u_d) \in \mathbb{R}^d$ be given and define

$$N_t = \langle u, \tilde{B}^{T-S}_t \rangle = u_1(B^1_{(t+S)\wedge T} - B^1_S) + \dots + u_d(B^d_{(t+S)\wedge T} - B^d_S)$$

Clearly, N_t is a continuous local $\tilde{\mathcal{F}}$-martingale such that $N_0 = 0$ and also has the quadratic variation

$$\begin{aligned}
[N]_t &= \sum_{i=1}^d \sum_{j=1}^d u_i u_j [(\tilde{B}^i)^{T-S}, (\tilde{B}^j)^{T-S}]_t \\
&= \sum_{i=1}^d \sum_{j=1}^d u_i u_j ([B^i, B^j]_{t+S} - [B^i, B^j]_S)^{T-S} = (t \wedge (T - S)) \left\| u \right\|_2^2
\end{aligned}$$

By Lemma 14.55 we know that $Z_t = \exp(iN_t + \frac{1}{2}[N]_t)$ is a continuous local $\tilde{\mathcal{F}}$-martingale that satisfies $Z_0 = 1$. Since $[N]_\infty = [N]_{T-S} = (T - S) \left\| u \right\|_2^2 < \infty$ we know that Z_t is bounded and therefore we can apply Lemma 14.4 to see that is a uniformly integrable martingale. For any $A \in \tilde{\mathcal{F}}_0 = \mathcal{F}_S$ we have by the martingale property of Z and the definition of N

$$\begin{aligned}
\mathbf{P}\{A\} &= \mathbf{E}\left[Z_0; A\right] = \mathbf{E}\left[Z_\infty; A\right] = \mathbf{E}\left[\exp(iN_\infty + \frac{1}{2}[N]_\infty); A\right] \\
&= \mathbf{E}\left[e^{i\langle u, \tilde{B}_{T-S}\rangle}; A\right] e^{\frac{1}{2}(T-S)\left\| u \right\|_2^2}
\end{aligned}$$

which shows us that $\mathbf{E}\left[e^{i\langle u, \tilde{B}_{T-S}\rangle} \mid \mathcal{F}_S\right] = e^{-\frac{1}{2}(T-S)\langle u, u\rangle}$. We may now apply uniqueness of conditional characteristic functions Lemma 8.38 and Theorem 7.18 to conclude that the conditional probability distribution $\mathbf{P}\{(B^1_T - B^1_S, \dots, B^d_T - B^d_S) \in \cdot \mid \mathcal{F}_S\}$ is centered Gaussian with covariance matrix $(T - S)$ times the identity. As the conditional probability distribution is deterministic and we see that $B_T - B_S$ is independent of $\mathcal{F}_S$ as well. $\square$

The Martingale Representation theorem shows that when we are given a Brownian motion, then all Brownian functionals (taken so broadly as to include continuous local martingales) are obtained as stochastic integrals. There is related question that shall prove important going forward; given a continuous local martingale can we construct a Brownian motion such that the local martingale is a stochastic integral. The following generalization of a theorem of Doob shows that this is possible. TODO: The assumption on the covariance matrix is all we need for our application to weak solutions of SDE's but it can be derived by from more primitive assumptions (e.g. see Karatzas and Shreve).

TODO: Define standard extension, prove that a continuous local martingale lifts with covariation preserved, progressive processes lift to progressive processes.

THEOREM 14.59. *Let $(\Omega, \mathcal{A}, P, \mathcal{F}_t)$ be a filtered probability space and let M be a continuous local $\mathcal{F}$-martingale in $\mathbb{R}^d$ such that $M_0 = 0$ and $[M]_t = \int_0^t VV^T \, ds$ where V is a $d \times n$ matrix valued $\mathcal{F}$-progressive process. There exists a standard extension $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{P}, \tilde{\mathcal{F}}_t)$ and an $\tilde{\mathcal{F}}$-Brownian motion B in reals^n such that M and $\int V_s \, dB_s$ are indistinguishable.*

PROOF. To clarify the notation in the hypothesis, we have $[M^{(i)}, M^{(j)}]_t = \sum_{k=1}^n \int_0^t V_s^{ik} V_s^{jk} \, ds$ where V^{ik} is an $\mathcal{F}$-progressive process for each $1 \leq i \leq d$ and $1 \leq k \leq n$.

TODO: Show that we can create an independent Brownian motion X in $\mathbb{R}^d$ on a standard extension

View the processes V^{ik} as an dXn matrix valued process that we call V. Working pointwise in Ω and $0 \leq t < \infty$ by Corollary 15.159 the singular value decomposition $V = A\Sigma B^T$ is a Borel measurable function of V. We let $V^+ = B\Sigma^{-1}A^T$ be the Moore-Penrose pseudoinverse of V and note that since the pseudoinverse is a Borel measurable function of (A, Σ, B), it is a Borel measurable function of V and therefore the $n \times d$ matrix valued process V^+ is also $\tilde{\mathcal{F}}$-progressive. Let $\pi_N = 1 - BB^T$ be constructed pointwise as the projection onto the nullspace of V and note that it too is an $d \times d$ matrix valued $\tilde{\mathcal{F}}$-progressive process. In a similar fashion we get $\tilde{\mathcal{F}}$-progressive projection processes $\pi_{N^\perp} = BB^T$, $\pi_R = AA^T$ and $\pi_{R^\perp} = 1 - AA^T$ onto the orthogonal complement of the null space of V, the range space of V and the orthogonal complement of the range space of V respectively. Now define a continuous local $\tilde{\mathcal{F}}$-martingale B in $\mathbb{R}^n$ as

$$B_t = \int_0^t V_s^+ \, dM_s + \int_0^t \pi_N(s) \, dX_s$$

TODO: Justify that the integrals are well defined; or at least note that the computations below show this implicitly.

TODO: Independence of M and X shows that $\left[\int_0^t V_s^+ \, dM_s, \int_0^t \pi_N(s) \, dB_s\right] = 0$.

Now applying independence of M and X, Exercise 58, the fact that $V^+V = \pi_{N^\perp}$ and the fact that projections satisfy $\pi\pi^T = \pi^2 = \pi$

$$\begin{aligned}
[B]_t &= \left[\int V_s^+ \, dM_s\right]_t + \left[\int_0^t \pi_N(s) \, dB_s\right]_t \\
&= \int_0^t (V^+VV^T(V^+)^T)_s \, ds + \int_0^t (\pi_N \pi_N^T)_s \, ds \\
&= \int_0^t (\pi_{N^\perp} \pi_{N^\perp}^T)_s \, ds + \int_0^t (\pi_N \pi_N^T)_s \, ds \\
&= \int_0^t (\pi_{N^\perp} + \pi_N)_s \, ds = t \, \mathrm{Id}
\end{aligned}$$

where Id is the $n \times n$ identity matrix. Now by Levy's Theorem 14.58 we conclude that X is a $\tilde{\mathcal{F}}$-Brownian motion in $\mathbb{R}^n$.

Note also that by Exercise 58, the hypothesis on the covariation matrix of M and the fact that $\pi_{R^\perp} V = 0$ we get

$$[\int \pi_{R^\perp} \, dM]_t = \int_0^t (\pi_{R^\perp} V V^t \pi_{R^\perp}^T)_s \, ds = 0$$

and therefore $\int \pi_{R^\perp} \, dM = 0$ almost surely. From these facts, Chain Rule Lemma 14.41 and $V\pi_N = 0$ we compute

$$\begin{aligned}
\int_0^t V_s \, dB_s &= \int_0^t V_s \, d \int_0^s V_u^+ \, dM_u + \int_0^t V_s \, d \int_0^s \pi_N(u) \, dX_u \\
&= \int_0^t V_s V_s^+ \, dM_s + \int_0^t V_s \pi_N(s) \, dX_s \\
&= \int_0^t \pi_R(s) \, dM_s \\
&= \int_0^t (\pi_R + \pi_R)_s \, dM_s = M_t
\end{aligned}$$

□

6.1. Girsanov Theory. We now begin an investigation of how continuous local martingales behave as the underlying probability measure is changed.

Definition 14.60. Let P and Q be probability measures on a measure space $(\Omega, \mathcal{A})$ with a filtration $\mathcal{F}_t$ with index set T. We say that Q is *locally absolutely continuous* with respect to P is for each $t \in T$ we have $Q \ll P$ on $\mathcal{F}_t$. If in addition P is locally absolutely continuous with respect to Q then say that P and Q are *locally equivalent*.

TODO: Are there any subtleties about the usual conditions? The fact that the usual conditions are tied to the probability measure by the assumption that each $\mathcal{F}_t$ contains all subsets of the null sets of the probability measure. When we pass to a new probability measure (even with the absolute continuity assumptions) we may be introducing new null sets and the filtration may no longer satisfy the usual conditions right? Most presentations restrict to the situation of equivalent probability measures so this problem doesn't arise but Kallenberg clearly doesn't intend to make this restriction at the outset. There is indeed something subtle about the usual conditions (see Bichteler). In the first place it is observed (and noted elsewhere) that the usual conditions mean that if $Q \ll P$ on $\mathcal{F}_0$ then $Q \ll P$ on all of $\mathcal{F}_\infty$ (i.e. there isn't a useful notion of being only locally absolutely continuous). On the other hand, Kallenberg uses the usual conditions to assume a cadlag version of the likelihood ratio process. Moreover, it seems that some of the more useful variants of Girsanov are not compatible with the usual conditions since they require that the change of measure is not absolutely continuous but merely locally so. It seems like Kallenberg has made a bit of a muddle of this. I'm still trying to distill the core issues. In the Brownian motion case a constant drift term illustrates the problem. Let B_t be a standard Brownian motion, let $\mu > 0$ be a constant and consider $\tilde{B}_t = B_t - \mu t$. Let $\mathcal{F}_t$ be the filtration generated by B_t and let $\tilde{\mathcal{F}}_t$ be the usual augmentation. If we define $Z_t = \exp\left[\mu B_t - \frac{1}{2}\mu^2 t\right]$ then we can define a new probability measure $\tilde{P}$ on each $\tilde{\mathcal{F}}_t$ by $\tilde{P}(A) = \mathbf{E}[Z_t; A]$. What is true is that

(i) $\tilde{B}_t$ is a $\tilde{\mathcal{F}}$-Brownian motion on $[0, t]$ with respect to $\tilde{P}$ for all $0 \le t < \infty$.

(ii) P and $\tilde{P}$ are mutually absolutely continous on $\tilde{\mathcal{F}}_t$ for all $0 \leq t < \infty$.
(iii) There is an extension of $\tilde{P}$ to all of $\mathcal{F}_\infty$ such that $\tilde{B}_t$ is an $\mathcal{F}$-Brownian motion

what is not true is that

(i) The extension of $\tilde{P}$ to all of $\mathcal{F}_\infty$ is not equal to $\mathbf{E}[Z_t; A]$ on all of $\tilde{\mathcal{F}}_t$ but only on $\mathcal{F}_t$. In particular $\{\lim_{t\to\infty} B_t/t = \mu\}$ is a P - null set but is $\tilde{P}$ almost sure.
(ii) The extension $\tilde{P}$ and P are not mutually absolutely continuous on $\mathcal{F}_\infty$.
(iii) $\tilde{B}_t$ is not a $\tilde{\mathcal{F}}$-Brownian motion on $[0, \infty)$ with respect to $\tilde{P}$.

TODO: Is there an obstruction to extending $\tilde{P}$ to a probability measure on $\tilde{\mathcal{F}}_\infty$ or is it just that $\tilde{B}_t$ will not be a Brownian motion on $[0, \infty)$ with respect to such an extension? From van der Vaart's notes, the Brownian motion example above shows that there is an obstruction. The key seems to be that if such an extension did exist then Girsanov would apply and we'd be able to conclude that $\tilde{B}_t$ would be a Brownian motion on it. That would cause a contradiction because the Brownianness of $\tilde{B}_t$ would imply that $\tilde{P}\{\lim_{t\to\infty} B_t/t = \mu\} = 1$ but the completeness of $\mathcal{F}_t$ would imply that $\tilde{P}\{\lim_{t\to\infty} B_t/t = \mu\} = P\{\lim_{t\to\infty} B_t/t = \mu\} = 0$. Does this mean that Kallenberg's Lemma 18.18 is incorrect?

TODO: Is it also that case that $\tilde{B}_t$ is not a continuous local $\tilde{P}$-martingale on $[0, \infty)$? This question only makes sense if the answer to the previous question is that there is such an extension to $\tilde{P}$ on all of $\tilde{\mathcal{F}}_\infty$.

One of the nasty things about developing the theory in this way (and having a result that doesn't hold for filtrations that satisfy the usual conditions) is the fact that we have all sorts of results that do assume the usual conditions and it is not terribly clear what the ramifications of losing the assumption are. Bichteler has identified an extension procedure that is more conservative than the imposition of the usual conditions (his *natural* conditions) that seems to preserve all the important results but also allows an extension of $\tilde{P}$ such that $\tilde{B}_t$ will be a Brownian motion on all of $[0, \infty)$ with respect to $\tilde{P}$. Note that the example of the event $\{\lim_{t\to\infty} B_t/t = \mu\}$ shows that such an extension will not be mutually absolutely continuous but it will be locally so.

We need a preliminary result that says when a non-negative cadlag supermartingale hits 0 it is absorbed. The reader should convince herself that this is expected: since a supermartingle is non-increasing on average, once it hits zero any attempt to return to a positive value would have to be offset by corresponding negative value. Making sense of the intuition requires an argument using optional times.

LEMMA 14.61. *Let $X \geq 0$ be a cadlag $\mathcal{F}$-supermartingale (with $\mathcal{F}$ not necessarily satisfying the usual conditions) and let $\tau = \inf\{t \mid X_{t-} \wedge X_t = 0\}$, then $X \equiv 0$ a.s. on $[\tau, \infty)$.*

PROOF. First note that X is also an $\mathcal{F}^+$-supermartingale. To see this, pick for any $s < t$ pick a sequence $s_m \downarrow s$ and use the Levy Downward Theorem 9.56 to conclude $\mathbf{E}[X_t \mid \mathcal{F}^+_s] = \lim_{n\to\infty} \mathbf{E}\left[X_t \mid \mathcal{F}^+_{s_n}\right] \leq X_t$. Now we use an approximation to the optional time τ. The idea the approximation is that $X_{t-} \wedge X_t = 0$ if and only if $X_{t-} \wedge X_t < 1/n$ for all $n \in \mathbb{N}$ so we look for the first point t for which $X_{t-} \wedge X_t < 1/n$, that is to say we consider the hitting time $\tau_n = \{t \mid X_t < 1/n\}$ and use the fact that $\tau_n \uparrow \tau$ (we'll actually show this carefully later in the proof but it isn't hard to believe). Note that τ_n is an $\mathcal{F}^+$-optional time due to the openness of $(-\infty, 1/n)$ and the right continuity of X (Lemma 9.70). Note that for all $n \in \mathbb{N}$ the right continuity of X implies that $X_{\tau_n} \leq 1/n$ (just pick a random sequence $t_m \downarrow \tau_n$ such that $X_{t_m} < 1/n$). Now pick $t \geq 0$ and $n \in \mathbb{N}$ and use the supermartingale property, the $\mathcal{F}^+_{\tau_n \wedge t}$-measurability of $\{\tau_n \leq t\}$ (TODO: where do we show this) and Optional Sampling Theorem 9.82

$$\begin{aligned}\mathbf{E}\left[X_t; \tau_n \leq t\right] &= \mathbf{E}\left[\mathbf{E}\left[X_t \mid \mathcal{F}^+_{\tau_n \wedge t}\right]; \tau_n \leq t\right] \leq \mathbf{E}\left[X_{\tau_n \wedge t}; \tau_n \leq t\right] \\ &= \mathbf{E}\left[X_{\tau_n}; \tau_n \leq t\right] \leq 1/n\end{aligned}$$

Using the non-negativity and integrability of X_t, the fact that $\tau_n \uparrow \tau$ and Dominated Convergence, we get

$$0 \leq \mathbf{E}\left[X_t; \tau \leq t\right] \leq \lim_{n \to \infty} 1/n = 0$$

and therefore $X_t = 0$ a.s. on the set $\{\tau \leq t\}$. Taking a countable union of null events we see that almost surely for all $q \in \mathbb{Q}_+$, $X_q = 0$ on the set $\{\tau \leq q\}$ and by right continuity we get that

$$\cap_{q \in \mathbb{Q}_+} \{X_q = 0 \text{ on } \{\tau \leq q\}\} = \cap_{0 \leq t < \infty} \{X_t = 0 \text{ on } \{\tau \leq t\}\}$$

(for $\omega \in \cap_{q \in \mathbb{Q}_+} \{X_q = 0 \text{ on } \{\tau \leq q\}\}$ and $t \geq \tau(\omega)$, pick $q_n \downarrow t$ and note that $X_{q_n}(\omega) = 0$ so by right continuity $X_t(\omega) = 0$). This yields the final result.

We now return to the deferred justification of the claim that $\tau_n \uparrow \tau$. It is clear that τ_n is non-decreasing. To see that $\tau_n \leq \tau$ note that if $X_{t-} \wedge X_t = 0$ the either $X_t = 0$ or we may find $s < t$ such that $X_s < 1/n$ and therefore $\tau_n \leq t$. In the opposite direction, we give ourselves an ϵ of room and pick an arbitrary $\epsilon > 0$. Pick a random $N > 0$ such that $\lim_{n \to \infty} \tau_n - \epsilon/2 \leq \tau_n$ for all $n \geq N$ and then for each $n \geq N$ we pick a t_n such that $X_{t_n} < 1/n$ and $t_n \leq \tau_n + \epsilon/2$. In this way we construct a sequence t_n in $[0, \lim_{n \to \infty} \tau_n + \epsilon]$ such that $X_{t_n} < 1/n$. By compactness we get a $t \in [0, \lim_{n \to \infty} \tau_n + \epsilon]$ and a convergent subsequence N' such that $t_n \to t$ along N' which by passing to another subsequence we may assume is either increasing or decreasing. From this and right continuity of X we conclude that $X_{t-} \wedge X_t = 0$ and therefore $\tau \leq \lim_{n \to \infty} \tau_n + \epsilon$ and since $\epsilon > 0$ was arbitrary we are done. □

Lemma 14.62. *Let P and Q be probability measures on a measure space $(\Omega, \mathcal{A})$ with a filtration $\mathcal{F}$ (not necessarily satisfying the usual conditions). Suppose that Q is locally absolutely continuous with respect to P and let Z_t be an $\mathcal{F}$-adapted process such that $Q = Z_t \cdot P$ on $\mathcal{F}_t$ for all $t \geq 0$, then*

(i) *An adapted process X is a Q-martingale if and only if XZ is a P-martingale. In particular, Z_t is a P-martingale. Moreover, Z is uniformly integrable if and only if $Q \ll P$ on $\mathcal{F}_\infty$.*

(ii) *If Z is a cadlag version then for any optional time*

$$Q = Z_\tau \cdot P \text{ on } \mathcal{F}_\tau \cap \{\tau < \infty\}$$

and an adapted cadlag process X is a local Q-martingale if and only if XZ is a local P-martingale.

(iii) *If τ_n is a sequence of optional times such that $\tau_n \uparrow \infty$ P-almost surely then $\tau_n \uparrow \infty$ Q-almost surely.*

(iv) *An adapted cadlag process X is a local Q-martingale if and only if XZ is a local P-martingale.*

(v) *If Z is a cadlag version then Q-almost surely for every $t > 0$ we have $\inf_{0 \leq s \leq t} Z_s > 0$. If Q and P are locally equivalent then this is true P-almost surely as well.*

PROOF. We start with proving (i). Note that since X_t is $\mathcal{F}_t$-measurable by Lemma 2.57 and non-negativity of Z_t we have $\mathbf{E}_Q [|X_t|] = \mathbf{E}_P [Z_t |X_t|] = \mathbf{E}_P [|Z_t X_t|]$ and therefore X_t is Q-integrable if and only if $Z_t X_t$ is P-integrable. If we let $A \in \mathcal{F}_s$ then if we assume $Z_t X_t$ is a P-martingale then for $t \geq s$,

$$\mathbf{E}_Q [X_t; A] = \mathbf{E}_P [Z_t X_t; A] = \mathbf{E}_P [Z_s X_s; A] = \mathbf{E}_Q [X_s; A]$$

and similarly if we assume that X_t is a Q-martingale then we just the run the logic in a slightly different order

$$\mathbf{E}_P [Z_t X_t; A] = \mathbf{E}_Q [X_t; A] = \mathbf{E}_Q [X_s; A] = \mathbf{E}_P [Z_s X_s; A]$$

Thus we see that X_t is a Q-martingale if and only if $Z_t X_t$ is a P-martingale. Since $X_t \equiv 1$ is obviously a Q-martingale we see that Z_t is a P-martingale. If we assume that Z is a uniformly integrable Q-martingale then by the Martingale Convergence Theorem 9.80 there exists Z_∞ such that $Z_t = \mathbf{E} [Z_\infty \mid \mathcal{F}_t]$ a.s. Therefore if we assume that $A \in \mathcal{F}_t$ we have

$$(Z_\infty \cdot P)(A) = \mathbf{E} [Z_\infty; A] = \mathbf{E} [\mathbf{E} [Z_\infty \mid \mathcal{F}_t] ; A] = \mathbf{E} [Z_t; A] = Q(A)$$

and since $\cup_{t \geq 0} \mathcal{F}_t$ is a π-system generating $\mathcal{F}_\infty$ we know that $Z_\infty \cdot P = Q$ on $\mathcal{F}_\infty$ by monotone classes (specifically Lemma 2.71). On the other hand suppose that $Q \ll P$ on $\mathcal{F}_\infty$ and write $Q = \xi \cdot P$. Then since $\mathcal{F}_t \subset \mathcal{F}_\infty$ we know that for all $t \geq 0$ and $A \in \mathcal{F}_t$ we have $Q(A) = \mathbf{E} [\xi; A] = \mathbf{E} [\mathbf{E} [\xi \mid \mathcal{F}_t] ; A]$ which shows $Z_t = \mathbf{E} [\xi \mid \mathcal{F}_t]$ by the P-almost sure uniqueness of the Radon-Nikodym derivative. This shows that Z_t is uniformly integrable.

We now show (ii). If we let τ be an optional time then we fix $t \geq 0$ and assume $A \in \mathcal{F}_{\tau \wedge t} \subset \mathcal{F}_t$ and apply Optional Sampling to see that

$$Q(A) = \mathbf{E} [Z_t; A] = \mathbf{E} [\mathbf{E} [Z_t \mid \mathcal{F}_{\tau \wedge t}] ; A] = \mathbf{E} [Z_{\tau \wedge t}; A]$$

Given an arbitrary $A \in \mathcal{F}_\tau$ we know from Proposition 9.32 that for all $t \geq 0$ we have $A \cap \{\tau \leq t\} \in \mathcal{F}_{\tau \wedge t}$. Therefore $Q(A; \tau \leq t) = \mathbf{E} [Z_\tau; A; \tau \leq t]$ and by continuity of measure and Monotone Convergence we get

$$Q(A; \tau < \infty) = \lim_{n \to \infty} Q(A; \tau \leq n) = \lim_{n \to \infty} \mathbf{E} [Z_\tau; A; \tau \leq n] = \mathbf{E} [Z_\tau; A; \tau < \infty]$$

To see (iii), we let $\tau = \sup_n \tau_n$ and note that τ is an optional time by Lemma 9.72. Therefore we may apply (ii) and the fact that $\mathbf{P}\{\tau < \infty\} = 0$ to conclude that $\mathbf{P}_Q\{\tau < \infty\} = \mathbf{E}_P [Z_\tau; \tau < \infty] = 0$. Note that the above argument is necessary since we don't necessarily have $Q \ll P$ on $\mathcal{F}_\infty$.

To see (iv), suppose that X is a local P-martingale. Let $\tau_n \uparrow \infty$ P-a.s. be a localizing sequence for X so that X^{τ_n} is a P-martingale for every $n \in \mathbb{N}$. By (i) we conclude that ZX^{τ_n} is a Q-martingale. It follows that $(ZX^{\tau_n})^{\tau_n} = Z^{\tau_n} X^{\tau_n} = (ZX)^{\tau_n}$ is a Q-martingale for every $n \in \mathbb{N}$. Since by (iii) it follows that $\tau_n \uparrow \infty$ Q-almost surely we conclude that ZX is a local Q-martingale.

To see (v), by the $\mathcal{F}_t$-measurability of Z_t we have $\mathbf{P}_Q\{Z_t = 0\} = \mathbf{E} [Z_t; Z_t = 0] = 0$ and therefore $Z_t > 0$ Q-almost surely for each $t \geq 0$. Since Z_t is a cadlag P-martingale we let $\tau = \inf\{t \mid Z_{t-} \wedge Z_t = 0\}$ and apply Lemma 14.61 to conclude that $Z_t \equiv 0$ P-almost surely on $[\tau, \infty)$: more formally written as $\mathbf{P}\{\cap_{0 \leq t < \infty} \{Z_t \mathbf{1}_{\tau \leq t} = 0\}\} = 1$. In particular, we have $\mathbf{P}\{Z_t \mathbf{1}_{\tau \leq t} = 0\} = 1$ for all

$t \geq 0$. Since $\{Z_t \mathbf{1}_{\tau \leq t} = 0\} \in \mathcal{F}_t$-measurable and $Q \ll P$ on $\mathcal{F}_t$, by taking complements we see that $\mathbf{P}_Q\{Z_t > 0; \tau \leq t\} = 0$ for all $t \geq 0$.. Putting these two facts together we conclude for all $t \geq 0$ that

$$\begin{aligned}\mathbf{P}_Q\{\tau \leq t\} &= \mathbf{P}_Q\{Z_t = 0; \tau \leq t\} + \mathbf{P}_Q\{Z_t > 0; \tau \leq t\} \\ &\leq \mathbf{P}_Q\{Z_t = 0\} + \mathbf{P}_Q\{Z_t > 0; \tau \leq t\} = 0\end{aligned}$$

By continuity of measure $\mathbf{P}_Q\{\tau < \infty\} = \lim_{t\to\infty} \mathbf{P}_Q\{\tau \leq t\} = 0$ and therefore $\tau = \infty$ Q-almost surely. and therefore $Z_{t-} \wedge Z_t > 0$ for all $t \geq 0$ Q-almost surely which shows the result. If we now assume that P and Q are locally equivalent, we also have $Z_t > 0$ P-almost surely for each $t \geq 0$. We have already shown that, indepedent of the assumption of local equivalence, we have $Z_t \equiv 0$ P-almost surely on $[\tau, \infty)$. By exactly the same argument as above, these two facts imply the result with respect to P. $\square$

TODO: Van der Vaart claims that it is not true that X is a local Q-martingale if and only if ZX is a local P-martingale unless we assume that P and Q are locally equivalent. Understand whether that is true and if so produce a counterexample and find the flaw in the proof from Kallenberg (which is very brief).

Theorem 14.63. *Let P and Q be locally equivalent probability measures on a measure space $(\Omega, \mathcal{A})$ with a filtration $\mathcal{F}$. Let Z_t be an $\mathcal{F}$-adapted process such that $Q = Z_t \cdot P$ on $\mathcal{F}_t$ for all $t \geq 0$ and assume that Z_t is almost surely continuous. Then if M is a local P-martingale, the process $\tilde{M}_t = M_t - \int_0^t Z_s^{-1} d[Z, M]_s$ is a local Q-martingale.*

Proof. The first thing to note is that the process $M_t - \int_0^t Z_s^{-1} d[Z, M]_s$ is well defined. From Lemma 14.62 we know that Q-almost surely and P-almost surely, for all $t \geq 0$ $\inf_{0 \leq s \leq t}\{t \mid Z_s\} > 0$; thus Q-a.s. and P-a.s. the process Z^{-1} is bounded on every $[0, t]$ and therefore $\int_0^t Z_s^{-1} d[Z, M]_s$ exists.

For each $n \in \mathbb{N}$, let $\tau_n = \inf\{t \mid Z_t < 1/n\}$ and define $\tilde{M}^n_t = M^{\tau_n}_t - \int_0^t \mathbf{1}_{[0,\tau_n]} Z_s^{-1}\, d[Z, M^{\tau_n}]_s$. Note that from the definition of τ_n, $\mathbf{1}_{[0,\tau_n]} Z_s^{-1}$ is bounded and therefore $\tilde{M}^n_t$ is well defined and moreover is a continuous $\mathcal{F}$-semimartingale. By integration by parts (Lemma 14.44) and the Chain Rule (Lemma 14.41) we get

$$\begin{aligned}\tilde{M}^n_t Z_t - \tilde{M}^n_0 Z_0 &= \int_0^t \tilde{M}^n\, dZ + \int_0^t Z\, d\tilde{M}^n + [\tilde{M}^n, Z]_t \\ &= \int_0^t \tilde{M}^n\, dZ + \int_0^t Z\, dM^{\tau_n} - \int_0^t Z_s\, d\int_0^s Z_u^{-1} [Z, M^{\tau_n}]_u + [M^{\tau_n}, Z]_t \\ &= \int_0^t \tilde{M}^n\, dZ + \int_0^t Z\, dM^{\tau_n} - \int_0^t d[Z, M]_s + [M^{\tau_n}, Z]_t \\ &= \int_0^t \tilde{M}^n\, dZ + \int_0^t Z\, dM^{\tau_n}\end{aligned}$$

which shows that $\tilde{M}^n_t Z_t$ is a local P-martingale and therefore $\tilde{M}^n_t$ is a local Q-martingale by Lemma 14.62.

TODO: Kallenberg states this Lemma without the assumption of local equivalence of P and Q (i.e. only assuming that Q is locally absolutely continuous with respect to P). The main point that I don't understand is showing that M is well defined both P-a.s. as well as Q-a.s.

TODO: How do we see that $\int_0^t Z_s^{-1} d[Z,M]_s$ is well defined in general? Perhaps there is an argument that shows that if we let $\tau = \inf\{t \mid Z_t = 0\}$ then $0 < Z^{-1} < \infty$ on $[0,\tau)$ and by Lemma 14.61 $Z^{-1} = \infty$ on $[\tau, \infty)$ P-a.s. If we can then show that $[Z,M] = 0$ on $[\tau,\infty)$. By then perhaps we can conclude that $\int_0^t Z_s^{-1} d[Z,M]_s$ is well defined for all $t \geq 0$ (the only issue is $t \geq \tau$. This is a question of Stieltjes integrals but seems possible as it is a $\infty \cdot 0$ type of situation. Perhaps this argument won't work either as we still would have to control the rate that $Z_t^{-1} \uparrow \infty$ as $t \to \tau$.

TODO: As an alternative to showing that Here I try to show that $\tilde{M}$ is well defined by defining it as a limit of the $\tilde{M}^n$; is this any better because we only know that $\tau_n \uparrow \tau$ Q-a.s? Now note that since $\inf_{0 \leq s \leq t} Z_t > 0$ Q-almost surely (Lemma 14.62) we know that $\tau_n \uparrow \tau$ Q-almost surely. We know that for any fixed $t \geq 0$ the sequence $\tilde{M}_t^n$ is Q-almost surely constant and therefore we can define $\tilde{M}_t = \lim_{n\to\infty} \tilde{M}_t^n$ and it follows that $\tilde{M}^n = (\tilde{M})^{\tau_n}$. Now we can apply Lemma 14.8 to conclude that $\tilde{M}$ is a local Q-martingale. $\square$

Note that in the result above the quadratic covariation $[Z,M]$ is taken with respect to the measure P. However we know from Lemma 14.48 that for each $t \geq 0$ we can find a sequence of partitions $0 = t_{n,0} < \cdots < t_{n,k_n} = t$ such that $\sum_{j=1}^{k_n} (Z_{t_{n,j}} - Z_{t_{n,j-1}})(M_{t_{n,j}} - M_{t_{n,j-1}}) \xrightarrow{P} [Z,M]_t$ with respect to P. Since P and Q are locally equivalent this implies that in fact $[Z,M]$ is the quadratic covariation of Z and M under Q as well.

If we specialize the previous result to the case in which M is a Brownian motion then it is easy to see that $\tilde{M}$ is also a Brownian motion; thus the family of Brownian motions is invariant under locally equivalent changes of measure.

Corollary 14.64. *Let P and Q be locally equivalent probability measures on a measure space $(\Omega, \mathcal{A})$ with a filtration $\mathcal{F}$. Let Z_t be an $\mathcal{F}$-adapted process such that $Q = Z_t \cdot P$ on $\mathcal{F}_t$ for all $t \geq 0$ and assume that Z_t is almost surely continuous. Then if M is a P-Brownian motion, then process $M_t - \int_0^t Z_s^{-1} d[Z,M]_s$ is a Q-Brownian motion.*

Proof. Since $\int_0^t Z_s^{-1} d[Z,M]_s$ has finite variation we know that $[M - \int Z_s^{-1} d[Z,M]_s]_t = [M]_t = t$ where we have used fact that M is a P-Brownian motion and the discussion preceeding the corollary to note that the quadratic variation of M with respect to Q is the same as the quadratic variation with respect to P. Since M is continuous local Q-martingale, it follows from Levy's Theorem 14.58 that $M - \int Z_s^{-1} d[Z,M]_s$ is in fact a Q-Brownian motion. $\square$

In some ways Theorem 14.63 is a deceptively clean result. In applications it is common that one is not given the measure Q rather one starts with a nonnegative process Z_t. Two things need to be addressed. First is that it is often easy to see that Z is a local martingale (e.g. by expressing Z as a stochastic integral) but the hypotheses of the theorem require that it is a true martingale. Therefore one should spend some time developing conditions that allow one to conclude that a nonnegative local martingale is a martingale; there are no necessary and sufficient conditions known but there are some useful sufficient conditions. The second, more subtle, issue is constructing the measure Q from the given nonnegative martingale Z_t. As in Lemma 14.62 this is easy if Z_t is uniformly integrable but in many important applications uniform integrability of Z_t will not hold. As it turns out, the existence of a Q such that $Q = Z_t \cdot P$ on every $\mathcal{F}_t$ is not guaranteed and depends on the properties of the underlying filtration $\mathcal{F}$. In particular, the usual conditions on $\mathcal{F}$ may be incompatible with the existence of Q. Theorem 14.63 has become such an important tool that this phenomenon is viewed as a deficiency of the usual conditions and has led some authors to propose that the usual conditions be replaced by a different extension procedure that is compatible with Theorem 14.63.

We examine conditions under which a positive local martingale is a martingale. First, we note that every positive continuous local martingale has a logarithm that is a continuous local martingale.

LEMMA 14.65. *A continuous process $Z > 0$ is a local martingale if and only if there exists a continuous local martingale M such that*

$$Z_t = \mathcal{E}(M)_t \equiv e^{M_t - \frac{1}{2}[M]_t} \text{ for all } t \geq 0$$

Such an M is almost surely unique and satifies $[M,N]_t = \int_0^t Z_s^{-1}\,[Z,N]_s$ for any continuous local martingale N.

PROOF. Suppose that M is a continuous local martingale then apply Itô's Lemma to the continuous semimartingale $M - \frac{1}{2}[M]$ to see

$$\mathcal{E}(M)_t = e^{M_0} + \int_0^t \mathcal{E}(M)\,d(M - \frac{1}{2}M) + \frac{1}{2}\int_0^t \mathcal{E}(M)_s\,d[M]_s = e^{M_0} + \int_0^t \mathcal{E}(M)\,dM$$

which shows that $\mathcal{E}(M)$ is a stochastic integral hence a continuous local martingale.

If we assume that $Z > 0$ is a continuous local martingale then again apply Itô's Lemma, Lemma 14.14 and the defining property of stochastic integrals to see (TODO: We need the extension to functions defined on an open subset of $\mathbb{R}^d$) to see

$$\begin{aligned}
\log(Z)_t - \log(Z)_0 &= \int_0^t Z^{-1}\,dZ - \frac{1}{2}\int_0^t Z_s^{-2}\,d[Z]_s \\
&= \int_0^t Z^{-1}\,dZ - \frac{1}{2}\int_0^t Z_s^{-1} d\int_0^s Z_s^{-1} d[Z]_s \\
&= \int_0^t Z^{-1} dZ - \frac{1}{2}\int_0^t Z_s^{-1} d[\int Z^{-1}\,dZ, Z]_s \\
&= \int_0^t Z^{-1} dZ - \frac{1}{2}[\int Z^{-1}\,dZ]_t
\end{aligned}$$

so the result holds with $M_t = \int_0^t Z^{-1}\,dZ$. From this expression for M it follows that for any continuous local martingale N, we have $[M,N]_t = [\int Z^{-1}\,dZ, N]_t = \int_0^t Z_s^{-1}\,d[Z,N]_s$. Uniqueness follows that if M and N are continuous local martingales with $M - \frac{1}{2}[M] = N - \frac{1}{2}[N]$ then we have $M - N = \frac{1}{2}[N] - \frac{1}{2}[M]$ is a continuous local martingale of finite variation hence is almost surely zero by Lemma 14.9. ∎

As a result of Lemma 14.65 we look for conditions on a continuous local martingale M that guarantee that $\mathcal{E}(M)$ is a continuous martingale. The following is a commonly used condition.

Lemma 14.66 (Novikov's Condition). *Let M be a continuous local martingale with $M_0 = 0$ such that $\boldsymbol{E}\left[e^{\frac{1}{2}[M]_t}\right] < \infty$ for all $t \geq 0$ then $\mathcal{E}(M)$ is a martingale. If in addition $\boldsymbol{E}\left[e^{\frac{1}{2}[M]_\infty}\right] < \infty$ then $\mathcal{E}(M)$ is a uniformly integrable martingale.*

Proof. TODO □

Example 14.67. Constructing an example of a Brownian motion with a drift term that can be removed with respect to a filtration generated by the Brownian motion but cannot be removed with respect to the completion of that filtration turns out not to be too hard. Let B_t be a standard Brownian motion

The following theorm clarifies when a deterministic drift term can be removed from a Brownian motion. Historically, this is one of the first results dealing with change of measure. TODO: The result as specified in Kallenberg states that we use the augmented filtration; I'm not sure if the result is true under these circumstances (just consider the limit event $B_t/t \to \mu$ as in the example). I suspect it is true if the filtration is that generated by B (or restricting to an arbitrary finite interval $[0, T]$).

Theorem 14.68 (Cameron-Martin Theorem). *Let $B = (B_1, \dots, B_d)$ be a d-dimensional Brownian motion and let $\mathcal{F}_t$ be the complete filtration generated by B. Let $h : \mathbb{R}_+ \to \mathbb{R}^d$ be a continuous function with $h(0) = 0$ and let P_h be distribution of $B + h$. Then $P_0 \sim P_h$ on $\mathcal{F}_t$ for all $t \geq 0$ if and only $h(t) = \int_0^t f(s)\,ds$ for some $f \in L^2_{loc}$. Moreover, in this case we have $P_h = \mathcal{E}(f \cdot B)_t \cdot P_0$ on $\mathcal{F}_t$.*

Proof. First assume that $P_0 \sim P_h$ on $\mathcal{F}_t$ for all $t \geq 0$. By Lemma 14.62 we know that there exists a P_0-martingale Z with $Z_t > 0$ and $P_h = Z_t \cdot P_0$ on $\mathcal{F}_t$ for all $t \geq 0$. Since Z is a Brownian martingale we may apply the Martingale Representation Theorem 14.57 to conclude that Z is almost surely continuous and therefore by Lemma 14.65 we may write $Z = \mathcal{E}(M)$ for some continuous local P_0-martingale M. Applying the Martingale Representation Theorem to M we know there are almost surely unique processes $V^j \in L(B^j)$ such that $M = M_0 + \sum_{j=1}^d \int V^j\,dB^j$. Note that $V^j \in L(B^j)$ implies $\int_0^t (V^j(s))^2\,ds < \infty$ for all $t \geq 0$ and therefore we have $V^j \in L^2_{loc}$. TODO: Finish □

7. Stochastic Differential Equations

In this section we consider *stochastic differential equations* of the form

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dB_t$$

involving a continuous stochastic process X_t and a Brownian motion B_t. This equation is a traditional short hand for the integral equation

$$X_t = X_0 + \int_0^t b(s, X_s)ds + \int_0^t \sigma(s, X_s)dB_s \tag{18}$$

As it turns out there a couple of different ways to think of a stochastic differential equation based on what one considers as known and what one is trying to solve for. In the first way we consider a probability space and a Brownian motion as given and one tries to find a process adapted to the Brownian motion for which Equation (18) holds; such solutions are referred to as strong solutions. In the second case one assumes that only b and σ are given and one seeks to construct a filtered probability space, a Brownian motion B and an adapted X such that Equation (18) holds; such solutions are called weak solutions. As we will see it is important to examine both notions of solution as there are problems for which weak solutions exists yet strong solutions don't and the existence of weak solution is sufficient, yet there are situations in which strong solutions are necessary.

In order to make sure that the stochastic differential equation makes sense it is required that the stochastic integrals is defined. We first dispense with a preliminary measurability result that tells us conditions under which the candidate integrand is progressive.

LEMMA 14.69. *Let $f : [0, \infty) \times C([0, \infty); \mathbb{R}^d) \to \mathbb{R}$, $\pi_{s,j}(f) = f_j(s)$ for $0 \leq s < \infty$, $1 \leq j \leq d$ and assume that f is progressive for the filtration $\mathcal{G}_t = \sigma(\pi_{s,j}; 0 \leq s \leq t, 1 \leq j \leq d)$. Let X be a continuous $\mathcal{F}$-adapted process in $\mathbb{R}^d$ then $Y_t = f(t, X_t)$ is $\mathcal{F}$-progressive.*

PROOF. This amounts to unwinding the definitions involved. First we recast the property of $\mathcal{F}$-adaptedness in terms of $\mathcal{G}$. X is $\mathcal{F}$-adapted if and only if $\pi_{s,j} \circ f$ is $\mathcal{F}_t$ measurable for all $0 \leq s \leq t$, $1 \leq j \leq d$ if and only if $X^{-1}(\pi_{s,j}^{-1}(A)) \in \mathcal{F}_t$ for all $0 \leq s \leq t$, $1 \leq j \leq d$, $A \in \mathcal{B}(\mathbb{R})$ if and only if $X^{-1}(B) \in \mathcal{F}_t$ for all $0 \leq t < \infty$, $B \in \mathcal{G}_t$; thus X being $\mathcal{F}$-adapted is equivalent to saying that X is $\mathcal{F}_t/\mathcal{G}_t$ for all $0 \leq t < \infty$.

The property of f being progressive means that $f \mid_{[0,t] \times C([0,\infty);\mathbb{R}^d)}$ is $\mathcal{B}([0,t]) \otimes \mathcal{G}_t$ measurable. Thus by writing $Y \mid_{[0,t]\times\Omega} (s, \omega) = f(s, X(\omega))$ we see that $Y \mid_{[0,t]\times\Omega}$ is $\mathcal{B}([0,t]) \otimes \mathcal{F}_t$ measurable for all $0 \leq t < \infty$ i.e. $\mathcal{F}$-progressive. $\square$

7.1. Strong Solutions. The first task at hand is to record a precise definition of a strong solution for a stochastic differential equation.

TODO: State clearly the assumptions on (b, σ); do we want each to be progressive as above?

DEFINITION 14.70. Suppose we are given a probability space $(\Omega, \mathcal{A}, P)$, a Brownian motion B_t with values in $\mathbb{R}^r$, a random vector ξ in $\mathbb{R}^d$ independent of B. Let $\mathcal{F}_t$ be the completion of $\sigma(\xi, B_s; 0 \leq s \leq t)$. We say than continuous stochastic process X_t with values in $\mathbb{R}^d$ is a *strong solution* if

(i) X_t is $\mathcal{F}$-adapted
(ii) $X_0 = \xi$ almost surely
(iii) $\int_0^t |b_i(s, X_s)| \, ds < \infty$ almost surely and $\int_0^t \sigma_{ij}^2(s, X_s) \, ds < \infty$ almost surely for all $i = 1, \dots, d$, $j = 1, \dots, r$ and for all $0 \leq t < \infty$
(iv)

$$X_t = X_0 + \int_0^t b(s, X_s) \, ds + \int_0^t \sigma(s, X_s) \, dB_s \text{ for all } 0 \leq t < \infty \text{ almost surely}$$

When one first learns about differential equations the most important tasks seem to be understanding existence and uniqueness results for solutions of differential equations and learning some techinques for finding closed form solutions for simple differential equations. The agenda here is similar.

The uniqueness theorems for strong solutions are a bit stronger than one might guess at first. The idea is related to the $\mathcal{F}$-adaptedness of a strong solution. By Lemma 2.23 we expect that there is a functional representation of a strong solution X_t in terms of the driving Brownian motion B_t and the initial condition ξ.

The existence of such a functional representation would obviously imply uniqueness. However, it seems reasonable to expect that the functional representation will only depend on the pair (b, σ) which would mean that the uniqueness would hold uniformly over all driving Brownian motions and initial conditions. We record a precise definition here.

DEFINITION 14.71. We say that *strong uniqueness* holds for a pair (b, σ) if for every probability space $(\Omega, \mathcal{A}, P)$, Brownian motion B, initial vector ξ any two strong solutions are indistinguishable.

The pair (b, σ) being locally Lipschitz continuous is sufficient for strong uniqueness to hold.

THEOREM 14.72. *Suppose that for every $n \in \mathbb{N}$ there exists a constant K_n with the property that for every $\|x\| \leq n$ and $\|y\| \leq n$ we have*

$$\sup_{0 \leq t < \infty} \|b(t,x) - b(t,y)\| + \|\sigma(t,x) - \sigma(t,y)\| \leq K_n \|x - y\|$$

then strong uniqueness holds for (b, σ).

We have a need for the following basic tool from the theory of ordinary differential equations.

PROPOSITION 14.73 (Gronwall's Inequality). *Suppose we have $0 \leq g(t) \leq \alpha(t) + \beta \int_0^t g(s)\, ds$ for all $0 \leq t \leq T$ with $\beta \geq 0$ and $\alpha : [0,T] \to \mathbb{R}$ integrable then we have*

$$g(t) \leq \alpha(t) + \beta \int_0^t \alpha(s) e^{\beta(t-s)}\, ds \text{ for all } 0 \leq t \leq T$$

PROOF. We massage the hypothesis a bit by writing $g(t) - \beta \int_0^t g(s)\, ds \leq \alpha(t)$. Multiplying by $e^{-\beta t}$ and using the product rule and the Fundamental Theorem of Calculus we get

$$\frac{d}{dt} e^{-\beta t} \int_0^t g(s)\, ds \leq e^{-\beta t} \alpha(t)$$

Now integrate both sides of the inequality to see that $\int_0^t g(s)\, ds \leq \int_0^t \alpha(s) e^{\beta(t-s)}\, ds$. Plug this inequality back into the original hypothesis to see

$$0 \leq g(t) \leq \alpha(t) + \beta \int_0^t \alpha(s) e^{\beta(t-s)}$$

□

Now we can give the proof of the theorem.

PROOF. Suppose that $(\Omega, \mathcal{A}, P$, B and ξ are given and let X and Y be strong solutions. Let $\tau_n^X = \inf\{t \geq 0 \mid \|X_t\| \geq n\}$, $\tau_n^Y = \inf\{t \geq 0 \mid \|Y_t\| \geq n\}$ and let $\tau_n = \tau_n^X \wedge \tau_n^Y$. By continuity of X on $[0, \infty)$ we know that $\lim_{n \to \infty} \tau_n^X = \infty$ a.s. (if $\liminf_{n \to \infty} \tau_n^X < \infty$ then we get a subsequence $\tau_{n_j}^X \to T < \infty$ with $\left\| X_{\tau_{n_j}^X} \right\| \geq n_j$ contradicting continuity of X at T). Since X and Y are strong solutions we have

$$X_{t \wedge \tau_n} - Y_{t \wedge \tau_n} = \int_0^{t \wedge \tau_n} (b(s, X_s) - b(s, Y_s))\, ds + \int_0^{t \wedge \tau_n} (\sigma(s, X_s) - \sigma(s, Y_s))\, dB_s$$

and by the using the inequality $\|x+y\|^2 \leq 2(\|x\|^2 + \|y\|^2)$, the fact that $\left\| \int f(s)\, ds \right\| \leq \int \|f(s)\|\, ds$, the Cauchy Schwartz formula, the Ito Isometry (Lemma 14.38), $[B^{(i)}, B^{(j)}]_t = \delta_{ij} t$ and the local Lipschitz continuity we get

$$
\begin{aligned}
&\mathbf{E}\left[\|X_{t\wedge\tau_n} - Y_{t\wedge\tau_n}\|^2\right] \\
&\leq 2\mathbf{E}\left[\left\|\int_0^{t\wedge\tau_n} (b(s,X_s) - b(s,Y_s))\, ds\right\|^2\right] + 2\mathbf{E}\left[\left\|\int_0^{t\wedge\tau_n} (\sigma(s,X_s) - \sigma(s,Y_s))\, dB_s\right\|^2\right] \\
&\leq 2\mathbf{E}\left[\left[\int_0^{t\wedge\tau_n} \|b(s,X_s) - b(s,Y_s)\|\, ds\right]^2\right] + 2\sum_{i=1}^d \mathbf{E}\left[\left[\sum_{j=1}^r \int_0^{t\wedge\tau_n} (\sigma_{ij}(s,X_s) - \sigma_{ij}(s,Y_s))\, dB_s^{(j)}\right]^2\right] \\
&\leq 2t\mathbf{E}\left[\int_0^{t\wedge\tau_n} \|b(s,X_s) - b(s,Y_s)\|^2\, ds\right] + 2\sum_{i=1}^d \mathbf{E}\left[\sum_{j=1}^r \int_0^{t\wedge\tau_n} (\sigma_{ij}(s,X_s) - \sigma_{ij}(s,Y_s))^2\, ds\right] \\
&= 2t\mathbf{E}\left[\int_0^{t\wedge\tau_n} \|b(s,X_s) - b(s,Y_s)\|^2\, ds\right] + 2\mathbf{E}\left[\int_0^{t\wedge\tau_n} \|\sigma(s,X_s) - \sigma(s,Y_s)\|^2\, ds\right] \\
&\leq 2(T+1)K_n^2 \mathbf{E}\left[\int_0^{t\wedge\tau_n} \|X_s - Y_s\|^2\, ds\right]
\end{aligned}
$$

Apply the Gronwall Inequality with $g(t) = \|X_{t\wedge\tau_n} - Y_{t\wedge\tau_n}\|^2$ and $\alpha(t) = 0$ to conclude that $\mathbf{E}\left[\|X_{t\wedge\tau_n} - Y_{t\wedge\tau_n}\|^2\right] = 0$ which implies that $X_{t\wedge\tau_n} = Y_{t\wedge\tau_n}$ a.s. Letting $n \to \infty$ and using the fact that $\lim_{n\to\infty} \tau_n = \infty$ a.s. we see that $X_t = Y_t$ a.s. for all $t \geq 0$. By continuity of X and Y it follows that X and Y are indistinguishable. □

The same proof technique used in demonstrating uniqueness can be used to show a closely related result that is worth calling out.

Proposition 14.74. *Let $(\Omega, \mathcal{A}, P)$, the filtration $\mathcal{F}$ and b and σ be given. Let ξ and η be $\mathcal{F}_0$-measurable and suppose that X is the strong solution with initial condition ξ and Y is the strong solution with initial condition η, then $X = Y$ almost surely on $\{\xi = \eta\}$.*

Proof. Let

$$
\tau = \begin{cases} \infty & \text{if } \xi = \eta \\ 0 & \text{if } \xi \neq \eta \end{cases}
$$

and note that by $\mathcal{F}_0$-measurability of $\{\xi = \eta\}$ it follows that τ is $\mathcal{F}$-optional. Now compute for all $0 \leq T < \infty$ and all $0 \leq t \leq T$ using the same steps as in Theorem 14.72

$$
\begin{aligned}
\mathbf{E}\left[\|(X_t - Y_t)\mathbf{1}_{\xi=\eta}\|^2\right] &= \mathbf{E}\left[\left\|\left(\int_0^t b(s,X_s) - b(s,Y_s)\, ds + \int_0^t \sigma(s,X_s) - \sigma(s,Y_s)\, dB_s\right)\mathbf{1}_{\xi=\eta}\right\|^2\right] \\
&= \mathbf{E}\left[\left\|\int_0^{t\wedge\tau} b(s,X_s) - b(s,Y_s)\, ds + \int_0^{t\wedge\tau} \sigma(s,X_s) - \sigma(s,Y_s)\, dB_s\right\|^2\right] \\
&\leq 2K^2(T+1)\mathbf{E}\left[\int_0^{t\wedge\tau} \|X_s - Y_s\|^2\, ds\right] = 2K^2(T+1)\mathbf{E}\left[\int_0^t \|(X_s - Y_s)\mathbf{1}_{\xi=\eta}\|^2\, ds\right]
\end{aligned}
$$

from which an application of the Gronwall inequality implies that $X_t = Y_t$ almost surely on $\{\xi = \eta\}$. Thus X_t and Y_t are indistinguishable on $\{\xi = \eta\}$. $\square$

Now we turn to proof of existence of strong solutions. Local Lipschitz continuity is not enough for existence; for that we must assume a global Lipschitz property plus a growth condition on the coefficients to make sure that solutions don't blow up in finite time.

THEOREM 14.75. *Suppose that there exists a constant K with the properties that for every $0 \leq t < \infty$, $x, y \in \mathbb{R}^d$ we have*

$$\|b(t,x) - b(t,y)\| + \|\sigma(t,x) - \sigma(t,y)\| \leq K \|x - y\| \tag{19}$$

$$\|b(t,x)\|^2 + \|\sigma(t,x)\|^2 \leq K^2(1 + \|x\|^2) \tag{20}$$

Let $(\Omega, \mathcal{A}, P)$ be a probability space, B a Brownian motion in $\mathbb{R}^r$, ξ be a random vector in $\mathbb{R}^d$ independent of B such that

$$\boldsymbol{E}\left[\|\xi\|^2\right] < \infty$$

and $\mathcal{F}_t$ be the completion of $\sigma(\xi, B_s; 0 \leq s \leq t)$. The there exists a continuous $\mathcal{F}$-adapted process X which is a strong solution. In addition it follows that for every $T > 0$ there exists a constant C depending only on K and T such that for all $0 \leq t \leq T$,

$$\boldsymbol{E}\left[\|X_t\|^2\right] \leq C(1 + \boldsymbol{E}\left[\|\xi\|^2\right])e^{Ct} \tag{21}$$

PROOF. The approach is to define an approximation scheme mimicing Picard iteration using for ordinary differential equations.

TODO: What about $\mathcal{F}$-adaptedness (this is more or less obvious but needs to be called out)

Let $X_t^{(0)} = \xi$ and then define for $k \in \mathbb{N}$

$$X_t^{(k)} = \xi + \int_0^t b(s, X_s^{(k-1)})\, ds + \int_0^t \sigma(s, X_s^{(k-1)})\, dB_s$$

Our first task is to show that the $X^{(k)}$ are well defined and provide some estimates on them. In order for the integrals to exist we observe that

$$\left\| \int_0^t b(s, X_s^{(k)})\, ds \right\| \leq \int_0^t \left\| b(s, X_s^{(k)}) \right\| ds \leq \left(t \int_0^t \left\| b(s, X_s^{(k)}) \right\|^2 ds \right)^{1/2} \leq \left(Kt \int_0^t (1 + \left\| X_s^{(k)} \right\|^2)\, ds \right)^{1/2}$$

and

$$\int_0^t \left\| \sigma(s, X_s^{(k)}) \right\|^2 ds \leq K^2 \int_0^t (1 + \left\| X_s^{(k)} \right\|^2)\, ds$$

and therefore the integrals will exist if we can show that $\sup_{0 \leq t \leq T} \mathbf{E}\left[\left\| X_t^{(k)} \right\|^2 \right] < \infty$ for all $T \geq 0$.

The proof of this is by induction. For $k = 0$ and $T \geq 0$, $\sup_{0 \leq t \leq T} \mathbf{E}\left[\left\| X_t^{(0)} \right\|^2 \right] = \mathbf{E}\left[\|\xi\|^2\right] < \infty$. The induction step follows using estimates very similar to those in the proof of uniqueness. Suppose $\sup_{0\leq t\leq T} \mathbf{E}\left[\left\|X_t^{(k)}\right\|^2\right] < \infty$ and let $0 \leq t \leq T$,

$$
\begin{aligned}
&\mathbf{E}\left[\left\|X_t^{(k+1)}\right\|^2\right] \\
&\leq 4\mathbf{E}\left[\|\xi\|^2\right] + 4\mathbf{E}\left[\left\|\int_0^t b(s, X_s^{(k)})\, ds\right\|^2\right] + 4\mathbf{E}\left[\left\|\int_0^t \sigma(s, X_s^{(k)})\, dB_s\right\|^2\right] \\
&\leq 4\mathbf{E}\left[\|\xi\|^2\right] + 4\mathbf{E}\left[\left[\int_0^t \left\|b(s, X_s^{(k)})\right\|\, ds\right]^2\right] + 4\sum_{i=1}^d \mathbf{E}\left[\left[\sum_{j=1}^r \int_0^t \sigma_{ij}(s, X_s^{(k)})\, dB_s^{(j)}\right]^2\right] \\
&\leq 4\mathbf{E}\left[\|\xi\|^2\right] + 4t\mathbf{E}\left[\int_0^t \left\|b(s, X_s^{(k)})\right\|^2\, ds\right] + 4\sum_{i=1}^d\sum_{j=1}^r \mathbf{E}\left[\int_0^t \sigma_{ij}^2(s, X_s^{(k)})\, ds\right] \\
&= 4\mathbf{E}\left[\|\xi\|^2\right] + 4t\mathbf{E}\left[\int_0^t \left\|b(s, X_s^{(k)})\right\|^2\, ds\right] + 4\mathbf{E}\left[\int_0^t \left\|\sigma(s, X_s^{(k)})\right\|^2\, ds\right] \\
&\leq 4\mathbf{E}\left[\|\xi\|^2\right] + 4(T+1)K^2\mathbf{E}\left[\int_0^t (1 + \left\|X_s^{(k)}\right\|^2)\, ds\right] \\
&\leq C(1 + \mathbf{E}\left[\xi\right]^2 + \int_0^t \mathbf{E}\left[\left\|X^{(k)}\right\|_s^2\right]\, ds) \\
&\leq C(1 + \mathbf{E}\left[\xi\right]^2 + \int_0^T \mathbf{E}\left[\left\|X^{(k)}\right\|_s^2\right]\, ds)
\end{aligned}
$$

where we have defined $C = 4(T+1)K^2 \vee 4(T+1)TK^2 \vee 4$ which gives us the result.

The estimate that we have just proven can also give a nice estimate of $\mathbf{E}\left[\left\|X_t^{(k)}\right\|^2\right]$ in terms of only C and ξ. Noting that the bounding constant C depends only on K and T, we can iterate the estimate.

Claim 14.75.1. For all $k \in \mathbb{Z}_+$ and $0 \leq t \leq T$, we have

$$
\mathbf{E}\left[\left\|X_t^{(k)}\right\|^2\right] \leq C(1 + \mathbf{E}\left[\|\xi\|^2\right])(1 + Ct + \cdots + \frac{C^k t^k}{k!})
$$

The proof of the claim is a simple induction, noting that it is trivially true for $k = 0$ given that $C > 1$. Now we just apply the estimate above and use the induction hypothesis to see that

$$\begin{aligned}
&\mathbf{E}\left[\left\|X_t^{(k+1)}\right\|^2\right] \\
&\leq C(1+\mathbf{E}[\xi]^2) + C\int_0^t \mathbf{E}\left[\left\|X^{(k)}\right\|_s^2\right] ds \\
&\leq C(1+\mathbf{E}[\xi]^2) + C^2(1+\mathbf{E}[\xi]^2)\int_0^t (1+Cs+\cdots+\frac{C^k t^k}{k!})\, ds \\
&= C(1+\mathbf{E}[\xi]^2) + C^2(1+\mathbf{E}[\xi]^2)(t+\frac{Ct^2}{2!}+\cdots+\frac{C^k t^{k+1}}{(k+1)!}) \\
&= C(1+\mathbf{E}[\xi]^2)(1+Ct+\frac{C^2t^2}{2!}+\cdots+\frac{C^{k+1}t^{k+1}}{(k+1)!})
\end{aligned}$$

From the previous claim we conclude that $\mathbf{E}\left[\left\|X_t^{(k)}\right\|^2\right] \leq C(1+\mathbf{E}[\xi]^2)e^{Ct}$ for all $0 \leq t \leq T$ and $k \in \mathbb{Z}_+$ where the constant C depends only on K and T.

Claim 14.75.2. The sequence $X^{(k)}$ is almost surely Cauchy on every interval $[0, T]$.

We need some estimates so that we can apply Borel Cantelli. Pick $k \in \mathbb{Z}_+$ and consider the difference

$$X_t^{(k+1)} - X_t^{(k)} = \int_0^t (b(s, X_s^{(k)}) - b(s, X_s^{(k-1)}))\, ds + \int_0^t (\sigma(s, X_s^{(k)}) - \sigma(s, X_s^{(k-1)}))\, dB_s$$

As for the first term here by the Lipschitz continuity assumption (19) and Cauchy Schwartz we get

$$\begin{aligned}
\sup_{0\leq s\leq t}\left\|\int_0^s (b(u, X_u^{(k)}) - b(u, X_u^{(k-1)}))\, du\right\|^2 &= \left\|\int_0^t (b(u, X_u^{(k)}) - b(u, X_u^{(k-1)}))\, du\right\|^2 \\
&\leq \left[\int_0^t \left\|b(s, X_s^{(k)}) - b(s, X_s^{(k-1)})\right\| ds\right]^2 \\
&\leq K^2 t \int_0^t \left\|X_s^{(k)} - X_s^{(k-1)}\right\|^2 ds
\end{aligned}$$

For the second term we have to work harder. Since $\int_0^t (\sigma(s, X_s^{(k)}) - \sigma(s, X_s^{(k-1)}))\, dB_s$ is a vector of martingales, we can apply the vector-valued BDG inequality (Corollary 14.33), the defining property of stochastic integrals and the fact that $[B^{(i)}, B^{(j)}]_t =$

$t\delta_{ij}$ to get

$$
\begin{aligned}
&\mathbf{E}\left[\sup_{0\le s\le t}\left\|\int_0^s (\sigma(u,X_u^{(k)})-\sigma(u,X_u^{(k-1)}))\,dB_u\right\|^2\right]\\
&\le d_2\mathbf{E}\left[\sum_{i=1}^d\left[\sum_{j=1}^r\int(\sigma_{ij}(s,X_s^{(k)})-\sigma_{ij}(s,X_s^{(k-1)}))\,dB_s^{(j)}\right]_t\right]\\
&= d_2\mathbf{E}\left[\sum_{i=1}^d\sum_{j=1}^r\int_0^t(\sigma_{ij}(s,X_s^{(k)})-\sigma_{ij}(s,X_s^{(k-1)}))^2\,ds\right]\\
&= d_2\mathbf{E}\left[\int_0^t\left\|\sigma(s,X_s^{(k)})-\sigma(s,X_s^{(k-1)})\right\|^2\,ds\right]\\
&\le d_2K^2\mathbf{E}\left[\int_0^t\left\|X_s^{(k)}-X_s^{(k-1)}\right\|^2\,ds\right]
\end{aligned}
$$

Putting these two estimates together we get for $0\le t\le T$,

$$
\begin{aligned}
\mathbf{E}\left[\sup_{0\le s\le t}\left\|X_s^{(k+1)}-X_s^{(k)}\right\|^2\right] &\le 2\mathbf{E}\left[\sup_{0\le s\le t}\left\|\int_0^t(b(u,X_u^{(k)})-b(u,X_u^{(k-1)}))\,du\right\|^2\right]+\\
&\qquad 2\mathbf{E}\left[\sup_{0\le s\le t}\left\|\int_0^s(\sigma(u,X_u^{(k)})-\sigma(u,X_u^{(k-1)}))\,dB_u\right\|^2\right]\\
&\le 2K^2(T+d_2)\int_0^t\left\|X_s^{(k)}-X_s^{(k-1)}\right\|^2\,ds
\end{aligned}
$$

where none of the constants depend on k. We iterate the estimate to yield

$$
\begin{aligned}
\mathbf{E}\left[\sup_{0\le s\le t}\left\|X_s^{(k+1)}-X_s^{(k)}\right\|^2\right] &\le \frac{(2K^2(T+d_2))^k}{k!}\sup_{0\le t\le T}\mathbf{E}\left[\left\|X_t^{(1)}-\xi\right\|^2\right]\\
&\le 2C(1+\mathbf{E}\left[\|\xi\|^2\right])e^{CT}\frac{(2K^2(T+d_2))^k}{k!}
\end{aligned}
$$

Now we can apply the Markov Inequality to see that

$$
\begin{aligned}
\mathbf{P}\{\sup_{0\le t\le T}\left\|X_t^{(k+1)}-X_t^{(k)}\right\|\ge\frac{1}{2^{k+1}}\} &= \mathbf{P}\{\sup_{0\le t\le T}\left\|X_t^{(k+1)}-X_t^{(k)}\right\|^2\ge\frac{1}{4^{k+1}}\}\\
&\le 4^{k+1}\mathbf{E}\left[\sup_{0\le t\le T}\left\|X_t^{(k+1)}-X_t^{(k)}\right\|^2\right]\\
&\le 8C(1+\mathbf{E}\left[\|\xi\|^2\right])e^{CT}\frac{(8K^2(T+d_2))^k}{k!}
\end{aligned}
$$

thus $\sum_{k=0}^\infty\mathbf{P}\{\sup_{0\le t\le T}\left\|X_t^{(k+1)}-X_t^{(k)}\right\|\ge\frac{1}{2^{k+1}}\}=8C(1+\mathbf{E}\left[\|\xi\|^2\right])e^{CT+8K^2(T+d_2)}<\infty$ and we may apply the Borel Cantelli Theorem 4.23 to conclude that with probability one there exists a random N such that $\sup_{0\le t\le T}\left\|X_t^{(k+1)}-X_t^{(k)}\right\|<\frac{1}{2^{k+1}}$

for all $k \geq N$. Thus it also follows that for all $k \geq N$ and $m \geq 1$,

$$\begin{aligned}\sup_{0\leq t\leq T}\left\|X_t^{(k+m)} - X_t^{(k)}\right\| &\leq \sum_{j=0}^{m-1} \sup_{0\leq t\leq T}\left\|X_t^{(k+j+1)} - X_t^{(k+j)}\right\| \\ &\leq \sum_{j=0}^{m-1} \frac{1}{2^{k+j+1}} \leq \frac{1}{2^k}\end{aligned}$$

and thus almost surely $X_t^{(k)}$ is uniformly Cauchy on $[0,T]$. Taking the intersection of a countable number of almost sure events, we see that almost surely $X_t^{(k)}$ is uniformly Cauchy on $[0,T]$ for $T \geq 0$. By completeness of $C([0,\infty);\mathbb{R}^d)$ we conclude that there is a continuous process X such that $X^{(k)} \stackrel{a.s.}{\to} X$ in $C([0,\infty);\mathbb{R}^d)$.

Claim 14.75.3. X is a strong solution.

Since $X^{(k)}$ is $\mathcal{F}$-adapted for every $k \in \mathbb{Z}_+$, it follows that X_t is $\mathcal{F}$-adapted. Since $X_0^{(k)} = \xi$ for all $k \in \mathbb{Z}_+$ and $X_0^{(k)} \stackrel{a.s.}{\to} X_0$ it follows that $X_0 = \xi$ almost surely. In addition, by Cauchy Schwartz, (20) and (21) we get

$$\begin{aligned}&\mathbf{P}\{\int_0^t |b_i(s,X_s)| + |\sigma_{ij}(s,X_s)|^2 \, ds < \infty\} \geq \mathbf{P}\{t\int_0^t \|b(s,X_s)\|^2 + \|\sigma(s,X_s)\|^2 \, ds < \infty\} \\ &\geq \mathbf{P}\{(t+1)\int_0^t (1+\|X_s\|^2)\,ds < \infty\} = \mathbf{P}\{\int_0^t \|X_s\|^2 \, ds < \infty\} \\ &\geq \mathbf{P}\{C(1+\mathbf{E}\left[\|\xi\|^2\right])\int_0^t e^{Cs}\,ds < \infty\} = 1\end{aligned}$$

By this we know that $\int_0^t b(s,X_s)\,ds$ and $\int_0^t \sigma(s,X_s)\,dB_s$ are defined and we claim that X solves the stochastic differential equation

$$X_t = X_0 + \int_0^t b(s,X_s)\,ds + \int_0^t \sigma(s,X_s)\,dB_s \text{ almost surely}$$

where we have just shown that all of the terms in the equation are well defined. Writing

$$X_t = \lim_{k\to\infty} X_t^{(k+1)} = \lim_{k\to\infty}\left[\xi + \int_0^t b(s,X_s^{(k)})\,ds + \int_0^t \sigma(s,X_s^{(k)})\,dB_s\right]$$

and using the fact that $X_0 = \xi$ a.s. it suffices to show that $\int_0^t b(s,X_s^{(k)})\,ds \stackrel{a.s.}{\to} \int_0^t b(s,X_s)\,ds$ and $\int_0^t \sigma(s,X_s^{(k)})\,dB_s \stackrel{a.s.}{\to} \int_0^t \sigma(s,X_s)\,dB_s$; in fact it suffices to show that some subsequence of converges almost surely so it suffices to show that the integrals converge in probability or some L^p.

The almost sure convergence of the first integral follows from the Lipschitz continuity (19) and the uniform convergence of $X^{(k)}$ on intervals $[0,t]$.

$$\begin{aligned}\left\|\int_0^t b(s,X_s) - b(s,X_s^{(k)})\,ds\right\| &\leq \int_0^t \left\|b(s,X_s) - b(s,X_s^{(k)})\right\| ds \\ &\leq K\int_0^t \left\|X_s - X_s^{(k)}\right\| ds \leq Kt \sup_{0\leq s\leq t}\left\|X_s - X_s^{(k)}\right\| \stackrel{a.s.}{\to} 0\end{aligned}$$

For the second integral we show L^2 convergence. we first note that for every fixed t, $X_t^k \stackrel{L^2}{\to} X_t$ as well as converging almost surely. To see that simply note that for all $k \in \mathbb{Z}_+$ and $m \in \mathbb{N}$ with $m > k$,

$$\begin{aligned}
\mathbf{E}\left[\left\|X_t^{(m)} - X_t^{(k)}\right\|^2\right] &\leq \sum_{j=k}^{m-1} \mathbf{E}\left[\left\|X_t^{(j+1)} - X_t^{(j)}\right\|^2\right] \\
&\leq \sum_{j=0}^{m-1} \mathbf{E}\left[\sup_{0\leq s\leq t}\left\|X_s^{(j+1)} - X_s^{(j)}\right\|^2\right] \\
&\leq 8C(1+\mathbf{E}\left[\|\xi\|^2\right])e^{CT}\sum_{j=k}^{m-1}\frac{(8K^2(T+d_2))^j}{j!} \\
&\leq 8C(1+\mathbf{E}\left[\|\xi\|^2\right])e^{CT}\sum_{j=k}^{\infty}\frac{(8K^2(T+d_2))^j}{j!} < \infty
\end{aligned}$$

Moreover we know that $\lim_{k\to\infty}\sum_{j=k}^{\infty}\frac{(8K^2(T+d_2))^j}{j!} = 0$ and therefore we have shown that $X_t^{(k)}$ is Cauchy in L^2. Since we know that $X_t^{(k)} \stackrel{a.s.}{\to} X_t$ it follows that $X_t^{(k)} \stackrel{L^2}{\to} X_t$.

Now we can estimate and use Fubini

$$\begin{aligned}
\mathbf{E}\left[\left\|\int_0^t \sigma(s, X_s) - \sigma(s, X_s^{(k)})\, dB_s\right\|^2\right] &= \mathbf{E}\left[\sum_{i=1}^{d}\left[\sum_{j=1}^{r}\int_0^t \sigma_{ij}(s, X_s) - \sigma_i j(s, X_s^{(k)})\, dB_s^{(j)}\right]^2\right] \\
&= \mathbf{E}\left[\int_0^t \left\|\sigma(s, X_s) - \sigma(s, X_s^{(k)})\right\|^2\, ds\right] \\
&\leq K^2\int_0^t \mathbf{E}\left[\left\|X_s - X_s^{(k)}\right\|^2\right]\, ds
\end{aligned}$$

In addition, for all $0 \leq s \leq t$, by (21)

$$\mathbf{E}\left[\left\|X_s - X_s^{(k)}\right\|^2\right] \leq 2\mathbf{E}\left[\|X_s\|^2\right] + 2\mathbf{E}\left[\left\|X_s^{(k)}\right\|^2\right] \leq 2C(1+\mathbf{E}\left[\|\xi\|^2\right])e^{Ct}$$

where the upper bound is integrable, thus by the fact that $X_s^k \stackrel{L^2}{\to} X_s$ for all $0 \leq s \leq t$ and Dominated Convergence

$$\lim_{k\to\infty}\mathbf{E}\left[\left\|\int_0^t \sigma(s, X_s) - \sigma(s, X_s^{(k)})\, dB_s\right\|^2\right] \leq K^2 \lim_{k\to\infty}\int_0^t \mathbf{E}\left[\left\|X_s - X_s^{(k)}\right\|^2\right]\, ds = 0$$

We now use a truncation argument to show that we can drop the assumption $\mathbf{E}\left[\|\xi\|^2\right] < \infty$. Let ξ be random vector in $\mathbb{R}^d$. For every $k \in \mathbb{N}$, $\xi\mathbf{1}_{\|\xi\|\leq k}$ is bounded so in particular L^2; let $X^{(k)}$ be strong solution with initial condition $\xi\mathbf{1}_{\|\xi\|\leq k}$. Let $m \in \mathbb{N}$ such that $k < m$, then because $\xi\mathbf{1}_{\|\xi\|\leq k} = \xi\mathbf{1}_{\xi\leq m}$ on $\{\|\xi\| \leq k\}$ so we may apply Proposition 14.74 to conclude that $X^{(k)}$ and $X^{(m)}$ are indistinguishable on $\|\xi\| \leq k$. Because of this we may take the limit $X = \lim_{k\to\infty} X^{(k)}$ with the property that $X = X^{(k)}$ almost surely on $\|\xi\| \leq k$.

It remains to show that X is a strong solution with initial condition ξ. $\mathcal{F}$-adaptedness of X follows immediately from that of $X^{(k)}$ as does the condition

$X_0 = \xi$ almost surely. For a fixed $t \geq 0$ we see

$$\begin{aligned}
\mathbf{P}\{\int_0^t \|b(s, X_s)\| \, ds = \infty\} &= \sum_{k=1}^\infty \mathbf{P}\{\int_0^t \|b(s, X_s)\| \, ds = \infty; k-1 < \|\xi\| \leq k\} \\
&= \sum_{k=1}^\infty \mathbf{P}\{\int_0^t \left\|b(s, X^{(k)}_s)\right\| \, ds = \infty; k-1 < \|\xi\| \leq k\} \\
&\leq \sum_{k=1}^\infty \mathbf{P}\{\int_0^t \left\|b(s, X^{(k)}_s)\right\| \, ds = \infty\} = 0
\end{aligned}$$

and similarly for $\int_0^t \|\sigma(s, X_s)\|^2 \, ds$.

Lastly note that $\{\xi \leq k\}$ is $\mathcal{F}_0$-measurable hence

$$\tau_k = \begin{cases} \infty & \text{if } \xi \leq k \\ 0 & \text{if } \xi > k \end{cases}$$

is an $\mathcal{F}$-optional time. Now we apply Lemma 14.38 to see that almost surely

$$\begin{aligned}
X_t \mathbf{1}_{\|\xi\| \leq k} &= X^{(k)}_t \mathbf{1}_{\|\xi\| \leq k} \\
&= \left(\xi + \int_0^t b(s, X^{(k)}_s) \, ds + \int_0^t \sigma(s, X^{(k)}_s) \, dB_s\right) \mathbf{1}_{\|\xi\| \leq k} \\
&= \mathbf{1}_{\|\xi\| \leq k} \xi + \int_0^{t \wedge \tau_k} b(s, X^{(k)}_s) \, ds + \int_0^{t \wedge \tau_k} \sigma(s, X^{(k)}_s) \, dB_s \\
&= \mathbf{1}_{\|\xi\| \leq k} \xi + \int_0^{t \wedge \tau_k} b(s, X_s) \, ds + \int_0^{t \wedge \tau_k} \sigma(s, X_s) \, dB_s \\
&= \left(\xi + \int_0^t b(s, X_s) \, ds + \int_0^t \sigma(s, X_s) \, dB_s\right) \mathbf{1}_{\|\xi\| \leq k}
\end{aligned}$$

Taking the intersection of countably many almost sure events and using the fact that $\|\xi\| < \infty$ almost surely we conclude that $X_t = \xi + \int_0^t b(s, X_s) \, ds + \int_0^t \sigma(s, X_s) \, dB_s$ almost surely for fixed $0 \leq t < \infty$. By Proposition 14.21 we conclude that X and $\xi + \int_0^t b(s, X_s) \, ds + \int_0^t \sigma(s, X_s) \, dB_s$ are indistinguishable. □

TODO: We have presented existence and uniqueness proofs for the "standard" case of coefficients that are functions of $\mathbb{R}_+ \times \mathbb{R}^d$. We already mentioned the case in which the coefficients are functions of $\mathbb{R}_+ \times C([0, \infty); \mathbb{R}^d)$ (in the context of weak solutions, Karatzas and Shreve call this the functional case). Kallenberg presents results in these terms so it is worth understanding them. The functional case comes up again in the context of weak solutions. It sounds like Rogers and Williams also covers the functional case but most other authors don't seem to bother in the strong case.

7.2. Weak Solutions and Martingale Problems. TODO: Note some of the motivations for weak solutions. Apparently the original motivation for stochatic differential equations came from the desire to come up with a probabilistic proof of the existence of a diffusion process with a given b and σ. For that particular application weak solutions are perfectly sufficient. What about stochastic control applications?

Covariations of solutions to stochastic differential equations. Suppose $b_i : [0,\infty) \times C([0,\infty),\mathbb{R}^d) \to \mathbb{R}$ and $\sigma_{ij} : [0,\infty) \times C([0,\infty),\mathbb{R}^d) \to \mathbb{R}$ are progressive. Suppose that $X_t = X_0 + \int_0^t b(s,X)\,ds + \int_0^t \sigma(s,X)\,dB_s$, then for all $1 \leq i,j \leq d$ we have

$$\begin{aligned}
[X^{(i)}, X^{(j)}]_t &= \sum_{k=1}^r \sum_{l=1}^r [\int_0^t \sigma_{ik}(s,X)\,dB_s^{(k)}, \int_0^t \sigma_{jl}(s,X)\,dB_s^{(l)}]_t \\
&= \sum_{k=1}^r \sum_{l=1}^r \int_0^t \sigma_{ik}(s,X)\sigma_{jl}(s,X)\,d[B^{(k)}, B^{(l)}]_t \\
&= \sum_{k=1}^r \int_0^t \sigma_{ik}(s,X)\sigma_{jk}(s,X)\,dt = \int_0^t a_{ij}(s,X)\,dt
\end{aligned}$$

where recall we have defined the diffusion matrix $a(t,X_t) = \sigma(t,X_t)\sigma^T(t,X_t)$.

TODO: Figure out what the deal is with filtrations and martingale problems; specifically in the canonical process scenario do we just use the natural filtration on $C([0,\infty);\mathbb{R}^d)$ or do we pass to the usual augmentation. Under what circumstances does it not matter (Karatzas and Shreve have some info on this).

Here are computations related to the martingale problem formulation of weak solutions.

Note that if a process solves the local martingale problem in the restricted sense that M^f is a local $\mathcal{F}$-martingale for $f \in C_0^2(\mathbb{R}^d)$ then in fact it is a solution to the local martingale problem. To see this suppose that $f \in C^2(\mathbb{R}^d)$ and for each $n \in \mathbb{N}$ find a $f_n \in C_0^2(\mathbb{R}^d)$ such that $f = f_n$ on $\{\|x\| < n+1\}$. If follows that the first and second derivatives of f and f_n also agree on $\{\|x\| \leq n\}$. Now define $\tau_n = \inf\{t \geq 0 \mid \|X_t\| \geq n\}$. By continuity of X_t and closedness of $\{\|x\| \geq n\}$ we know by Lemma 9.70 that τ_n is an optional time. Moreover by almost sure continuity of X_t we know that $\tau_n \uparrow \infty$ almost surely. Note that for all $0 \leq t < \infty$

$$\begin{aligned}
(M^f)_{\tau_n \wedge t} &= f(X_{\tau_n \wedge t}) - f(X_0) - \int_0^t \sum_{i=1}^d b_i(s,X) \frac{\partial f}{\partial x_i}(X_{\tau_n \wedge t} + \frac{1}{2} \sum_{i,j=1}^d a_{ij}(s,X) \frac{\partial^2 f}{\partial x_i \partial x_j}(X_{\tau_n \wedge t}\,ds \\
&= f_n(X_{\tau_n \wedge t}) - f_n(X_0) - \int_0^t \sum_{i=1}^d b_i(s,X) \frac{\partial f_n}{\partial x_i}(X_{\tau_n \wedge t} + \frac{1}{2} \sum_{i,j=1}^d a_{ij}(s,X) \frac{\partial^2 f_n}{\partial x_i \partial x_j}(X_{\tau_n \wedge t}\,ds \\
&= M^{f_n}_{\tau_n \wedge t}
\end{aligned}$$

and therefore since M^{f_n} it follows that $(M^{f_n})^{\tau_n} = (M^f)^{\tau_n}$ is a local $\mathcal{F}$-martingale for all $n \in \mathbb{N}$. By Lemma 14.8 it follows that M^f is a local martingale.

An immediate corollary of the above observation is the reassuring fact that an X that solves the martingale problem also solves the corresponding local martingale problem.

Suppose that $X_t = X_0 + \int_0^t b(s,X)\,ds + \int_0^t \sigma(s,X)\,dB_s$. Thus we have By Ito's Formula

$$f(t,X_t) - f(0,X_0)$$

$$= \int_0^t \frac{\partial f}{\partial t}(s,X_s)\,ds + \sum_{i=1}^d \int_0^t \frac{\partial f}{\partial x_i}(s,X_s)\,dX_s^{(i)} + \frac{1}{2}\sum_{i=1}^d\sum_{j=1}^d \int_0^t \frac{\partial^2 f}{\partial x_i \partial x_j}(s,X_s)\,d[X^{(i)},X^{(j)}]_s$$

$$= \int_0^t \frac{\partial f}{\partial t}(s,X_s)\,ds + \sum_{i=1}^d \int_0^t b_i(s,X)\frac{\partial f}{\partial x_i}(s,X_s)\,ds + \sum_{i=1}^d\sum_{i=1}^r \int_0^t \sigma_{ij}(s,X)\frac{\partial f}{\partial x_i}(s,X_s)\,dB_s^{(j)} +$$

$$\frac{1}{2}\sum_{i=1}^d\sum_{j=1}^d \int_0^t a_{ij}(s,X)\frac{\partial^2 f}{\partial x_i \partial x_j}(s,X_s)\,ds$$

from which we see that

$$M_t^f =$$

$$f(t,X_t) - f(0,X_0) - \int_0^t \frac{\partial f}{\partial t}(s,X_s) - \sum_{i=1}^d b_i(s,X)\frac{\partial f}{\partial x_i}(s,X_s) - \frac{1}{2}\sum_{i=1}^d\sum_{j=1}^d a_{ij}(s,X)\frac{\partial^2 f}{\partial x_i \partial x_j}(s,X_s)\,ds$$

$$= \sum_{i=1}^d\sum_{j=1}^r \int_0^t \sigma_{ij}(s,X)\frac{\partial f}{\partial x_i}(s,X_s)\,dB_s^{(j)}$$

is a local martingale. Moreover from this formula, given f and g we can find the quadratic covariation of M^f and M^g by explicit calculation

$$[M^f,M^g]_t = \sum_{i=1}^d\sum_{j=1}^r\sum_{k=1}^d\sum_{l=1}^r [\int \sigma_{ij}(s,X)\frac{\partial f}{\partial x_i}(s,X_s)\,dB_s^{(j)}, \int \sigma_{kl}(s,X)\frac{\partial g}{\partial x_k}(s,X_s)\,dB_s^{(l)}]_t$$

$$\sum_{i=1}^d\sum_{j=1}^r\sum_{k=1}^d\sum_{l=1}^r \int_0^t \sigma_{ij}(s,X)\sigma_{kl}(s,X)\frac{\partial f}{\partial x_i}(s,X_s)\frac{\partial g}{\partial x_k}(s,X_s)\,d[B^{(j)},B^{(l)}]_s$$

$$\sum_{i=1}^d\sum_{j=1}^r\sum_{k=1}^d \int_0^t \sigma_{ij}(s,X)\sigma_{kj}(s,X)\frac{\partial f}{\partial x_i}(s,X_s)\frac{\partial g}{\partial x_k}(s,X_s)\,ds$$

$$\sum_{i=1}^d\sum_{k=1}^d \int_0^t a_{ik}(s,X)\frac{\partial f}{\partial x_i}(s,X_s)\frac{\partial g}{\partial x_k}(s,X_s)\,ds$$

Now we want to go the other way; show that if we have a X solves the local martingale problem for a suitable class of functions f then X is a weak solution to the SDE. Actually it suffices to show that M^f is a local martingale for $f(x) = x_i$ and $f(x) = x_i x_j$ with $1 \leq i,j \leq d$.

For each $1 \leq i \leq d$ let $f(x) = x_i$ and define

$$M_t^{(i)} = M_t^f = X_t^{(i)} - X_0^{(i)} - \int_0^t b_i(s,X)\,ds$$

so that $M^{(i)}$ is a continuous local martingale by hypothesis.

Claim 14.75.4. $[M^{(i)},M^{(j)}]_t = \int_0^t a_{ij}(s,X)\,ds$

Let $f(x) = x_i x_j$ so that

$$M_t^{(i,j)} = M_t^f = X_t^{(i)} X_t^{(j)} - X_0^{(i)} X_0^{(j)} - \int_0^t \left(b_i(s,X) X_s^{(j)} + b_j(s,X) X_s^{(i)} + a_{ij}(s,X) \right) \, ds$$

is a continuous local martingale by hypothesis. If we now apply integration by parts (Lemma 14.44) and substitute using the definition of $M^{(i)}$ we get

$$M_t^{(i,j)} =$$
$$\int_0^t X_s^{(i)} \, dX_s^{(j)} + \int_0^t X_s^{(j)} \, dX_s^{(i)} + [X^{(i)}, X^{(j)}]_t - \int_0^t \left(b_i(s,X) X_s^{(j)} + b_j(s,X) X_s^{(i)} + a_{ij}(s,X) \right) \, ds$$
$$\int_0^t X_s^{(i)} \, dM_s^{(j)} + \int_0^t X_s^{(j)} \, dM_s^{(i)} + [M^{(i)}, M^{(j)}]_t - \int_0^t a_{ij}(s,X) \, ds$$

which shows us that $[M^{(i)}, M^{(j)}]_t - \int_0^t a_{ij}(s,X) \, ds$ is a continuous local martingale of bounded variation hence is almost surely zero by Lemma 14.9. The claim is shown.

Now we may apply Theorem 14.59 to conclude that there exists a Brownian motion on an extension of $(\Omega, \mathcal{A}, P, \mathcal{F})$ such that $\int \sigma(s,X) \, dB_s$ and M are indistinguishable. Plugging this into the definition of M we get

$$X_t = X_0 + \int_0^t b(s,X) \, ds + \int_0^t \sigma(s,X) \, dB_s$$

and we are done.

CHAPTER 15

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).

1. Metric Spaces

2. Topological Spaces

LEMMA 15.1. *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 15.2. 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 15.3. 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 15.4. *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 15.1 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 15.1 we have shown that $f^{-1}(V)$ is open. □

DEFINITION 15.5. A *base* of a topology $\mathcal{T}$ at a point $x \in X$ is a collection of sets $\mathcal{B}$ such that for every open set $U \in \mathcal{T}$ such that $x \in U$ there exists a $B \in \mathcal{B}$ such that $x \in B \subset U$. A base of a topology is a collection of sets that is a base at all points $x \in X$.

LEMMA 15.6. *A set $\mathcal{B}$ of sets $B \subset X$ is a base of a topology if and only if for every $x \in X$ there exists $B \in \mathcal{B}$ such that $x \in B$ and for every $A, B \in \mathcal{B}$ and $x \in A \cap B$ there exists $C \in \mathcal{B}$ such that $x \in C \subset A \cap B$.*

PROOF. Suppose $\mathcal{B}$ satisfies the hypothesized conditions and let

$$\tau = \{U \subset X \mid \text{ for every } x \in U \text{ there exists } B \in \mathcal{B} \text{ such that } x \in B \subset U\}$$

It is certainly the case that $\mathcal{B} \subset \tau$ and we claim that τ is a topology. Certainly $\emptyset \in \tau$. Let U_α for $\alpha \in \Lambda$ are sets in τ. Then if $x \in \cup_{\alpha\in\Lambda} U_\alpha$ there exists an $\alpha \in \Lambda$ such that $x \in U_\alpha$ and by hypothesis we pick B such that $x \in B \subset U_\alpha \subset \cup_{\alpha\in\Lambda} U_\alpha$. If $U_1, \dots, U_n \in \tau$ and $x \in U_1 \cap \dots \cap U_n$ then there exists $B_1, \dots, B_n$ such that $x \in B_j \subset U_j$ for $j = 1, \dots, n$ and therefore $x \in B_1 \cap \dots \cap B_n \subset U_1 \cap \dots \cap U_n$. A simple induction on the hypothesis shows that $B_1 \cap \dots \cap B_n \in \mathcal{B}$. Because $\mathcal{B}$ is cover of X we have $X = \cup_{B\in\mathcal{B}} B \in \tau$ and therefore τ is a topology. By the definition of τ it is immediate that $\mathcal{B}$ is a base of the topology. □

DEFINITION 15.7. (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 15.8. *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: Fun fact, there is a non-first countable topological space which is compact but not sequentially compact! In first countable spaces sequential compactness is equivalent to

Separation axioms tells us that we have enough open sets in a topology to distinguish features of the the underlying set (e.g. distinguishing points from points or closed sets from closed sets). Another way of thinking about the size of a topology is by considering the number of continuous functions that the topology allows. The following theorem shows that in normal topological spaces we have enough continuous functions to approximate indicator functions of closed sets.

THEOREM 15.9 (Uryshon's Lemma). *Let X be a topological space, then following are equivalent*
(i) *X is normal*
(ii) *Given a closed set $F \subset X$ and an open neighborhood $F \subset U$ there is an open set V such that $F \subset V \subset \overline{V} \subset U$.*
(iii) *Given disjoint closed sets F and G there exists a continuous function $f : X \to [0,1]$ such that $f \equiv 1$ on F and $f \equiv 0$ on G.*
(iv) *Given a closed set F with an open neighborhood U there is a continuous function f such that $\mathbf{1}_F(x) \leq f(x) \leq \mathbf{1}_U(x)$ for all $x \in X$.*

PROOF. (i) $\implies$ (ii): Since U^c is and $F \cap U^c = \emptyset$ we use normality to find disjoint open sets V and O such that $F \subset V$ and $U^c \subset O$. Note that $\overline{V} \cap U^c = \emptyset$; if $x \in U^c$ then O is an open neighborhood x such that $O \cap V$ which implies $x \notin \overline{V}$. Therefore we have $F \subset V \subset \overline{V} \subset U$.

(ii) $\implies$ (i): Let F and G be closed subsets of X, it follows that G^c is open and $F \subset G^c$. Find an open set V such that $F \subset V \subset \overline{V} \subset G^c$ and observe that if we define $U = \overline{V}^c$ then we have $V \cap U = \emptyset$ and $F \subset V$ and $G \subset U$.

(iii) $\implies$ (iv): Construct continuous $f : X \to [0,1]$ such that f equals 1 on F and f equals 0 on U^c. Clearly $\mathbf{1}_F \leq f$ and $\mathbf{1}_{U^c} \leq 1 - f$. The latter is equivalent to $f \leq \mathbf{1}_U$ since $\mathbf{1}_{U^c} = 1 - \mathbf{1}_U$.

(iv) $\implies$ (iii): Note that $F \subset G^c$ and construct f such that $\mathbf{1}_F \leq f \leq \mathbf{1}_{G^c}$. The first inequality implies that $f \equiv 1$ on F while the second implies that $f \equiv 0$ on $(G^c)^c = G$.

(iii) $\implies$ (i): Given F and G and a continuous function $f : X \to [0,1]$ such that $F \subset f^{-1}(1)$ and $G \subset f^{-1}(0)$, simply define $U = f^{-1}(2/3, 1]$ and $V = f^{-1}[0, 1/3)$ and note that by continuity of f both U and V are open.

(ii) $\implies$ (iv): We construct f as a limit of (discontinuous) indicator functions. Suppose that F and U are given as in the hypothesis in (iv). Define $F_1 = F$ and $U_0 = U$. Using (ii) we find an open neighborhood V such that $F_1 \subset V \subset \overline{V} \subset U$. Define $F_{1/2} = \overline{V}$ and $U_{1/2} = V$ so we may rewrite our inclusions as

$$F_1 \subset U_{1/2} \subset F_{1/2} \subset U_0$$

Now we iterate this construction. To make it clear and to set the notation for the iteration we turn the crank one more time we apply (ii) to the pair $F_1 \subset U_{1/2}$ to construct an open set $U_{3/4}$ and closed set $F_{3/4}$ and to the pair $F_{1/2} \subset U_0$ to construct an open set $U_{1/4}$ and closed set $F_{1/4}$ yielding the inclusions

$$F_1 \subset U_{3/4} \subset F_{3/4} \subset U_{1/2} \subset F_{1/2} \subset U_{1/4} \subset F_{1/4} \subset U_0$$

Now we induct over the dyadic rationals $\mathcal{D} = \{a/2^n \mid a \in \mathbb{N} \text{ and } n \in \mathbb{N}\} \cap (0,1)$ so that we create a sequence of open and closed sets U_q and F_q satisfying

(i) $U_q \subset F_q$ for all $q \in \mathcal{D}$
(i) $F_r \subset U_q$ for all $r, q \in \mathcal{D}$ with $r > q$.

Now let $f(x) = \inf\{q \mid x \in U_q\}$. TODO: Show that f works... □

Theorem 15.10 (Tietze's Extension Theorem). *Let F be a closed subset of a normal topological space, let $a < b$ be real numbers and let $f : F \to [a,b]$ be a continuous function. There exists a continuous function $g : X \to [a,b]$ such that $g \mid_F = f$. If $f : F \to \mathbb{R}$ is a continuous function then there exists a continuous function $g : X \to \mathbb{R}$ such that $g \mid_F = f$.*

Proof. We begin with the case of f with bounded range. We construct g via an iterative procedure. TODO: □

Definition 15.11. 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 15.12. *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 5.41) we know that $g(x) = d(x, F)$ is continuous (in fact Lipschitz) and by Lemma 5.42 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. $\square$

3. Borel Spaces

TODO: The goal of the next set of results is to show that separable complete metric spaces (actually Polish spaces which are those with a separable topology which can be metrized by a complete metric) are Borel.

The following appears in Royden as Theorem 8.11 (with proof delgated to exercises)

Lemma 15.13. *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 15.1 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 $x = z$. $\square$

Theorem 15.14 (Tychonoff's Theorem). *Let I be index set and let $(X_i, \mathcal{T}_i)$ be a topological space for each $i \in I$, the cartesian product $\prod_{i \in I} X_i$ with the product topology is compact.*

Proof. TODO: $\square$

Theorem 15.15. *A Polish space is Borel.*

PROOF. TODO ☐

4. Prohorov's Theorem

The theory of probability measures on separable metric spaces is simpler in many ways than its general counterpart for non-separable metric spaces. The simplicity derives from the fact that a separable metric space is not too far being compact so in a sense doesn't have too many unbounded continuous functions with respect to which probability measures can misbehave.

One way in which to understand the way in which a separable metric space is close to being compact is to consider the real line. Through any number of homeomorphisms, the real line is homeomorhpic to the open unit interval $(0, 1)$. In this way, distances on the real line may be rescaled so as to make the real line bounded. Then by completing the open interval we see that real line is a couple of points away from being compact.

DEFINITION 15.16. Let (S, d) be a metric space then we denote by $U^d(S)$ the set of uniformly continuous functions from S to $\mathbb{R}$ and by $U_b^d(S)$ the set of bounded uniformly continuous functions from S to $\mathbb{R}$.

TODO: Show that $U_b^d(S)$ is a Banach space under the supremum norm.

LEMMA 15.17. *Let (S, d) be a separable metric space, then X is homeomorphic to a subset of $[0, 1]^{\mathbb{Z}_+}$ and furthermore*

(i) *S has a metric making it totally bounded*
(ii) *If S is compact then $C(S; \mathbb{R})$ with the uniform topology is separable.*
(iii) *If $\hat{d}$ is a totally bounded metric on S then $U^{\hat{d}}(S) = U_b^{\hat{d}}(S)$ and $U_b^{\hat{d}}(S)$ is separable*

PROOF. Let ρ be the product metric $\rho(x, y) = \sum_{n=1}^{\infty} \frac{|x_n - y_n|}{2^n}$ on the space $[0, 1]^{\mathbb{Z}_+}$. Pick a countable dense subset $x_1, x_2, \dots$ of S and define $f : S \in [0, 1]^{\mathbb{Z}_+}$ by

$$f(x) = \left(\frac{d(x_1, x)}{1 + d(x_1, x)}, \frac{d(x_2, x)}{1 + d(x_2, x)}, \dots \right)$$

CLAIM 15.17.1. $f(x)$ is continuous.

By definition of the product topology $f(x)$ is continuous if and only if each coordinate is. For any given fixed x_j, we know that $d(x_j, x)$ is continuous (in fact Lipschitz by Lemma 5.41) and thus the result follows from the continuity of $x/(1 + x)$ on $\mathbb{R}_+$.

CLAIM 15.17.2. $f(x)$ is injective.

For any $z \neq y$ we find $\epsilon > 0$ such that $B(z; \epsilon) \cap B(y; \epsilon) = \emptyset$ and then using density of $x_1, x_2, \dots$ to pick an x_n such that $d(z, x_n) < \epsilon$ and $d(y, x_n) \geq \epsilon$ showing $f(z) \neq f(y)$.

CLAIM 15.17.3. The inverse of $f(x)$ is continuous.

Fix an $x \in S$ and let $\epsilon > 0$ be given. Pick x_n such that $d(x_n, x) < \epsilon/2$. If we let $g(x) : [0, 1) \to \mathbb{R}_+$ be defined by $g(x) = x/(1 - x)$ then $g(x)$ is the inverse of $x/(1+x)$ and by continuity of $g(x)$ at the point $\frac{d(x_n, x)}{1+d(x_n, x)}$ we know that there exists a $\delta > 0$ such that $\left|\frac{d(x_n,x)}{1+d(x_n,x)} - \frac{d(x_n,y)}{1+d(x_n,y)}\right| < \delta$ implies $|d(x_n,x) - d(x_n,y)| < \epsilon/2$. Then if $f(y) \in B(f(x), \frac{\delta}{2^n})$ we have

$$\left|\frac{d(x_n,x)}{1+d(x_n,x)} - \frac{d(x_n,y)}{1+d(x_n,y)}\right| \leq 2^n \rho(f(x), f(y)) < \delta$$

$d(x,y) \leq d(x_n,x) + |d(x_n,x) - d(x_n,y)| < \epsilon$.

Now to see (i) we simply pull back the metric ρ via the embedding $f(x)$ and use the facts that ρ generates the product topology, $[0,1]^{\mathbb{Z}_+}$ is compact in product topology (by Tychonoff's Theorem 15.14; alternatively one can avoid the use of Tychonoff's Theorem for it is easy to see with a diagonal subsequence argument that a countable product of sequentially compact metric spaces is sequentially compact) hence totally bounded (Theorem 1.29).

Here is the argument that ρ generates the product topology; TODO: put this in a separate lemma. To see that the topology generated by ρ is finer than the product topology, suppose U is open in the topology generated by ρ. Pick $x \in U$ and select $N > 0$ such that $B(x, \epsilon) \subset U$. Then pick $N > 0$ such that $2^{-N-1} < \epsilon$ and consider $B = B(x_1, \epsilon/2) \times \cdots \times B(x_{2^N}, \epsilon/2) \times S \times \cdots$ which is open in the product topology. If $y \in B$ then

$$\rho(x,y) = \sum_{n=1}^{\infty} \frac{|x_n - y_n|}{2^n} = \sum_{n=1}^{2^N} \frac{|x_n - y_n|}{2^n} + \sum_{n=2^N+1}^{\infty} \frac{|x_n - y_n|}{2^n} \leq \frac{\epsilon}{2} \sum_{n=1}^{2^N} \frac{1}{2^n} + \sum_{n=2^N+1}^{\infty} \frac{1}{2^n} < \epsilon$$

To see that the product topology is finer than the metric topology, suppose $n > 0$ is an integer, $U \subset [0,1]$ is open and consider $\pi_n^{-1}(U)$. Let $x \in \pi_n^{-1}(U)$ and find an $\epsilon > 0$ such that $B(x_n, \epsilon) \subset U$. Note that if $y \in B(x, \frac{\epsilon}{2^n})$ then $|x_n - y_n| < 2^n \rho(x,y) \leq \epsilon$ and therefore $B(x, \frac{\epsilon}{2^n}) \subset \pi_n^{-1}(B(x_n, \epsilon)) \subset U$.

To see (ii), if S is compact then $f(S) \subset [0,1]^{\mathbb{Z}_+}$ is compact (Lemma 1.31). Observe that

$$A = \{\Pi_{i=1}^n p_i(x_i) \mid n \in \mathbb{N} \text{ and } p_i \in \mathbb{Q}[x]\}$$

is a subalgebra of $C([0,1]^{\mathbb{Z}_+}; \mathbb{R})$ and A separates points (given $x \neq y \in [0,1]$, pick n such that $x_n \neq y_n$ and pick the function $g(x) = x_n$). By the Stone-Weierstrass Theorem 1.44 we know that A is dense in $C([0,1]^{\mathbb{Z}_+}; \mathbb{R})$; now pullback A under $f(x)$ to a countable dense subset of $C(S; \mathbb{R})$.

To see (iii), suppose $\hat{\rho}$ is a totally bounded metric on S. Let $\hat{S}$ be the completion of S with respect to this metric.

Claim 15.17.4. $\hat{\rho}$ extends to a totally bounded metric on $\hat{S}$.

Let $\epsilon > 0$ be given and cover S by ball $B(x_i, \epsilon/2)$; we show that $B(x_i, \epsilon)$ covers $\hat{S}$. Biven $y \in \hat{S}$ we can find $x \in S$ such that $\hat{\rho}(x,y) < \epsilon/2$. Since $x \in S$ there exists an x_i such that $x \in B(x_i, \epsilon/2)$ and therefore $\hat{\rho}(x_i, y) \leq \hat{\rho}(x, x_i) + \hat{\rho}(x_i, y) < \epsilon$.

Because $(\hat{S}, \hat{\rho})$ is complete and totally bounded we know it is compact (Theorem 1.29) and we have just shown that $C(\hat{S}; \mathbb{R})$ has a countable dense subset.

Claim 15.17.5. $f|_S : C(\hat{S}; \mathbb{R}) \to U_b^{\hat{\rho}}(S; \mathbb{R})$ is a well defined, continuous and surjective.

Being well defined in this context means that restriction to S results in a bounded uniformly continuous function. This follows from the fact that any continuous function of a compact set is bounded and uniformly continuous (Theorem

1.31 and Theorem 1.35 respectively) and these properties are preserved upon restriction. To see surjectivity, let $g : S \to \mathbb{R}$ be uniformly continuous. We may apply Proposition 1.40 to see that g has a unique extension to a continuous function from the closure of S to $\mathbb{R}$. Since the closure of S in $\hat{S}$ is $\hat{S}$ be are done with the claim. Note that we did not need boundedness of g in order to prove the existence of the extension; therefore we have shown $U^{\hat{d}}(S) = U_b^{\hat{d}}(S)$.

Now the continuous image of a dense set under a surjective map is also dense. This is easily seen by picking a point $f(x)$ in the image; picking a sequence x_n such that $x_n \to x$ and then considering the image $f(x_n) \to f(x)$. Thus the result is proven. ∎

LEMMA 15.18 (Dini's Theorem). *Let K be a compact topological space and let $f_n : K \to \mathbb{R}$ be a sequence of continuous functions such that $f_n \downarrow 0$ pointwise on K, then $f_n \to 0$ uniformly.*

PROOF. Given $\epsilon > 0$ define $U_n = f_n^{-1}((-\infty, \epsilon))$. Then each U_n is open, $U_1 \subset U_2 \subset \cdots$ (since the f_n are decreasing) and the U_n form an open cover of K. We can extract a finite subcover which since the U_n are nested implies that $K = U_N$ for some $N > 0$. This is exactly the statement that $\sup_{x \in K} |f_n(x)| < \epsilon$ for all $n \geq N$ hence the result proven. ∎

LEMMA 15.19. *Let (S, d) be a separable metric space and let $\Lambda : U_b^d(S; \mathbb{R}) \to \mathbb{R}$ be a linear map such that*

(i) *Λ is non-negative (i.e. if $f \geq 0$ then $\Lambda(f) \geq 0$)*
(ii) *$\Lambda(1) = 1$*
(iii) *for all $\epsilon > 0$ there exists a compact set $K \subset S$ such that for all $f \in U_b^d(S; \mathbb{R})$,*

$$|\Lambda(f)| \leq \sup_{x \in K} |f(x)| + \epsilon \|f\|_u$$

then there exists a Borel probability measure μ on S such that $\Lambda(f) = \int f \, d\mu$. Whenever such a probability measure exists it is unique.

PROOF. We construct μ by use of the Daniell-Stone Theorem 2.142. It is clear that $U_b^d(S; \mathbb{R})$ is closed under max and min and contains the constant functions so $U_b^d(S; \mathbb{R})$ is a Stone Lattice. It remains to show that Λ obeys the "montone convergence" property: if $f_n \downarrow 0$ pointwise then $\Lambda(f_n) \downarrow 0$. This property is a corollary of Dini's Theorem 15.18 since by that result, if f_n are continuous and $f_n \downarrow 0$ pointwise on a compact set then the converge uniformly to 0 on the compact set. In particular, pick an $\epsilon > 0$ and let $K \subset S$ be compact as in the hypothesis. By Dini's Theorem there exists $N > 0$ such that $\sup_{x \in K} |f_n(x)| < \epsilon$ for all $n \geq N$. Therefore for all $n \geq N$,

$$\begin{aligned} |\Lambda(f_n)| &\leq \sup_{x \in K} |f_n(x)| + \epsilon \|f_n\|_\infty \\ &\leq \epsilon(1 + \|f_1\|_\infty) \end{aligned}$$

thus $\lim_{n \to \infty} \Lambda(f_n) = 0$ and we can apply Theorem 2.142.

Uniqueness follows because a probability measure is determined by its integrals over $U_b^d(S; \mathbb{R})$ (in fact over the subset of bounded Lipschitz functions). This follows because for any closed $F \subset S$ we can define $f_n(x) = nd(x, F) \wedge 1$ so that $f_n \downarrow \mathbf{1}_F$ and apply Montone Convergence (see the proof of the Portmanteau Theorem 5.43 for complete details on this argument). ∎

TODO: Apparently tight implies relatively compact does not require that S be separable, find a proof for that (Ethier and Kurtz have one).

THEOREM 15.20 (Prohorov's Theorem). *Let (S,d) be a separable metric space, then a tight set of probability measures on S is weakly relatively compact. If S is also complete then a weakly relatively compact set is tight.*

PROOF. By the Portmanteau Theorem 5.43 we know that a set of measures is tight if and only if its weak closure is tight (compact sets are closed hence can only gain mass in a weak limit). Thus it suffices to assume that we have a closed tight set M of measures. Put a totally bounded metric $\hat{d}$ on S so that $U_b^{\hat{d}}(S;\mathbb{R})$ is separable (Lemma 15.17); let $f_1, f_2, \ldots$ be a countable uniformly dense subset.

Pick a sequence μ_n from M; we must show that it has a weakly convergent subsequence. For every fixed f_m we know that $\left|\int f_m\, d\mu_n\right| \leq \|f_m\|_u < \infty$ so there is a subsequence $N \subset \mathbb{N}$ such that $\int f_m\, d\mu_n$ converges along N. Since this is true for every $m > 0$, a diagonalization argument shows there is a subsequence $\hat{\mu}_k$ such that $\lim_{k\to\infty} \int f_m\, d\hat{\mu}_k$ exists for every $m > 0$. Define $\Lambda(f_m) = \lim_{k\to\infty} \int f_m\, d\hat{\mu}_k$ for every such f_m. Our next goal is to extend Λ to all of $U_b^{\rho}(S;\mathbb{R})$. Since Λ is uniformly continuous on a dense subset we know that a continuous extension is defined; however we need a little bit more information.

CLAIM 15.20.1. $\lim_{k\to\infty} \int f\, d\hat{\mu}_k$ exists for every $f \in U_b^{\rho}(S;\mathbb{R})$; moreover

$$\lim_{k\to\infty} \int f\, d\hat{\mu}_k = \lim_{m\to\infty} \Lambda(\hat{f}_m)$$

where $\hat{f}_m$ is any subsequence of f_m that converges uniformly to f.

Pick a subsequence of the f_m that converges to f. Let that subsequence be donoted $\hat{f}_m$ so that $\lim_{m\to\infty} \left\|\hat{f}_m - f\right\|_\infty = 0$. For every $m > 0$ we have

$$\int \hat{f}_m\, d\hat{\mu}_k - \left\|\hat{f}_m - f\right\|_\infty \leq \int f\, d\hat{\mu}_k \leq \int \hat{f}_m\, d\hat{\mu}_k + \left\|\hat{f}_m - f\right\|_\infty$$

and therefore taking limits in k and using the definition of Λ at the points f_m,

$$\Lambda(\hat{f}_m) - \left\|\hat{f}_m - f\right\|_\infty \leq \liminf_{k\to\infty} \int f\, d\hat{\mu}_k \leq \limsup_{k\to\infty} \int f\, d\hat{\mu}_k \leq \Lambda(\hat{f}_m) + \left\|\hat{f}_m - f\right\|_\infty$$

Now letting m go to infinity we get $\lim_{m\to\infty} \Lambda(\hat{f}_m) = \lim_{k\to\infty} \int f\, d\hat{\mu}_k$.

As a result of the claim, we now define $\Lambda(f) = \lim_{k\to\infty} \int f\, d\hat{\mu}_k$ for every f and it is clearly linear (by linearity of integral and limits), nonnegative (by monotonicity of integral) and satisfies $\Lambda(1) = 1$ (since each $\hat{\mu}_k$ is a probability measure).

To show that Λ defines a probability measure, we bring the tightness hypothesis to the table. Pick $\epsilon > 0$ and by tightness take a compact set $K \subset S$ such that $\sup_{\mu\in M} \mu(K) > 1 - \epsilon$. For any $f \in U_b^{\hat{d}}(S;\mathbb{R})$ we have

$$|\Lambda(f)| = \lim_{k\to\infty} \left|\int f\, d\hat{\mu}_k\right| = \lim_{k\to\infty} \left|\int f\mathbf{1}_K\, d\hat{\mu}_k + \int f\mathbf{1}_{S\setminus K}\, d\hat{\mu}_k\right| \leq \sup_{x\in K} |f(x)| + \epsilon \|f\|_\infty$$

so we may apply Lemma 15.19 to conclude there exists a probability measure μ such that for all $f \in U_b^{\hat{d}}(S;\mathbb{R})$ we have $\Lambda(f) = \lim_{k\to\infty} \int f\, d\hat{\mu}_k = \int f\, d\mu$. Since $U_b^{\hat{d}}(S;\mathbb{R})$ contains all bounded Lipschitz functions by the Portmanteau Theorem 5.43 we conclude $\hat{\mu}_k$ converges weakly to μ.

Now assume that S is complete and separable and let M be a weakly relatively compact set of measures. Let $x_1, x_2, \dots$ be a countable dense subset of S. For every integer $n > 0$ we have $S = \cup_{k=1}^{\infty} B(x_k, 1/n)$. Thus $\cap_{N=1}^{\infty} \cap_{k=1}^{N} B(x_k, 1/n)^c = \emptyset$ so by continuity of measure (Lemma 2.30) for any fixed probability measure μ we can find an $N_{n,\mu} > 0$ such that $\mu(\cap_{k=1}^{N_{n,\mu}} B(x_k, 1/n)^c) < \epsilon/2^n$. We claim that, because M is compact, we can find an N_n for which this is true uniformly over the measures in M.

Claim 15.20.2. For every $n > 0$ there exists $N_n > 0$ such that $\mu(\cap_{k=1}^{N_n} B(x_k, 1/n)^c) < \epsilon/2^n$ for all $\mu \in M$.

We argue by contraction by reducing the case where M is a singleton set (where we have already shown the claim holds). If Claim 15.20.2 is not true then there exists n such that for every integer $N > 0$ we have some $\mu_N \in M$ such that $\mu_N(\cap_{k=1}^{N} B(x_k, 1/n)^c) \geq \epsilon/2^n$. By sequential compactness of M we know that there is a weakly convergent subsequence μ_{N_j} such that $\mu_{N_j} \xrightarrow{w} \mu$ for some probability measure μ. For every $N > 0$ we have $\cap_{k=1}^{N} B(x_k, 1/n)^c$ is closed and therefore by the Portmanteau Theorem 5.43

$$\begin{aligned}
\epsilon/2^n &\leq \limsup_{j \to \infty} \mu_{N_j}(\cap_{k=1}^{N_j} B(x_k, 1/n)^c) \\
&\leq \limsup_{j \to \infty} \mu_{N_j}(\cap_{k=1}^{N} B(x_k, 1/n)^c) \\
&\leq \mu(\cap_{k=1}^{N} B(x_k, 1/n)^c)
\end{aligned}$$

where in the second inequality we have used the fact that the limit only depends on the tail of the sequence of sets $\cap_{k=1}^{N_j} B(x_k, 1/n)^c$ and by a union bound for sufficiently large N_j we have $\mu_{N_j}(\cap_{k=1}^{N_j} B(x_k, 1/n)^c) \leq \mu_{N_j}(\cap_{k=1}^{N} B(x_k, 1/n)^c)$. To finish we get a contradiction by taking the limit and using continuity of measure

$$0 < \epsilon/2^n \leq \lim_{N \to \infty} \mu(\cap_{k=1}^{N} B(x_k, 1/n)^c) = 0$$

With Claim 15.20.2 proven we mimic the proof of Ulam's Theorem. Let

$$K = \cap_{m=1}^{\infty} \cup_{j=1}^{N_m} \overline{B}(x_j, \frac{1}{m})$$

which is easily seen to be closed (hence complete) and by construction is totally bounded thus is compact (Theorem 1.29) and furthermore for all $\mu \in M$,

$$\begin{aligned}
\mu(K^c) &\leq \mu((\cap_{m=1}^{\infty} \cup_{j=1}^{N_m} B(x_j, \frac{1}{m}))^c) \\
&= \mu(\cup_{m=1}^{\infty} \cap_{j=1}^{N_m} B(x_j, \frac{1}{m})^c) \\
&= \sum_{m=1}^{\infty} \mu(\cap_{j=1}^{N_m} B(x_j, \frac{1}{m})^c) \\
&\leq \sum_{m=1}^{\infty} \frac{\epsilon}{2^m} = \epsilon
\end{aligned}$$

□

LEMMA 15.21. *For $f, g \in C([0,\infty);\mathbb{R})$ define*

$$\rho(f,g) = \sum_{n=1}^{\infty} \frac{1}{2^n} \sup_{0 \leq t \leq n} (|f(t) - g(t)| \wedge 1)$$

then ρ is a metric on $C([0,\infty);\mathbb{R})$ and $C([0,\infty);\mathbb{R})$ is complete and separable with respect to this metric.

PROOF. It is clear that $\rho(f,f) = 0$ and furthermore if $\rho(f,g) = 0$ then $f = g$ on every interval $[0,n]$ and therefore $f = g$. Symmetry and the triangle inequality of ρ is immediate from the corresponding properties of the absolute value (TODO: OK the triangle inequality may need a bit more of an argument).

We claim that the set of polynomials with rational coefficients is dense in $C([0,\infty);\mathbb{R})$. Pick $f \in C([0,\infty);\mathbb{R})$ and let $\epsilon > 0$ be given. Now take $m > 0$ sufficiently large so that $1/2^m < \epsilon/2$ and by the Stone Weierstrass Theorem 1.44 we pick a polynomial with rational coefficients p such that $\sup_{0 \leq t \leq m} |f(t) - p(t)| < \epsilon/2$ then we have

$$\begin{aligned}
\rho(f,p) &\leq \sum_{n=1}^{m} \frac{1}{2^n} \sup_{0 \leq t \leq n} |f(t) - p(t)| + \sum_{n=m+1}^{\infty} \frac{1}{2^n} \\
&\leq \sup_{0 \leq t \leq m} |f(t) - p(t)| \sum_{n=1}^{m} \frac{1}{2^n} + \epsilon/2 < \epsilon
\end{aligned}$$

Completeness follows from arguing over intervals $[0,n]$. Suppose f_n is a Cauchy sequence in $C([0,\infty);\mathbb{R})$. Given $\epsilon > 0$ and $n > 0$ we can find $N > 0$ such that $\rho(f_m, f_N) < \epsilon/2^n$ for all $m \geq N$. Thus $\sup_{0 \leq t \leq n} |f_m(t) - f_N(t)| < \epsilon$ for all $m \geq N$ so we see that f_n is uniformly Cauchy on every interval $[0,n]$. By completeness of $C([0,n];\mathbb{R})$ we know that the pointwise limit of f_n exists on every $[0,n]$ and is a continuous function. Therefore we have a limit f defined on $[0,\infty)$ and since continuity is a local property $f \in C([0,\infty);\mathbb{R})$. It remains to show that f_n converges to f in the metric ρ. This follows arguing as we have above. Let $\epsilon > 0$ be given and choose $n > 0$ such that $\frac{1}{2^n} < \epsilon/2$ and choose $N > 0$ such that $\sup_{0 \leq t \leq n} |f_m(t) - f_N(t)| < \epsilon/2$ and then observe

$$\rho(f_m, f_N) \leq \sum_{k=1}^{n} \frac{1}{2^k} \sup_{0 \leq t \leq k} |f_m(t) - f_N(t)| + \sum_{k=n+1}^{\infty} \frac{1}{2^k} < \epsilon$$

□

The topology defined by ρ is often refered to as the topology of uniform convergence on compact sets by virtue of the following lemma.

LEMMA 15.22. *A sequence f_n converges to f in $C([0,\infty),\mathbb{R})$ if and only if f_n converges to f uniformly on every interval $[0,T]$ for $T > 0$.*

PROOF. TODO: This is elementary. □

DEFINITION 15.23. Given a function $f : [0,T] \to \mathbb{R}$ the *modulus of continuity* is the function

$$m(T, f, \delta) = \sup_{\substack{|s-t|<\delta \\ 0 \leq s,t \leq T}} |f(s) - f(t)|$$

LEMMA 15.24. *For fixed $T > 0$ and $\delta > 0$, $m(T, f, \delta)$ is a continuous function on $C([0,\infty);\mathbb{R})$. For fixed $T > 0$ and function $f : \mathbb{R} \to \mathbb{R}$, $m(T, f, \delta)$ is nonincreasing in δ and*

$$\lim_{\delta \to 0} m(T, f, \delta) = 0$$

provided $f \in C([0,\infty);\mathbb{R})$.

PROOF. To see continuity on $C([0,\infty);\mathbb{R})$ let $f \in C([0,\infty);\mathbb{R})$, $T > 0$, $\delta > 0$ and $\epsilon > 0$ be given and pick g that $\rho(f, g) < \epsilon/2^{\lceil T \rceil + 1}$. From the definition of the metric ρ for any $n > 0$, $\sup_{0 \leq t \leq n} |f(t) - g(t)| \wedge 1 \leq 2^n \epsilon$, so for any $T > 0$,

$$\sup_{0 \leq t \leq T} |f(t) - g(t)| \wedge 1 \leq \sup_{0 \leq t \leq \lceil T \rceil} |f(t) - g(t)| \wedge 1 \leq \epsilon/2$$

Therefore by the triangle inequality,

$$\begin{aligned} \sup_{\substack{|s-t|<\delta \\ 0 \leq s,t \leq T}} |g(s) - g(t)| \wedge 1 &\leq \sup_{\substack{|s-t|<\delta \\ 0 \leq s,t \leq T}} (|g(s) - f(s)| + |f(s) - f(t)| + |f(t) - g(t)|) \wedge 1 \\ &\leq \epsilon/2 + \sup_{\substack{|s-t|<\delta \\ 0 \leq s,t \leq T}} |f(s) - f(t)| \wedge 1 + \epsilon/2 \end{aligned}$$

and therefore arguing with the roles of f and g reversed shows $|m(T, f, \delta) - m(T, g, \delta)| \leq \epsilon$.

The fact that $m(T, f, \delta)$ is decreasing in δ is clear because the definition shows that for $\delta_1 \leq \delta_2$ we have

$$\{|f(t) - f(s)| \mid 0 \leq s,t \leq T \text{ and } |s-t| < \delta_1\} \subset \{|f(t) - f(s)| \mid 0 \leq s,t \leq T \text{ and } |s-t| < \delta_2\}$$

and therefore $m(T, f, \delta_2) \leq m(T, f, \delta_1)$.

Lastly if we suppose $f \in C([0,\infty);\mathbb{R})$ then f is uniformly continuous on $[0, T]$ for every $T > 0$ (Theorem 1.35). Thus given an $\epsilon > 0$ there exists $\delta > 0$ such that

$$\sup_{\substack{|s-t|<\delta \\ 0 \leq s,t \leq T}} |f(s) - f(t)| < \epsilon$$

which shows $\lim_{\delta \to 0} m(T, f, \delta) = 0$. □

The following Theorem is a version of the Arzela-Ascoli Theorem of real analysis.

THEOREM 15.25 (Arzela-Ascoli Theorem). *A set $A \subset C([0,\infty);\mathbb{R})$ is relatively compact if and only if*

(i) $\sup_{f \in A} |f(0)| < \infty$
(ii) $\lim_{\delta \to 0} \sup_{f \in A} m(T, f, \delta) = 0$ *for all $T > 0$.*

PROOF. To see the necessity of condition (i), observe that $\overline{A}$ is compact and by completeness of $C([0,\infty);\mathbb{R})$ we know that $\overline{A}$ comprises continuous functions. Therefore we know that $A \subset \overline{A} \subset \cup_{n=1}^{\infty} \{f \in C([0,\infty) \mid |f(0)| < n\}$. Since each $\{f \in C([0,\infty) \mid |f(0)| < n\}$ is easily seen to be an open set, by compactness of $\overline{A}$ we have a finite subcover which implies there exists an N such that $A \subset \overline{A} \subset \{f \in C([0,\infty) \mid |f(0)| < N\}$.

To see the necessity of condition (ii), fix $\epsilon > 0$, $T > 0$ and define for each $\delta > 0$ the set

$$F_\delta = \{f \in \overline{A} \mid m(T, f, \delta) \geq \epsilon\}$$

By continuity of $m(T, f, \delta)$ we know that F_δ is closed. Since $F_\delta \subset \overline{A}$ with $\overline{A}$ compact we conclude that F_δ is compact. Furthermore since for fixed $f \in \overline{A}$ continuity (more specifically uniform continuity on compact sets) implies $\lim_{\delta\to 0} m(T, f, \delta) = 0$, we know that $\cap_{\delta>0} F_\delta = \emptyset$. By nestedness and compactness of the F_δ we know that there is some specific $\delta > 0$ for which $F_\delta = \emptyset$ (Lemma 1.36) and (ii) is established.

To see the sufficiency of conditions (i) and (ii), we first construct the limiting subsequence on a the set of rationals $\mathbb{Q}_+ \subset [0,\infty)$. To do this, we first claim that for any $T \in \mathbb{Q}_+$, (in fact any $T \in [0,\infty)$), the set $\{|f(T)| \mid f \in A\}$ is bounded. The claim follows for $T > 0$ by using (ii) to select a $\delta > 0$ such that $\sup_{f\in A} m(T, f, \delta) < 1$. Picking the integer $m \geq 0$ such that $m\delta < T \leq (m+1)\delta$ and considering the grid $0, \delta, 2\delta, \ldots, m\delta, T$ we can write the telescoping sum

$$f(T) - f(0) = f(T) - f(m\delta) + \sum_{k=1}^{m} f(k\delta) - f((k-1)\delta)$$

and use the triangle inequality to conclude that $|f(T)| \leq |f(0)| + m + 1$ for every $f \in A$. Coupled with (i) this shows that $\sup_{f\in A} |f(T)| < \infty$.

We now enumerate the rationals $\mathbb{Q}_+$ and use compactness in $\mathbb{R}$ and a diagonal subsequence argument to pick a sequence f_n with $f \in A$ such that $f_n(T)$ converges for every $T \in \mathbb{Q}_+$. Define $f : \mathbb{Q}_+ \to \mathbb{R}$ by $f(T) = \lim_{n\to\infty} f_n(T)$.

Having selected a convergent subsequence f_n and defined f on $\mathbb{Q}_+$ we proceed to see that f is uniformly continuous on the interval $[0, T]$ for every $T \in \mathbb{Q}_+$. This follows by using (ii) to see that for every f_n, $T > 0$ and $\epsilon > 0$ there is $\delta > 0$ such that $|f_n(s) - f_n(t)| < \epsilon$ when $0 \leq s, t \leq T$ and $|s - t| < \delta$. From this we have for every $n > 0$, and $s, t \in \mathbb{Q}$, $0 \leq s, t \leq T$ and $|s - t| < \delta$

$$\begin{aligned} |f(s) - f(t)| &\leq |f(s) - f_n(s)| + |f_n(s) - f_n(t)| + |f_n(t) - f(t)| \\ &\leq |f(s) - f_n(s)| + \epsilon + |f_n(t) - f(t)| \end{aligned}$$

Taking the limit as $n \to \infty$ using pointwise convergence of f_n to f shows uniform continuity on every $[0, T] \cap \mathbb{Q}$ hence on $\mathbb{Q}_+$. Since f is uniformly continuous on $\mathbb{Q}_+$ it follows that f has a continuous extension to $f : [0,\infty) \to \mathbb{R}$. Moreover we have shown that $|f(s) - f(t)| < \epsilon$ when $|s - t| < \delta$.

It remains to prove that $f_n \to f$ in $C([0,\infty); \mathbb{R})$. It suffices (Lemma 15.22) to show that $f_n \to f$ uniformly on every interval $[0, T]$. Let $T > 0$ be given. Pick $\epsilon > 0$ and let $\delta > 0$ be such that $m(T, f_n, \delta) < \epsilon$ (hence $m(T, f, \delta) < \epsilon$ by the above comment). Pick $N > 0$ such that $|f_n(k\delta) - f(k\delta)| < \epsilon/3$ for all $k = 0, 1, \ldots, \lceil T/\delta \rceil$ and $n \geq N$. Then for every $0 \leq t \leq T$ and $n \geq N$ let $k \geq 0$ be such that $k\delta \leq t < (k+1)\delta$

$$|f_n(t) - f(t)| \leq |f_n(t) - f_n(k\delta)| + |f_n(k\delta) - f(k\delta)| + |f(k\delta) - f(t)| < \epsilon$$

and we are done. □

5. Donsker's Theorem

Provided with a characterization of compact sets in $C([0,\infty); \mathbb{R})$ we can now state the probabilistic analogue that characterizes tightness.

Lemma 15.26. *A sequence of Borel probability measures μ_n on $C([0,\infty); \mathbb{R})$ is tight if and only if*

(i) $\lim_{\lambda\to\infty} \sup_{n\geq 1} \boldsymbol{P}_{\mu_n}\{|f(0)| \geq \lambda\} = 0$.
(ii) $\lim_{\delta\to 0} \sup_{n\geq 1} \boldsymbol{P}_{\mu_n}\{m(T, f, \delta) \geq \lambda\} = 0$ *for all* $\lambda > 0$ *and* $T > 0$.

PROOF. Let μ_n be a tight sequence. Let $\epsilon > 0$ be given and pick $K \subset C([0,\infty);\mathbb{R})$ compact with $\mu_n(K) > 1-\epsilon$ for all n. Then by Theorem 15.25 we know that $\sup_{f \in K} |f(0)| < \infty$ and therefore $\mathbf{P}_{\mu_n}\{|f(0)| \geq \lambda\} \leq \mu_n(K^c) < \epsilon$ for any $\lambda > \sup_{f\in K} |f(0)|$. Thus (i) is shown. Similarly applying Theorem 15.25 we know that for every $T > 0$ and $\lambda > 0$ there exists $\delta > 0$ such that $\sup_{f \in K} m(T, f, \delta) < \lambda$. Therefore $\{f \mid m(T, f, \delta) \geq \lambda\} \subset K^c$ and by a union bound, for every $n > 0$ we have $\mathbf{P}_{\mu_n}\{m(T, f, \delta) \geq \lambda\} \leq \mathbf{P}_{\mu_n}\{K^c\} < \epsilon$. Therefore we have shown (ii).

Now assume that (i) and (ii) hold and suppose that $\epsilon > 0$ is given. By (i) there exists $\lambda > 0$ such that $\sup_{n\geq 1} \mathbf{P}_{\mu_n}\{|f(0)| \geq \lambda\} < \epsilon/2$. By (ii) for every integer $T > 0$ and $k > 0$, there exists a $\delta_{T,k}$ such that $\sup_{n\geq 1} \mathbf{P}_{\mu_n}\{m(T, f, \delta_{T,k}) \geq 1/k\} < \epsilon/2^{T+k+1}$. If we define

$$A_T = \{f \mid m(T, f, \delta_{T,k}) < 1/k \text{ for all } k \geq 1\}$$

so that $A_T^c \subset \cup_{k=1}^\infty \{f \mid m(T, f, \delta_{T,k}) \geq 1/k\}$ then by a union bound

$$\begin{aligned} \sup_{n\geq 1} \mu_n(A_T) &= \sup_{n\geq 1} (1 - \mu_n(A_T^c)) \\ &\geq \sup_{n \geq 1} \left(1 - \sum_{k=1}^\infty \mathbf{P}_{\mu_n}\{m(T, f, \delta_{T,k}) \geq 1/k\} \right) \\ &\geq 1 - \epsilon/2^{T+1} \end{aligned}$$

If we define $K = \{f \mid |f(0)| < \lambda\} \cap \cap_{T=1}^\infty A_T$ then another union bound shows $\sup_{n\geq 1} \mu_n(K) > 1-\epsilon$ and by construction the set K satisfies the conditions of Theorem 15.25 so is proven compact. $\square$

To prove that the rescaled and linearly interpolated random walk converges we need prove tightness. To prove tightness we need to show equicontinuity. The following Lemma begins the process by demonstrating equicontinuity at 0. Keep in mind the picture of the scaling of the random walk at level n which places the value of S_j at the point j/n scaled by the factor $1/\sigma\sqrt{n}$. With this geometry in mind note that what we are proving is a bound for each of the sequence of rescaled random walks on the interval $[0, \delta]$.

TODO: Replace ϵ by λ in the following Lemma?

LEMMA 15.27. *Let ξ_n be i.i.d. with mean* 0 *and finite variance σ^2 and define $S_n = \sum_{k=1}^n \xi_k$. Then for all $\epsilon > 0$*

$$\lim_{\delta \to 0} \limsup_{n\to\infty} \frac{1}{\delta} \boldsymbol{P}\{ \max_{1 \leq j \leq \lfloor n\delta \rfloor + 1} \frac{|S_j|}{\sigma\sqrt{n}} \geq \epsilon\} = 0$$

PROOF. The idea of the proof is to leverage the Central Limit Theorem and Gaussian tail bounds to control behavior at the right endpoint of the interval under consideration. Then independence of increments and finite variance can be used to control the behavior over the entire interval.

The sequence of random variables $\frac{1}{\sigma\sqrt{\lfloor n\delta \rfloor + 1}} S_{\lfloor n\delta\rfloor + 1}$ is a subsequence of $\frac{1}{\sigma\sqrt{n}} S_n$ and therefore converges in distribution to $N(0,1)$ by the Central Limit Theorem. Furthermore, $\lim_{n\to\infty} \frac{\sqrt{\lfloor n\delta\rfloor+1}}{\sqrt{n\delta}} = 1$ so by Slutsky's Lemma we also have

$\frac{1}{\sigma\sqrt{n\delta}} S_{\lfloor n\delta \rfloor + 1} \xrightarrow{d} Z$ where Z is an $N(0,1)$ Gaussian random variable. By the Portmanteau Theorem (Theorem 5.43) and a Markov bound (Lemma 10.1) we have

$$\limsup_{n\to\infty} \mathbf{P}\{\left|\frac{1}{\sigma\sqrt{n\delta}} S_{\lfloor n\delta \rfloor + 1}\right| \geq \lambda\} \leq \mathbf{P}\{|Z| \geq \lambda\} \leq \frac{\mathbf{E}\left[|Z|^3\right]}{\lambda^3}$$

We want to leverage this bound to create a maximal inequality that controls the entire interval of values of the rescaled random walk the approach being to leverage the fact that either the final point is in the tail (in which case the Central Limit Theorem bound just proven applies) or the final point is outside the tail and some interior point is in the tail providing us with an amount of variation whose probability can be controlled by use of a second moment bound. With $\epsilon > 0$ fixed as in the hypothesis of the Lemma, define the random variable $\tau = \min\{j \geq 1 \mid \left|\frac{S_j}{\sigma\sqrt{n}}\right| > \epsilon\}$ (this is a stopping time though we make no use of the concept here). Pick $\delta > 0$ satisfying $0 < \delta < \epsilon^2/2$.

$$\begin{aligned}
&\mathbf{P}\{\max_{1\leq j\leq \lfloor n\delta \rfloor+1} \left|\frac{S_j}{\sigma\sqrt{n}}\right| \geq \epsilon\} \\
&= \mathbf{P}\{\max_{1\leq j\leq \lfloor n\delta \rfloor+1} \left|\frac{S_j}{\sigma\sqrt{n}}\right| \geq \epsilon; \left|\frac{S_{\lfloor n\delta \rfloor+1}}{\sigma\sqrt{n}}\right| \geq \epsilon - \sqrt{2\delta}\} \\
&+ \mathbf{P}\{\max_{1\leq j\leq \lfloor n\delta \rfloor+1} \left|\frac{S_j}{\sigma\sqrt{n}}\right| \geq \epsilon; \left|\frac{S_{\lfloor n\delta \rfloor+1}}{\sigma\sqrt{n}}\right| < \epsilon - \sqrt{2\delta}\} \\
&\leq \mathbf{P}\{\left|\frac{S_{\lfloor n\delta \rfloor+1}}{\sigma\sqrt{n}}\right| \geq \epsilon - \sqrt{2\delta}\} + \sum_{j=1}^{\lfloor n\delta \rfloor} \mathbf{P}\{\left|\frac{S_{\lfloor n\delta \rfloor+1}}{\sigma\sqrt{n}}\right| < \epsilon - \sqrt{2\delta}; \tau = j\} \\
&= \mathbf{P}\{\left|\frac{S_{\lfloor n\delta \rfloor+1}}{\sigma\sqrt{n}}\right| \geq \epsilon - \sqrt{2\delta}\} + \sum_{j=1}^{\lfloor n\delta \rfloor} \mathbf{P}\{\left|\frac{S_{\lfloor n\delta \rfloor+1}}{\sigma\sqrt{n}} - \frac{S_j}{\sigma\sqrt{n}}\right| > \sqrt{2\delta}; \tau = j\} \\
&\leq \mathbf{P}\{\left|\frac{S_{\lfloor n\delta \rfloor+1}}{\sigma\sqrt{n}}\right| \geq \epsilon - \sqrt{2\delta}\} + \frac{1}{2\delta}\sum_{j=1}^{\lfloor n\delta \rfloor} \mathbf{E}\left[\left(\frac{S_{\lfloor n\delta \rfloor+1}}{\sigma\sqrt{n}} - \frac{S_j}{\sigma\sqrt{n}}\right)^2 \mathbf{1}_{\tau=j}\right] \\
&= \mathbf{P}\{\left|\frac{S_{\lfloor n\delta \rfloor+1}}{\sigma\sqrt{n}}\right| \geq \epsilon - \sqrt{2\delta}\} + \frac{1}{2\delta}\sum_{j=1}^{\lfloor n\delta \rfloor} \mathbf{E}\left[\left(\sum_{i=j+1}^{\lfloor n\delta \rfloor+1} \frac{\xi_i}{\sigma\sqrt{n}}\right)^2\right] \mathbf{P}\{\tau = j\} \\
&= \mathbf{P}\{\left|\frac{S_{\lfloor n\delta \rfloor+1}}{\sigma\sqrt{n}}\right| \geq \epsilon - \sqrt{2\delta}\} + \frac{\lfloor n\delta \rfloor}{2n\delta}\sum_{j=1}^{\lfloor n\delta \rfloor} \mathbf{P}\{\tau = j\} \\
&= \mathbf{P}\{\left|\frac{S_{\lfloor n\delta \rfloor+1}}{\sigma\sqrt{n}}\right| \geq \epsilon - \sqrt{2\delta}\} + \frac{1}{2}\mathbf{P}\{\max_{1\leq j\leq \lfloor n\delta \rfloor+1} \left|\frac{S_j}{\sigma\sqrt{n}}\right| \geq \epsilon\}
\end{aligned}$$

Therefore we have shown that

$$\mathbf{P}\{\max_{1\leq j\leq \lfloor n\delta \rfloor+1} \left|\frac{S_j}{\sigma\sqrt{n}}\right| \geq \epsilon\} \leq 2\mathbf{P}\{\left|\frac{S_{\lfloor n\delta \rfloor+1}}{\sigma\sqrt{n}}\right| \geq \epsilon - \sqrt{2\delta}\}$$

and we can use our tail bound derived from the Central Limit Theorem (with $\lambda = \frac{\epsilon - \sqrt{2\delta}}{\sqrt{\delta}}$) to see that

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \frac{1}{\delta} \mathbf{P}\{ \max_{1 \leq j \leq \lfloor n\delta \rfloor + 1} \left| \frac{S_j}{\sigma\sqrt{n}} \right| \geq \epsilon \} \leq \lim_{\delta \to 0} \frac{2}{\delta} \mathbf{E}\left[\left| Z \right|^3 \right] \left(\frac{\sqrt{\delta}}{\epsilon - \sqrt{2\delta}} \right)^3 = 0$$

□

The next step is to extend the estimate that provides equicontinuity at 0 to prove equicontinuity of the random walk on all finite intervals.

Lemma 15.28. *Let ξ_n be i.i.d. with mean 0 and finite variance σ^2 and define $S_n = \sum_{k=1}^n \xi_k$. Then for all $\epsilon > 0$ and $T > 0$*

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \boldsymbol{P}\{ \max_{\substack{1 \leq j \leq \lfloor n\delta \rfloor + 1 \\ 0 \leq k \leq \lfloor nT \rfloor + 1}} \frac{|S_{j+k} - S_k|}{\sigma\sqrt{n}} \geq \epsilon \} = 0$$

Proof. Pick $0 \leq \delta \leq T$ and let $m \geq 2$ be the integer such that $T/m < \delta \leq T/(m-1)$. Since

$$\lim_{n \to \infty} \frac{\lfloor nT \rfloor + 1}{\lfloor n\delta \rfloor + 1} = \frac{T}{\delta} < m$$

we know that for sufficiently large n we have $\lfloor nT \rfloor + 1 < (\lfloor n\delta \rfloor + 1)m$. For any such n, suppose $\frac{|S_{j+k} - S_k|}{\sigma\sqrt{n}} > \epsilon$ for some k with $0 \leq k \leq \lfloor nT \rfloor + 1$ and some j with $0 \leq j \leq \lfloor n\delta \rfloor + 1$. Now let p be the integer such that $0 \leq p \leq m-1$ and

$$(\lfloor n\delta \rfloor + 1)p \leq k < (\lfloor n\delta \rfloor + 1)(p+1)$$

Since $0 \leq j \leq \lfloor n\delta \rfloor + 1$ either

$$(\lfloor n\delta \rfloor + 1)p \leq k + j < (\lfloor n\delta \rfloor + 1)(p+1)$$

or

$$(\lfloor n\delta \rfloor + 1)(p+1) \leq k + j < (\lfloor n\delta \rfloor + 1)(p+2)$$

In the first case by the triangle inequality we have

$$|S_{j+k} - S_k| \leq \left| S_k - S_{(\lfloor n\delta \rfloor + 1)p} \right| + \left| S_{j+k} - S_{(\lfloor n\delta \rfloor + 1)p} \right|$$

and therefore we know that either $\frac{\left| S_k - S_{(\lfloor n\delta \rfloor + 1)p} \right|}{\sigma\sqrt{n}} \geq \epsilon/2 > \epsilon/3$ or $\frac{\left| S_{k+j} - S_{(\lfloor n\delta \rfloor + 1)p} \right|}{\sigma\sqrt{n}} \geq \epsilon/2 > \epsilon/3$. In the second case by the triangle inequality we have

$$|S_{j+k} - S_k| \leq \left| S_k - S_{(\lfloor n\delta \rfloor + 1)p} \right| + \left| S_{(\lfloor n\delta \rfloor + 1)(p+1)} - S_{(\lfloor n\delta \rfloor + 1)p} \right| + \left| S_{j+k} - S_{(\lfloor n\delta \rfloor + 1)(p+1)} \right|$$

and therefore we know that either $\frac{\left| S_k - S_{(\lfloor n\delta \rfloor + 1)p} \right|}{\sigma\sqrt{n}} \geq \epsilon/3$, $\frac{\left| S_{(\lfloor n\delta \rfloor + 1)(p+1)} - S_{(\lfloor n\delta \rfloor + 1)p} \right|}{\sigma\sqrt{n}} \geq \epsilon/3$ or $\frac{\left| S_{k+j} - S_{(\lfloor n\delta \rfloor + 1)(p+1)} \right|}{\sigma\sqrt{n}} \geq \epsilon/3$. Therefore we have the inclusion of events

$$\left\{ \max_{\substack{1 \leq j \leq \lfloor n\delta \rfloor + 1 \\ 0 \leq k \leq \lfloor nT \rfloor + 1}} \frac{|S_{j+k} - S_k|}{\sigma\sqrt{n}} \geq \epsilon \right\} \subset \bigcup_{p=0}^m \left\{ \max_{1 \leq j \leq \lfloor n\delta \rfloor + 1} \frac{\left| S_{j + (\lfloor n\delta \rfloor + 1)p} - S_{(\lfloor n\delta \rfloor + 1)p} \right|}{\sigma\sqrt{n}} \geq \epsilon/3 \right\}$$

By the i.i.d. nature of ξ_n and the fact that $S_0 = 0$ we know that

$$\mathbf{P}\{ \max_{1 \leq j \leq \lfloor n\delta \rfloor + 1} \frac{\left| S_{j + (\lfloor n\delta \rfloor + 1)p} - S_{(\lfloor n\delta \rfloor + 1)p} \right|}{\sigma\sqrt{n}} \geq \epsilon/3 \} = \mathbf{P}\{ \max_{1 \leq j \leq \lfloor n\delta \rfloor + 1} \frac{|S_j|}{\sigma\sqrt{n}} \geq \epsilon/3 \}$$

and therefore

$$\mathbf{P}\{\max_{\substack{1\leq j\leq \lfloor n\delta \rfloor +1 \\ 0\leq k \leq \lfloor nT \rfloor +1}} \frac{|S_{j+k} - S_k|}{\sigma\sqrt{n}}\} \leq (m+1)\mathbf{P}\{\max_{1\leq j\leq \lfloor n\delta \rfloor +1} \frac{|S_j|}{\sigma\sqrt{n}} \geq \epsilon/3\}$$

Since $\lim_{\delta\to 0}(m+1)\delta < \lim_{\delta \to 0}(T/\delta + 2)\delta = T < \infty$ we can apply Lemma 15.27 to get the result. ☐

By Prohorov's Theorem 15.20 we know that a tight sequence of probability measures on a separable metric space has a convergent subsequence. What is often required is some way of proving that a particular measure is indeed the limit of that subsequence. Recalling Lemma 9.6 we know that finite dimensional distributions characterize the laws of stochastic processes which leads one to the following general procedure for proving convergence of a sequence of processes.

TODO: Kallenberg (Chapter 16) has general results here for $C(T;S)$ with T a $lcscH$-space and S metric. Of course there are also results for spaces of discontinuous functions for use in proving convergence of empirical distribution functions. Kallenberg also has results for point process/spaces of measures.

We are taking the point of view of Brownian motion and the linearly interpolated random walk as being a random element in $C([0,\infty);\mathbb{R})$. On the other hand we have thus far treated a stochastic process as a random element in a subset of a path space $(S^T, \mathcal{S}^{\otimes T})$ *equipped with the product σ-algebra*. It is tempting to gloss over this point, however to tie in the general definition of stochastic processes with the random elements of $C([0,\infty);\mathbb{R})$ we are dealing with it is important to understand the relationship between the Borel σ-algebra on $C([0,\infty);\mathbb{R})$ and the product σ-algebra $\mathcal{B}(\mathbb{R})^{\otimes [0,\infty)}$ used in the definition of processes.

Lemma 15.29. *For every $t \in [0,\infty)$ let $\pi_t : C([0,\infty);\mathbb{R}) \to \mathbb{R}$ be the evaluation map $\pi_t(f) = f(t)$. The Borel σ-algebra on $C([0,\infty);\mathbb{R})$ is equal to $\sigma(\{\pi_t \mid t \in [0,\infty)\})$ and therefore $\mathcal{B}(C([0,\infty);\mathbb{R})) = C([0,\infty);\mathbb{R}) \cap \mathcal{B}(\mathbb{R})^{\otimes[0,\infty)}$.*

Proof. Since each π_t is a continuous function, it is Borel measurable and therefore the Borel σ-algebra contains $\sigma(\{\pi_t \mid t \in [0,\infty)\})$.

TODO: The following proof looks incorrect; I am using the supremum on $[0,\infty)$ which is not the metric on $C([0,\infty);\mathbb{R})$. What follows is valid on $C([0,T];\mathbb{R})$ but not on $C([0,\infty);\mathbb{R})$.

On the other hand, we know that $C([0,\infty);\mathbb{R})$ is separable so we may pick a countable dense set $f_1, f_2, \dots$. If we let $U \subset C([0,\infty);\mathbb{R})$ be open then for every $f_j \in U$ there exists $r_j > 0$ such that $B(f_j, r_j) \subset U$ and U is the union of such $B(f_j, r_j)$ (indeed, any $y \in U$ not in the union of balls can't be the limit of the f_j that are in U; on the other hand it can't be the limit of the f_j that are in U^c since the latter set is closed; thus the existence of such a y would contradict the density of $f_1, f_2, \dots$). To show $U \in \sigma(\{\pi_t \mid t \in [0,\infty)\})$ it suffices to show that $B(f,r) \in \sigma(\{\pi_t \mid t \in [0,\infty)\})$ for every $f \in C([0,\infty);\mathbb{R})$ and $r > 0$.

Let $B(f,r)$ be given and note that by continuity of the elements of $C([0,\infty);\mathbb{R})$ the closed ball

$$\begin{aligned}\overline{B(f,r)} &= \{g \mid \sup_{x\in[0,\infty)} |f(x)-g(x)| \leq r\} \\ &= \{g \mid \sup_{\substack{x\in[0,\infty)\\ x\in\mathbb{Q}}} |f(x)-g(x)| \leq r\} \\ &= \cap_{\substack{x\in[0,\infty)\\ x\in\mathbb{Q}}} \pi_x^{-1}([f(x)-r, f(x)+r])\end{aligned}$$

which shows that $\overline{B(f,r)} \in \sigma(\{\pi_t \mid t\in[0,\infty)\})$ and $B(f,r) = \cap_{n=1}^\infty \overline{B(f,r+1/n)}$ which shows that $B(f,r)\in\sigma(\{\pi_t \mid t\in[0,\infty)\})$. ∎

THEOREM 15.30. *Let X_n be a tight sequence of continuous processes such that for all $d>0$ and $0\leq t_1<\cdots<t_d<\infty$ the sequence $(X_{n,t_1},\dots,X_{n,t_d})$ converges in distribution, then the laws X_n converge to a Borel probability distribution μ on $C([0,\infty);\mathbb{R})$ for which the canonical process $W_t(\omega)=\omega(t)$ satisfies*

$$(X_{n,t_1},\dots,X_{n,t_d}) \xrightarrow{d} (W_{t_1},\dots,W_{t_d})$$

PROOF. By tightness and Prohorov's Theorem 15.20 we know that X_n has a weakly convergent subsequence. Our first claim is that any two weakly convergent subsequences of X_n have the same limiting distribution. Let $\check{X}_n$ and $\hat{X}_n$ be two such subsequences and suppose that $P\circ\check{X}_n^{-1}\to\check{\mu}$ and $P\circ\hat{X}_n^{-1}\to\hat{\mu}$ respectively. Fix $0\leq t_1<\cdots<t_d<\infty$ and note that by the Continuous Mapping Theorem 5.45 we know that $P\circ(\check{X}_{n,t_1},\dots,\check{X}_{n,t_d})^{-1}\xrightarrow{d}\check{\mu}\circ(\pi_{t_1},\dots,\pi_{t_d})^{-1}$ and $P\circ(\hat{X}_{n,t_1},\dots,\hat{X}_{n,t_d})^{-1}\xrightarrow{d}\hat{\mu}\circ(\pi_{t_1},\dots,\pi_{t_d})^{-1}$. By hypothesis we conclude that $\check{\mu}\circ(\pi_{t_1},\dots,\pi_{t_d})^{-1}=\hat{\mu}\circ(\pi_{t_1},\dots,\pi_{t_d})^{-1}$ and therefore by Lemma 15.29 we can apply Lemma 9.6 to conclude $\check{\mu}=\hat{\mu}$ which we now refer to as μ.

Now suppose that the distributions of X_n do not converge weakly to μ. Then there exists a bounded continuous f such that either $\lim_{n\to\infty}\mathbf{E}[f(X_n)]$ does not exist or exists and is different from $\int f\,d\mu$. In either case by the boundedness of f we know that

$$-\infty < -\|f\|_\infty \leq \liminf_{n\to\infty}\mathbf{E}[f(X_n)] \leq \limsup_{n\to\infty}\mathbf{E}[f(X_n)] \leq \|f\|_\infty < \infty$$

and we can extract a subsequence $\check{X}_n$ such that $\lim_{n\to\infty}\mathbf{E}\left[f(\check{X}_n)\right]$ exists and $\lim_{n\to\infty}\mathbf{E}\left[f(\check{X}_n)\right]\neq\int f\,d\mu$. This is a contradiction since by tightness we know that $\check{X}_n$ has a weakly convergent subsequence and we have already just shown that the limiting distribution is μ. ∎

The power of this Theorem is that it is often not too difficult to prove weak convergence of finite dimensional distributions because we have the power of a rich theory available (e.g. the Central Limit Theorem, Slutsky's Theorem, characteristic functions).

LEMMA 15.31. *Let ξ_n be i.i.d. with mean 0 and finite variance σ^2, define $S_n=\sum_{k=1}^n\xi_k$, $S_n^*(t)=S_{\lfloor t\rfloor}+(t-\lfloor t\rfloor)\xi_{\lfloor t\rfloor+1}$ and $X_n(t)=\frac{1}{\sigma\sqrt{n}}S_n^*(nt)$ where the latter are interpreted as random elements of the Borel measurable space $C([0,\infty);\mathbb{R})$. For every $d>0$ and real numbers $0\leq t_1<\cdots<t_d<\infty$ we have*

$$(X_n(t_1),\dots,X_n(t_d))\xrightarrow{d}(B_{t_1},\dots,B_{t_d})$$

where B_t is a standard Brownian motion.

PROOF. Let $0 \leq t_1 < \cdots < t_n < \infty$ be given. The basic point is that the result follows by the Central Limit Theorem; however due to the linear interpolation there is a bit of extra work to do.

First note that by definition

$$\left| X_n(t) - \frac{1}{\sigma\sqrt{n}} S_{\lfloor nt \rfloor} \right| \leq \frac{1}{\sigma\sqrt{n}} \left|\xi_{\lfloor nt \rfloor + 1}\right|$$

so by a Chebyshev bound (Lemma 10.2) we have

$$\lim_{n\to\infty} \mathbf{P}\{\left| X_n(t) - \frac{1}{\sigma\sqrt{n}} S_{\lfloor nt \rfloor} \right| > \epsilon\} \leq \lim_{n\to\infty} \frac{1}{n\epsilon^2} = 0$$

thus $X_n(t) \xrightarrow{P} \frac{1}{\sigma\sqrt{n}} S_{\lfloor nt \rfloor}$ and by Lemma 5.13 we have $(X_n(t_1), \ldots, X_n(t_d)) \xrightarrow{P} (\frac{1}{\sigma\sqrt{n}} S_{\lfloor nt_1 \rfloor}, \ldots, \frac{1}{\sigma\sqrt{n}} S_{\lfloor nt_d \rfloor})$. Our result will follow by Slutsky's Theorem 5.46 if we can show that

$$(\frac{1}{\sigma\sqrt{n}} S_{\lfloor nt_1 \rfloor}, \ldots, \frac{1}{\sigma\sqrt{n}} S_{\lfloor nt_d \rfloor}) \xrightarrow{d} (B_{t_1}, \ldots, B_{t_d})$$

Application of the Continuous Mapping Theorem 5.45 lets us reduce further to showing that

$$(\frac{1}{\sigma\sqrt{n}}(S_{\lfloor nt_1 \rfloor} - S_{\lfloor nt_0 \rfloor}), \ldots, \frac{1}{\sigma\sqrt{n}}(S_{\lfloor nt_d \rfloor} - S_{\lfloor nt_{d-1} \rfloor})) \xrightarrow{d} (B_{t_1} - B_{t_0}, \ldots, B_{t_d} - B_{t_{d-1}})$$

where for uniformity of notation we have defined $t_0 = 0$. Since the ξ_n are independent this implies that the $S_{\lfloor nt_j \rfloor} - S_{\lfloor nt_{j-1} \rfloor}$ are independent for $j = 1, \ldots, d$ and by definition of independent increments property of Brownian motion we know that $B_{t_j} - B_{t_{j-1}}$ are independent, thus by Lemma 4.5 it suffices to show that $\frac{1}{\sigma\sqrt{n}}(S_{\lfloor nt_j \rfloor} - S_{\lfloor nt_{j-1} \rfloor}) \xrightarrow{d} N(0, t_j - t_{j-1})$. We shall prove this fact for an arbitrary $0 \leq s < t < \infty$.

By the definition of S_n we write $\frac{1}{\sigma\sqrt{n}}(S_{\lfloor nt \rfloor} - S_{\lfloor ns \rfloor}) = \frac{1}{\sigma\sqrt{n}} \sum_{i=\lfloor ns \rfloor+1}^{\lfloor nt \rfloor} \xi_i$. For every $\epsilon > 0$ we have by another Chebyshev bound

$$\begin{aligned}
&\lim_{n\to\infty} \mathbf{P}\{\left| \frac{1}{\sigma\sqrt{n}} \sum_{i=\lfloor ns \rfloor+1}^{\lfloor nt \rfloor} \xi_i - \frac{\sqrt{t-s}}{\sigma\sqrt{\lfloor nt \rfloor - \lfloor ns \rfloor}} \sum_{i=\lfloor ns \rfloor+1}^{\lfloor nt \rfloor} \xi_i \right| > \epsilon\} \\
&\leq \lim_{n\to\infty} \frac{1}{\epsilon^2} \mathbf{Var}\left(\left(\frac{1}{\sigma\sqrt{n}} - \frac{\sqrt{t-s}}{\sigma\sqrt{\lfloor nt \rfloor - \lfloor ns \rfloor}} \right) \sum_{i=\lfloor ns \rfloor+1}^{\lfloor nt \rfloor} \xi_i \right) \\
&= \lim_{n\to\infty} \frac{1}{\epsilon^2} \left(\frac{1}{\sigma\sqrt{n}} - \frac{\sqrt{t-s}}{\sigma\sqrt{\lfloor nt \rfloor - \lfloor ns \rfloor}} \right)^2 (\lfloor nt \rfloor - \lfloor ns \rfloor)\sigma^2 \\
&= \lim_{n\to\infty} \frac{1}{\epsilon^2} \left(\frac{\sqrt{\lfloor nt \rfloor - \lfloor ns \rfloor}}{\sqrt{n}} - \sqrt{t-s} \right)^2 = 0
\end{aligned}$$

Therefore we have $\frac{1}{\sigma\sqrt{n}} \sum_{i=\lfloor ns \rfloor + 1}^{\lfloor nt \rfloor} \xi_i \xrightarrow{P} \frac{\sqrt{t-s}}{\sigma\sqrt{\lfloor nt \rfloor - \lfloor ns \rfloor}} \sum_{i=\lfloor ns \rfloor + 1}^{\lfloor nt \rfloor} \xi_i$ and one last appeal to Slutsky's Theorem 5.46 implies that it suffices to show

$$\frac{\sqrt{t-s}}{\sigma\sqrt{\lfloor nt \rfloor - \lfloor ns \rfloor}} \sum_{i=\lfloor ns \rfloor + 1}^{\lfloor nt \rfloor} \xi_i \xrightarrow{d} N(0, t-s)$$

which is just the Central Limit Theorem (and to be precise the Continuous Mapping Theorem 5.45 to account for the multiplication by $\sqrt{t-s}$). □

The last step we make is in extending the equicontinuity of the random walk to equicontinuity of the linearly interpolated random walk which are honest elements of $C([0,\infty);\mathbb{R})$. This equicontinuity will prove tightness and weak convergence of the linearly interpolated random walk. One of the elements of proving the equicontinuity of the linearly interpolated random walk is a general fact about the modulus of continuity of a class of piecewise linear functions which we prove as a separate lemma.

Lemma 15.32. *Let $f(t)$ be a continuous function that is linear on every interval $[j, j+1]$ for $j = 0, 1, \dots$. For every integer $M > 0$ and $N > 0$, we have*

$$\sup_{\substack{|s-t| \leq M \\ 0 \leq s,t \leq N}} |f(s) - f(t)| \leq \sup_{\substack{1 \leq j \leq M \\ 0 \leq k \leq N}} |f(k+j) - f(k)|$$

Proof. Pick $0 \leq s < t \leq M$. If there exists $j < N$ such that $j \leq s < t \leq j+1$ then it is clear from linearity that $|f(s) - f(t)| \leq |f(j) - f(j+1)|$ so it suffices to consider the case in which $j \leq s < j+1 < \cdots < j+k < t \leq j+k+1$ for some $j \geq 0$ and $k > 0$. If we let $f(t)$ has slope a_j on the interval $[j, j+1]$ then we can write $f(t) - f(s) = a_j(j+1-s) + \cdots + a_{j+k}(t-j-k)$. Note that if $f(t) - f(s)$ has a different sign than a_j then $|f(t) - f(s)| \leq |f(t) - f(j+1)|$ and similarly with a_{j+k} so if suffices to assume that a_j and a_{j+k} have the same sign as $f(t) - f(s)$. Now if $|a_j| \leq |a_{j+k}|$ then we slide the pair (s,t) to the right until either s or t hits an integer. More formally if $j+1-s \leq j+k+1-t$ then we get $|f(t) - f(s)| \leq |f(t+j+1-s) - f(j+1)|$ and if $j+k+1-t \leq j+1-s$ we get the bound $|f(t) - f(s)| \leq |f(j+k+1) - f(s+j+k+1-t)|$. If we $|a_j| \geq |a_{j+k}|$ we slide to the left in an analogous way. The point is that we are reduced to the case in which either $s = j-1$ or $t = j+k+1$.

Once we know that either $s = j-1$ or $t = j+k+1$, because M is integer we know that in fact $k \leq M$ and therefore we get a final bound $|f(t) - f(s)| \leq |f(j+k+1) - f(j-1)|$ which proves the result.

TODO: This proof is grotesque. Try to do better! □

We are finally ready to put all of the pieces together to prove Donsker's Theorem on the convergence of random walks to Brownian motion. Note that we have not used the existence of Brownian motion anywhere in the proof so this Theorem is among other things an existence proof for Brownian motion.

Theorem 15.33 (Donsker's Invariance Principle for Random Walks). *Let ξ_n be i.i.d. with mean 0 and finite variance σ^2, define $S_n = \sum_{k=1}^n \xi_k$, $S_n^*(t) = S_{\lfloor t \rfloor} + (t - \lfloor t \rfloor)\xi_{\lfloor t \rfloor + 1}$ and $X_n(t) = \frac{1}{\sigma\sqrt{n}} S_n^*(nt)$ where the latter are interpreted as random elements of the Borel measurable space $C([0,\infty);\mathbb{R})$. Then the law of X_n converges* *weakly to a probability measure under which the coordinate mapping* $(f,t) \to f(t)$ *is a standard Brownian motion.*

Proof. Lemma 15.31 shows that finite dimensional distributions of the linearly interpolated and rescaled random walk converge to the finite dimensional distributions of Brownian motion. Therefore by Theorem 15.30 it remains to show that X_n is a tight sequence of processes. By Lemma 15.26 we must show for all $X_n(t)$,

(i) $\lim_{\lambda\to\infty} \sup_{n\geq 1} \mathbf{P}\{|X_n(0)| \geq \lambda\} = 0$.
(ii) $\lim_{\delta\to 0} \sup_{n\geq 1} \mathbf{P}\{m(T, X_n, \delta) \geq \lambda\} = 0$ for all $\lambda > 0$ and $T > 0$.

Since $X_n(0) = 0$ the condition (i) holds trivially. As for condition (ii) we first argue that it suffices to show $\lim_{\delta\to 0} \limsup_{n\geq 1} \mathbf{P}\{m(T, X_n, \delta) \geq \lambda\} = 0$. This follows from the fact that for fixed $n > 0$, $\lim_{\delta\to 0} \mathbf{P}\{m(T, X_n, \delta) \geq \lambda\} = 0$ (continuity of X_n) and $\mathbf{P}\{m(T, X_n, \delta) \geq \lambda\}$ is a decreasing function of δ. Indeed, if we let $\epsilon > 0$ be given pick $\Delta > 0$ such that $\limsup_{n\geq 1} \mathbf{P}\{m(T, X_n, \delta) \geq \lambda\} < \epsilon$ for all $\delta \leq \Delta$. Then pick $N > 0$ is such that $\sup_{n\geq N} \mathbf{P}\{m(T, X_n, \Delta) \geq \lambda\} < \epsilon$ and note that because $\mathbf{P}\{m(T, X_n, \delta) \geq \lambda\}$ is decreasing in fact we have $\sup_{n\geq N} \mathbf{P}\{m(T, X_n, \delta) \geq \lambda\} < \epsilon$ for all $\delta \leq \Delta$. Since $\lim_{\delta\to 0} \mathbf{P}\{m(T, X_n, \delta) \geq \lambda\} = 0$ for every $n > 0$ we can find $\hat{\Delta} < \Delta$ such that $\mathbf{P}\{m(T, X_n, \delta) \geq \lambda\} < \epsilon$ for all $n = 1, \ldots, N-1$ and $\delta \leq \hat{\Delta}$ and thus $\sup_{n\geq 1} \mathbf{P}\{m(T, X_n, \Delta) \geq \lambda\} < \epsilon$ for all $\delta < \hat{\Delta}$.

With this reduction in hand, we can estimate

$$\begin{aligned}
\mathbf{P}\{m(T, X_n, \delta) \geq \lambda\} &= \mathbf{P}\{\sup_{\substack{|s-t|\leq\delta \\ 0\leq s,t\leq T}} |X_n(s) - X_n(t)| \geq \lambda\} \\
&\leq \mathbf{P}\{\sup_{\substack{|s-t|\leq\lfloor n\delta\rfloor+1 \\ 0\leq s,t\leq\lfloor T\delta\rfloor+1}} |S_n^*(s) - S_n^*(t)| \geq \sigma\sqrt{n}\lambda\} \\
&\leq \mathbf{P}\{\sup_{\substack{1\leq j\leq\lfloor n\delta\rfloor+1 \\ 0\leq k\leq\lfloor T\delta\rfloor+1}} |S_n(k+j) - S_n(k)| \geq \sigma\sqrt{n}\lambda\}
\end{aligned}$$

where the last inequality follows Lemma 15.32. Now we can apply Lemma 15.28 to conclude $\lim_{\delta\to 0} \limsup_{n\to\infty} \mathbf{P}\{m(T, X_n, \delta) \geq \lambda\}$ and tightness is shown. □

6. Banach Spaces

We start with some of the standard examples of Banach spaces.

Definition 15.34. Let $(\Omega, \mathcal{A})$ be a measurable space, let $B(\Omega)$ be space of bounded measurable functions $f : \Omega \to \mathbb{R}$ and define $\|f\|_\infty = \sup_{\omega\in\Omega} |f(\omega)|$.

Proposition 15.35. *The space* $B(\Omega)$ *is a Banach space.*

Proof. It is elementary that $B(\Omega)$ is a vector space : if f is bounded and measurable with $\|f\|_\infty$ then clearly af is bounded measurable with bound $|a| \|f\|_\infty$ for all $a \in \mathbb{R}$ and if f and g are bounded and measurable with bounds M and N respectively then $f + g$ is bounded and measurable with bound $M + N$.

To see that $\|\cdot\|_\infty$ is in fact a norm, first note that it is immediate from the definition that $\|f\|_\infty \geq 0$ and $\|f\|_\infty$ if and only if $f = 0$. Let $f \in B(\Omega)$ and let $a \in \mathbb{R}$ with $a \neq 0$. Then for every $\epsilon > 0$ we may find $\omega \in \Omega$ such that $\|f\|_\infty - \epsilon/|a| \leq |f(\omega)|$. It follows that $|a| \|f\|_\infty - \epsilon \leq |af(\omega)| \leq \|af\|_\infty$. Since $\epsilon > 0$

was arbitrary we may take $\epsilon \downarrow 0$ to conclude $|a| \|f\|_\infty \leq \|af\|_\infty$. In a similar way we may find an $\omega \in \Omega$ such that

$$\|af\|_\infty - \epsilon \leq |af(\omega)| = |a| |f(\omega)| \leq |a| \|f\|_\infty$$

and since $\epsilon > 0$ was arbitrary we conclude the opposite inequality $\|af\|_\infty \leq |a| \|f\|_\infty$. To see the triangle inequality let $f, g \in B(\Omega)$ be given. For every $\omega \in \Omega$ we have $|(f+g)(\omega)| \leq |f(\omega)| + |g(\omega)| \leq \|f\|_\infty + \|g\|_\infty$. Now take the supremum over all ω to conclude $\|f+g\|_\infty \leq \|f\|_\infty + \|g\|_\infty$.

Lastly we have to show that $B(\Omega)$ is complete. Let f_n be a Cauchy sequence in $B(\Omega)$. Thus for every $\epsilon > 0$ there exists and $n \in \mathbb{N}$ such that for all $m \geq n$ we have $\|f_n - f_m\| < \epsilon$. In particular, for every $\omega \in \Omega$ we have $|f_n(\omega) - f_m(\omega)| \leq \|f_n - f_m\| < \epsilon$ so that $f_n(\omega)$ is a Cauchy sequence in $\mathbb{R}$ for every $\omega \in \Omega$. By completeness of $\mathbb{R}$ we know that f_n converges pointwise and by Lemma 2.14 we know that there is a measurable function $f : \Omega \to \mathbb{R}$ such that $f_n \to f$.

Now note that for any $\epsilon > 0$ we can pick $N \in \mathbb{N}$ such that for $m, n \geq N$ we have $\|f_n - f_m\|_\infty < \epsilon$ so by using the triangle inequality and taking limits over $m \in \mathbb{N}$ we get for every $\omega \in \Omega$ and $n \geq N$,

$$|f(\omega) - f_n(\omega)|_\infty \leq \lim_{m\to\infty} |f_m(\omega) - f_n(\omega)| + \lim_{m\to\infty} |f(\omega) - f_m(\omega)| \leq \lim_{m\to\infty} \|f_m - f_n\|_\infty \leq \epsilon$$

Now using the fact with the triangle inequality we get $|f(\omega)| \leq |f(\omega) - f_n(\omega)| + |f_n(\omega)| \leq \epsilon + \|f_n\|_\infty$ which shows f is bounded. Taking the supremum over ω shows that $\|f - f_n\| \leq \epsilon$ for all $n \geq N$ which shows that f_n converges to f in $B(\Omega)$. □

DEFINITION 15.36. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and let $f : \Omega \to \mathbb{R}$ be a measurable function. We say that $a \in \mathbb{R}$ is an *essential upper bound of f* if $\mu(f^{-1}(a, \infty)) = 0$ (i.e. $f(x) \leq a\mu$-almost everywhere). The *essential supremum of f* is defined as the infimum of the set of essential upper bounds of f:

$$\operatorname{ess\,sup} f = \inf\{a \in \mathbb{R} \mid \mu(f^{-1}(a, \infty)) = 0\}$$

A measurable function f is said to be *essentially bounded* if $\operatorname{ess\,sup} |f| < \infty$.

The set of essentially bounded functions on a measure space can be made into a Banach space under the essential supremum norm. As is standard we have to pass to equivalence classes to do this so we first note when a function has an essential supremum of 0.

PROPOSITION 15.37. *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and let $f : \Omega \to \mathbb{R}$ be a measurable function then $\operatorname{ess\,sup} |f| = 0$ if and only if $f = 0$ μ-almost everywhere.*

PROOF. Clearly if f is almost everywhere 0 then $|f|^{-1}(0, \infty) = \{f \neq 0\}$ hence by monotonicity of measure, $\mu(|f|^{-1}(0, \infty)) = 0$. Thus $\operatorname{ess\,sup} |f| \leq 0$. On the other hand for every $a < 0$ we know that $|f|^{-1}(a, \infty) \supset \{f = 0\}$ and therefore $\mu(|f|^{-1}(a, \infty)) = \mu(\Omega) > 0$ and therefore $\operatorname{ess\,sup} |f| \geq 0$.

On the other hand if f is not almost everywhere zero then by Fatou's Lemma $0 < \mu(|f|^{-1}(0, \infty)) \leq \liminf_{n\to\infty} \mu(|f|^{-1}(1/n, \infty))$ which implies that $\mu(|f|^{-1}(1/n, \infty)) > 0$ for some $n \in \mathbb{N}$ and therefore $\operatorname{ess\,sup} |f| \geq 1/n > 0$. □

DEFINITION 15.38. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space then $L^\infty(\Omega, \mathbb{R})$ is space of equivalence classes of essentially bounded measurable functions under the equivalence relation of μ-almost everywhere equality.

THEOREM 15.39. *$L^\infty(\Omega,\mathbb{R})$ with essential supremum as a norm is a Banach space.*

PROOF. We have already seen that $\|f\|_\infty = 0$ implies $f = 0$ in L^∞. Let $a \in \mathbb{R}$ with $a \neq 0$ and $f \in L^\infty$ then we note that b is an essential upper bound for $|f|$ if and only if $|a|\, b$ is an essential upper bound for $|af|$. This follows from the set identity that

$$|f|^{-1}(b,\infty) = \{\omega \mid b < |f(\omega)|\} = \{\omega \mid |a|\, b < |af(\omega)|\} = |af|^{-1}(|a|\, b,\infty)$$

Thus we have

$$\|af\|_\infty = \inf\{|a|\, b \mid \mu(|f|^{-1}(b,\infty)) = 0\} = |a|\, \|f\|_\infty$$

If we are given $f, g \in L^\infty$ then if a is an essential upper bound for f and b is an essential upper bound for g then by the triangle inequality in $\mathbb{R}$, $|f+g|^{-1}(a+b,\infty) \subset |f|^{-1}(a,\infty) \cup |g|^{-1}(b,\infty)$ and by subadditivity of μ we know that $\mu(|f+g|^{-1}(a+b,\infty)) = 0$. Thus $a+b$ is an essential upper bound of $|f+g|$. For any $\epsilon > 0$ we can find essential upper bounds a and b such that $a \leq \|f\|_\infty + \epsilon/2$ and $b \leq \|g\|_\infty + \epsilon/2$ and therefore $\|f+g\|_\infty \leq a+b \leq \|f\|_\infty + \|g\|_\infty + \epsilon$. As $\epsilon > 0$ was arbitrary the triangle inequality follows.

To see completeness, suppose that f_n is a Cauchy sequence. We choose representatives of the equivalence class that are everywhere bounded on Ω; thus there exist constants M_n such that $\sup|f| \leq M_n$. For each $\epsilon > 0$ there exists an $N \in \mathbb{N}$ such that $\operatorname{ess\,sup}|f_n - f_m| < \epsilon$ for all $n, m \geq N$. For each such pair n, m if follows that $|f_n - f_m| < \epsilon$ except on a set of μ null set. Taking the union of such null sets over all rational $\epsilon > 0$ and all pairs n, m we can take a countable union of null sets and find a single null set $\mathcal{N}$ such that for every $\epsilon \in \mathbb{Q}_+$ there exists an $N \in \mathbb{N}$ such that $|f_n(\omega) - f_m(\omega)| < \epsilon$ for all $n, m \geq N$ and $\omega \notin \mathcal{N}$. Without changing equivalence classes of or the uniform bounds on the f_n we may redefine f_n to be zero on $\mathcal{N}$ and then it follows that for each $\epsilon \in \mathbb{Q}_+$ there exists an $N \in \mathbb{Z}$ such that $|f_n(\omega) - f_m(\omega)| < \epsilon$ for all $n, m \geq N$ and all $\omega \in \Omega$. Thus f_n is pointwise Cauchy and by completeness of $\mathbb{R}$ we may define f as the pointwise limit of f_n and we know that f is a measurable function. In fact the convergence is easily seen to be uniform. For every $\epsilon \in \mathbb{Q}_+$ there exists an $n \in \mathbb{N}$ such that $|f_n(\omega) - f_m(\omega)| \leq \epsilon$ for all $m \geq n$ and $\omega \in \Omega$ and therefore using the triangle inequality and pointwise convergence to f we have

$$|f_n(\omega) - f(\omega)| \leq \lim_{m\to\infty}(|f_n(\omega) - f_m(\omega)| + |f_m(\omega) - f(\omega)|) \leq \epsilon$$

for all $\omega \in \Omega$. In particular by chosing $\epsilon = 1$, by the triangle inequality we have $|f(\omega)| \leq |f_n(\omega)| + |f_n(\omega) - f(\omega)| \leq M_n + 1$ for all $\omega \in \Omega$ so $f \in L^\infty$. $\square$

DEFINITION 15.40. Let X be a topological space the $C_c(X)$ is the set of all continuous function $f : X \to \mathbb{R}$ with compact support (i.e. *supp*$(f) = \overline{\{x \in X \mid f(x) \neq 0\}}$ is compact).

DEFINITION 15.41. Let X be a topological space the $C_0(X)$ is the set of all continuous function $f : X \to \mathbb{R}$ which vanish at infinity in the sense that for every $\lambda > 0$ the set $\{x \in X \mid |f(x)| \geq \lambda\}$ is compact.

PROPOSITION 15.42. *If X is a topological space, then for each $f \in C_0(X)$ define $\|f\| = \sup_{x\in X}|f(x)|$ then $\|f\|$ is a norm on $C_0(X)$ and $C_0(X)$ is a Banach space. Furthermore $C_c(X)$ is dense in $C_0(X)$.*

PROOF. TODO: □

Note that if X is compact then every continuous function vanishes at infinity so it follows that the space of all continuous functions on X is a Banach space under the uniform norm. In the case that X is locally compact and Hausdorff but not necessarily compact we can embed X in its one point compactification $\tilde{X}$. It is useful to understand the relationship between $C_0(X)$ and $C(\tilde{X})$; in fact the following justifies the use of the phrase "vanishing at infinity".

PROPOSITION 15.43. *Let X be a locally compact Hausdorff space and let $\tilde{X}$ be the one point compactification of X. Every $f \in C_0(X)$ extends uniquely to an element $\tilde{f} \in C(\tilde{X})$ by defining $\tilde{f}(\Delta) = 0$. This defines an isometry of $C_0(X)$ with the subspace $\{g \in C(\tilde{X}) \mid g(\Delta) = 0\}$.*

PROOF. We first show that defining $\tilde{f}(\Delta) = 0$ makes $\tilde{f} : \tilde{X} \to \mathbb{R}$ continuous. Suppose $U \subset \mathbb{R}$ be open. If $0 \notin U$ then $\tilde{f}^{-1}(U) = f^{-1}(U) \subset U$ which is open X by continuity of f; it is also open in $\tilde{X}$ by the definition of the topology on $\tilde{X}$. If $0 \in U$ then there exists $\epsilon > 0$ such that $(-\epsilon, \epsilon) \subset U$ and it follows from the fact that f vanishes at infinity that $K = \{|f(x)| \geq \epsilon\}$ is compact. Since we can write $\tilde{f}^{-1}(U) = f^{-1}(U) \cup (X \setminus K) \cup \{\Delta\}$ which is open in $\tilde{X}$ we see that $\tilde{f}$ is continuous.

Since $\tilde{X}$ is Hausdorff it follows that the value of $\tilde{f}$ at Δ is determined by the restriction to X and thus the uniqueness of the continuous extension $\tilde{f}$ is determined. It is elementary that the assignment $f \mapsto \tilde{f}$ is linear and it is trivial that $\sup_{v \in X} |f(v)| \leq \sup_{v \in \tilde{X}} \left|\tilde{f}(v)\right|$. On the other hand for arbitrary $\epsilon > 0$ we have

$$\sup_{v \in \tilde{X}} \left|\tilde{f}(v)\right| \leq \sup_{\substack{v \in X \\ |f(v)| \geq \epsilon}} |f(v)| + \epsilon \leq \sup_{v \in X} |f(v)| + \epsilon$$

Since $\epsilon > 0$ was arbitrary we let $\epsilon \to 0$ and see that we have an isometry; in particular the mapping is injective.

It remains to characterize the range. For that suppose that $g \in C(\tilde{X})$ is such that $g(\Delta) = 0$, it suffices to show that restriction of g to X vanishes at infinity. For every $\epsilon > 0$ by continuity of g there exists an open neighborhood V of Δ such that $g(V) \subset (-\epsilon, \epsilon)$. By definition of the topology on $\tilde{X}$ we can write $V = U \cap (X \setminus K) \cup \{\Delta\}$ with $U \subset X$ open and $K \subset X$ compact. In particular it follows that $|g(v)| < \epsilon$ for all $v \in X \setminus K$ which is to say that $\{v \in X \mid |g(v)| \geq \epsilon\} \subset K$. As a closed subset of a compact set in a Hausdorff space it follows that $\{v \in X \mid |g(v)| \geq \epsilon\}$ is compact and thus $g\mid_X$ vanishes at infinity. □

As an example of the utility of this result consider the following

COROLLARY 15.44. *Let X be locally compact Hausdorff and suppose that $f \in C_0(X)$ then it follows that there exists $x_0 \in X$ such that $|f(x_0)| = \|f\|$.*

PROOF. If $f = 0$ the result is trivial so assume that f is non-zero. Apply Proposition 15.43 to extend f by zero to $\tilde{f}$. We know that $\left|\tilde{f}(x)\right|$ is a continuous positive function on $\tilde{X}$ such that $\left|\tilde{f}(\Delta)\right| = 0$. Since $\tilde{X}$ is compact we know that it attains its supremum (TODO: Where do we show this???) so that there exists $x_0 \in \tilde{X}$ such that $\left|\tilde{f}(x_0)\right| = \left\|\tilde{f}\right\| = \|f\| > 0$; it follows that $x_0 \in X$.

To see that f attains its supremum, first assume that $f(x) \geq 0$ for all $x \in S$, then $\sup_{x\in X} f(x) = \|f\|$. Now we apply the result for norm attainment to f. For general f we simply consider $f - \inf_{x\in S} f(x)$. The result for infimums follow from that for supremums since there exists x_0 such that $-f(x_0) = \sup_{x\in X} -f(x) = -\inf_{x\in X} f(x)$. □

We will also have reason to use the following simple fact.

Corollary 15.45. *Let X be locally compact Hausdorff then $f \mapsto \inf_{x\in X} f(x)$ and $f \mapsto \sup_{x\in X} f(x)$ are continuous.*

Proof. Since $\inf_{x\in X} f(x) = -\sup_{x\in X}(-f(x))$ it suffices to handle the case of $\sup_{x\in X} f(x)$. Let $f \in C_0(X)$ and $\epsilon > 0$ be given. By Corollary 15.44 we see that there exists $x_0 \in X$ such that $f(x_0) = \sup_{x\in X} f(x)$ therefore if $\|g - f\| < \epsilon$ we have

$$\sup_{x\in X} g(x) \geq g(x_0) \geq f(x_0) - \epsilon = \sup_{x\in X} f(x) - \epsilon$$

and if we pick x_1 such that $g(x_1) = \sup_{x\in X} g(x)$ then

$$\sup_{x\in X} g(x) = g(x_1) \leq f(x_1) + \epsilon \leq \sup_{x\in X} f(x) + \epsilon$$

TODO: Prefer the following argument using sequences? In a metric space (or any first countable topological space) convergence is determined by sequences right? Using Corollary 15.44 for each $n \in \mathbb{N}$ we select $x_n \in S$ such that $f_n(x_n) = \inf_{x\in S} f_n(x)$ and select $x_0 \in S$ such that $f(x_0) = \inf_{x\in S} f(x)$. From the definition of x_n and $\lim_{n\to\infty} f_n = f$ we see that $\lim_{n\to\infty} f_n(x_n) \leq \lim_{n\to\infty} f_n(x_0) = f(x_0)$. On the other hand for every $\epsilon > 0$ there exists N such that $\sup_{x\in S} |f_n(x) - f(x)| < \epsilon$ for $n \geq N$ and therefore $f(x_0) - \epsilon \leq f(x_n) - \epsilon \leq f_n(x_n)$ for $n \geq N$. This implies $f(x_0) - \epsilon \leq \lim_{n\to\infty} f_n(x_n)$ and since $\epsilon > 0$ was arbitrary we get $f(x_0) \leq \lim_{n\to\infty} f_n(x_n)$. □

Corollary 15.46. *Let X be locally compact separable Hausdorff then $C_0(X)$ is separable.*

Proof. We know that $C(\tilde{X})$ is separable (Lemma 15.17) so we may take a countable dense set f_n. We claim that $f_n - f_n(\Delta)$ is dense in $C_0(X)$. To see this, note that for any $f \in C_0(X)$ we know that $f_n \to f$ in $C(\tilde{X})$ along some subsequence N. On the other hand, $f \mapsto f - f(\Delta)$ is a continuous map from $C(\tilde{X})$ to $C_0(X)$ and therefore $f_n - f_n(\Delta) \to f$ along N. □

Question: is there a compact non-Hausdorff space that is not locally compact?

Theorem 15.47 (Principle of Uniform Boundedness). *Let X be a Banach space, Y be a normed vector space and $A_\alpha : X \to Y$ be a family of bounded linear maps. If for all $v \in X$ we have $\sup_\alpha \|A_\alpha v\| < \infty$ then $\sup_\alpha \|A_\alpha\| < \infty$.*

Proof. For each $n \in \mathbb{N}$ let

$$V_n = \{x \in X \mid \sup_\alpha \|A_\alpha x\| > n\} = \cup_\alpha \{x \in X \mid \|A_\alpha x\| > n\}$$

and note that V_n is open. Furthermore by our hypothesis we know that $\cap_{n=1}^\infty V_n = \emptyset$. If follows from the Baire Category Theorem 1.47 that some V_n is not dense. Thus there exists an $n \in \mathbb{N}$, $x \in X$ and $r > 0$ such that $\overline{B}(x, r) \cap V_n = \emptyset$; that is to say for all $y \in X$ such that $\|x - y\| \leq r$ we have $\sup_\alpha \|A_\alpha y\| \leq n$. Now let $u \in X$ with $\|u\| \leq 1$ and observe that by linearity for all α

$$\|A_\alpha u\| = r^{-1} \|A_\alpha r u\| \leq r^{-1}(\|A_\alpha(x + ru)\| + \|Ax\|) \leq 2r^{-1}n$$

Therefore $\sup_\alpha \|A_\alpha\| = \sup_\alpha \sup_{\|u\| \leq 1} \|A_\alpha u\| \leq 2r^{-1}n < \infty$. □

7. Riesz Representation

We saw in the Daniell-Stone Theorem 2.142 that one may recapture a part of integration theory by considering certain linear functionals on a space of functions. There is an analogue to that result that applies in the case of measure on topological spaces and allows one to bring the machinery of functional analysis to bear on problems of measure theory.

Definition 15.48. Let X be a locally compact Hausdorff topological space then a Borel measure μ is said to be a *Radon measure* if $\mu(K) < \infty$ for every compact set K and μ is inner regular i.e. if for every Borel set A we have

$$\mu(A) = \sup\{\mu(K) \mid K \subset A \text{ and } K \text{ compact}\}$$

N.B. Other definitions of Radon measure are in use. For example Folland uses the following definition:

(i) (locally finite) $\mu(K) < \infty$ for all compact K
(ii) (outer regularity) $\mu(A) = \inf\{\mu(U) \mid A \subset U \text{ and } U \text{ open}\}$ for every Borel set A
(iii) (inner regularity on open sets) $\mu(U) = \inf\{\mu(K) \mid K \subset U \text{ and } K \text{ compact}\}$ for every open set U

The combination of outer regularity and inner regularity on open sets is sometimes referred to as *quasi-regularity*. For the case of σ-compact locally compact Hausdorff spaces the definitions coincide but dropping σ-compactness the definitions are no longer equivalent. The definition present here is chosen so that the Riesz Representation Theorem below holds. A Riesz Representation Theorem holds with the alternative definition as well. Note that we will not be concerned with Radon measures on non-locally compact spaces.

To illustrate

Example 15.49. Let $\mathbb{R}_d$ be the real line with the discrete topology and let $X = \mathbb{R} \times \mathbb{R}_d$. TODO: Show that X is locally compact. Show that the compact subsets of X are all of the form $\cup_{j=1}^n K_j \times \{y_j\}$ with K_j compact. Show that if $A \subset X$ is Borel then $A \cap (\mathbb{R}\{y\}) = A_y \times \{y\}$ with A_y a Borel subset of $\mathbb{R}$ with the usual topology. Define

$$\mu(A) = \sum_{y \in \mathbb{R}} \lambda(A_y)$$

Show that μ is locally finite and inner regular. Define

$$\nu(A) = \begin{cases} \sum_{y \in \mathbb{R}} \lambda(A_y) & \text{if at most countably many } A_y \neq \emptyset \\ \infty & \text{otherwise} \end{cases}$$

Show that ν is locally finite and quasi-regular. Show that μ and ν both induce the same continuous linear functional on the compactly supported functions.

The Riesz representation theorem is actually a class of different theorems with different hypotheses made about the measures involved and the topology on the underlying space. Here we concentrate the reasonable general case of Hausdorff locally compact spaces. Other presentations may treat the slightly simpler cases in which either second countability, compactness or σ-compactness are added as hypotheses on the topological space. More general presentations may drop the the assumption of local compactness and treat arbitrary Hausdorff spaces.

Definition 15.50. A topological space X is said to be *locally compact* if every point in X has a compact neighborhood (i.e. for every $x \in X$ there exists an open set U and a compact set K such that $x \in U \subset K$).

Lemma 15.51. *Let X be a Hausdorff topological space then the following are equivalent*

(i) *X is locally compact*
(ii) *Every point in X has an open neighborhood with compact closure*
(iii) *X has a base of relatively compact neighborhoods*

Proof. (i) implies (ii): If X is Hausdorff then a closed subset of a compact set is compact and therefore if X is locally compact and $x \in X$ we take U open and K compact such that $x \in U \subset K$ and then it follows that $\overline{U}$ is compact hence (ii) follows.

The fact that (ii) implies (i) is immediate.

(ii) implies (iii): For each $x \in X$ pick a relatively compact neighborhood U_x, let $\mathcal{B}_x = \{U \subset U_x \mid U \in \mathcal{T}\}$ and let $\mathcal{B} = \cup_{x \in X} \mathcal{B}_x$. It is clear that $\mathcal{B}$ is a base for the topology $\mathcal{T}$. Moreover for each $U \in \mathcal{B}$ there exists $x \in X$ such that $U \subset U_x$ with $\overline{U}_x$ compact and then since X is Hausdorff we know that $\overline{U}$ is compact.

(iii) implies (ii) is immediate. $\square$

A locally compact space can be embedded in a compact space by adding a single point.

Definition 15.52. Let X be a locally compact Hausdorff space with topology τ. The *one point compactification* of X is the set $X \cup \{\Delta\}$ where $\Delta \notin X$ and topology given by τ and sets of the form $X \setminus K \cup \{\Delta\}$ where $K \subset X$ is compact. The point Δ is called the *point at infinity.*

The one point compactification is indeed compact.

Theorem 15.53. *The one point compactification is a compact space.*

Proof. First we show that $\tilde{\tau} = \tau \cup \cup_{\substack{K \in \tau \\ K \text{ is compact}}} X \setminus K \cup \{\Delta\}$ is a topology. It is clear that $\emptyset \in \tilde{\tau}$ since $\emptyset \in \tau$ and moreover since $\emptyset$ is compact we see that $X \cup \{\Delta\} = (X \setminus \emptyset) \cup \{\Delta\} \in \tilde{\tau}$. Let $U_1, \dots, U_n$ be open and and $K_1, \dots, K_m$ be compact then if $n > 0$

$$\begin{aligned} &U_1 \cap \dots \cap U_n \cap (X \setminus K_1 \cup \{\Delta\}) \cap \dots \cap (X \setminus K_m \cup \{\Delta\}) \\ &= U_1 \cap \dots \cap U_n \cap (X \setminus K_1 \cap \dots \cap X \setminus K_m \in \tau \subset \tilde{\tau} \end{aligned}$$

and if $n = 0$ then

$$(X \setminus K_1 \cup \{\Delta\}) \cap \dots \cap (X \setminus K_m \cup \{\Delta\}) = (X \setminus K_1 \cap \dots \cap X \setminus K_m) \cup \{\Delta\} \in \tilde{\tau}$$

If A and B are arbitrary index sets and we have U_α for $\alpha \in A$ open in X and K_β for $\beta \in B$ compact in X then if we define $U = \cup_{\alpha \in A} U_\alpha$ and $K = \cap_{\beta \in B} K_\beta$ note that $U \in \tau$ and K is a closed subset of a compact set hence compact. Furthermore, $U^c \cap K$ is compact for the same reason. We compute using De Morgan's Law

$$\begin{aligned}\cup_{\alpha\in A} U_\alpha \cup \cup_{\beta\in B}(X \setminus K_\beta) \cup \{\Delta\} &= [U \cup \cup_{\beta\in B}(X \setminus K_\beta)] \cup \{\Delta\} \\ &= X \setminus (U^c \cap K) \cup \{\Delta\} \in \tilde{\tau}\end{aligned}$$

Lastly note that $X \cup \{\Delta\}$ is compact. If U_α and $(X \setminus K_\beta) \cup \{\Delta\}$ cover $X \cup \{\Delta\}$. As before $\cap_{\beta\in B} K_\beta$ is compact and it follows that the U_α for $\alpha \in A$ form a cover. Thus we may find a finite subcover $U_{\alpha_1}, \dots, U_{\alpha_n}$ and it follows that $\left\{^{U_{\alpha_1},\dots,U_{\alpha_n},(X\setminus\cap_{\beta\in B}K_\beta)\cup\{\Delta}\right\}$ is a finite subcover of $X \cup \{\Delta\}$. □

Proposition 15.54. *A locally compact Hausdorff space X is completely regular (i.e. for every $x \in X$ and closed set $F \subset X$ such that $x \notin F$ there is are disjoint open sets U and V such that $x \in U$ and $F \subset V$).*

Proof. Let F be a closed set and pick $x \in X \setminus F$. By Lemma 15.51 and the openness of $X \setminus F$ we can find a relatively compact neighborhood U_0 of x such that $x \in U_0 \subset X \setminus F$. The set $\overline{U_0} \cap F$ is a closed subset of a compact set hence is compact. For each $y \in \overline{U_0} \cap F$ by the Hausdorff property we may find open neighborhoods $x \in U_y$ and $y \in V_y$ such that $U_y \cap V_y = \emptyset$. By compactness of $\overline{U} \cap F$ we get a finite subcover $V_{y_1}, \dots, V_{y_n}$ of $\overline{U_0} \cap F$. Now define $U = U_0 \cap U_{y_1} \cap \cdots U_{y_n}$. This is an open neighborhood of x and moreover $\overline{U} \cap F = \emptyset$. Define $V = X \setminus \overline{U}$. □

TODO: Show that a locally compact separable Hausdorff space is metrizable (in fact Polish).

Definition 15.55. Let X be a topological space, then a subset $A \subset X$ is said to be *bounded* if there exists a compact set K such that $A \subset K \subset X$. A subset $A \subset X$ is said to be *σ-bounded* if there exists a sequence of compact sets $K_1, K_2, \dots$ such that $A \subset \cup_{i=1}^\infty K_i \subset X$.

Proposition 15.56. *A set A is σ-bounded Borel set if and only if there exist disjoint bounded Borel sets $A_1, A_2, \dots$ such that $A = \cup_{i=1}^\infty A_i$.*

Proof. Suppose A is a σ-bounded Borel set and let $K_1, K_2, \dots$ be compact sets such that $A \subset \cup_{i=1}^\infty K_i$. Define $A_1 = A \cap K_1$ and for $n > 1$ let $A_n = A \cap K_n \setminus \cup_{j=1}^{n-1} A_j$ Trivially each A_n is bounded (it is contained in K_n), $A = \cup_{i=1}^\infty A_i$ (by construction $A_n \subset A$ and for any $x \in A$ we can find n such that $x \in K_n$; it follows that $x \in A_n$). Moreover by construction it is clear that the A_n are Borel. On the other hand, if $A = \cup_{i=1}^\infty A_i$ with A_i bounded and Borel and disjoint, then take K_i compact such that $A_i \subset K_i$ and it follows that $A \subset \cup_{i=1}^\infty K_i$. A is clearly Borel as it is a countable union of Borel sets. □

Lemma 15.57. *Let K be a compact set in a locally compact Hausdorff topological space X, then there exists a bounded open set U such that $K \subset U$. Moreover if V is a open set such that $K \subset V$ then there is a bounded open set U such that $K \subset U \subset \overline{U} \subset V$.*

Proof. By taking $V = X$ we see the second assertion implies the first so it suffices to prove the second assertion. By complete regularity of X (Proposition 15.54) and local compactness of X for each $x \in K$ we may find a relatively compact open neighborhood $x \in U_x$ such that $\overline{U}_x \cap V^c = \emptyset$. By compactness of K we may take a finite subcover $U_{x_1}, \dots, U_{x_n}$. Then $U = U_{x_1} \cup \cdots \cup U_{x_n}$ is an open set with $K \subset U$ and $\overline{U} = \overline{U}_{x_1} \cup \cdots \cup \overline{U}_{x_n}$ is a finite union of compact sets and is therefore compact. Lastly $\overline{U} \cap V^c = (\overline{U}_{x_1} \cap V^c) \cup \cdots \cup (\overline{U}_{x_n} \cap V^c) = \emptyset$ and therefore $K \subset U \subset \overline{U} \subset V$. □

For our purposes the reason for bringing up σ-bounded sets is the fact that the properties of inner and outer regularity are essentially equivalent on them.

LEMMA 15.58. *Let X be a locally compact Hausdorff topological space and let μ be a measure that is finite on compact sets. Then μ is inner regular on σ-bounded Borel sets if and only if μ is outer regular on σ-bounded Borel sets.*

PROOF. Suppose that μ is inner regular on σ-bounded sets. Let A be a bounded Borel set and suppose $\epsilon > 0$ is given. First, note that $\overline{A}$ is compact so may apply Lemma 15.57 to find a bounded open set U such that $\overline{A} \subset U$. Therefore $\overline{U} \setminus A$ is a bounded Borel set so by inner regularity we may find a compact set $K \subset \overline{U} \setminus A$ such that

$$\mu(\overline{U} \setminus A) - \epsilon < \mu(K) \leq \mu(\overline{U} \setminus A)$$

Let $V = U \cap K^c$. Then V is an open set and $A \subset V$. Moreover,

$$\mu(V) = \mu(U) - \mu(K) \leq \mu(\overline{U}) - \mu(\overline{U} \setminus A) + \epsilon = \mu(A) + \epsilon$$

Since $\epsilon > 0$ was arbitrary we see that μ is outer regular on bounded Borel sets. Now we need to extend to outer regularity on σ-bounded sets. Let A be a σ-bounded Borel set and let $\epsilon > 0$ be given. Apply Lemma 15.56 to find disjoint bounded Borel sets A_i such that $A = \cup_{i=1}^{\infty} A_i$. By the just proven outer regularity on bounded Borel sets we may find open sets U_i such that $\mu(U_i) \leq \mu(A_i) + \epsilon/2^i$. Then clearly $A \subset U$, U is open and

$$\mu(U) \leq \sum_{i=1}^{\infty} \mu(U_i) \leq \epsilon + \sum_{i=1}^{\infty} \mu(A_i) = \epsilon + \mu(A)$$

Again, as $\epsilon > 0$ is arbitrary we see that μ is outer regular on σ-bounded Borel sets.

Now we assume that μ is outer regular on σ-bounded Borel sets. As before we start with the bounded case. Let A be a bounded Borel set and suppose that $\epsilon > 0$ is given. Let L be a compact set such that $A \subset L$. Since $L \setminus A$ is also a bounded Borel set, we may apply outer regularity to find an open set U such that $L \setminus A \subset U$ and

$$\mu(U) - \epsilon < \mu(L \setminus A) \leq \mu(U)$$

Define $K = L \setminus U = L \cap U^c$. As K is a closed subset of the compact set L it is compact. Also

$$\mu(K) = \mu(L) - \mu(L \cap U) \geq \mu(L) - \mu(U) = \mu(A) + \mu(L \setminus A) - \mu(U) > \mu(A) - \epsilon$$

As $\epsilon > 0$ was arbitrary we see that μ is inner regular on bounded Borel sets.

Lastly we extend inner regularity to σ-bounded Borel sets. Let A be σ-bounded Borel and write $A = \cup_{i=1}^{\infty} A_i$ with the A_i disjoint and each A_i bounded Borel (Lemma 15.56). Let $\epsilon > 0$ be given. As each A_i is bounded and μ is finite on compact sets it follows that $\mu(A_i) < \infty$ for all $i \in \mathbb{N}$. Now by the just proven inner regularity on bounded Borel sets we find $L_i \subset A_i$ with L_i compact and

$$\mu(A_i) - \epsilon/2^i < \mu(L_i) \leq \mu(A_i)$$

The disjointness of the A_i implies that the L_i are disjoint as well. Let $K_n = L_1 \cup \cdots \cup L_n$ and note that

$$\mu(K_n) = \sum_{i=1}^{n} \mu(L_i) > \sum_{i=1}^{n} \left(\mu(A_i) - \epsilon/2^i\right) > \sum_{i=1}^{n} \mu(A_i) - \epsilon$$

Now take the limit as $n \to \infty$ to conclude that

$$\sup\{\mu(K) \mid K \subset A \text{ and } K \text{ is compact}\} \geq \sup_n \mu(K_n) \geq \mu(A) - \epsilon$$

and as $\epsilon > 0$ was arbitrary inner regularity of μ on σ-bounded Borel sets is proven. □

The difficult part of the Riesz-Markov Theorem is the construction of a Radon measure that corresponds to a positive functional. The tradition is to break that construction into two pieces: first the construction of a set function on a smaller class of sets than the full σ-algebra and secondly the extension of that set function to a full blown Radon measure. In many developments the set function is defined on the compact subsets of the locally compact Hausdorff space X and are called *contents*. Following Arveson, we choose a set function is one that is defined on just the open subsets of X.

The description of the desireable properties of the set function and the process of extending the set function to a Radon measure is dealt with in the following Lemma.

LEMMA 15.59. *Let X be a locally compact Hausdorff space and let m be a function from the open set of X to $[0, \infty]$ satisfying:*

(i) *$m(U) < \infty$ if $\overline{U}$ is compact*
(ii) *if $U \subset V$ then $m(U) \leq m(V)$*
(iii) *$m(\cup_{i=1}^{\infty} U_i) \leq \sum_{i=1}^{\infty} m(U_i)$ for all open sets $U_1, U_2, \dots$.*
(iv) *if $U \cap V = \emptyset$ then $m(U \cup V) = m(U) + m(V)$*
(v) *$m(U) = \sup\{m(V) \mid V$ is open, $\overline{V} \subset U$ and $\overline{V}$ is compact$\}$*

then there is a unique Radon measure μ such that $\mu(U) = m(U)$ for all open sets U. Moreover every Radon measure satisfies properties (i) through (v) when restricted to the open subsets of X.

PROOF. First we show that a Radon measure satisfies properties (i) through (v) on the open sets of X. In fact, properties (ii), (iii) and (iv) follow for all measures and (i) follows from the fact that μ is finite on compact subsets and monotonicity of measure. Property (v) requires a bit more justification. If we let U is an open set and $\epsilon > 0$ is given then by inner regularity of μ we may find a compact set K such that $K \subset U$ and $\mu(U) \geq \mu(K) > \mu(U) - \epsilon$. By Lemma 15.57 we may find a relatively compact open set V such that $K \subset V \subset \overline{V} \subset U$. Then by monotonicity we have $\mu(U) \geq \mu(V) > \mu(U) - \epsilon$ and since ϵ was arbitrary (v) follows.

Next we prove uniqueness of the extension of m to a Radon measure μ. Since a Radon measure is inner regular on all Borel sets Lemma 15.58 implies that any extension μ is outer regular on all σ-bounded Borel sets. Since the values of μ are determined on all open sets this implies that the values of μ are determined on all σ-bounded Borel sets; in particular the values of μ are determined on all compact sets. Clearly a Radon measure is determined uniquely by its values on compact sets.

Now we turn to proving existence of the extension μ. The proof goes in a few steps. First we define an outer measure from m and observe that Borel sets are measurable with respect to it; though the Caratheordory restriction of the outer measure is outer regular it is not necessarily inner regular. The second step is to modify the Caratheodory restriction to make it inner regular.

We begin by defining the outer measure in a standard way. Let A be an arbitrary subset of X and define

$$\mu^*(A) = \inf\{m(U) \mid A \subset U \text{ and } U \text{ is open}\}$$

Note that $\mu^*(U) = m(U)$ for all open sets.

Claim 15.59.1. μ^* is an outer measure

Note that because the emptyset is relatively compact we know from (i) that $m(\emptyset) < \infty$ and thus from (iv) we see that $m(\emptyset) = 2m(\emptyset)$. Thus $m(\emptyset) = 0$ and it follows that $\mu^*(\emptyset) = 0$. If $A \subset B$ then it is trival that

$$\{m(U) \mid A \subset U \text{ and } U \text{ is open }\} \subset \{m(U) \mid B \subset U \text{ and } U \text{ is open }\}$$

which implies $\mu^*(A) \leq \mu^*(B)$. If we let $A_1, A_2, \dots$ be given and define $A = \cup_{i=1}^\infty A_i$. If any $\mu^*(A_i) = \infty$ it follows that $\mu(A) \leq \sum_{i=1}^\infty \mu^*(A_i) = \infty$. If on the other hand every $\mu^*(A_i) < \infty$ then let $\epsilon > 0$ be given and find an open set $U_i \subset A_i$ such that $m(U_i) \leq \mu^*(A_i) + \epsilon/2^i$. Clearly $\cup_{i=1}^\infty U_i$ is an open subset of A and it follows from (iii) and the definition of μ^* that

$$\mu^*(A) \leq m(\cup_{i=1}^\infty U_i) \leq \sum_{i=1}^\infty m(U_i) \leq \sum_{i=1}^\infty m(A_i) + \epsilon$$

Since $\epsilon > 0$ is arbitrary we see that μ^* is countably subadditive and is therefore proven to be an outer measure.

Claim 15.59.2. Borel sets are μ^*-measurable.

The μ^*-measurable sets form a σ-algebra by Lemma 2.66 and therefore it suffices to show that open sets are μ^*-measurable. Let U be open subset and A be an arbitrary subset of X, by subadditivity of μ^* we only have to show the inequality

$$\mu^*(A) \geq \mu^*(A \cap U) + \mu^*(A \cap U^c)$$

Obviously we may assume that $\mu^*(A) < \infty$ since otherwise the inequality is trivially satisfied. We first assume that A is an open set. Since μ^* and m agree on open sets we have to show

$$m(A) \geq m(A \cap U) + \mu^*(A \cap U^c)$$

Let $\epsilon > 0$ be given and use property (v) so we can find an relatively compact open set V such that $\overline{V} \subset A \cap U$ and $m(V) \geq m(A \cap U) - \epsilon$. Then $A \cap \overline{V}^c$ is an open set containing $A \cap U^c$ disjoint from V and it follows from (ii), (iv) and the definition of μ^* that

$$m(A) \geq m(V \cup A \cap \overline{V}^c) = m(V) + m(A \cap \overline{V}^c) \geq m(A \cap U) - \epsilon + \mu^*(A \cap U^c)$$

As $\epsilon > 0$ was arbitrary we are done with the case of open sets A. Now suppose that A is an arbitrary set with $\mu^*(A) < \infty$ and let $\epsilon > 0$ be given. We find an open set V such that $A \subset V$ and $m(V) \leq \mu^*(A) + \epsilon$. From what we have just proven of open sets and the monotonicity of μ^*

$$\mu^*(A) + \epsilon \geq \mu^*(V) \geq \mu^*(V \cap U) + \mu^*(V \cap U^c) \geq \mu^*(A \cap U) + \mu^*(A \cap U^c)$$

The claim follows by observing that $\epsilon > 0$ was arbitrary.

Now by Caratheodory Restriction (Lemma 2.66) we may restrict μ^* to a Borel measure $\overline{\mu}$ that is outer regular by definition and that satisfies $\overline{\mu}(U) = m(U)$ for all open sets U. Moreover $\overline{\mu}(K) < \infty$ for all compact sets since by Lemma 15.57 we may find a relatively compact open neighborhood U such that $K \subset U$; monotonicity and (i) tell us that

$$\overline{\mu}(K) \leq \overline{\mu}(U) = m(U) < \infty$$

Since $\overline{\mu}$ is outer regular on all Borel sets *a fortiori* it is outer regular on all σ-bounded Borel sets. By Lemma 15.58 it follows that $\overline{\mu}$ is inner regular on all σ-bounded Borel sets. Note that if we assume that X is σ-compact (i.e. all Borel sets are σ-bounded) then we already know that $\overline{\mu}$ is a Radon measure. In the general case it is not necessarily true and we must make a further modification to $\overline{\mu}$ to make it inner regular.

For an arbitrary Borel set A we define

$$\mu(A) = \sup\{\overline{\mu}(B) \mid B \subset A \text{ and } B \text{ is a } \sigma\text{-bounded Borel set }\}$$

Clearly, $\mu(\emptyset) = \overline{\mu}(\emptyset) = 0$. It is also immediate from the definition that $\mu(A) = \overline{\mu}(A)$ for all σ-bounded Borel sets A and therefore that $\mu(U) = m(U)$ for all σ-bounded open sets U. In fact more is true.

Claim 15.59.3. $\mu(U) = m(U)$ for all open sets U.

Let V be a relatively compact open set with $\overline{V} \subset U$. We have

$$m(U) = \overline{\mu}(U) \geq \mu(U) \geq \mu(V) = m(V)$$

Now we take the supremum over all such V and by property (v)

$$m(U) \geq \mu(U) \sup\{m(V) \mid V \text{ is relatively compact and } \overline{V} \subset U\} = m(U)$$

and therefore $\mu(U) = m(U)$.

Claim 15.59.4. μ is a measure.

To see that μ is a measure it remains to show countable additivity. Let $A_1, A_2, \ldots$ be disjoint Borel sets. First we show countable subadditivity. Let B be a σ-bounded Borel subset of $\cup_{i=1}^\infty A_i$ and define $B_i = B \cap A_i$. Clearly the B_i are disjoint σ-bounded Borel measures, thus using the countable additivity of $\overline{\mu}$ we get

$$\overline{\mu B} = \sum_{i=1}^\infty \overline{\mu}(B_i) \leq \sum_{i=1}^\infty \mu(A_i)$$

Taking the supremum over all such B subadditivity follows.

We need to show the opposite inequality. Suppose that some $\mu(A_j) = \infty$ for some j. Then we may find a sequence of σ-bounded Borel sets B_n such that $\overline{\mu}(B_n) \geq n$. Since $B_n \subset \cup_{i=1}^\infty A_i$ we also see that $\mu(\cup_{i=1}^\infty A_i) = \infty$. Thus we may now assume that $\mu(A_i) < \infty$ for all i. Let $\epsilon > 0$ be given and for each i find a σ-bounded Borel set B_i such that $B_i \subset A_i$ and $\overline{\mu}(B_i) \geq \mu(A_i) - \epsilon/2^i$. For each n define

$C_n = \cup_{j=1}^n B_j$ and note that C_n is a σ-bounded Borel set such that $C_n \subset \cup_{i=1}^\infty A_i$. Also, for every n,

$$\mu(\cup_{i=1}^\infty A_i) \geq \mu(C_n) = \overline{\mu}(C_n) = \sum_{j=1}^n \overline{\mu}(B_j) \geq \sum_{j=1}^n \mu(A_j) - \epsilon/2^j \geq \sum_{j=1}^n \mu(A_j) - \epsilon$$

Now take the limit as $n \to \infty$ and using the fact that $\epsilon > 0$ was arbitrary, we get $\sum_{j=1}^\infty \mu(A_j) \leq \mu(\cup_{j=1}^\infty A_j)$.

CLAIM 15.59.5. μ is a Radon measure.

The fact that $\mu(K) < \infty$ for all compact sets follows from the fact that μ and $\overline{\mu}$ agree on σ-bounded sets and the fact that $\overline{\mu}(K) < \infty$. To see inner regularity, let A be a Borel set and let $\epsilon > 0$ be given. By the definition of μ we find a σ-bounded Borel set $B \subset A$ such that $\overline{\mu}(B) \geq \mu(A) - \epsilon/2$. Then by the fact that $\overline{\mu}$ is inner regular on σ-bounded sets we find a compact set K such that $\overline{\mu}(K) \geq \overline{\mu}(B) - \epsilon/2$. Combining the two inequalities and using the fact that μ and $\overline{\mu}$ agree on compact sets we get $\mu(K) \geq \mu(A) - \epsilon$. Since $\epsilon > 0$ was arbitrary we are done. ☐

Given a Radon measure on a locally compact Hausdorff space, all compactly supported continuous functions are integrable: $\int |f| \, d\mu \leq \|f\|_\infty \mu(supp(f)) < \infty$. Thus such a measure yields a linear functional on $C_c(X)$. Such functionals possess another simple property in addition to linearity.

DEFINITION 15.60. A linear functional Λ on $C_c(X)$ is said to be *positive* if $f \geq 0$ implies $\Lambda(f) \geq 0$.

The reader should be careful that we have not been using any type of topology on $C_c(X)$ at this point and in particular we are not claiming that the function defined by a Radon measure is continuous with respect to any underlying topology. In fact as we will see later, $C_c(X)$ is a normed vector space (not complete) under the sup norm but it is only for finite Radon measures that the corresponding functional is continuous.

It is clear that the linear functional defined by integration with respect to a Radon measure is positive. The Riesz-Markov Theorem tells us that the positive linear functionals are precisely those generated by integration with respect to a Radon measure. To prove the result we will need to figure out how to define a measure from a positive linear functional. As a warm up let's first answer that question in the case of integration with respect to a Radon measure.

LEMMA 15.61. *Let X be a locally compact Hausdorff space and let μ be a Radon measure on X, then for every open set U we have*

$$\mu(U) = \sup\{\int f \, d\mu \mid 0 \leq f \leq 1, f \in C_c(X), supp(f) \subset U\}$$

PROOF. For the inequality $\geq$, suppose that $f \in C_c(X)$ satisfies $0 \leq f \leq 1$ and $supp(f) \subset U$, then observe the hypotheses imply that $f \leq \mathbf{1}_U$ so that

$$\int f(x) \, d\mu(x) \leq \int \mathbf{1}_U(x) \, d\mu(x) = \mu(U)$$

and the inequality follows by taking the supremum over all such f.

For the inequality $\leq$ we leverage the inner regularity of μ. Let $K \subset U$ be a compact set. By Lemma 15.57 we find an relatively compact open set V with

$K \subset V \subset \overline{V} \subset U$. Since $\overline{V}$ is a compact Hausdorff space, it is normal and we may apply the Tietze Extension Theorem 15.10 to find a continuous function $g : \overline{V} \to [0, 1]$ such that $g \equiv 1$ on K. Applying Urysohn's Lemma 15.9 we construct a continuous function $h : X \to [0, 1]$ such that $h = 1$ on K and $h = 0$ on V^c. We define

$$f(x) = \begin{cases} h(x)g(x) & \text{if } x \in \overline{V} \\ 0 & \text{if } x \notin \overline{V} \end{cases}$$

By the corresponding properties of g and h, it is clear that $0 \leq f \leq 1$ and that $f = 1$ on K. We claim that f is continuous on all of X. Since h restricts to a continuous function on $\overline{V}$ it is clear that the restriction of f to $\overline{V}$ is continuous. Let $O \subset \mathbb{R}$ be an open set. If $0 \notin O$ then it follows that $f^{-1}(O) \subset V$ and is therefore open by the continuity of f restricted to V. If on the other hand, $0 \in O$ then $f^{-1}(O) \cap \overline{V}$ is open in $\overline{V}$ hence is of the form $Z \cap \overline{V}$ for some open subset $Z \subset X$. Because $0 \in O$ if follows that $Z \subset f^{-1}(O)$ and therefore by the definition of f we we may write $f^{-1}(O) = Z \cup \overline{V}^c$ which is an open set.

TODO: This is a locally compact Hausdorff version of Tietze, factor it out into a separate result.

With the extension f in hand we see that

$$\mu(K) \leq \int f(x)\, d\mu(x) \leq \sup\{\int f\, d\mu \mid 0 \leq f \leq 1, f \in C_c(X), supp(f) \subset U\}$$

Now taking the supremum over all compact subsets $K \subset U$ and using the inner regularity of μ the result follows. $\square$

Before we state an prove the Riesz-Markov theorem we need the existence of finite partitions of unity on compact sets in an LCH: a standard bit of general topology.

LEMMA 15.62. *Let X be a locally compact Hausdorff space, K be a compact subset of X and $\{U_\alpha\}$ an open covering of K. There exists a finite subset $\alpha_1, \dots, \alpha_n$ and continuous functions with compact support $f_{\alpha_1}, \dots, f_{\alpha_n}$ such that $supp(f_{\alpha_j}) \subset U_{\alpha_j}$ and $f_{\alpha_1} + \cdots + f_{\alpha_n} = 1$ on K.*

PROOF. Pick an $x \in K$, pick an U_{α_x} such that $x \in U_{\alpha_x}$ and using complete regularity of X, construct a continuous function g_x from X to $[0, 1]$ such that $g_x(x) = 1$ and $g_x \equiv 0$ on $U^c_{\alpha_x}$. Thus $g_x^{-1}(0, 1] \subset U_{\alpha_x}$ and the $g_x^{-1}(0, 1]$ form an open cover of K. By compactness of K we extract a finite subcover $g_{x_1}^{-1}(0, 1], \dots, g_{x_n}^{-1}(0, 1]$. If we clean up notation by denoting $U_{\alpha_{x_j}} = U_{\alpha_j}$ and $g_{\alpha_j} = g_{\alpha_{x_j}}$, it follows that $U_{\alpha_1}, \dots, U_{\alpha_n}$ is an open cover of K and $g = \sum_{j=1}^n g_{\alpha_j}$ is strictly positive on K. Moreover by compactness of K we know that g has a minimum value $C > 0$ on K. Define $h = g \vee C$ so that h is continuous, $h = g$ on K and $h \geq C > 0$ everywhere on X. By continuity and strict positivity of h we can define $f_{\alpha_j} = g_{\alpha_j}/h$ so that f_{α_j} is continuous and moreover $f_{\alpha_1} + \cdots + f_{\alpha_n} = g/h = 1$ on K. $\square$

THEOREM 15.63 (Riesz-Markov Theorem). *Let X be a locally compact Hausdorff space and let $\Lambda : C_c(X) \to \mathbb{R}$ be a positive linear functional then there exists a unique Radon measure μ such that $\Lambda(f) = \int f\, d\mu$ for all $f \in C_c(X)$.*

PROOF. The uniqueness part of the result is straightforward. By Lemma 15.61 we know that the values of μ on open sets are determined by Λ. By Lemma 15.58

we conclude that the values of μ on σ-bounded Borel sets are determined by Λ, in particular the values on compact sets are determined by Λ. The inner regularity of μ implies that the values on all Borel sets are determined by Λ.

For existence we follow the lead of Lemma 15.61 and define the set function on the open sets of X

$$m(U) = \sup\{\Lambda(f) \mid 0 \leq f \leq 1, f \in C_c(X), supp(f) \subset U\}$$

We proceed by showing that $m(U)$ satisfies properties (i) through (v) from Lemma 15.59 and that if μ is the Radon measure constructed by that result that we indeed have $\Lambda(f) = \int f \, d\mu$.

Claim 15.63.1. m satisfies (i)

Let U be a relatively compact open set. By Lemma 15.57 we can find another relatively compact open set V such that $\overline{U} \subset V$. By the Tietze Extension Theorem argument of Lemma 15.59 we can find a continuous function $g : X \to [0,1]$ such that $g = 1$ on $\overline{U}$ and $g = 0$ on V^c. Since $g \in C_c(X)$ we have $\Lambda(g) < \infty$. Now suppose that $f \in C_c(X)$ satisfies $0 \leq f \leq 1$ and $f \in C_c(X), supp(f) \subset U$. It follows that $0 \leq f \leq g$ and linearity and positivity of Λ we know that $\Lambda(f) \leq \Lambda(g)$. Taking the supremum over all such f we get

$$m(U) = \sup\{\Lambda(f) \mid 0 \leq f \leq 1, f \in C_c(X), supp(f) \subset U\} \leq \Lambda(g) < \infty$$

Claim 15.63.2. m satsifies (ii)

This is immediate since $U \subset V$ implies

$$\{\Lambda(f) \mid 0 \leq f \leq 1, f \in C_c(X), supp(f) \subset U\} \subset \{\Lambda(f) \mid 0 \leq f \leq 1, f \in C_c(X), supp(f) \subset V\}$$

Claim 15.63.3. m satisfies (iii)

Let $U_1, U_2, \dots$ be open sets and let $f \in C_c(X)$ satisfy $0 \leq f \leq 1$ and $supp(f) \subset \cup_{n=1}^{\infty} U_n$. By compactness of $supp(f)$ and Lemma 15.62 we may find an N and continuous functions g_i for $i = 1, \dots, N$ such that $0 \leq g_i \leq 1$, $supp(g_i) \subset U_i$ and $\sum_{i=1}^{N} g_i = 1$ on $supp(f)$ and therefore $f = f \cdot \sum_{i=1}^{N} g_i$. It also follows that $0 \leq f g_i \leq 1$ and $supp(f g_i) \subset U_i$ for $i = 1, \dots, N$ and thus

$$\Lambda(f) = \sum_{i=1}^{N} \Lambda(f g_i) \leq \sum_{i=1}^{N} m(U_i) \leq \sum_{i=1}^{\infty} m(U_i)$$

Now we take the supremum over all such f to conclude that $m(\cup_{i=1}^{\infty} U_i) \leq \sum_{i=1}^{\infty} m(U_i)$.

Claim 15.63.4. m satisfies (iv)

Suppose U and V are disjoint open sets. We only need to show that $m(U \cup V) \geq m(U) + m(V)$ since the opposite inequality follows from (iii). Let $f, g \in C_c(X)$ such that $0 \leq f, g \leq 1$, $supp(f) \subset U$ and $supp(g) \subset V$. By disjointness of U and V it follows that $f + g \in C_c(X)$, $0 \leq f + g \leq 1$ and $supp(f+g) \subset supp(f) \cup supp(g) \subset U \cup V$. Therefore

$$\Lambda(f) + \Lambda(g) = \Lambda(f+g) \leq m(U \cup V)$$

Now take the supremum over all f and g to get the result.

Claim 15.63.5. m satisfies (v)

Let U be an open set and let $f \in C_c(X)$ such that $0 \leq f \leq 1$ and $supp(f) \subset U$. By compactness of $supp(f)$ and Lemma 15.57 we may find a relatively compact open set V such that $supp(f) \subset V \subset \overline{V} \subset U$. It follows that

$$\Lambda(f) \leq m(V) \leq \sup\{m(V) \mid V \text{ is open, } \overline{V} \subset U \text{ and } \overline{V} \text{ is compact}\}$$

Now take the supremum over all such f.

We may now apply Lemma 15.59 to construct a Radon measure μ such that $\mu(U) = m(U)$. We need to show that for every $f \in C_c(X)$ we have $\Lambda(f) = \int f\, d\mu$. By linearity we know that $\Lambda(0) = \int 0\, d\mu = 0$ so we may assume that $f \neq 0$. We may write $f = f_+ - f_-$ with $f_+ = f \vee 0 \in C_c(X)$ and $f_- = (-f) \vee 0 \in C_c(X)$. Since both Λ and the integral are linear it suffices to show the result for $f \geq 0$. Since f is continuous with compact support, it follows that f is bounded and since $f \neq 0$ we have $0 < \|f\|_\infty < \infty$. Again, by linearity of Λ and integration it suffices to prove the result of $f / \|f\|_\infty$ and thus we may assume that $0 \leq f \leq 1$.

We proceed by constructing a generalized upper and lower sum approximation to the integral of f. Once again apply Lemma 15.57 to find a relatively compact open neighborhood U_0 with $supp(f) \subset U_0$. Let $\epsilon > 0$ be given and choose $n \in \mathbb{N}$ such that $\mu(U_0) < \epsilon n$. For $j = 1, \dots, n$ define $U_j = f^{-1}(j/n, \infty)$. Because f is continuous and of compact support, each U_j is a relatively compact open set and it is trivial from the definitions that we have $\emptyset = U_n \subset U_{n-1} \subset \cdots \subset U_0$. In fact by the continuity of f it is also true that $\overline{U}_j \subset U_{j+1}$. Define the lower and upper approximations to f

$$u(x) = \begin{cases} \frac{j}{n} & \text{if } x \in U_j \setminus U_{j+1} \text{ for some } j = 1, \dots, n-1 \\ 0 & \text{if } x \notin U_1 \end{cases}$$

and similarly

$$v(x) = \begin{cases} \frac{j}{n} & \text{if } x \in U_{j-1} \setminus U_j \text{ for some } j = 1, \dots, n \\ 0 & \text{if } x \notin U_0 \end{cases}$$

Note that we have the property that $u \leq f \leq v$ and moreover we have the useful alternative defintion of u and v

$$u(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{U_j}(x)$$
$$v(x) = \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}_{U_{j-1}}(x)$$

which shows that $u, v \in C_c(X)$.

Claim 15.63.6. $\int (v - u)\, d\mu < \epsilon$

This follows by observing that $v - u = \frac{1}{n}(\mathbf{1}_{U_0} - \mathbf{1}_{U_n}) = \frac{1}{n}\mathbf{1}_{U_0}$ since $U_n = \emptyset$.

Claim 15.63.7. $\int u\, d\mu \leq \Lambda(f) \leq \int v\, d\mu + \epsilon$

To see this claim we decompose f into a representation that is adapted to the nested sequence $U_n \subset \cdots \subset U_0$. For $j = 1, \ldots, n$ define

$$\begin{aligned} \phi_j(x) &= \begin{cases} 1/n & \text{if } x \in U_j \\ f(x) - \frac{j-1}{n} & \text{if } x \in U_{j-1} \setminus U_j \\ 0 & \text{if } x \notin U_{j-1} \end{cases} \\ &= [(f(x) - \frac{j-1}{n}) \vee 0] \wedge \frac{1}{n} \end{aligned}$$

where the second representation shows that $\phi_j \in C_c(X)$ and $0 \leq \phi_j \leq \frac{1}{n}$. Note also that if we are given $x \in U_{j-1} \setminus U_j$ for some $j = 1, \ldots, n$ then $\phi_i(x) = 0$ for $j < i \leq n$ and $\phi_i(x) = \frac{1}{n}$ for $1 \leq i < j$. Therefore we have

$$\begin{aligned} \phi_1(x) + \cdots + \phi_n(x) &= \phi_1(x) + \cdots + \phi_j(x) \\ &= \frac{j-1}{n} + f(x) - \frac{j-1}{n} = f(x) \end{aligned}$$

It is clear that for $x \notin U_0$ we have $f(x) = 0$ and $\phi_j(x) = 0$ for all $j = 1, \ldots, n$ so we have $f = \phi_1 + \cdots + \phi_n$ on all of X.

We now need to bound $\Lambda(\phi_j)$ in terms of $\mu(U_k) = m(U_k)$ for suitable $k = 1, \ldots, n$. First we get a lower bound on $\Lambda(\phi_j)$. Let $1 \leq j \leq n$ be given. Suppose that we have a $g \in C_c(X)$ with $0 \leq g \leq 1$ and $supp(g) \subset U_j$. Then by positivity of ϕ_j and the fact that $\phi_j(x) = \frac{1}{n}$ on U_j we see that $g \leq \mathbf{1}_{U_j} \leq n\phi_j$ and therefore $\Lambda(g) \leq n\Lambda(\phi_j)$. Taking the supremum over all such g we see that $\frac{1}{n}\mu(U_j) \leq \Lambda(\phi_j)$. Taking the sum over all $j = 1, \ldots, n$ and using linearity of Λ we get

$$\int u \, d\mu = \frac{1}{n} \sum_{j=1}^{n} \mu(U_j) \leq \sum_{j=1}^{n} \Lambda(\phi_j) = \Lambda(f)$$

Now we get an upper bound on $\Lambda(\phi_j)$. For $j = 2, \ldots, n$ we that $nphi_j \in C_c(X)$, $0 \leq n\phi_j \leq 1$ and $supp(n\phi_j) \subset \overline{U_{j-1}} \subset U_{j-2}$. From the definition of $\mu(U_{j-2}) = m(U_{j-2})$ it follows that $\Lambda(\phi_j) \leq \frac{1}{n}\mu(U_{j-2})$. As for ϕ_1, we have $n\phi_1 \in C_c(X)$ and $0 \leq n\phi_1 \leq 1$ by exactly the same argument as for $j \geq 2$. We also have $supp(n\phi_1) \subset supp(f) \subset U_0$ so that $\Lambda(\phi_1) \leq \frac{1}{n}\mu(U_0)$. If we define $U_{-1} = U_0$ the we get have $\Lambda(\phi_j) \leq \frac{1}{n}\Lambda(U_{j-2})$ for $j = 1, \ldots, n$. Again we sum and use linearity of Λ,

$$\begin{aligned} \Lambda(f) = \sum_{j=1}^{n} \Lambda(\phi_j) &\leq \frac{1}{n} \sum_{j=1}^{n} \mu(U_{j-2}) = \frac{1}{n}\mu(U_{-1}) + \frac{1}{n} \sum_{j=1}^{n-1} \mu(U_{j-1}) \\ &\leq \frac{1}{n}\mu(U_0) + \frac{1}{n} \sum_{j=1}^{n} \mu(U_{j-1}) \leq \epsilon + \int v \, d\mu \end{aligned}$$

It remains to stitch together the previous claims to show that $\Lambda(f) = \int f \, d\mu$. Integrating the inequality $u \leq f \leq v$ we get $\int u \, d\mu \leq f \leq \int v \, d\mu$. Now using this fact and previous two claims we get

$$\Lambda(f) - \int f \, d\mu \leq \Lambda(f) - \int v \, d\mu \leq \int u \, d\mu - \int u \, d\mu + \epsilon \leq 2\epsilon$$

and

$$\Lambda(f) - \int f\,d\mu \geq \int u\,d\mu - \int f\,d\mu \geq \int u\,d\mu - \int v\,d\mu \geq -\epsilon$$

from which we conclude that $\left|\Lambda(f) - \int f\,d\mu\right| \leq 2\epsilon$. Since $\epsilon > 0$ was arbitrary we are done. $\square$

Given a finite signed measure μ and any $|\mu|$-integrable f we define

$$\int f\,d\mu = \int f\,d\mu_+ - \int f\,d\mu_-$$

Note that this is well defined since $\mu_\pm$ are uniquely determined by μ.

Theorem 15.64. *Let $(\Omega, \mathcal{A})$ be a measurable space space then the space of bounded signed measures is a Banach space with norm given by*

$$\|\mu\| = |\mu|(\Omega) = \mu_+(\Omega) + \mu_-(\Omega)$$

Moreover $\|\mu\| = \sup_{|f|\leq 1} \int f(x)\,\mu(dx)$.

Proof. First we show that $\mu_+(\Omega) + \mu_-(\Omega) = \sup_{|f|\leq 1} \int f(x)\,\mu(dx)$. Taking the Hahn Decomposition $\Omega = A_+ \cup A_-$ so that $\mu_\pm(A) = \pm\mu(A_\pm \cap A)$, we see that

$$\begin{aligned}
\mu_+(\Omega) + \mu_-(\Omega) &= \mu_+(A_+) + \mu_-(A_-) = \mu_+(A_+) - \mu_+(A_-) - \mu_-(A_+) + \mu_-(A_-) \\
&= \int (\mathbf{1}_{A_+} - \mathbf{1}_{A_-})(x)\,\mu_+(dx) - \int (\mathbf{1}_{A_+} - \mathbf{1}_{A_-})(x)\,\mu_-(dx) \\
&= \int (\mathbf{1}_{A_+} - \mathbf{1}_{A_-})(x)\,\mu(dx) \leq \sup_{|f|\leq 1} \int f(x)\,\mu(dx)
\end{aligned}$$

Given an arbitrary measurable f with $\|f\| \leq 1$ we let $f_\pm = \pm f \mathbf{1}_{A_\pm}$ then $f \leq 1$ implies $f_+ \leq \mathbf{1}_{A_+}$ and $-1 \leq f$ implies $-f_- \leq \mathbf{1}_{A_-}$. Furthermore $f_\pm = 0$ is $\mu_\mp$-almost everywhere therefore $\int f_\pm(x)\,\mu_\mp(x) = 0$ and we get

$$\begin{aligned}
\int f(x)\,\mu(dx) &= \int (f_+(x) - f_-(x))\,\mu(dx) = \int f_+(x)\,\mu_+(dx) + \int f_-(x)\,\mu_-(dx) \\
&\leq \mu_+(A_+) + \mu_-(A_-) = \|\mu\|
\end{aligned}$$

Now we show that $\|\mu\|$ is in fact a norm. If $\|\mu\| = 0$ then we have $\mu_+ = \mu_- = 0$ and therefore $\mu = \mu_+ - \mu_- = 0$. From the proof of Theorem 2.109 we see that $\mu_+(\Omega) = \sup_{A\in\mathcal{A}} \mu(A)$ and $\mu_-(\Omega) = -\inf_{A\in\mathcal{A}} \mu(A)$. If $a > 0$ it follows that $(a\mu)_\pm(\Omega) = a\mu_\pm(\Omega)$ and that if $a < 0$ it follows that $(a\mu)_\pm(\Omega) = -a\mu_\mp(\Omega)$; therefore we have $\|a\mu\| = |a|\,\|\mu\|$. Moreover given two finite signed measures ν and μ we have

$$\begin{aligned}
\|\mu + \nu\| &= \sup_{A\in\mathcal{A}} (\mu(A) + \nu(A)) - - \inf_{A\in\mathcal{A}} (\mu(A) + \nu(A)) \\
&\leq \sup_{A\in\mathcal{A}} \mu(A) + \sup_{A\in\mathcal{A}} \nu(A)) - \inf_{A\in\mathcal{A}} \mu(A) - \inf_{A\in\mathcal{A}} \nu(A) \\
&= \|\mu\| + \|\nu\|
\end{aligned}$$

Now we show completeness. Suppose that we have a Cauchy sequence μ_n. There exists an $n \in naturals$ such that $\|\mu_{n+m} - \mu_n\| < 1$ for all $m \in \mathbb{N}$ and from this it follows that $\sup_j \|\mu_j\| \leq \|\mu_1\| \vee \cdots \vee \|\mu_{n-1}\| \vee \|\mu_n\| + 1 < \infty$. Thus we may define

$$\nu = \sum_{j=1}^{\infty} 2^{-j} |\mu_j|$$

to get a bounded measure on Ω. From $\mu_{n,\pm}(A) \leq |\mu_n|(A) \leq 2^n \nu(A)$ it follows that $\mu_{n,\pm} \ll \nu$ and therefore we may apply the Radon-Nikodym Theorem 2.110 to construct a positive ν-integrable $g_{n,\pm}$ such that $\mu_{n,\pm}(A) = \int_A g_{n,\pm}(x)\, \nu(dx)$. If we define $g_n = g_{n,+} - g_{n,-}$ then g_n is integrable and $\mu_n(A) = \int_A g_n(x)\, \nu(dx)$. Observe that

$$\begin{aligned}
&\int |g_n(x) - g_m(x)|\; \nu(dx) \\
&= \int (g_n(x) - g_m(x))\mathbf{1}_{g_n \geq g_m}(x)\, \nu(dx) + \int (g_m(x) - g_n(x))\mathbf{1}_{g_m > g_n}(x)\, \nu(dx) \\
&= \int \mathbf{1}_{g_n \geq g_m}(x)\, (\mu_n - \mu_m)(dx) + \int \mathbf{1}_{g_m > g_n}(x)\, (\mu_m - \mu_n)(dx) \\
&= \int (\mathbf{1}_{g_n \geq g_m} - \mathbf{1}_{g_m > g_n})(x)\, (\mu_n - \mu_m)(dx) \\
&\leq \sup_{f \leq 1} \int f(x)\, (\mu_n - \mu_m)(dx) = \|\mu_n - \mu_m\|
\end{aligned}$$

which shows g_n is Cauchy in $L^1(\Omega, \mathcal{A}, \nu)$. By completeness of $L^1(\Omega, \mathcal{A}, \nu)$ we know that there is a ν-integrable g such that $g_n \xrightarrow{L^1} g$. Define $\mu = (g \cdot \nu)$. Then $|\mu|(\Omega) = \int |g|(x)\, \nu(dx) < \infty$ which implies that μ is a bounded signed measure.

$$\begin{aligned}
\|\mu_n - \mu\| &= \sup_{|f| \leq 1} \int f(x)\, (\mu_n - \mu)(dx) = \sup_{|f| \leq 1} \int f(x)(g_n(x) - g(x))\, \nu(dx) \\
&\leq \int |g_n(x) - g(x)|\; \nu(dx)
\end{aligned}$$

which shows that $\lim_{n \to \infty} \mu_n = \mu$. ☐

Definition 15.65. Let X be a locally compact Hausdorff space then a bounded signed measure μ is said to be a *finite signed Radon measure* if its total variation $|\mu| = \mu_+ + \mu_-$ is a Radon measure on X.

Theorem 15.66. *Let X be a locally compact Hausdorff space then the space of finite signed Radon measures is a Banach space with norm given by $\|\mu\| = \mu_+(X) + \mu_-(X)$. Moreover $\|\mu\| = \sup_{|f| \leq 1} \int f(x)\, \mu(dx)$.*

Proof. By Theorem 15.64 it suffices to show that the finite signed Radon measures are a closed subspace of the Banach space of bounded signed measures. Suppose μ_n converges to μ in total variation and each μ_n is a finite Radon measure. Since we know that $|\mu|$ is in fact a finite measure we only have to show inner regularity. TODO: Finish and show that Radon measures are a closed subset in total variation norm. ☐

TODO: The Riesz Representation Theorem itself; namely the dual space of $C_0(X)$ for a locally compact Hausdorff X is isomorphic to the Banach space of finite signed Radon measures with the total variation norm.

THEOREM 15.67. *Let X be a locally compact Hausdorff space then the dual space of $C_0(X)$ is isomorphic to the space of finite signed Radon measures under the total variation norm.*

PROOF. TODO: □

8. Regularity Properties of Borel Measures

In this section we turn to the approximation properties of Borel measures on topological spaces. In particular we concern ourselves with the ability to approximate arbitrary measurable sets by compact, closed and open sets.

DEFINITION 15.68. Let μ be a measure on the Borel σ-algebra of a Hausdorff 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).

PROPOSITION 15.69. *Let μ be a measure on the Borel σ-algebra of a locally compact Hausdorff space S. Then μ is locally finite if and only if $\mu(K) < \infty$ for all compact sets $K \subset S$.*

PROOF. If $\mu(K) < \infty$ for all compact sets K we let $x \in S$ and pick a relatively compact neighborhood U of x. Then $\mu(U) \leq \mu(\overline{U}) < \infty$ which shows μ is locally finite. On the other hand, suppose μ is locally finite and let K be a compact set. For each $x \in K$ we take an open neighborhood U_x such that $\mu(U_x) < \infty$ and then extract a finite subcover $U_{x_1}, \dots, U_{x_n}$. By subadditivity, we have $\mu(K) \leq \mu(U_{x_1}) + \cdots + \mu(U_{x_n}) < \infty$. □

DEFINITION 15.70. Let μ be a Borel measure on a Hausdorff topological space. A 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?) I think this use of closed inner regularity is a bit non-standard should probably get rid of it.

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 (I think we prove this more generally above in the context of LCH spaces; note that every set in $\mathbb{R}^n$ is σ-bounded). 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...).

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 15.71. *Let X be a Hausdorff topological space, $\mathcal{A}$ a σ-algebra on X and μ a finite tight measure. Then*

$$\mathcal{R} = \{A \in \mathcal{A} \mid A \textit{ and } A^c \textit{ are } \mu\textit{-inner regular}\}$$

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 $A_1, A_2, \dots \in \mathcal{R}$ and let $\epsilon > 0$ be given. By finiteness of μ, $\mu(\cup_{n=1}^\infty A_n) < \infty$ and continuity of measure (Lemma 2.30) 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 15.72. *Let (S, d) be a metric space and μ be a Borel measure on $(S, \mathcal{B}(S))$, then 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 2.30) 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 15.71 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 closed inner regularity to inner regularity.

Lemma 15.73. *Let (S, d) be a separable metric space and μ be a finite 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 15.71 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 15.72 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 15.74 (Ulam's Theorem). *Let (S, d) be a complete separable metric space and μ be a finite Borel measure on $(S, \mathcal{B}(S))$, then μ is inner regular.*

PROOF. By Lemma 15.73 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 2.30) 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 1.29 we know K is compact.

To finish the result we claim $\mu(S \setminus K) < \epsilon$:

$$\begin{aligned}
\mu(S \setminus K) &= \mu(S \cap \left(\cap_{m=1}^\infty \cup_{j=1}^{N_m} \overline{B}(x_j, \frac{1}{m})\right)^c) \\
&= \mu(S \cap \cup_{m=1}^\infty \left(\cup_{j=1}^{N_m} \overline{B}(x_j, \frac{1}{m})\right)^c) \\
&= \mu(\cup_{m=1}^\infty S \setminus \cup_{j=1}^{N_m} \overline{B}(x_j, \frac{1}{m})) \\
&\leq \sum_{m=1}^\infty \mu(S \setminus \cup_{j=1}^{N_m} \overline{B}(x_j, \frac{1}{m})) \\
&< \epsilon
\end{aligned}$$

□

THEOREM 15.75. *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? Finite ones are I believe.

Tao proves Riesz representation under assumption of local compactness, Hausdorff and σ-compactness.

Kallenberg proves Riesz representation under assumption of LCH and second countability (this is less general than the Tao result as lcscH implies LCH σ-compact) and targets Radon measures. Our results taken from Arveson are more general as the remove the second countability assumption and the σ-compactness assumptions.

Evans and Gariepy prove Riesz representation only on $\mathbb{R}^n$ using Radon outer measures. This is probably subsumed by our results taken from Arveson but I need to understand whether the use of outer measures adds anything to the picture in general.

Arveson has some well known lecture notes that prove Riesz on general LCH spaces and emphasizes Radon measures (it also explores how Baire measures figure in the picture). I have chosen to follow these notes. Note that Arveson mentions that the lcscH approach of Kallenberg avoids the distinction between Baire and Borel sets (sounds like the σ-algebras agree in this case).

Fremlin probably has some very general account of Reisz representation (of course).

Dudley proves Riesz representation of compact Hausdorff spaces and phrases things in terms of Baire measures. Dudley does not really discuss Radon measures. Arveson discusses the relationship between the use of Baire and Radon measures.

9. Hausdorff Measure

9.1. 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.

9.2. Construction of Hausdorff Measure. 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 15.76. *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 15.77. Let (S, d) be a metric space and $A \subset S$, the *diameter* of A is

$$\operatorname{diam}(A) = \sup\{d(x, y) \mid x, y \in A\}$$

and we use the convention that $\operatorname{diam}(\{x\}) = 0$.

Definition 15.78. 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 we use the convention that $0^0 = 1$ and

$$\alpha(s) = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}$$

For A and s as above define

$$\mathcal{H}^s(A) = \lim_{\delta \to 0} \mathcal{H}^s_\delta(A) = \sup_{\delta > 0} \mathcal{H}^s_\delta(A)$$

The definition of Hausdorff measures considers covering by arbitrary sets but it is not hard to see that the same values occur if we limit covering to open or closed sets.

Proposition 15.79. *Let (S,d) be a metric space, $0 \leq s < \infty$ and $0 < \delta$ then*

$$\mathcal{H}^s_\delta(A) = \inf\{\sum_{n=1}^\infty \alpha(s) \left(\frac{\operatorname{diam}(C_n)}{2} \right)^s \mid A \subset \cup_{n=1}^\infty C_n \textit{ where } C_n \textit{ is open and } \operatorname{diam}(C_n) \leq \delta \textit{ for all } n\}$$

and

$$\mathcal{H}^s_\delta(A) = \inf\{\sum_{n=1}^\infty \alpha(s) \left(\frac{\operatorname{diam}(C_n)}{2} \right)^s \mid A \subset \cup_{n=1}^\infty C_n \textit{ where } C_n \textit{ is closed and } \operatorname{diam}(C_n) \leq \delta \textit{ for all } n\}$$

Theorem 15.80. *$\mathcal{H}^s$ is a Borel regular outer measure.*

Proof. TODO ☐

In some case Hausdorff measure is equivalent to a well-known object; here is a simple case of this phenomenon.

Proposition 15.81. *Let (S,d) be a metric space then $\mathcal{H}^0$ is counting measure.*

Proof. Let $x \in S$ and $\delta > 0$, then the set $\{x\}$ covers itself and therefore $\mathcal{H}^0(\{x\}) \leq (\operatorname{diam}(\{x\}))^0 = 1$. Since every term $(\operatorname{diam}(C))^0 \geq 1$ it follows from the definition that $\mathcal{H}^0_\delta \geq 1$. Thus $\mathcal{H}^0_\delta(\{x\}) = 1$ and the result follows by letting $\delta \to 0$. ☐

Proposition 15.82. *Let (S,d) and (T,d') be metric spaces and let $f : S \to T$ be an isometric embedding then for any $0 \leq s < \infty$ and $\delta > 0$ and $A \subset S$*

$$\mathcal{H}^s_\delta(f(A)) = \mathcal{H}^s_\delta(A)$$

In particular $\mathcal{H}^s(f(A)) = \mathcal{H}^s(A)$

Proof. Let $\delta > 0$ and $\epsilon > 0$ be arbitrary. First we show $\mathcal{H}^s_\delta(f(A)) \leq \mathcal{H}^s_\delta(A)$. This is trivial if $\mathcal{H}^s_\delta(A) = \infty$ so we assume that $\mathcal{H}^s_\delta(A) < \infty$ and pick $U_1, U_2, \dots$ with $\operatorname{diam}(U_n) < \delta$, $A \subset \cup_{n=1}^\infty U_n$ and $\sum_{n=1}^\infty (\operatorname{diam}(C_n))^s \leq \mathcal{H}^s_\delta(A) + \epsilon$. It is elementary that $f(A) \subset \cup_{n=1}^\infty f(U_n)$ and since f is an isometry $\operatorname{diam}(U_n) = \operatorname{diam}(f(U_n)) \leq \delta$ hence

$$\mathcal{H}^s_\delta(f(A)) \leq \sum_{n=1}^\infty (\operatorname{diam}(f(U_n)))^s = \sum_{n=1}^\infty (\operatorname{diam}(U_n))^s \leq \mathcal{H}^s_\delta(A) + \epsilon$$

Now let $\epsilon \to 0$ to conclude $\mathcal{H}^s_\delta(f(A)) \leq \mathcal{H}^s_\delta(A)$.

To see that $\mathcal{H}^s_\delta(A) \leq \mathcal{H}^s_\delta(f(A))$ arguing as above it suffices to assume $\mathcal{H}^s_\delta(f(A)) < \infty$ and to pick $U_1, U_2, \dots$ with $\mathrm{diam}(U_n) < \delta$, $f(A) \subset \cup_{n=1}^\infty U_n$ and $\sum_{n=1}^\infty (\mathrm{diam}(C_n))^s \leq \mathcal{H}^s_\delta(f(A)) + \epsilon$. It is elementary that $A \subset \cup_{n=1}^\infty f^{-1}(U_n)$ (if $a \in A$ then $f(a) \in U_n$ for some n hence $a \in f^{-1}(U_n)$) and since f is an isometry $\mathrm{diam}(f^{-1}(U_n)) \leq \mathrm{diam}(U_n) \leq \delta$ and therefore

$$\mathcal{H}^s_\delta(A) \leq \sum_{n=1}^\infty \left(\mathrm{diam}(f^{-1}(U_n))\right)^s \leq \sum_{n=1}^\infty (\mathrm{diam}(U_n))^s \leq \mathcal{H}^s_\delta(f(A)) + \epsilon$$

Now let $\epsilon \to 0$ to conclude $\mathcal{H}^s_\delta(A) \leq \mathcal{H}^s_\delta(f(A))$ and therefore $\mathcal{H}^s_\delta(A) = \mathcal{H}^s_\delta(f(A))$.

To see that $\mathcal{H}^s(f(A)) = \mathcal{H}^s(A)$ simply take the limit as $\delta \to 0$. $\square$

Proposition 15.83. *Let (S, d) and (T, d') be metric spaces and let $f : S \to T$ be a Lipschitz function then for any $0 \leq s < \infty$ and $A \subset S$*

$$\mathcal{H}^s(f(A)) \leq (\mathrm{Lip}(f))^s \, \mathcal{H}^s(A)$$

Proof. If $\mathcal{H}^s(A) = \infty$ then the result is trivial so assume that $\mathcal{H}^s(A) < \infty$ Let $\delta > 0$ be given and pick $\epsilon > 0$, sets $C_1, C_2, \dots$ such that $\mathrm{diam}(C_n) \leq \delta$, $A \subset \cup_{n=1}^\infty C_n$ and $\sum_{n=1}^\infty (\mathrm{diam}(C_n))^s < \mathcal{H}^s_\delta(A) + \epsilon$. It is elementary that $f(A) \subset \cup_{n=1}^\infty f(C_n)$ and by the Lipschitz property of f we know

$$\mathrm{diam}(f(C_n)) \leq \mathrm{Lip}(f) \, \mathrm{diam}(C_n) \leq \mathrm{Lip}(f)\delta$$

Thus

$$\begin{aligned}\mathcal{H}^s_{\mathrm{Lip}(f)\delta}(f(A)) &\leq \sum_{n=1}^\infty (\mathrm{diam}(f(C_n)))^s \leq (\mathrm{Lip}(f))^s \sum_{n=1}^\infty (\mathrm{diam}(C_n))^s \\ &\leq (\mathrm{Lip}(f))^s \left(\mathcal{H}^s_\delta(A) + \epsilon\right)\end{aligned}$$

Now let $\epsilon \to 0$ and then $\delta \to 0$. $\square$

Theorem 15.84. *For any $d \in \mathbb{N}$ we have*

$$\lambda^d(A) \leq \alpha(d) \left(\frac{\mathrm{diam}(A)}{2}\right)^d$$

Proof. TODO $\square$

Theorem 15.85. *For any $d \in \mathbb{N}$ we have $\mathcal{H}^d = \lambda^d$ (or $2^{-d}\alpha(d)\mathcal{H}^d = \lambda^d$ depending on normalization) on $\mathbb{R}^d$.*

Proof. TODO $\square$

10. Covering Theorems in $\mathbb{R}^d$

Since our purposes have been to understand probability theory we have hitherto avoided making assumptions that we are dealing with $\mathbb{R}^d$. While this decision has benefits, it has drawbacks as well. Among those drawbacks are that we lose sight of some history and also some very beautiful and deep understanding of the measure theory of the reals. TODO: Vitali and Besicovich.

THEOREM 15.86 (Besicovich's Covering Theorem). *Let $\mathcal{I}$ be an arbitrary index set and for each $\alpha \in \mathcal{I}$ let $\overline{B}(x_\alpha, r_\alpha)$ be a closed ball in $\mathbb{R}^d$ such that $r_\alpha > 0$. If $\sup_\alpha r_\alpha < \infty$ then there exists a constant N_d depending only on d and countable collections $\mathcal{J}_1, \dots, \mathcal{J}_{N_d}$ such that for each $j = 1, \dots, N_d$ the balls $\overline{B}(x_\alpha, r_\alpha)$ with $\alpha \in \mathcal{J}_j$ are disjoint and*

$$\{x_\alpha \mid \alpha \in \mathcal{I}\} \subset \cup_{j=1}^{N_d} \cup_{\alpha \in \mathcal{J}_j} \overline{B}(x_\alpha, r_\alpha)$$

PROOF. TODO: See Evans and Gariepy for the proof until I get around copying it down here (I have little to add to their proof). □

COROLLARY 15.87. *Let μ be a Borel outer measure on $\mathbb{R}^d$, $\mathcal{I}$ be an arbitrary index set and for each $\alpha \in \mathcal{I}$ $\overline{B}(x_\alpha, r_\alpha)$ is a closed ball with $r_\alpha > 0$. Define $A = \{x_\alpha \mid \alpha \in \mathcal{I}\}$ and assume that $\mu(A) < \infty$ and for every $x \in A$ we have $\inf\{r_\alpha \mid x_\alpha = x\} = 0$. Then for each open set $U \subset \mathbb{R}^d$ there exists a countable set $\mathcal{J} \subset \mathcal{I}$ such that*

(i) $\overline{B}(x_\alpha, r_\alpha) \cap \overline{B}(x_\beta, r_\beta) = \emptyset$ *for each* $\alpha, \beta \in \mathcal{J}$ *with* $\alpha \neq \beta$
(ii) $\cup_{\alpha \in \mathcal{J}} \overline{B}(x_\alpha, r_\alpha) \subset A$
(iii) $\mu\left((A \cap U) \setminus \cup_{\alpha \in \mathcal{J}} \overline{B}(x_\alpha, r_\alpha)\right) = 0$

PROOF. Let N_d be the constant in Theorem 15.86. Pick θ such that $1 - N_d^{-1} < \theta < 1$.

CLAIM 15.87.1. There exist a finite collection $\overline{B}(x_{\alpha_1}, r_{\alpha_1}), \dots, \overline{B}(x_{\alpha_{M_1}}, r_{\alpha_{M_1}})$ with $\alpha_j \in \mathcal{I}$, $\overline{B}(x_{\alpha_j}, r_{\alpha_j}) \subset U$ for $j = 1, \dots, M$ such that the $\overline{B}(x_{\alpha_j}, r_{\alpha_j})$ are disjoint and

$$\mu\left((A \cap U) \setminus \cup_{j=1}^{M_1} \overline{B}(x_{\alpha_j}, r_{\alpha_j})\right) \leq \theta \mu(A \cap U)$$

Let $\mathcal{K} = \{\alpha \in \mathcal{I} \mid r_\alpha \leq 1 \text{ and } \overline{B}(x_\alpha, r_\alpha) \subset U\}$. By the assumptions on r_α and the openness of U we know that $A \cap U = \{x_\alpha \mid \alpha \in \mathcal{K}\}$. Apply Theorem 15.86 to the collection $\mathcal{K}$ to construct countable sets $\mathcal{J}_1, \dots, \mathcal{J}_{N_d}$ with

$$A \cap U \subset \cup_{k=1}^{N_d} \cup_{\alpha \in \mathcal{J}_k} \overline{B}(x_\alpha, r_\alpha) = \cup_{k=1}^{N_d} A \cap U \cap \cup_{\alpha \in \mathcal{J}_k} \overline{B}(x_\alpha, r_\alpha)$$

By countable subadditivity

$$\mu(A \cap U) \leq \sum_{k=1}^{N_d} \mu\left(A \cap U \cap \cup_{\alpha \in \mathcal{J}_k} \overline{B}(x_\alpha, r_\alpha)\right) \leq N_d \max_{1 \leq k \leq N_d} \mu\left(A \cap U \cap \cup_{\alpha \in \mathcal{J}_k} \overline{B}(x_\alpha, r_\alpha)\right)$$

from which it follows that there exists $k \in \mathbb{N}$ with $1 \leq k \leq N_d$ and

$$\mu\left(A \cap U \cap \cup_{\alpha \in \mathcal{J}_k} \overline{B}(x_\alpha, r_\alpha)\right) \geq N_d^{-1} \mu(A \cap U)$$

By continuity of regular outer measures (Proposition 2.97) and the fact that $1 - \theta < N_d^{-1}$ we know that there is a finite subset $\alpha_1, \dots, \alpha_M \in \mathcal{J}_k$ such that

$$\mu\left(A \cap U \cap \cup_{j=1}^{M} \overline{B}(x_{\alpha_j}, r_{\alpha_j})\right) \geq (1 - \theta) \mu(A \cap U)$$

and by the measurability of $\cup_{j=1}^{M} \overline{B}(x_{\alpha_j}, r_{\alpha_j})$ we conclude

$$\begin{aligned} \mu\left(A \cap U \setminus \cup_{j=1}^{M} \overline{B}(x_{\alpha_j}, r_{\alpha_j})\right) &= \mu(A \cap U) - \mu\left(A \cap U \cap \cup_{j=1}^{M} \overline{B}(x_{\alpha_j}, r_{\alpha_j})\right) \\ &\leq \theta \mu(A \cap U) \end{aligned}$$

The disjointness of the $\overline{B}(x_{\alpha_j}, r_{\alpha_j})$ follows from the disjointness of the larger family of balls $\overline{B}(x_\alpha, r_\alpha)$ with $\alpha \in \mathcal{J}_k$.

Let $U_1 = U$ and we apply the claim inductively to define M_n, $\alpha_1^n, \dots, \alpha_{M_n}^n$ such that $\overline{B}(x_{\alpha_j^n}, r_{\alpha_j^n}) \subset U_n$ for all $j = 1, \dots, M_n$,

$$U_{n+1} = U_n \setminus \cup_{j=1}^{M_n} \overline{B}(x_{\alpha_j^n}, r_{\alpha_j^n}) = U \setminus \cup_{k=1}^{n} \cup_{j=1}^{M_k} \overline{B}(x_{\alpha_j^k}, r_{\alpha_j^k})$$

and

$$\mu(A \cap U_{n+1}) = \mu\left((A \cap U_n) \setminus \cup_{j=1}^{M_n} \overline{B}(x_{\alpha_j^n}, r_{\alpha_j^n})\right) \leq \theta \mu(A \cap U_n)$$

For fixed n and $1 \leq i < j \leq M_n$ we have $\overline{B}(x_{\alpha_i^n}, r_{\alpha_i^n}) \cap \overline{B}(x_{\alpha_j^n}, r_{\alpha_j^n}) = \emptyset$ by the claim and for $n, m \in \mathbb{N}$ with $m < n$, $1 \leq i \leq M_m$ and $1 \leq j \leq M_n$ we have $\overline{B}(x_{\alpha_i^m}, r_{\alpha_i^m}) \cap \overline{B}(x_{\alpha_j^n}, r_{\alpha_j^n}) = \emptyset$ since

$$\overline{B}(x_{\alpha_j^n}, r_{\alpha_j^n}) \subset U_n = U \setminus \cup_{k=1}^{n-1} \cup_{j=1}^{M_k} \overline{B}(x_{\alpha_j^k}, r_{\alpha_j^k})$$

Let $\mathcal{J} = \{\alpha_j^n \mid n \in \mathbb{N}, j = 1, \dots, M_n\}$ and note that by continuity of regular outer measures (TODO: Where do we prove this the descending case???)

$$\begin{aligned}
\mu\left((A \cap U) \setminus \cup_{\alpha \in \mathcal{J}} \overline{B}(x_\alpha, r_\alpha)\right) &= \mu\left((A \cap U) \setminus \cup_{k=1}^{\infty} \cup_{j=1}^{M_k} \overline{B}(x_{\alpha_j^k}, r_{\alpha_j^k})\right) \\
&= \lim_{n \to \infty} \mu\left((A \cap U) \setminus \cup_{k=1}^{n} \cup_{j=1}^{M_k} \overline{B}(x_{\alpha_j^k}, r_{\alpha_j^k})\right) \\
&= \lim_{n \to \infty} \mu\left((A \cap U_n) \setminus \cup_{j=1}^{M_n} \overline{B}(x_{\alpha_j^n}, r_{\alpha_j^n})\right) \\
&\leq \lim_{n \to \infty} \theta \mu(A \cap U_n) \\
&\leq \lim_{n \to \infty} \theta^{n-1} \mu(A \cap U) = 0
\end{aligned}$$

□

10.1. Derivatives of Radon Outer Measures on $\mathbb{R}^d$. In this section we apply the Besicovich covering theorem as a key ingredient in some formulas for computing Radon-Nikodym derivatives of Radon measures on $\mathbb{R}^d$.

Definition 15.88. Let μ and ν be Radon outer measures on $\mathbb{R}^d$ then for each $x \in \mathbb{R}^d$ we define

$$\overline{D}_\mu \nu(x) = \begin{cases} \limsup_{r \to 0} \frac{\nu(B(x,r))}{\mu(B(x,r))} & \text{if } \mu(B(x,r)) > 0 \text{ for all } r > 0 \\ +\infty & \text{if } \mu(B(x,r)) = 0 \text{ for some } r > 0 \end{cases}$$

$$\underline{D}_\mu \nu(x) = \begin{cases} \liminf_{r \to 0} \frac{\nu(B(x,r))}{\mu(B(x,r))} & \text{if } \mu(B(x,r)) > 0 \text{ for all } r > 0 \\ +\infty & \text{if } \mu(B(x,r)) = 0 \text{ for some } r > 0 \end{cases}$$

If $\underline{D}_\mu \nu(x) = \overline{D}_\mu \nu(x) < \infty$ then we say that ν is *differentiable* with respect to μ at x and in this case we write $D_\mu \nu(x) = \underline{D}_\mu \nu(x) = \overline{D}_\mu \nu(x)$.

Our first goal is to show that derivatives of Radon measures exist almost everywhere, then we can show that the derivative in the above sense is the Radon-Nikodym derivative.

Lemma 15.89. *Let $0 < \lambda < \infty$ be given then for an arbitrary subset $A \subset \mathbb{R}^d$*

(i) *if $A \subset \{x \mid \underline{D}_\mu \nu(x) \leq \lambda\}$ then $\nu(A) \leq \lambda \mu(A)$.*
(ii) *if $A \subset \{x \mid \overline{D}_\mu \nu(x) \geq \lambda\}$ then $\nu(A) \geq \lambda \mu(A)$.*

PROOF. First assume that $\mu(\mathbb{R}^d) < \infty$ and $\nu(\mathbb{R}^d) < \infty$. Suppose that $A \subset \{x \mid \underline{D}_\mu \nu(x) \leq \lambda\}$. Let $\epsilon > 0$ and pick an open set U such that $A \subset U$. Now we define

$$\mathcal{A} = \{\overline{B}(a,r) \mid x \in A, \overline{B}(a,r) \subset U, \nu(\overline{B}(a,r)) \leq (\alpha+\epsilon)\mu(\overline{B}(a,r))\}$$

Observe that for each $a \in A$ there exists a sequence $r_n > 0$ such that $r_n \downarrow 0$ and $\lim_{n\to\infty} \frac{\nu(\overline{B}(a,r_n))}{\mu(\overline{B}(a,r_n))} \leq \alpha$. Thus for sufficiently large n we have $\nu(\overline{B}(a,r_n)) \leq (\alpha+\epsilon)\mu(\overline{B}(a,r_n))$ and it follows that $\inf\{r > 0 \mid \overline{B}(a,r) \in \mathcal{A} = 0$. Therefore we can apply Corollary 15.87 to get a countable collection of disjoint balls $\overline{B}(a_1,r_1), \overline{B}(a_2,r_2), \ldots$ in $\mathcal{A}$ such that

$$\nu\left(A \setminus \cup_{n=1}^\infty \overline{B}(a_n,r_n)\right) = 0$$

From this it follows from measurability of the balls, subadditivity and the definition of $\mathcal{A}$

$$\begin{aligned}\nu(A) = \nu(A \cap \cup_{n=1}^\infty \overline{B}(a_n,r_n)) &\leq \sum_{n=1}^\infty \nu(\overline{B}(a_n,r_n)) \leq (\alpha+\epsilon)\sum_{n=1}^\infty \mu(\overline{B}(a_n,r_n)) \\ &\leq (\alpha+\epsilon)\mu(U)\end{aligned}$$

Now take the infimum over all open sets U such that $A \subset U$ and use the outer regularity of μ to conclude that $\nu(A) \leq (\alpha+\epsilon)\mu(A)$. Since $\epsilon > 0$ we arbitrary we let $\epsilon \to 0$ to get $\nu(A) \leq \alpha\mu(A)$.

TODO: Prove (ii) using the same argument and put as an exercise.

TODO: Remove the finiteness assumption □

THEOREM 15.90. *Let μ and ν be Radon outer measures on $\mathbb{R}^d$ then $D_\mu \nu$ exists μ-almost everywhere and $D_\mu \nu$ is μ-measurable.*

PROOF. First assume that $\mu(\mathbb{R}^d) < \infty$ and $\nu(\mathbb{R}^d) < \infty$. Let $A = \{\underline{D}_\mu \nu = \infty$ and note that for every λ we have $A \subset \{\underline{D}_\mu \nu \geq \lambda$ and therefore by Lemma 15.89 we have $\mu(A) \leq \lambda^{-1}\nu(A)$. Since $\nu(A) < \infty$ by assumption we may let $\lambda \to \infty$ to conclude that $\mu(A) = 0$.

Now for every $0 \leq p < q < \infty$ with $p, q \in \mathbb{Q}$ consider $A_{p,q} = \{\underline{D}_\mu \nu \leq p < q \leq \overline{D}_\mu \nu < \infty$. By two applications of Lemma 15.89 we see that

$$q\mu(A_{p,q}) < \nu(A_{p,q}) < p\mu(A_{p,q})$$

and therefore since $p < q$ we conclude that $\mu(A_{p,q}) = 0$. Since $\{D_\mu \nu \text{ not exists}\} = A \cup \cup_{q,p \in \mathbb{Q}_+} A_{p,q}$ we conclude that $D_\mu \nu$ exists μ-almost everywhere.

CLAIM 15.90.1. Let μ be an arbitrary Radon measure on $\mathbb{R}^d$. For every $r > 0$ and $x \in \mathbb{R}^d$ we have $\limsup_{y\to x} \mu(\overline{B}(y,r)) \leq \mu(\overline{B}(x,r))$ (i.e. $\mu(\overline{B}(x,r))$ is an upper semicontinuous function of x).

Pick a sequence y_n such that $\lim_{n\to\infty} y_n = x$. Suppose that $\mathbf{1}_{\overline{B}(x,r)}(y) = 0$ i.e. we are given y such that $\|y - x\| > r$. We can pick N such that $\|y_n - x\| < \|y - x\| - r$ for all $n \geq N$; it follows that $\|y - y_n\| \geq \|y - x\| - \|y_n - x\| > r$ for all $n \geq N$ which is to say that $\mathbf{1}_{\overline{B}(y_n,r)}(y) = 0$ which shows $\lim_{n\to\infty} \mathbf{1}_{\overline{B}(y_n,r)}(y) = 0 = \mathbf{1}_{\overline{B}(x,r)}(y)$. If on the other hand $\mathbf{1}_{\overline{B}(x,r)}(y) = 1$ it follows trivially that $\limsup_{n\to\infty} \mathbf{1}_{\overline{B}(y_n,r)}(y) \leq \mathbf{1}_{\overline{B}(x,r)}(y)$. We conclude that

$$\limsup_{n\to\infty} \mathbf{1}_{\overline{B}(y_n,r)} \leq \mathbf{1}_{\overline{B}(x,r)}$$

which is to say

$$1 - \liminf_{n\to\infty} \mathbf{1}_{\overline{B}(y_n,r)} \geq 1 - \mathbf{1}_{\overline{B}(x,r)}$$

Now we apply Fatou's Lemma to conclude

$$\begin{aligned}
\mu(\overline{B}(x,2r)) - \mu(\overline{B}(x,r)) &= \int_{\overline{B}(x,2r)} (1 - \mathbf{1}_{\overline{B}(x,r)})\, d\mu \leq \int_{\overline{B}(x,2r)} (1 - \liminf_{n\to\infty} \mathbf{1}_{\overline{B}(y_n,r)})\, d\mu \\
&\leq \liminf_{n\to\infty} \int_{\overline{B}(x,2r)} (1 - \mathbf{1}_{\overline{B}(y_n,r)})\, d\mu \\
&\leq \liminf_{n\to\infty} (\mu(\overline{B}(x,2r)) - \mu(\overline{B}(y_n,r))) \\
&= \mu(\overline{B}(x,2r)) - \limsup_{n\to\infty} \mu(\overline{B}(y_n,r))
\end{aligned}$$

Since μ is a Radon measure we know that $\mu(\overline{B}(x,2r)) < \infty$ so we can cancel terms and the claim follows.

For each $r > 0$ define

$$f_r(x) = \begin{cases} \frac{\nu(\overline{B}(x,r))}{\mu(\overline{B}(x,r))} & \text{if } \mu(\overline{B}(x,r)) > 0 \\ \infty & \text{if } \mu(\overline{B}(x,r)) = 0 \end{cases}$$

and observe from the previous claim and the Borel measurability of upper semicontinous functions (TODO: Where do we show this????) we know that f_r is Borel measurable (hence μ-measurable). Since $D_\mu \nu = \lim_{n\to\infty} f_{1/n}$ μ-almost everywhere it follows that $D_\mu \nu$ is μ-measurable (TODO: Something subtle about concluding μ-measurability but not Borel measurability; understand it).

TODO: Remove the finiteness assumption ☐

Now we are ready to show that the derivative defined above is indeed the Radon-Nikodym derivative.

THEOREM 15.91. *Let μ and ν be Radon outer measure on $\mathbb{R}^d$ then for all μ-measurable sets A we have*

$$\nu(A) = \int_A D_\mu \nu \, d\mu$$

PROOF. Let A be a μ-measurable set. Since Radon outer measures are Borel regular we can find a Borel set B such that $A \subset B$ and $\mu(B) = \mu(A)$.

TODO: Finish ☐

10.2. Differentiation of Lipschitz Functions.

THEOREM 15.92 (Rademacher's Theorem). *Let $f : \mathbb{R}^d \to \mathbb{R}^m$ be a locally Lipschitz map then f is Frechet differentiable λ^d-a.e. Moreover the Frechet derivative Df is Borel measurable.*

PROOF. Since each coordinate function $f_i : \mathbb{R}^d \to \mathbb{R}$ is locally Lipschitz the case of general m follows from the case of $m = 1$ by Proposition 15.144.

The outline of the proof is to show that successively that f has directional derivatives almost surely, is Gâteaux differentiable almost surely and then that it is Frechet differentiable almost surely.

Gâteaux differentiability in the case of $d = 1$ is essentially a corollary of the Fundamental Theorem of Calculus.

CLAIM 15.92.1. Let $f : \mathbb{R} \to \mathbb{R}$ be a locally Lipschitz map then f is differentiable λ-a.e.

Since f is locally Lipschitz it is absolutely continuous and therefore of bounded variation (Lemma 2.124); we know that $g(t)$ is differentiable λ-almost everywhere (by Theorem 2.119 write $g(t)$ as a difference of monotone functions then apply the Fundamental Theorem of Calculus 2.113).

CLAIM 15.92.2. Let v be a unit vector in $\mathbb{R}^d$ then the directional derivative $df(x, v)$ exists for λ^d-almost all x.

Let A be a rotation of $\mathbb{R}^d$ such that $v = Ae_1$ is the first standard basis vector and define $f_A(x) = f(Ax)$. Note that

$$|f_A(x) - f_A(y)| = |f(Ax) - f(Ay)| \leq \mathrm{Lip}(f)\,|Ax - Ay| = \mathrm{Lip}(f)\,|x - y|$$

and therefore f_A is Lipschitz. Also

$$df(x, v) = \lim_{t\to 0} \frac{f(x + tv) - f(x)}{t} = \lim_{t\to 0} \frac{f(A(A^{-1}x + te_1)) - f(A(A^{-1}x))}{t} = df_A(A^{-1}x, e_1)$$

Since λ^d is rotation invariant (Corollary 2.91) it follows that it suffices to consider the case of $v = e_1$.

By the continuity of f we have

$$\limsup_{t\to 0} \frac{f(x + te_1) - f(x)}{t} = \lim_{n\to\infty} \sup_{0<t<1/n} \frac{f(x + te_1) - f(x)}{t} = \lim_{n\to\infty} \sup_{\substack{0<t<1/n \\ t\in\mathbb{Q}}} \frac{f(x + te_1) - f(x)}{t}$$

hence by Lemma 2.14 we know that $\limsup_{t\to 0} \frac{f(x+te_1)-f(x)}{t}$ is Borel measurable and the same argument shows $\liminf_{t\to 0} \frac{f(x+te_1)-f(x)}{t}$ is Borel measurable as well. It follows that $A = \{\liminf_{t\to 0} \frac{f(x+te_1)-f(x)}{t} < \limsup_{t\to 0} \frac{f(x+te_1)-f(x)}{t}\}$ is a Borel measurable set (and shows that when the limit exists it is Borel measurable). For a fixed $x \in \mathbb{R}^d$ we consider the function $g(t) = f(x + te_1)$ which is Lipschitz and therefore differentiable λ-a.e. by the first claim. Writing $\mathbb{R}^d = \mathbb{R} \times \mathbb{R}^{d-1}$ and letting $\pi : \mathbb{R}^d \to \mathbb{R}^{d-1}$ be the projection on to the last $d - 1$ coordinates we see that for every $y \in \mathbb{R}^{d-1}$, $\lambda(A \cap \{\pi_2(x) = y\}) = 0$. Now by Tonelli's Theorem 2.88 we see that

$$\lambda^d(A) = \int \lambda(A \cap \{\pi_2(x) = y\})\,\lambda^{d-1}(dy) = 0$$

By the previous claim the partial derivatives $\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_d}$ exist λ^d-almost everywhere and therefore the gradient

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_d}\right)$$

exists λ^d-almost everwhere.

CLAIM 15.92.3. f is almost everywhere Gâteaux differentiable and for $v \in \mathbb{R}^d$ we have $df(x, v) = \langle v, \nabla f(x)\rangle$.

Let $g(x)$ be a compactly supported C^∞ function, by change of variables and the translation invariance of Lebesgue measure we get

$$\begin{aligned}\int \left(\frac{f(x+tv)-f(x)}{t}\right) g(x)\,dx &= t^{-1}\int f(x+tv)g(x)\,dx - t^{-1}\int f(x)g(x)\,dx \\ &= t^{-1}\int f(x)g(x-tv)\,dx - t^{-1}\int f(x)g(x)\,dx \\ &= -\int f(x)\left(\frac{g(x)-g(x-tv)}{t}\right)\,dx\end{aligned}$$

Let K be a compact set containing the support of g so that $\|g\|_\infty = \sup_{x\in K}|g(x)| < \infty$ and note that by continuity of f we have $\sup_{x\in K}|f(x)|<\infty$

$$\left|\left(\frac{f(x+tv)-f(x)}{t}\right) g(x)\right| \leq t^{-1}\operatorname{Lip}(f)\,|x+tv-x|\,\|g\|_\infty = \operatorname{Lip}(f)\,|v|\,\|g\|_\infty$$

and by the Mean Value Theorem (more specifically the corollary Proposition 15.143)

$$\begin{aligned}\left|f(x)\left(\frac{g(x)-g(x-tv)}{t}\right)\right| &\leq t^{-1}\|dg(x,v)\|_\infty\,|x+tv-x|\sup_{x\in K}|f(x)| \\ &= \|dg(x,v)\|_\infty \sup_{x\in K}|f(x)| < \infty\end{aligned}$$

and therefore we may use Dominated Convergence, Proposition 15.146 and the previous claim to see

$$\begin{aligned}\int df(x,v)g(x)\,dx &= \lim_{t\to 0}\int\left(\frac{f(x+tv)-f(x)}{t}\right)g(x)\,dx \\ &= -\lim_{t\to 0}\int f(x)\left(\frac{g(x)-g(x-tv)}{t}\right)\,dx \\ &= -\int f(x)dg(x,v)\,dx \\ &= -\sum_{i=1}^d v_i\int f(x)\frac{\partial g}{\partial x_i}\,dx \\ &= -\sum_{i=1}^d v_i\int f(x)\frac{\partial g}{\partial x_i}\,dx \\ &= -\sum_{i=1}^d v_i\lim_{t\to 0}\int f(x)\left(\frac{g(x)-g(x-te_i)}{t}\right)\,dx \\ &= \sum_{i=1}^d v_i\lim_{t\to 0}\int\left(\frac{f(x+te_i)-f(x)}{t}\right)g(x)\,dx \\ &= \sum_{i=1}^d v_i\int\frac{\partial f}{\partial x}g(x)\,dx \\ &= \int\langle v,\nabla f(x)\rangle g(x)\,dx\end{aligned}$$

Now if we choose non-negative compactly supported C^∞ functions g_n such that $g_n \uparrow 1$ (e.g. Lemma 2.128) by Monotone Convergence we get

$$\int |df(x,v) - \langle v, \nabla f(x)\rangle| \, dx = \lim_{n\to\infty} \int |df(x,v) - \langle v, \nabla f(x)\rangle| \, g_n(x) \, dx = 0$$

and it follows that $df(x,v) = \langle v, \nabla f(x)\rangle$ λ^n-a.e. by Lemma 2.50.

Pick a countable dense subset of the unit sphere $v_1, v_2, \dots$ and let

$$A_n = \{x \in \mathbb{R}^d \mid df(x,v_n), \nabla f(x) \text{ exist and } df(x,v_n) = \nabla f(x)\}$$

From the previous claims, we know that $\lambda^d(A_n^c) = 0$ and therefore if we define $A = \cap_{n=1}^\infty A_n$ we have $\lambda^d(A^c) = 0$.

Claim 15.92.4. f is Frechet differentiable on A and $Df(x)v = \langle v, \nabla f(x)\rangle$ for all $x \in A$ and $v \in \mathbb{R}^d$.

Let $x \in A$ and $\epsilon > 0$ be given. Since the unit sphere in $\mathbb{R}^d$ is compact it is totally bounded (Theorem 1.29) hence there exists an $N > 0$ such that for every $v \in \mathbb{R}^d$ with $\|v\| = 1$ there exists a v_i with $i = 1, \dots, N$ and $\|v - v_i\| < \epsilon/2(1+\sqrt{d}) \operatorname{Lip}(f)$ (of course the terms in the denominator are just there so we a nice clean ϵ at the end; their source will be made clear). By the previous claim and the finiteness of N there exists a $\delta > 0$ such that

$$\left| \frac{f(x+tv_i) - f(x)}{t} - \langle v_i, \nabla f(x)\rangle \right| \leq \epsilon/2$$

for all $i = 1, \dots, N$ and $0 < |t| < \delta$. Now for an arbitrary $h \in \mathbb{R}^d$ with $0 < \|h\| < \delta$ we pick v_i with $i = 1, \dots, N$ and $\|v_i - h/\|h\|\| < \epsilon/2(1+\sqrt{d}) \operatorname{Lip}(f)$ and compute

$$\begin{aligned}
&\frac{|f(x+h) - f(x) - \langle h, \nabla f(x)\rangle|}{\|h\|} \\
&= \frac{|f(x+h) - f(x+\|h\| v_i) + f(x+\|h\| v_i) - f(x) - \langle \|h\| v_i, \nabla f(x)\rangle - \langle h - \|h\| v_i, \nabla f(x)\rangle|}{\|h\|} \\
&\leq \left| \frac{f(x+\|h\| v_i) - f(x)}{\|h\|} - \langle v_i, \nabla f(x)\rangle \right| + \frac{|f(x+h) - f(x+\|h\| v_i)|}{\|h\|} + \frac{|\langle h - \|h\| v_i, \nabla f(x)\rangle|}{\|h\|} \\
&\leq \epsilon/2 + \operatorname{Lip}(f) \|h/\|h\| - v_i\| + \|\nabla f(x)\| \|h/\|h\| - v_i\| \\
&\leq \epsilon/2 + (\operatorname{Lip}(f) + \|\nabla f(x)\|)\epsilon/2(1+\sqrt{d}) \operatorname{Lip}(f) \leq \epsilon
\end{aligned}$$

where in the last line we have used the fact that $\|\nabla f(x)\| \leq \sqrt{d} \operatorname{Lip}(f)$ which follows from

$$\left| \frac{\partial f}{\partial x_i} \right| = \lim_{t\to 0} \frac{|f(x+te_i) - f(x)|}{t} \leq \operatorname{Lip}(f)$$

∎

10.3. The Area Formula.

Definition 15.93. Let $L : \mathbb{R}^d \to \mathbb{R}^m$ be a linear map then the *Jacobian* of L is defined to be

$$[\![A]\!] = \begin{cases} \det(L^T L) & \text{if } d \leq m \\ \det(L L^T) & \text{if } d > m \end{cases}$$

It is very useful to be able to compute the Jacobian of a matrix using the singular value decomposition

LEMMA 15.94. *Let $L : \mathbb{R}^d \to \mathbb{R}^m$ be a linear map and let $L = U\Sigma V^T$ be a singular value decomposition of L then*

$$[\![L]\!] = \prod_{j=1}^{d\wedge m} \Sigma_{jj} = \det(\Sigma[1 : d \wedge m, 1 : d \wedge m])$$

PROOF. TODO □

The reason why we bring up Jacobians of matrices is the following result which is a generalization of Corollary 2.92 which shows how Lebesgue measure transforms under linear maps.

LEMMA 15.95. *Let $d \leq m$ and $L : \mathbb{R}^d \to \mathbb{R}^m$ be a linear map then for all $A \subset \mathbb{R}^d$*

$$\mathcal{H}^d(L(A)) = \frac{2^d [\![L]\!]}{\alpha(d)} \lambda^d(A)$$

PROOF. TODO: the proof in Evans and Gariepy seems a bit more complicated than it has to be (argues with balls and then uses Radon derivatives to generalize); presumably this is because they aren't assuming the change of measure formula for Lebesgue measure under linear change of coordinates???? Can't we argue like this: Let $L = U\Sigma V^T$ be a singular value decomposition and write

$$\Sigma = \begin{bmatrix} \mathrm{Id}_d \\ 0_{m-d} \end{bmatrix} \hat{\Sigma}$$

where $\hat{\Sigma}$ is $d \times d$ diagonal matrix with $\det(\hat{\Sigma}) = [\![L]\!]$. Since

$$\hat{U} = U \begin{bmatrix} \mathrm{Id}_d \\ 0_{m-d} \end{bmatrix}$$

is an isometric embedding of $\mathbb{R}^d$ into $\mathbb{R}^m$ we have $\mathcal{H}^d(\hat{U}(A)) = \mathcal{H}^d(A) = \frac{2^d}{\alpha(d)}\lambda^d(A)$ by Proposition 15.82 and Theorem 15.85.

Now by Corollary 2.92

$$\begin{aligned} \mathcal{H}^d(L(A)) = \mathcal{H}^d(\hat{U}\hat{\Sigma}V^T(A)) &= \frac{2^d}{\alpha(d)} \lambda^d(\hat{\Sigma}V^T(A)) \\ &= \frac{2^d}{\alpha(d)} \det(\hat{\Sigma}) \det(V^T) \lambda^d(A) = \frac{2^d [\![L]\!]}{\alpha(d)} \lambda^d(A) \end{aligned}$$

□

The following quantity is fundamental to the change of variables formula.

DEFINITION 15.96. Let $f : \mathbb{R}^d \to \mathbb{R}^m$ be function then we call the map $y \mapsto \mathcal{H}^0(A \cap f^{-1}(y))$ the *multiplicity function*.

We need to establish some basic properties of multiplicity functions before we proceed.

The first lemma is completely elementary.

LEMMA 15.97. *The support of the multiplicity function is $f(A)$ and $\mathbf{1}_{f(A)} \leq \mathcal{H}^0(A \cap f^{-1}(y))$.*

PROOF. Suppose $y \in f(A)$, then there exists $x \in A$ such that $f(x) = y$ hence $A \cap f^{-1}(y) \neq \emptyset$. Since $\mathcal{H}^0$ is the counting measure (Proposition 15.81) we see that $\mathcal{H}^0(A \cap f^{-1}(y)) \geq 1$ which shows $\mathbf{1}_{f(A)}(y) \leq \mathcal{H}^0(A \cap f^{-1}(y)) \geq 1$ for all y (and shows that $f(A)$ is contained in the support of $\mathcal{H}^0(A \cap f^{-1}(y))$). Now we suppose that $\mathcal{H}^0(A \cap f^{-1}(y)) \neq 0$; since $\mathcal{H}^0$ is counting measure this implies $A \cap f^{-1}(y) \neq 0$ hence there is an $x \in A$ such that $f(x) = y$. □

We also need some measurability properties and estimates for Lipschitz f.

LEMMA 15.98. *Let $A \subset \mathbb{R}^d$ be λ^d-measurable and let $f : \mathbb{R}^d \to \mathbb{R}^m$ with $d \leq m$ be Lipschitz then*

(i) *$f(A)$ is $\mathcal{H}^d$-measurable*
(ii) *the multiplicity function $\mathcal{H}^0(A \cap f^{-1}(y))$ is a $\mathcal{H}^d$-measurable on $\mathbb{R}^m$.*
(iii) *$\int \mathcal{H}^0(A \cap f^{-1}(y))\, \mathcal{H}^d(dy) \leq (\mathrm{Lip}(f))^d \lambda^d(A)$.*

PROOF. We first prove (i). First we assume that A is bounded. By the inner regularity of Lebesgue measure we find compact sets $K_n \subset A$ such that $\lambda^d(K_n) \geq \lambda^d(A) - 1/n$. Since $\lambda^d(A) < \infty$ and finite additivity it follows that $\lambda^d(A \setminus K_n) \leq 1/n$. Since f is continuous we know that $f(K_n)$ is compact (Theorem 1.31) and therefore since $\mathcal{H}^d$ is Borel regular (Theorem 15.80) we know that $f(K_i)$ is $\mathcal{H}^d$-measurable and therefore $\cup_{n=1}^\infty f(K_n)$ is $\mathcal{H}^d$-measurable. On the other hand by subadditivity of $\mathcal{H}^d$, Proposition 15.83 and Theorem 15.85 we get

$$\begin{aligned}
\mathcal{H}^d\left(f(A) \setminus \cup_{n=1}^\infty f(K_n)\right) &\leq \mathcal{H}^d\left(f\left(A \setminus \cup_{n=1}^\infty K_n\right)\right) \\
&\leq (\mathrm{Lip}(f))^d\, \mathcal{H}^d\left(A \setminus \cup_{n=1}^\infty K_n\right) = (\mathrm{Lip}(f))^d\, \lambda^d\left(A \setminus \cup_{n=1}^\infty K_n\right) = 0
\end{aligned}$$

which shows us that $f(A) \setminus \cup_{n=1}^\infty f(K_n)$ is $\mathcal{H}^d$-measurable. To remove the assumption that A is bounded simply write $f(A) = \cup_{n=1} f(A \cap B(0,n))$ and apply (i) to each $A \cap B(0,n)$.

To see (ii), for each $n \in \mathbb{N}$ we write $\mathbb{R}^d$ as the disjoint union of cubes with sides of length $1/n$:

$$\mathcal{C}_n = \{(k_1/n, (k_1+1)/n] \times \cdots \times (k_d/n, (k_d+1)/n] \mid k_1, \ldots, k_d \in \mathbb{Z}\}$$

By (i) for each cube $C \in \mathcal{C}_n$ we have $f(A \cap C)$ is $\mathcal{H}^d$-measurable so it follows that the function

$$g_n(y) = \sum_{C \in \mathcal{C}_n} \mathbf{1}_{f(A \cap C)}(y)$$

is $\mathcal{H}^d$-measurable.

CLAIM 15.98.1. $\lim_{n \to \infty} g_n(y) = \mathcal{H}^0(A \cap f^{-1}(y))$

To see this, first observe that $g_n(y)$ is equal to the number of cubes $C \in \mathcal{C}_n$ such that $f^{-1}(y) \cap A \cap C \neq \emptyset$. If $\mathcal{H}^0(A \cap f^{-1}(y)) = 0$ then $A \cap f^{-1}(y)$ is empty and we see that $g_n(y) = 0$ for all n. If $\mathcal{H}^0(A \cap f^{-1}(y)) < \infty$ let $x_1, \ldots, x_{\mathcal{H}^0(A \cap f^{-1}(y))}$ be an enumeration of $A \cap f^{-1}(A)$ and if $\mathcal{H}^0(A \cap f^{-1}(y)) < \infty$ let $x_1, x_2, \ldots$ be an enumeration of an arbitrarily chosen countable subset of $A \cap f^{-1}(y)$. Let $M \in \mathbb{N}$ be given then define

$$D_M = \inf\{|x_i - x_j| \mid 1 \leq i < j \leq M \wedge \mathcal{H}^0(A \cap f^{-1}(y))\}$$

and observe that $0 < D_M \leq \infty$ (as usual $\inf \emptyset = \infty$). Now we can pick N large enough so that $\mathrm{diam}(C) < D_M$ for every $n \geq N$ and $C \in \mathcal{C}_n$ (concretely since

$\operatorname{diam}(C) = \sqrt{d}/n$ for $C \in \mathcal{C}_n$ we may pick $N = \lceil \sqrt{d}/D_M \rceil$ if $D_M < \infty$ or $N = 1$ if $D_M = \infty$). It follows that for every $n \geq N$ we have $g_n(y) = M$. This shows that $g_n(y) \uparrow \mathcal{H}^0(A \cap f^{-1}(y))$.

$\mathcal{H}^d$-measurability of $\mathcal{H}^0(A \cap f^{-1}(y))$ follows from Lemma 2.14.

To see (iii) we use Monotone Convergence, Proposition 15.83, Theorem 15.85 and countable additivity of Lebesgue measure to compute

$$\begin{aligned}
\int \mathcal{H}^0(A \cap f^{-1}(y))\, \mathcal{H}^d(dy) &= \lim_{n\to\infty} \int g_n(y)\, \mathcal{H}^d(dy) \\
&= \lim_{n\to\infty} \sum_{C \in \mathcal{C}_n} \mathcal{H}^d(f(A \cap C)) \\
&\leq \lim_{n\to\infty} (\operatorname{Lip}(f))^d \sum_{C \in \mathcal{C}_n} \mathcal{H}^d(A \cap C) \\
&= \limsup_{n\to\infty} (\operatorname{Lip}(f))^d \sum_{C \in \mathcal{C}_n} \lambda^d(A \cap C) \\
&= (\operatorname{Lip}(f))^d \lambda^d(A)
\end{aligned}$$

□

Next we need an kind of inverse function theorem for Lipschitz maps (TODO: Check the variational analysis literature to see if there are related results; in variational analysis, things are expressed in terms of generalized derivatives which in the Lipschitz case can be defined as limits of the a.e. defined Frechet derivative. Clarke has a paper on this and I'll bet Rockafellar and Wets has something).

LEMMA 15.99. *Let $d \leq m$, $f : \mathbb{R}^d \to \mathbb{R}^m$ be a locally Lipschitz map, $t > 1$ and*

$$A = \{x \in \mathbb{R}^d \mid Df(x) \text{ exists and } Jf(x) > 0\}$$

There exists a countable set of Borel sets $A_1, A_2, \ldots$ and symmetric invertible linear map $T_n : \mathbb{R}^d \to \mathbb{R}^d$ such that

(i) $A = \cup_{n=1}^\infty A_n$
(ii) *$f \mid_{A_n}$ is injective for each $n \in \mathbb{N}$*
(iii) *For each $n \in$ naturals*

$$\operatorname{Lip}((f \mid_{A_n}) \circ T_n^{-1}) \leq t, \qquad \operatorname{Lip}(T_n \circ (f \mid_{A_n})^{-1}) \leq t \tag{22}$$

(iv) *For each $n \in \mathbb{N}$*

$$t^{-d} \left|\det(T_n)\right| \leq Jf \mid_{A_n} \leq t^d \left|\det(T_n)\right| \tag{23}$$

PROOF. Pick $\epsilon > 0$ small enough so that

$$\frac{1}{t} + \epsilon < 1 < t - \epsilon$$

First recall that every subset of a separable metric space is separable in the relative topology; it is obviously true that every subset of a second countable topological space is second countable in the relative topology (every base of a topology is a base of the relative topology as well) and in metric spaces separability and second countability are equivalent (Lemma 15.8). Therefore we may select B, a countable dense subset of A, and $\mathbf{S}$, a countable dense subset of the set of symmetric invertible linear maps of $\mathbb{R}^d$. For each $i \in \mathbb{N}$, $x \in B$ and $T \in \mathbf{S}$ we let $A(x, T, i)$ be the set of $y \in A \cap B(x, 1/i)$ such that

$$\left(\frac{1}{t} + \epsilon\right) \|Tv\| \leq \|Df(y)v\| \leq (t - \epsilon) \|Tv\| \text{ for all } v \in \mathbb{R}^d$$

and

$$\|f(z) - f(y) - Df(x) \cdot (z - y)\| \leq \epsilon \|T(z - y)\| \text{ for all } z \in B(y, 2/i)$$

By continuity of f, the linear maps T, $Df(y)$ and $Df(a)$ we may equivalently define as the countable intersection

$$\begin{aligned} A(x, T, i) = \cap_{v \in \mathbb{Q}^d} \{y \mid \left(\frac{1}{t} + \epsilon\right) \|Tv\| \leq \|Df(y)v\| \leq (t - \epsilon) \|Tv\|\} \bigcap \\ \cap_{z \in B(0, 2/i) \cap \mathbb{Q}^d} \{y \mid \|f(z + y) - f(y) - Df(y) \cdot z\| \leq \epsilon \|Tz\|\} \end{aligned}$$

Continuity of f and Borel measurability of Df (Theorem 15.92) then shows that $A(x, T, i)$ is a Borel subset of $\mathbb{R}^d$.

We now show (iv) holds on the sets $A(x, T, i)$.

Claim 15.99.1. For every $y \in A(x, T, i)$ we have

$$\left(\frac{1}{t} + \epsilon\right)^d |\det T| \leq Jf(y) \leq (t - \epsilon)^d |\det(T)|$$

Take the polar decomposition $Df(y) = U \circ S$ with $U : \mathbb{R}^d \to \mathbb{R}^m$ orthogonal and $S : \mathbb{R}^d \to \mathbb{R}^d$ symmetric. Therefore $Jf(y) = [\![Df(y)]\!] = |\det(S)|$ and we have

$$\left(\frac{1}{t} + \epsilon\right) \|Tv\| \leq \|Df(y)v\| = \|(U \circ S)v\| = \|Sv\| \leq (t - \epsilon) \|Tv\| \text{ for all } v \in \mathbb{R}^d$$

and therefore

$$\left(\frac{1}{t} + \epsilon\right) \|v\| \leq \left\|(S \circ T^{-1})v\right\| \leq (t - \epsilon) \|v\| \text{ for all } v \in \mathbb{R}^d$$

From the right hand inequality here we conclude $(S \circ T^{-1})B(0, 1) \subset B(0, t - \epsilon)$ and therefore by Corollary 15.163 and Corollary 2.92

$$\begin{aligned} |\det(S)| \, |\det(T)|^{-1} \lambda^d(B(0, 1)) &= \left|\det(S \circ T^{-1})\right| \lambda^d(B(0, 1)) \\ &= \lambda^d((S \circ T^{-1})B(0, 1)) \leq \lambda^d(B(0, t - \epsilon)) = (t - \epsilon)^d \lambda^d(B(0, 1)) \end{aligned}$$

Thus

$$|\det(S)| \leq (t - \epsilon)^d |\det(T)|$$

From the left hand inequality we conclude $B(0, \frac{1}{t} + \epsilon) \subset (S \circ T^{-1})B(0, 1)$ and by the same reasoning

$$\begin{aligned} \left(\frac{1}{t} + \epsilon\right)^d \lambda^d(B(0, 1)) &= \lambda^d(B(0, \frac{1}{t} + \epsilon)) \leq \lambda^d((S \circ T^{-1})B(0, 1)) \\ &= \left|\det(S \circ T^{-1})\right| \lambda^d(B(0, 1)) = |\det(S)| \, |\det(T)|^{-1} \lambda^d(B(0, 1)) \end{aligned}$$

which shows $\left(\frac{1}{t} + \epsilon\right)^d |\det T| \leq |\det(S)|$.

Now we show (i)

Claim 15.99.2. $\cup_{i=1}^{\infty} \cup_{T \in \mathbf{S}} \cup_{x \in B} A(x, T, i) = A$.

Note that for any linear map T we have

$$\mathrm{Lip}(T) = \sup_{x \neq y} \frac{\|Ty - Tx\|}{\|y - x\|} = \sup_{x \neq y} \frac{\|T(y-x)\|}{\|y - x\|} = \sup_{v \neq 0} \frac{\|Tv\|}{\|v\|} = \|T\|$$

Therefore if $S, S_1, S_2, \ldots$ are invertible linear maps such that $S_n \to S$ in the operator norm then from

$$\left| \left\| S_n \circ S^{-1} \right\| - 1 \right| \leq \left| \left\| S_n \circ S^{-1} - \mathrm{Id} \right\| + \|\mathrm{Id}\| - 1 \right| = \left\| S_n \circ S^{-1} - \mathrm{Id} \right\| \leq \left\| S^{-1} \right\| \|S_n - S\|$$

it follows that $\mathrm{Lip}(S_n \circ S^{-1}) \to 1$. Similarly since $S_n \to S$ implies $\|S_n\| \to \|S\| \neq 0$ we get

$$\left| \left\| S \circ S_n^{-1} \right\| - 1 \right| \leq \left\| S_n^{-1} \right\| \|S - S_n\|$$

from which it follows that $\mathrm{Lip}(S \circ S_n^{-1}) \to 1$. Let $x \in A$ and write the polar decomposition $Df(x) = U \circ S$. By the above argument and the density of $\mathbf{S}$ we may find a $T \in \mathbf{S}$ such that

$$\mathrm{Lip}(T \circ S^{-1}) \leq \left(\frac{1}{t} + \epsilon \right)^{-1}, \qquad \mathrm{Lip}(S \circ T^{-1}) \leq t - \epsilon$$

which gives us

$$\left(\frac{1}{t} + \epsilon \right) \|Tv\| \leq \left(\frac{1}{t} + \epsilon \right) \mathrm{Lip}(T \circ S^{-1}) \|Sv\| \leq \|Sv\| = \|Df(x)v\|$$

and

$$\|Df(x)v\| = \|Sv\| \leq \mathrm{Lip}(S \circ T^{-1}) \|Tv\| \leq (t - \epsilon) \|Tv\|$$

By definition of the Frechet derivative pick $i \in \mathbb{N}$ large enough so that

$$|f(y) - f(x) - Df(x) \cdot (y - x)| \leq \frac{\epsilon}{\mathrm{Lip}(T^{-1})} \|y - x\| \leq \epsilon \|T(y - x)\| \text{ for all } y \in B(x, 2/i)$$

Lastly by density of B in A we may pick $z \in B$ such $\|z - y\| < 1/i$. Now note that we have all three conditions that show $x \in A(z, T, i)$.

To prove (ii) and (iii) we need the following estimates

Claim 15.99.3. For all $y, z \in A(x, T, i)$

$$\frac{1}{t} \|T(z - y)\| \leq \|f(z) - f(y)\| \leq t \|T(z - y)\|$$

From the definition of $A(x, T, i)$ if we suppose $y \in A(x, T, i)$ and $z \in B(y, 2/i)$ then

$$\begin{aligned} \|f(z) - f(y)\| &\leq \|f(z) - f(y) - Df(y) \cdot (z - y)\| + \|Df(y) \cdot (z - y)\| \\ &\leq \epsilon \|T(z - y)\| + (t - \epsilon) \|T(z - y)\| \\ &= t \|T(z - y)\| \end{aligned}$$

and

$$\begin{aligned} \left(\frac{1}{t} + \epsilon \right) \|T(z - y)\| &\leq \|Df(y) \cdot (z - y)\| \\ &\leq \|f(z) - f(y) - Df(y) \cdot (z - y)\| + \|f(z) - f(y)\| \\ &\leq \epsilon \|T(z - y)\| + \|f(z) - f(y)\| \end{aligned}$$

which we summarize as

$$\frac{1}{t}\left\|T(z-y)\right\| \leq \left\|f(z)-f(y)\right\| \leq t\left\|T(z-y)\right\|$$

However by construction $A(x,T,i) \subset B(x,1/i) \subset B(y,2/i)$ for all $y \in A(x,T,i)$ and therefore

$$\frac{1}{t}\left\|T(z-y)\right\| \leq \left\|f(z)-f(y)\right\| \leq t\left\|T(z-y)\right\| \text{ for all } y,z \in A(x,T,i)$$

To see (ii) simply note that injectivity $f\mid_{A(x,T,i)}$ follows immediately from the left hand inequality of the previous claim. As for (iii) if $T^{-1}y, T^{-1}z \in A(x,T,i)$ then by the right hand inequality of the previous claim

$$\frac{\left\|f(T^{-1}z)-f(T^{-1}y)\right\|}{\left\|z-y\right\|} \leq t\frac{\left\|T(T^{-1}z-T^{-1}y)\right\|}{\left\|z-y\right\|} = t$$

so $\mathrm{Lip}(f\mid_{A(x,T,i)} \circ T^{-1}) \leq t$. If $y,z \in f(A(x,T,i))$ then by the left hand inequality of the previous claim

$$\frac{\left\|T(f^{-1}(z)-T(f^{-1}y)\right\|}{\left\|z-y\right\|} \leq t\frac{\left\|f(f^{-1}z)-f(f^{-1}y)\right\|}{\left\|z-y\right\|} = t$$

so $\mathrm{Lip}(T\circ(f\mid_{A(x,T,i)})^{-1}) \leq t$ ∎

Theorem 15.100 (Area Formula). *Let $d \leq m$ and $f : \mathbb{R}^d \to \mathbb{R}^m$ be Lipschitz. For each λ^d-measurable set $A \subset \mathbb{R}^d$*

$$\int_A Jf(x)\,\lambda^d(dx) = \int_{\mathbb{R}^m} \mathcal{H}^0(A\cap f^{-1}(y))\,\mathcal{H}^d(dy)$$

Proof. Let

$$B = \{x \in \mathbb{R}^d \mid Df(x) \text{ exists}\}$$

By Rademacher's Theorem 15.92 we know that $\lambda^d(B^c)=0$ and therefore $\int_A Jf(x)\,\lambda^d(dx) = \int_{A\cap B} Jf(x)\lambda^d(dx)$. Moreover we know since $\mathcal{H}^0$ is counting measure (Proposition 15.81) and by Lemma 15.98

$$\begin{aligned}
&\int_{\mathbb{R}^m} \mathcal{H}^0(A\cap B\cap f^{-1}(y))\,\mathcal{H}^d(dy)\\
&\leq \int_{\mathbb{R}^m} \mathcal{H}^0(A\cap f^{-1}(y))\,\mathcal{H}^d(dy)\\
&= \int_{\mathbb{R}^m} \mathcal{H}^0(A\cap B\cap f^{-1}(y))\,\mathcal{H}^d(dy) + \int_{\mathbb{R}^m} \mathcal{H}^0(A\cap B^c\cap f^{-1}(y))\,\mathcal{H}^d(dy)\\
&\leq \int_{\mathbb{R}^m} \mathcal{H}^0(A\cap B\cap f^{-1}(y))\,\mathcal{H}^d(dy) + (\mathrm{Lip}(f))^d\,\lambda^d(A\cap B^c)\\
&= \int_{\mathbb{R}^m} \mathcal{H}^0(A\cap B\cap f^{-1}(y))\,\mathcal{H}^d(dy)
\end{aligned}$$

Therefore $\int_{\mathbb{R}^m} \mathcal{H}^0(A\cap f^{-1}(y))\,\mathcal{H}^d(dy) = \int_{\mathbb{R}^m} \mathcal{H}^0(A\cap B\cap f^{-1}(y))\,\mathcal{H}^d(dy)$ and it suffices to assume that Df (hence Jf) exists everywhere on A.

We first assume that $A \subset \{Jf > 0\}$. Let $t > 1$ be arbitrary and choose Borel sets $A_1, A_2, \dots$ and symmetric invertible linear maps $T_1, T_2, \dots$ as in Lemma 15.99.

Let $\mathcal{C}_n$ be as in the proof of Lemma 15.98 so that the sets $A \cap A_i \cap C$ for $i \in \mathbb{N}$ and $C \in \mathcal{C}_n$ are disjoint and for any $n \in \mathbb{N}$

$$A = \cup_{i=1}^{\infty} \cup_{C \in \mathcal{C}_n} A \cap A_i \cap C$$

Since $\operatorname{diam}(A_i \cap C) \leq \operatorname{diam}(C) \to 0$ as $n \to \infty$ by the same argument as in Lemma 15.98 we conclude

$$\lim_{n\to\infty} \sum_{i=1}^{\infty} \sum_{C \in \mathcal{C}_n} \mathcal{H}^d(A \cap A_i \cap C) = \int \mathcal{H}^0(A \cap f^{-1}(y))\, \mathcal{H}^d(dy) \tag{24}$$

TODO: Maybe factor that argument out into a separate lemma.

By choice of A_i and (22) of Lemma 15.99, Proposition 15.83 and Theorem 15.85 we have

$$\begin{aligned}\mathcal{H}^d(f(A \cap A_i \cap C)) = \mathcal{H}^d((f\mid_{A_i} \circ T_i^{-1} \circ T_i)(A \cap A_i \cap C)) &\leq t^d \mathcal{H}^d(T_i(A \cap A_i \cap C)) \\ &= t^d \lambda^d(T_i(A \cap A_i \cap C))\end{aligned}$$

and similarly

$$\lambda^d(T_i(A \cap A_i \cap C) = \mathcal{H}^d((T_i \circ (f\mid_{A_i})^{-1} \circ f\mid_{A_i})(A \cap A_i \cap C) \leq t^d \mathcal{H}^d(f(A \cap A_i \cap C)$$

Using these two facts together with (23) of Lemma 15.99 we get

$$\begin{aligned}t^{-2d}\mathcal{H}^d(f(A \cap A_i \cap C)) &\leq t^{-d}\lambda^d(T_i(A \cap A_i \cap C)) \\ &= t^{-d}\left|\det(T_i)\right| \lambda^d(A \cap A_i \cap C) \\ &\leq \int_{A \cap A_i \cap C} Jf(x)\, \lambda^d(dx) \\ &\leq t^d \left|\det(T_i)\right| \lambda^d(A \cap A_i \cap C) \\ &= t^d \lambda^d(T_i(A \cap A_i \cap C)) \\ &\leq t^{2d}\mathcal{H}^d(f(A \cap A_i \cap C)\end{aligned}$$

Summing over $i \in \mathbb{N}$ and $C \in \mathcal{C}_n$ yields

$$\begin{aligned}t^{-2d}\sum_{i=1}^{\infty}\sum_{C \in \mathcal{C}_n} \mathcal{H}^d(f(A \cap A_i \cap C)) &\leq \sum_{i=1}^{\infty}\sum_{C \in \mathcal{C}_n} \int_{A \cap A_i \cap C} Jf(x)\, \lambda^d(dx) \\ &= \int_A Jf(x)\, \lambda^d(dx) \\ &\leq t^{2d} \sum_{i=1}^{\infty}\sum_{C \in \mathcal{C}_n} \mathcal{H}^d(f(A \cap A_i \cap C)\end{aligned}$$

Now let $n \to \infty$ and use (24) then let $t \downarrow 1$.

Now we assume that $A \subset \{Jf = 0\}$. Somewhat surprisingly we can reduce this case to case in which $A \subset \{Jf = 0\}$. In order to do that we let $0 < \epsilon < 1$ be arbitrary and define $g : \mathbb{R}^d \to \mathbb{R}^m \times \mathbb{R}^d$ by $g(x) = (f(x), \epsilon x)$. If we let $\pi : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}^d$ be the projection on the second coordinate then if follows that $f = \pi \circ g$. Furthermore since

$$\begin{aligned}\|g(y) - g(x)\| = \sqrt{\|f(y) - f(x)\|^2 + \epsilon^2 \|y - x\|^2} &\leq \sqrt{(\operatorname{Lip}(f))^2 \|y - x\|^2 + \epsilon^2 \|y - x\|^2} \\ &= \sqrt{(\operatorname{Lip}(f))^2 + \epsilon^2}\, \|y - x\|\end{aligned}$$

we see that g is Lipschitz. It also follows from Proposition 15.146 that g is Frechet differentiable whenever f is Frechet differentiable and moreover

$$Dg(x) = \begin{bmatrix} Df(x) \\ \epsilon \operatorname{Id}_d \end{bmatrix}$$

From the Cauchy-Binet Theorem 15.162 we see that

$$(Jg(x))^2 = \sum_{\psi \in Psi} \det((Dg(x))_\psi)^2$$

where Ψ is the set of all strictly increasing functions $\{1, \dots, d\}$ to $\{1, \dots, m+d\}$. To get a lower bound for $Jg(x)$ note the function $\psi(i) = i + m$ is in Ψ and for this ψ we have $(Dg(x))_\psi = \epsilon \operatorname{Id}$ and it follows that $Jg(x) \geq \epsilon^d > 0$. For an upper bound on $Jg(x)$ first note that by a second application of Cauchy-Binet we have

$$\begin{aligned}
(Jg(x))^2 &= \sum_{\psi \in Psi} \det((Dg(x))_\psi)^2 = \sum_{\substack{\psi \in Psi \\ \psi(d) \leq m}} \det((Dg(x))_\psi)^2 + \sum_{\substack{\psi \in Psi \\ \psi(d) > m}} \det((Dg(x))_\psi)^2 \\
&= \sum_{\substack{\psi \in Psi \\ \psi(d) \leq m}} \det((Df(x))_\psi)^2 + \sum_{\substack{\psi \in Psi \\ \psi(d) > m}} \det((Dg(x))_\psi)^2 \\
&= (Jf(x))^2 + \sum_{\substack{\psi \in Psi \\ \psi(d) > m}} \det((Dg(x))_\psi)^2
\end{aligned}$$

CLAIM 15.100.1. There is a constant C independent of $x \in A$ such that $Jg(x) \leq \epsilon C$.

By the above calculation using the Cauchy-Binet Theorem and the finiteness of Ψ it suffices to show that $|\det((Dg(x))_\psi| \leq \epsilon C$ for every $\psi \in \Psi$ with $\psi(d) > m$ and some C independent of $x \in A$. We need a few simple facts about determinants and operator norms of matrices. Note for a general $d \times d$ matrix M we have $|\det(M)| \leq \|M\|^d$; for example this follows from taking a Singular Value Decomposition (Theorem 15.157) $M = U\Sigma V^T$, recalling that $\Sigma_{11} = \|M\|$ and noting that

$$|\det(M)| = \det(\Sigma) = \prod_{i=1}^d \Sigma_{ii} \leq \Sigma_{11}^d = \|M\|^d$$

Also M is a $d \times m$ matrix and N is a $c \times m$ matrix then $\left\| \begin{bmatrix} M \\ N \end{bmatrix} \right\| \leq \|M\| + \|N\|$. For any ψ with $\psi(d) > m$ let $k \geq 1$ be largest integer such that $\psi(d-k+1) > m$ then we can write $(Dg(x))_\psi = \epsilon^k \pi \circ \left\| \begin{bmatrix} Df(x) \\ \operatorname{Id} \end{bmatrix} \right\|$ for an $m \times d$ projection matrix P. Using the linear algebra facts mentioned, the fact that $0 < \epsilon \leq 1$, $\|\pi\| = 1$ and $\|Df(x)\| \leq \operatorname{Lip}(f)$ we get the bound

$$\begin{aligned}
|\det((Dg(x))_\psi)| &\leq \|(Dg(x))_\psi\| = \epsilon^k \left\| P \begin{bmatrix} Df(x) \\ \operatorname{Id} \end{bmatrix} \right\| \\
&\leq \epsilon \|P\| (\|Df(x)\| + \|\operatorname{Id}\|) \leq \epsilon(\operatorname{Lip}(f) + 1)
\end{aligned}$$

Since $\pi : \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}^d$ is a projection operator it is Lipschitz and $\operatorname{Lip}(\pi) = \|\pi\| = 1$. Thus from Proposition 15.83, Lemma 15.97 and the part of this theorem already proven

$$\begin{aligned}\mathcal{H}^d(f(A)) &= \mathcal{H}^d(\pi(g(A))) \le \mathcal{H}^d(g(A)) \\ &\le \int \mathcal{H}^0(A \cap g^{-1}(y,z))\,\mathcal{H}^d(dy,dz) \\ &= \int_A Jg(x)\,\lambda^d(dx) \\ &\le \epsilon C \lambda^d(A)\end{aligned}$$

Since $\epsilon > 0$ was arbitrary we let $\epsilon \to 0$ and conclude $\mathcal{H}^d(f(A)) = 0$. Since the support of $\mathcal{H}^0(A \cap f^{-1}(y))$ equals $f(A)$ (Lemma 15.97) we also get

$$\int \mathcal{H}^0(A \cap f^{-1}(y))\,\mathcal{H}^d(dy) = 0 = \int_A Jf(x)\,\lambda^d(dx)$$

The proof is finished by taking a general A, writing $A = (A \cap \{Jf > 0\}) \cup (A \cap \{Jf = 0\})$ and using the fact that both $\int_A Jf(x)\,\lambda^d(dx)$ and $\int \mathcal{H}^0(A \cap f^{-1}(y))\,\mathcal{H}^d(dy)$ are finitely additive over disjoint sets. □

10.4. The Coarea Formula. The following generalizes the Fundamental Theorem of Calculus 2.113 to Radon measures on $\mathbb{R}^d$.

THEOREM 15.101. *Let μ be a Radon measure on $\mathbb{R}^d$ and f be locally μ integrable then*

$$\lim_{r \to 0} \frac{1}{\mu(B(x,r))} \int_{B(x,r)} f(y)\,\mu(dy) = f(x)$$

for μ-almost every $x \in \mathbb{R}^d$.

PROOF. For every Borel set $B \subset \mathbb{R}^d$ let $\nu_\pm(B) = \int_B f_\pm(y)\mu(dy)$. Then create the outer measures

$$\nu_\pm(A) = \inf\{\nu_\pm(B) \mid B \subset \mathbb{R}^d \text{ Borel}\}$$

These are Radon outer measures (Borel regularity is immediate from the construction and local finiteness follows from the local integrability of f). It is also clear that each of $\nu_\pm$ is absolutely continuous with respect to μ and therefore we have by the Radon-Nikodym Theorem 15.91

$$\nu_\pm(A) = \int_A D_\mu \nu_\pm\,d\mu = \int_A f_\pm\,d\mu \text{ for all measurable } A \subset \mathbb{R}^d$$

From this it follows that $D_\mu \nu_\pm = f_\pm$. Therefore by the definition of $D_\mu \nu_\pm$ we get for λ^d almost every $x \in \mathbb{R}^d$

$$\begin{aligned}\lim_{r \to 0} \frac{1}{\mu(B(x,r))} \int_{B(x,r)} f(y)\,\mu(dy) &= \lim_{r \to 0} \frac{1}{\mu(B(x,r))} \left(\nu_+(B(x,r)) - \nu_-(B(x,r))\right) \\ &= \lim_{r \to 0} (D_\mu \nu_+(B(x,r)) - D_\mu \nu_-(B(x,r))) \\ &= f_+(x) - f_-(x) = f(x)\end{aligned}$$

□

COROLLARY 15.102. *Let $A \subset \mathbb{R}^d$ be measurable then*

$$\lim_{r \to 0} \frac{\lambda^d(A \cap B(x,r))}{\lambda^d(B(x,r))} = 1 \textit{ for } \lambda^d \textit{ almost every } x \in A$$

and

$$\lim_{r \to 0} \frac{\lambda^d(A \cap B(x,r))}{\lambda^d(B(x,r))} = 0 \textit{ for } \lambda^d \textit{ almost every } x \in \mathbb{R}^d \setminus A$$

PROOF. Simply apply Theorem 15.101 to the measurable function $\mathbf{1}_A$ and the Radon measure λ^d. □

LEMMA 15.103. (i) *Let $f : \mathbb{R}^d \to \mathbb{R}^m$ be locally Lipschitz and $Z = \{x \in \mathbb{R}^d \mid f(x) = 0\}$ then $Df(x) = 0$ λ^d-a.e. on Z.*

(ii) *Let $f, g : \mathbb{R}^d \to \mathbb{R}^d$ are locally Lipschitz and $Y = \{x \in \mathbb{R}^d \mid g(f(x)) = 0\}$ then $Dg(f(x))Df(x) = \mathrm{Id}$ λ^d-a.e. on Y.*

PROOF. TODO: Can we derive the first part from Lemma 2.115? The only trick seems to be use of separability to get a countable intersection of null sets. Below we give the Evans-Gariepy proof.

To see (i), by Rademacher's Theorem and Corollary 15.102 we know that λ^d-almost everywhere on Z, $Df(x)$ exists and

$$\lim_{r \to 0} \frac{\lambda^d(A \cap B(x,r))}{B(x,r)} = 1$$

(i.e. x is in the measure theoretic interior of Z). Suppose that x is in the measure theoretic interior of Z and $Df(x) \neq 0$ exists; we will derive a contradiction. By definition of $Df(x)$ there exists $\delta > 0$ such that for all $w \in \mathbb{R}^d$ with $0 < \|w\| < \delta$ we have

$$|f(x+w) - \langle w, Df(x) \rangle| \leq \|w\| \, \|Df(x)\| \, 4$$

Let S be the unit vectors v such that the angle between v and $Df(x)$ is between -60° and 60° (i.e. $\langle v, Df(x) \rangle \geq \|Df(x)\| / 2$). It follows that for any $0 < r < \delta$,

$$|f(x+rv)| \geq r \, |\langle v, Df(x) \rangle| - |f(x+rv) - r\langle v, Df(x) \rangle| \quad \geq r \, \|Df(x)\| \, 2 - r \, \|Df(x)\| \, 4 = r \, \|Df(x)\| \, 4 > 0$$

and therefore $rS \cap Z = \emptyset$ for all $0 < r < \delta$. This yields the bound

$$\lim_{r \to 0} \frac{\lambda^d(Z \cap B(x,r))}{\lambda^d(B(x,r))} \leq \lim_{r \to 0} \frac{\lambda^d(B(x,r) \setminus rS)}{\lambda^d(B(x,r))} < 1$$

which contradicts the fact that x is in the measure theoretic interior of Z.

To see (ii) we reduce to the case of (i) by considering $g \circ f - \mathrm{Id}$ but we have to be careful to track a few λ^d null sets. We consider the points in Y where $Dg(f(x))Df(x)$ is defined; let

$$X = Y \cap \mathcal{D}(Df) \cap f^{-1}(\mathcal{D}(Dg))$$

CLAIM 15.103.1. $Y \setminus X \subset (\mathbb{R}^d \setminus \mathcal{D}(Df)) \cup g(\mathbb{R}^d \setminus \mathcal{D}(Dg))$

If $x \in Y \setminus X$ then either $x \notin \mathcal{D}(Df)$ or $x \notin f^{-1}(\mathcal{D}(Dg))$. Since the first case is equivalent to $x \in \mathbb{R}^d \setminus \mathcal{D}(Df)$ it suffices to assume $x \in Y \setminus f^{-1}(\mathcal{D}(Dg))$. In this case $f(x) \notin \mathcal{D}(Dg)$ hence $g(f(x)) \in g(\mathbb{R}^d \setminus \mathcal{D}(Dg))$; since $x \in Y$ we know that $x = g(f(x))$ and the claim follows.

From the claim and Rademacher's Theorem applied to f and g we know that $\lambda^d(Y \setminus X) = 0$. Now if we assume $x \in X$ then, by definition of X, $Dg(f(x))$ and $Df(x)$ exist and therefore by the Chain Rule (Proposition 15.137) we know that $D(g \circ f)(x)$ exists and $D(g \circ f)(x) = Dg(f(x))Df(x)$. Now we simply note that Y is contained in the zero set of $g \circ f - \mathrm{Id}$ and apply (i). □

LEMMA 15.104. *Let $A \subset \mathbb{R}^d$ and $f : A \to \mathbb{R}^m$ be a Lipschitz map then there exists a Lipschitz map $\overline{f} : \mathbb{R}^d \to \mathbb{R}^m$ such that*

(i) $\overline{f}\,|_A = f$
(ii) $\operatorname{Lip}(\overline{f}) \le \sqrt{m}\operatorname{Lip}(f)$

PROOF. First assume that $m = 1$. Define

$$\overline{f}(x) = \inf_{a \in A} (f(a) + \operatorname{Lip}(f) \|x - a\|)$$

CLAIM 15.104.1. If $x \in A$ then $\overline{f}(x) = f(x)$

Note that for all $a \in A$

$$f(x) - f(a) \le |f(x) - f(a)| \le \operatorname{Lip}(f) \|x - a\|$$

hence $f(x) \le f(a) + \operatorname{Lip}(f) \|x - a\|$ and taking the infimum over all $a \in A$ we conclude $f(x) \le \overline{f}(x)$. On the other hand since $x \in A$ it is immediate from the definition of $\overline{f}$ that $\overline{f} \le f(x)$.

CLAIM 15.104.2. $\operatorname{Lip}(\overline{f}) = \operatorname{Lip}(f)$

Given $x, y \in \mathbb{R}^d$ we get

$$\begin{aligned}\overline{f}(x) &= \inf_{a \in A} (f(a) + \operatorname{Lip}(f) \|x - a\|) \le \inf_{a \in A} (f(a) + \operatorname{Lip}(f) \|y - a\| + \operatorname{Lip}(f) \|x - y\|) \\ &= \overline{f}(y) + \operatorname{Lip}(f) \|x - y\|\end{aligned}$$

By symmetry we also have $\overline{f}(y) \le \overline{f}(x) + \operatorname{Lip}(f) \|x - y\|$ and it follows that $|f(x) - f(y)| \le \operatorname{Lip}(f) \|x - y\|$ hence $\operatorname{Lip}(\overline{f}) \le \operatorname{Lip}(f)$. The fact that $\operatorname{Lip}(f) \le \operatorname{Lip}(\overline{f})$ follows from the fact that $\overline{f}\,|_A = f$.

Now to handle the case of general m let $f = (f_1, \dots, f_m)$ and construct $\overline{f} = (\overline{f}_1, \dots, \overline{f}_m)$ and note that

$$\left\|\overline{f}(x) - \overline{f}(y)\right\|^2 = \sum_{j=1}^m \left|\overline{f}_j(x) - \overline{f}_j(y)\right|^2 \le m(\operatorname{Lip}(f))^2 \|x - y\|$$

□

The first lemma proves the coarea formula for linear maps.

LEMMA 15.105. *Let $d \ge m$, $L : \mathbb{R}^d \to \mathbb{R}^m$ be linear and $A \subset \mathbb{R}^d$ be λ^d-measurable then*

(i) *The map $y \to \mathcal{H}^{d-m}(A \cap L^{-1}(y))$ is λ^m-measurable*
(ii) $\int \mathcal{H}^{d-m}(A \cap L^{-1}(y))\, \lambda^m(dy) = \frac{2^{d-m}}{\alpha(d-m)} [\![L]\!] \lambda^d(A)$.

PROOF. First suppose that $\dim(L) < m$. In this case $\lambda^m(L(\mathbb{R}^d)) = 0$ and $A \cap L^{-1}(y) = \emptyset$ for all $y \in \mathbb{R}^m \setminus L(\mathbb{R}^d)$. From this it follows that $\mathcal{H}^{d-m}(A \cap L^{-1}(y)) = 0$ λ^m-almost everywhere which implies $\mathcal{H}^{d-m}(A \cap L^{-1}(y))$ is λ^m-measurable and $\int \mathcal{H}^{d-m}(A \cap L^{-1}(y))\, \lambda^m(dy) = 0$. On the other hand $[\![L]\!]^2 = \det(L \circ L^T) = 0$ and (ii) follows.

We begin by handling the case of the orthogonal projection $P(x_1, \dots, x_d) = (x_1, \dots, x_m)$; i.e. if we write $\mathbb{R}^d = \mathbb{R}^m \times \mathbb{R}^{d-m}$ then P is just the first coordinate projection. In this case $A \cap P^{-1}(y)$ is just $\{y\} \times A_y$ where A_y is the section $\{x \in \mathbb{R}^{d-m} \mid (x, y) \in A\}$. Furthermore for any $y \in \mathbb{R}^m$ the map $x \mapsto (y, x)$ from

$\mathbb{R}^{d-m} \to \mathbb{R}^d$ is an isometric embedding hence by Proposition 15.82 (TODO: the σ-algebra of λ^d-measurable sets is actually bigger than the product σ-algebra yet we have only proven Fubini for the product σ-algebra; we must address the completion! Note that sections of λ^d-measurable sets are only almost surely measurable; as an example let A be a non-measurable set in $\mathbb{R}$ and then consider $\{0\} \times A \subset \mathbb{R}^2$ it is a Lebesgue null set hence λ^2-measurable but has a non-measurable section) and Theorem 15.85 we know that

$$\mathcal{H}^{d-m}(A \cap P^{-1}(y)) = \mathcal{H}^{d-m}(\{y\} \times A_y) = \mathcal{H}^{d-m}(A_y) = \frac{2^{d-m}}{\alpha(d-m)} \lambda^{d-m}(A_y)$$

The λ^m-measurability of $\mathcal{H}^{d-m}(A \cap P^{-1}(y))$ follows from Lemma 2.87 and Tonelli's Theorem shows us that

$$\int \mathcal{H}^{d-m}(A \cap P^{-1}(y)) \, \lambda^m(dy) = 2^{d-m}\alpha(d-m) \int \lambda^{d-m}(A_y) \, \lambda^m(dy) = \frac{2^{d-m}}{\alpha(d-m)} \lambda^d(A)$$

Since $[\![P]\!] = 1$ in this case the result is proven.

Now we will handle the case of a general linear map L in which $\dim(L) = m$. Take the singular value decomposition and use the fact that $d \geq m$ to write $L = U\Sigma P V^T$ where U and V are orthogonal, P is the projection matrix above and Σ is the $m \times m$ diagonal matrix of singular values (in particular $\det(\Sigma) = [\![L]\!]$ by Lemma 15.94). Now using Corollary 2.92, the case of the current lemma for the matrix P and Proposition 15.82

$$\begin{aligned}
\frac{2^{d-m}}{\alpha(d-m)} \lambda^d(A) &= \frac{2^{d-m}}{\alpha(d-m)} \lambda^d(V^T(A)) \\
&= \int \mathcal{H}^{d-m}(V^T(A) \cap P^{-1}(y)) \, \lambda^m(dy) \\
&= \int \mathcal{H}^{d-m}(V^T(A \cap (V \circ P^{-1})(y))) \, \lambda^m(dy) \\
&= \int \mathcal{H}^{d-m}(A \cap (V \circ P^{-1})(y)) \, \lambda^m(dy)
\end{aligned}$$

Now we use this equality and the Area Formula with the transformation $U \circ \Sigma : \mathbb{R}^m \to \mathbb{R}^m$ and the function $\mathcal{H}^{d-m}(A \cap (V \circ P^{-1})(y))$ to calculate (TODO: get all the constants right)

$$\begin{aligned}
\int \mathcal{H}^{d-m}(A \cap L^{-1}(z)) \, \lambda^m(z) &= \int \mathcal{H}^{d-m}(A \cap (V \circ P^{-1} \circ \Sigma^{-1} \circ U^T)(z)) \, \lambda^m(z) \\
&= 2^{-m}\alpha(m) \int \sum_{y \in (\Sigma^{-1} \circ U^T)(z)} \mathcal{H}^{d-m}(A \cap (V \circ P^{-1})(y)) \, \mathcal{H}^m(z) \\
&= |\det(U \circ \Sigma)| \int \mathcal{H}^{d-m}(A \cap (V \circ P^{-1})(y)) \, \lambda^m(y) \\
&= \frac{2^{d-m}}{\alpha(d-m)} [\![L]\!] \lambda^d(A)
\end{aligned}$$

(Here we use the invertibility of $U \circ \Sigma$ to see that $(\Sigma^{-1} \circ U^T)(z)$ is a singleton set).

TODO: The measurability of $\mathcal{H}^{d-m}(A \cap L^{-1}(y))$ for general full-rank L is also supposed to follow from measurability of sections by essentially the same argument as that for P. $\square$

Now we address the measurability of the mapping $y \mapsto \mathcal{H}^{d-m}(A \cap f^{-1}(y))$ for Lipschitz f.

LEMMA 15.106. *Let $d \geq m$, $f : \mathbb{R}^d \to \mathbb{R}^m$ be Lipschitz and $A \subset \mathbb{R}^d$ be λ^d-measurable. Then*

(i) *The set $A \cap f^{-1}(y)$ is $\mathcal{H}^{d-m}$-measurable for λ^m almost everywhere.*
(ii) *The function $y \mapsto \mathcal{H}^{d-m}(A \cap f^{-1}(y))$*
(iii) $\int \mathcal{H}^{d-m}(A \cap f^{-1}(y))\, \lambda^m(dy) \leq 2^{d-m}\alpha(m)(\mathrm{Lip}(f))^m \lambda^d(A)$.

PROOF. We first prove a variant of (iii) in which the integral is replaced by an outer integral (use of the latter means we can avoid showing measurability). Now by a basic property of Lebesgue outer measure (TODO: Where do we show this; perhaps this will come up in the proof of Theorem 15.85???? In fact follows from Theorem 15.85 once we show that Hausdorff measure can be calculated using coverings by closed balls) we know that for every $n \in \mathbb{N}$ there exist closed balls $B_1^n, B_2^n, \ldots$ such that $A \subset \cup_{j=1}^\infty B_j^n$, $\mathrm{diam}(B_j^n) \leq 1/n$ and $\lambda^d(A) \leq \sum_{j=1}^n \lambda^d(B_j^n) \leq \lambda^d(A) + 1/n$. This covering of A shows that for each $y \in \mathbb{R}^m$ we have

$$\mathcal{H}^{d-m}_{1/n}(A \cap f^{-1}(y)) \leq \sum_{\substack{1 \leq j < \infty \\ B_j^n \cap f^{-1}(y) \neq \emptyset}} (\mathrm{diam}(B_j^n))^{d-m} = \sum_{j=1}^\infty (\mathrm{diam}(B_j^n))^{d-m} \mathbf{1}_{f(B_j^n)}(y)$$

Since the ball B_j^n is compact and f is continuous it follows that $f(B_j^n)$ is compact hence λ^m-measurable. Taking the limit of both sides we get

$$\mathcal{H}^{d-m}(A \cap f^{-1}(y)) = \lim_{n \to \infty} \mathcal{H}^{d-m}_{1/n}(A \cap f^{-1}(y)) \quad \leq \liminf_{n \to \infty} \sum_{j=1}^\infty (\mathrm{diam}(B_j^n))^{d-m} \mathbf{1}_{f(B_j^n)}(y)$$

thus we have a measurable majorant of the mapping $y \mapsto \mathcal{H}^{d-m}(A \cap f^{-1}(y))$. Applying the above majorant, Fatou's Lemma, Monotone Convergence (specifically Corollary 2.44), the Isodiametric Inequality (Theorem 15.84) and the fact that

$\lambda^d(B^n_j) = (\mathrm{diam}(B^n_j)/2)^d$ we get

$$\begin{aligned}
\int^* \mathcal{H}^{d-m}(A \cap f^{-1}(y))\,\lambda^m(dy) &\leq \int \liminf_{n\to\infty} \sum_{j=1}^\infty (\mathrm{diam}(B^n_j))^{d-m} \mathbf{1}_{f(B^n_j)}(y)\,\lambda^m(dy) \\
&\leq \liminf_{n\to\infty} \int \sum_{j=1}^\infty (\mathrm{diam}(B^n_j))^{d-m} \mathbf{1}_{f(B^n_j)}(y)\,\lambda^m(dy) \\
&= \liminf_{n\to\infty} \sum_{j=1}^\infty (\mathrm{diam}(B^n_j))^{d-m} \lambda^m(f(B^n_j)) \\
&\leq \liminf_{n\to\infty} \sum_{j=1}^\infty (\mathrm{diam}(B^n_j))^{d-m} \alpha(m) \left(\frac{\mathrm{diam}(f(B^n_j))}{2}\right)^m \\
&\leq \liminf_{n\to\infty} \sum_{j=1}^\infty (\mathrm{diam}(B^n_j))^{d} 2^{-m} (\mathrm{Lip}(f))^m \alpha(m) \\
&= 2^{d-m} (\mathrm{Lip}(f))^m \alpha(m) \liminf_{n\to\infty} \sum_{j=1}^\infty \lambda^d(B^n_j) \\
&= 2^{d-m} (\mathrm{Lip}(f))^m \alpha(m) \lambda^d(A)
\end{aligned}$$

We first prove (ii). We break this down into several different cases depending on the characteristics of A.

Case 15.106.1. A is compact

Fix $t \geq 0$. For every $n \in \mathbb{N}$ let U_n be the set of $y \in \mathbb{R}^m$ such that there exists $k \in \mathbb{N}$ and open sets $V_1, \dots, V_k$ satisfying

(i) $A \cap f^{-1}(y) \subset \cup_{j=1}^k V_j$

(ii) $\mathrm{diam}(V_j) \leq 1/n$ for $j = 1, \dots, k$

(iii) $\sum_{j=1}^k (\mathrm{diam}(V_j))^{d-m} \leq t + 1/n$

Claim 15.106.1. U_n is open

In fact given $y \in U_n$ and open sets $V_1, \dots, V_k$ there is an open neighborhood W of y such that $A \cap f^{-1}(z) \subset \cup_{j=1}^k V_j$ for all $z \in W$. If this were not true then set could find a sequence $y_i \to y$ such that $A \cap f^{-1}(y_i) \setminus \cup_{j=1}^k V_j \neq \emptyset$. By using the compactness of A to pass to a subsequence if necessary we may pick $x_i \in A \cap f^{-1}(y_i) \setminus \cup_{j=1}^k V_j$ such that $x_i \to x$ with $x \in A \setminus \cup_{j=1}^k V_j$. By continuity of f we know that

$$f(x) = \lim_{i\to\infty} f(x_i) = \lim_{i\to\infty} y_i = y$$

hence $x \in A \cap f^{-1}(y) \subset \cup_{j=1}^k V_j$ which is a contradiction.

Claim 15.106.2. The map $y \mapsto \mathcal{H}^{d-m}(A \cap f^{-1}(y))$ is Borel measurable in fact

$$\{y \in \mathbb{R}^m \mid \mathcal{H}^{d-m}(A \cap f^{-1}(y)) \leq t\} = \cap_{n=1}^\infty U_n$$

We first show the inclusion $\{y \in \mathbb{R}^m \mid \mathcal{H}^{d-m}(A \cap f^{-1}(y)) \leq t\} = \cap_{n=1}^\infty U_n$. Since $\mathcal{H}^{d-m}(A \cap f^{-1}(y)) \leq t$ and $\mathcal{H}^{d-m}(A \cap f^{-1}(y)) = \sup_{\delta > 0} \mathcal{H}^{d-m}_\delta(A \cap f^{-1}(y))$ we know that $\mathcal{H}^{d-m}_\delta(A \cap f^{-1}(y)) \leq t$ for all $\delta > 0$. By Proposition 15.79 for any $n \in \mathbb{N}$ we pick $0 < \delta < 1/n$ and some open sets $V_1, V_2, \dots$ such that

(i) $A \cap f^{-1}(y) \subset \cup_{j=1}^\infty V_j$
(ii) $\operatorname{diam}(V_j) \leq \delta < 1/n$ for $j \in \mathbb{N}$
(iii) $\sum_{j=1}^\infty (\operatorname{diam}(V_j))^{d-m} < t + 1/n$

Since f is continuous we know that $f^{-1}(y)$ is closed hence $A \cap f^{-1}(y)$ is a closed subset of a compact set hence compact (Corollary 1.30). Thus we may find a finite subcover $V_1, \dots, V_k$ of $A \cap f^{-1}(y)$ which shows $y \in U_n$.

Now suppose that $y \in \cap_{n=1}^\infty U_n$. By the definitions of U_n and $\mathcal{H}^{d-m}$ we know that $y \in U_n$ implies $\mathcal{H}^{d-m}_{1/n}(A \cap f^{-1}(y)) \leq t + 1/n$; therefore we know that $\mathcal{H}^{d-m}_{1/n}(A \cap f^{-1}(y)) \leq t + 1/n$ for all $n \in \mathbb{N}$. Now take the limit as $n \to \infty$ to conclude that $\mathcal{H}^{d-m}(A \cap f^{-1}(y)) \leq t$.

Case 15.106.2. A is open.

For each $n \in \mathbb{N}$ define

$$K_n = ((A^c))1/n)^c \cap \overline{B}(0,n) = \{x \mid d(x, A^c) \geq 1/n\} \cap \overline{B}(0,n)$$

which is a closed bounded set and therefore compact. By the definition of K_n and the openness of A we have $K_1 \subset K_2 \subset \cdots \subset A$ and $\cup_{n=1}^\infty K_n = A$. By Borel regularity of $\mathcal{H}^{d-m}$ and continuity of regular outer measures (Proposition 2.97) we have

$$\mathcal{H}^{d-m}(A \cap f^{-1}(y)) = \lim_{n \to \infty} \mathcal{H}^{d-m}(K_n \cap f^{-1}(y))$$

which shows that $y \mapsto \mathcal{H}^{d-m}(A \cap f^{-1}(y))$ is open.

Case 15.106.3. $\lambda^d(A) < \infty$.

By the outer regularity of λ^d (or the Lebesgue measurability of A) we can find open sets $V_1 \supset V_2 \supset \cdots \supset A$ such that $\lim_{n\to\infty} \lambda^d(V_n \setminus A) = 0$ and $\lambda^d(V_1) < \infty$. By the subadditivity of $\mathcal{H}^{d-m}$ we get

$$\mathcal{H}^{d-m}(V_n \cap f^{-1}(y)) \leq \mathcal{H}^{d-m}(A \cap f^{-1}(y)) + \mathcal{H}^{d-m}((V_n \setminus A) \cap f^{-1}(y))$$

Now put this bound together with the already proven version of (iii) with outer integrals

$$\begin{aligned}
0 &\leq \limsup_{n\to\infty} \int^* \left| \mathcal{H}^{d-m}(V_n \cap f^{-1}(y)) - \mathcal{H}^{d-m}(A \cap f^{-1}(y)) \right| \lambda^m(dy) \\
&\leq \limsup_{n\to\infty} \int^* \mathcal{H}^{d-m}((V_n \setminus A) \cap f^{-1}(y)) \, \lambda^m(dy) \\
&\leq 2^{d-m} \alpha(m) (\operatorname{Lip}(f))^m \limsup_{n\to\infty} \lambda^m(V_n \setminus A) = 0
\end{aligned}$$

which shows (TODO: be careful to go through the details of this argument since we have an outer integral) $\lim_{n\to\infty} \mathcal{H}^{d-m}(V_n \cap f^{-1}(y)) = \mathcal{H}^{d-m}(A \cap f^{-1}(y))$ λ^m-almost everywhere. Thus by the Borel measurablility of $\mathcal{H}^{d-m}(V_n \cap f^{-1}(y))$ for V_n open we conclude that $\mathcal{H}^{d-m}(A \cap f^{-1}(y))$ is λ^m-measurable. Furthermore this shows

$$\mathcal{H}^{d-m}((\cap_{n=1}^\infty V_n \setminus A) \cap f^{-1}(y)) = \lim_{n\to\infty} \mathcal{H}^{d-m}((V_n \setminus A) \cap f^{-1}(y)) = 0$$

so $(\cap_{n=1}^\infty V_n \setminus A) \cap f^{-1}(y)$ is $\mathcal{H}^{d-m}$-measurable. Since we can write

$$A \cap f^{-1}(y) = \cap_{n=1}^\infty V_n \setminus (\cap_{n=1}^\infty V_n \setminus A) \cap f^{-1}(y) = \cap_{n=1}^\infty V_n \cap f^{-1}(y) \setminus (\cap_{n=1}^\infty V_n \setminus A) \cap f^{-1}(y)$$

and $\cap_{n=1}^\infty V_n \cap f^{-1}(y)$ is Borel measurable (hence $\mathcal{H}^{d-m}$-measurable) we see that $A \cap f^{-1}(y)$ is also $\mathcal{H}^{d-m}$ measurable for λ^m-almost every y.

CASE 15.106.4. $\lambda^d(A) = \infty$.

We write $A = \cup_{n=1}^\infty (A \cap B(0, n))$ and use the previous case to see that $A \cap f^{-1}(y)$ is $\mathcal{H}^{d-m}$ measurable. From continuity of regular outer measures (Proposition 2.97) we have

$$\mathcal{H}^{d-m}(A \cap f^{-1}(y)) = \lim_{n\to\infty} \mathcal{H}^{d-m}(A \cap B(0, n) \cap f^{-1}(y))$$

which in turns shows the $\mathcal{H}^{d-m}$ measurability of $y \mapsto \mathcal{H}^{d-m}(A \cap f^{-1}(y))$.

To complete the proof note that (i) and (ii) together with the version of (iii) proven for outer integrals shows (iii) (TODO: where do we show that the outer integral of a measurable function equals the integral). □

LEMMA 15.107. *Let $f : \mathbb{R}^d \to \mathbb{R}^d$ be a Lipschitz map, $t > 1$ and*

$$A = \{x \in \mathbb{R}^d \mid Df(x) \text{ exists and } Jf(x) > 0\}$$

There exists a countable set of Borel sets $A_1, A_2, \dots$ and symmetric invertible linear map $T_n : \mathbb{R}^d \to \mathbb{R}^d$ such that

(i) $\lambda^d(A \setminus \cup_{n=1}^\infty A_n)$
(ii) *$f\mid_{A_n}$ is injective for each $n \in \mathbb{N}$*
(iii) *For each $n \in naturals$*

$$\operatorname{Lip}(T_n^{-1} \circ (f\mid_{A_n})) \le t, \qquad \operatorname{Lip}((f\mid_{A_n})^{-1} \circ T_n) \le t \tag{25}$$

(iv) *For each $n \in \mathbb{N}$*

$$t^{-d} \left|\det(T_n)\right| \le Jf\mid_{A_n} \le t^d \left|\det(T_n)\right| \tag{26}$$

PROOF. We apply Lemma 15.99 to f, t and A to construct Borel sets $B_1, B_2, \dots$ and symmetric invertible maps $S_1, S_2, \dots$ such that $A = \cup_{n=1}^\infty B_n$, $f\mid_{B_n}$ is injective for each $n \in \mathbb{N}$ and for all $n \in \mathbb{N}$

$$\operatorname{Lip}((f\mid_{B_n}) \circ S_n^{-1}) \le t, \qquad \operatorname{Lip}(S_n \circ (f\mid_{B_n})^{-1}) \le t$$
$$t^{-d} \left|\det(S_n)\right| \le Jf\mid_{A_n} \le t^d \left|\det(S_n)\right|$$

Note that we have for all $n \in \mathbb{N}$ and $x, y \in f(B_n)$

$$\begin{aligned}\left\|(f\mid_{B_n})^{-1}(x) - (f\mid_{B_n})^{-1}(y)\right\| &= \left\|S_n^{-1} \circ S_n \circ (f\mid_{B_n})^{-1}(x) - S_n^{-1} \circ S_n \circ (f\mid_{B_n})^{-1}(y)\right\| \\ &\le \left\|S_n^{-1}\right\| \operatorname{Lip}(S_n \circ (f\mid_{B_n})^{-1}) \left\|x - y\right\|\end{aligned}$$

hence $(f\mid_{B_n})^{-1}$ is a Lipschitz map. By Lemma 15.104 we may let $g_n : \mathbb{R}^d \to \mathbb{R}^d$ be a Lipschitz extension of $(f\mid_{B_n})^{-1}$ (so in particular $g_n(f(x)) = x$ for all $x \in B_n$.

CLAIM 15.107.1. $Jg_n > 0$ λ^d-almost everywhere on $f(B_n)$.

We know that $g_n \circ f = \mathrm{Id}$ on B_n hence by Lemma 15.103 we know that $Dg_n(f(x))Df(x) = \mathrm{Id}$ λ^d almost everywhere on B_n. Therefore by multiplicativity of determinants (Corollary 15.163) we know $Jg_n(f(x))Jf(x) = 1$ λ^d almost everywhere on B_n and in particular $Jg_n(f(x)) > 0$ λ^d almost everywhere on B_n. Since f is Lipschitz the claim follows by Proposition 15.83 (TODO: Is there an obvious

Lebesgue measure version of this or do we need to appeal to Theorem 15.85 as well?).

For each $n \in \mathbb{N}$ we apply Lemma 15.99 to g_n and t to construct Borel sets $C_1^n, C_2^n, \dots$ and symmetric invertible maps $R_1^n, R_2^n, \dots$ such that

$$\{x \in \mathbb{R}^d \mid Dg_n(x) \text{ exists and } Jg_n(x) > 0\} = \cup_{j=1}^\infty C_j^n$$

$g_n \mid_{C_j^n}$ is injective and

$$\mathrm{Lip}((g_n \mid_{C_j^n}) \circ (R_j^n)^{-1}) \leq t, \qquad \mathrm{Lip}(R_j^n \circ (g_n \mid_{R_j^n})^{-1}) \leq t$$
$$t^{-d} \left|\det(R_j^n)\right| \leq Jg_n \mid_{C_j^n} \leq t^d \left|\det(R_j^n)\right|$$

for $j \in \mathbb{N}$. By the previous claim

$$\lambda^d \left(g(B_n) \setminus \cup_{j=1}^\infty C_j^n\right) = 0$$

Now we are ready to define our final sets and injective maps. For $j, n \in \mathbb{N}$ let $A_j^n = B_n \cap f^{-1}(C_j^n)$ and $T_j^n = (R_j^n)^{-1}$.

Claim 15.107.2. For all $n, j \in \mathbb{N}$

(i) $x \in A_j^n$ we have $f(x) = (g_n \mid_{C_j^n})^{-1}(x)$

(ii) for all $y \in f(A_j^n)$ we have $g_n(y) = (f \mid_{A_j^n})^{-1}(y)$

To see (i) first note that the statement makes sense. In particular we need to make sure the domain of $(g_n \mid_{C_j^n})^{-1}$ of the domain contains A_j^n. If $x \in A_j^n$ then $x \in B_n$ and $f(x) \in C_j^n \cap f(B_n)$. Since by construction $g_n \mid_{f(B_n)} = (f \mid_{B_n})^{-1}$ we know that $g_n(f(x)) = (f \mid_{B_n})^{-1}(f(x)) = x$ which shows us that $x \in g_n(C_j^n) = \mathcal{D}((g_n \mid_{C_j^n})^{-1})$. By injectivity of g_n on C_j^n the fact that $g_n(f(x)) = x$ also shows us that $f(x) = (g_n \mid_{C_j^n})^{-1}(x)$ on A_j^n.

The claim (ii) follows from (i) by simply writing $y = f(x)$ for $x \in A_j^n$ to see that

$$g_n(y) = g_n(f(x)) = x = (f \mid_{A_j^n})^{-1}(y)$$

Claim 15.107.3. $\lambda^d \left(A \setminus \cup_{n=1}^\infty \cup_{j=1}^\infty A_j^n\right) = 0$.

For each $n \in \mathbb{N}$

$$\begin{aligned} g_n \left(f(B_n) \setminus \cup_{j=1}^\infty C_j^n\right) &= (f \mid_{B_n})^{-1} \left(f(B_n) \setminus \cup_{j=1}^\infty C_j^n\right) \\ &= B_n \setminus \cup_{j=1}^\infty f^{-1}(C_j^n) \\ &= B_n \setminus \cup_{j=1}^\infty A_j^n \end{aligned}$$

Therefore

$$\lambda^d \left(B_n \setminus \cup_{j=1}^\infty A_j^n\right) \leq \mathrm{Lip}(g_n)\lambda^d \left(f(B_n) \setminus \cup_{j=1}^\infty C_j^n\right) = 0$$

and

$$\begin{aligned} \lambda^d \left(A \setminus \cup_{n=1}^\infty \cup_{j=1}^\infty A_j^n\right) &= \lambda^d \left(\cup_{n=1}^\infty B_n \setminus \cup_{n=1}^\infty \cup_{j=1}^\infty A_j^n\right) \\ &\leq \sum_{n=1}^\infty \lambda^d \left(B_n \setminus \cup_{k=1}^\infty \cup_{j=1}^\infty A_j^k\right) \\ &\leq \sum_{n=1}^\infty \lambda^d \left(B_n \setminus \cup_{j=1}^\infty A_j^n\right) = 0 \end{aligned}$$

Claim 15.107.4. For every $n, j \in \mathbb{N}$ $f \mid_{A_j^n}$ is injective

This follows immediately from the fact that $f \mid_{B_n}$ is injective and $A_j^n \subset B_n$.

Claim 15.107.5. For every $n, j \in \mathbb{N}$ we have

$$\operatorname{Lip}((T_j^n)^{-1} \circ (f \mid_{A_j^n})) \leq t, \qquad \operatorname{Lip}((f \mid_{A_j^n})^{-1} \circ T_j^n) \leq t$$
$$t^{-d} \left|\det(T_j^n)\right| \leq Jf \mid_{A_j^n} \leq t^d \left|\det(T_j^n)\right|$$

Using Claim 15.107.2 we know that $(g_n \mid_{C_j^n})^{-1}$ is an extension of $f \mid_{A_j^n}$ hence

$$\operatorname{Lip}((T_j^n)^{-1} \circ (f \mid_{A_j^n})) = \operatorname{Lip}(R_j^n \circ (f \mid_{A_j^n})) \leq \operatorname{Lip}(R_j^n \circ (g_n \mid_{C_j^n})^{-1}) \leq t$$

and similarly since $g_n \mid_{C_j^n}$ is an extension of $(f \mid_{A_j^n})^{-1}$

$$\operatorname{Lip}((f \mid_{A_j^n})^{-1} \circ T_j^n) = \operatorname{Lip}((f \mid_{A_j^n})^{-1} \circ (R_j^n)^{-1}) \leq \operatorname{Lip}((g_n \mid_{C_j^n} \circ (R_j^n)^{-1}) \leq t$$

From the proof of claim 15.107.1 we know that $Jg_n(f(x))Jf(x) = 1$ λ^d-almost everywhere on A_j^n therefore

$$t^{-n} \left|\det(S_j^n)\right| = t^{-n} \left|\det(R_j^n)\right|^{-1} \leq (Jg_n \mid_{C_j^n})^{-1} = Jf \mid_{A_j^n}$$

and

$$Jf \mid_{A_j^n} = (Jg_n \mid_{C_j^n})^{-1} \leq t^n \left|\det(R_j^n)\right|^{-1} = t^n \left|\det(S_j^n)\right|$$

$\square$

Theorem 15.108 (Coarea Formula). *Let $d \geq m$, $f : \mathbb{R}^d \to \mathbb{R}^m$ be a Lipschitz function and $A \subset \mathbb{R}^d$ be a λ^d-measurable set then*

$$\int_A Jf(x)\, \lambda^d(dx) = \int \mathcal{H}^{d-m}(A \cap f^{-1}(y))\, \lambda^m(dy)$$

Proof.

Claim 15.108.1. If suffices to prove the theorem for A on which $Df(x)$ and $Jf(x)$ exist everwhere and for which $\lambda^d(A) < \infty$.

By Rademacher's Theorem we know that $\lambda^d(\mathbb{R}^d \setminus \{Df(x) \text{ exists }\}) = 0$. From this and the additivity of λ^d we know that

$$\int_A Jf(x)\, \lambda^d(dx) = \int_{A \cap \{Df(x) \text{ exists }\}} Jf(x)\, \lambda^d(dx) + \int_{A \setminus \{Df(x) \text{ exists }\}} Jf(x)\, \lambda^d(dx) = \int_{A \cap \{Df(x) \text{ exists }\}} Jf(x)\, \lambda^d($$

and by the subadditivity of $\mathcal{H}^{d-m}$ and Lemma 15.106

$$\begin{aligned}\int \mathcal{H}^{d-m}(A \cap \{Df(x) \text{ exists }\} \cap f^{-1}(y))\, \lambda^m(dy) &\leq \int \mathcal{H}^{d-m}(A \cap f^{-1}(y))\, \lambda^m(dy) \\ &\leq \int \mathcal{H}^{d-m}(A \cap \{Df(x) \text{ exists }\} \cap f^{-1}(y))\, \lambda^m(dy) + \int \mathcal{H}^{d-m}(\\ &\leq \int \mathcal{H}^{d-m}(A \cap \{Df(x) \text{ exists }\} \cap f^{-1}(y))\, \lambda^m(dy) + 2^{d-m}\alpha(m) \\ &= \int \mathcal{H}^{d-m}(A \cap \{Df(x) \text{ exists }\} \cap f^{-1}(y))\, \lambda^m(dy)\end{aligned}$$

hence $\int \mathcal{H}^{d-m}(A \cap f^{-1}(y))\, \lambda^m(dy) = \int \mathcal{H}^{d-m}(A \cap \{Df(x) \text{ exists }\} \cap f^{-1}(y))\, \lambda^m(dy)$. From these two facts it follows that we may replace A by $A \cap \{Df(x) \text{ exists }\}$ and assume that $Df(x)$ and $Jf(x)$ exists everywhere on A.

To see the reduction to the case in which $\lambda^d(A) < \infty$ assume the result holds for such sets and define $A_n = A \cap B(0,n)$. Applying Proposition 2.97 and Montone Convergence we see

$$\int_A Jf(x)\,\lambda^d(dx) = \lim_{n\to\infty}\int_{A_n} Jf(x)\,\lambda^d(dx) = \lim_{n\to\infty}\int \mathcal{H}^{d-m}(A_n \cap f^{-1}(y))\,\lambda^m(dy) = \int \mathcal{H}^{d-m}(A \cap f^{-1}(y))\,\lambda^m(dy$$

TODO: □

11. Integration in Banach Spaces

Our prior development of measure and integration theory made use of the special properties of the reals in various places and as a result the theory does not hold for functions with values in arbitrary vector spaces. As we shall soon see it is useful to be able to integrate functions with vector space values (in particular Banach spaces) so we need an integration theory. As it turns out there are a couple of directions that one can go. In the simplest case that will suffice for many of our needs, we simply develop the theory of Riemann integrals. The primary loss of generality is that the domains of functions in the Riemann integral case must be functions of a real variable. For our purposes we shall only be requiring the Riemann theory for a single real variable so that shall suit us fine. For problems in which the domain in an arbitrary measurable space we need a Lebesgue-like theory that was developed by Bochner. The reader may want to be made aware that in addition to these integrals there is also an integral due to Gelfand and Pettis that we shall not discuss.

11.1. Riemann Integrals. As mentioned we shall only bother to develop the Riemann integral for a single real variable.

Definition 15.109. Let $a \leq b$ be real numbers then a *partition* of the interval $[a,b]$ is a finite sequence of real numbers $a = a_0 \leq a_1 \leq \cdots \leq a_n = b$. Let X be a Banach space then a map $f : [a,b] \to X$ is said to be a *step map with repsect to* P if there exists a partition $P = \{a_j\}_{j=0}^n$ and elements $w_1, \dots, w_n \in X$ such that $f(t) = w_j$ for $a_{j-1} < t < a_j$. A *step map* is any map f such that for which there exists a partition P for which f is a step map with respect to P. The *integral* of a step map with repsect to P is

$$I_P(f) = \sum_{j=1}^n (a_j - a_{j-1}) w_j$$

Note that a step map has it values constrained on the open intervals (a_{j-1}, a_j) but not at the points a_j.

With all of these elementary definitions in hand we come to our first task which is to show that the integral of a step map is well defined.

Proposition 15.110. *Let X be a Banach space and let $f : [a,b] \to X$ be a step map with respect to partitions P and Q then it follows that $I_P(f) = I_Q(f)$.*

Proof. Given a partition P of the form $a = a_0 \leq \cdots \leq a_n = b$ let $c \in [a,b]$ and let the refinement P_c represent the partition obtained by adding c to the set of a_j. It is clear that f is still a step map with respect to P_c and that $I_P(f) = I_{P_c}(f)$. A partition R is said to be a refinement of P if it is a subset of P; by induction we see that $I_P(f) = I_R(f)$ whenever R is a refinement of P. Now given arbitrary partitions P and Q as in the hypotheses we simply find a common refinement (e.g. take the union of P and Q) and the result follows. $\square$

Now we extend the integral by a limiting procedure. To do this we use somewhat abstract language of Banach space theory. First let us set up the Banach space in which we operate.

PROPOSITION 15.111. *Let X be a normed vector space, let S be an arbitrary set and let $\mathfrak{B}(S,X)$ represent the set of bounded functions $f : S \to X$. Let $|x|$ denote the norm on X. If we define $\|f\| = \sup_{s\in S} |f(s)|$ then $\|f\|$ makes $\mathfrak{B}(S,X)$ into a normed vector space.*

PROOF. We first observe that $\mathfrak{B}(S,X)$ is a vector space. This follows from the fact that if f is bounded by C then for all $a \in \mathbb{R}$ we have af is bounded by $|a| C$ if both f and g are bounded by C_1 and C_2 respectively then using the triangle inequality in X we see that $f + g$ is bounded by $C_1 + C_2$.

Next we prove that we have defined a norm. The fact that $\|f\| \geq 0$ and $\|0\| = 0$ follow immediately from the definition and the fact that $|\cdot|$ is a norm on X. If $\|f\| = 0$ the it follows that $|f(s)| = 0$ for all $s \in S$ and therefore $f = 0$. Let $c \in \mathbb{R}$ then since $|cf(s)| = |c| |f(s)|$ it follows that $\|cf\| \leq |c| \|f\|$. On the other hand, let $\epsilon > 0$ be given then we may find an $s \in S$ such that $\|f\| - \epsilon < |f(s)|$. It follows that

$$|c| \|f\| - |c| \epsilon < |c| |f(s)| = |cf(s)|$$

Now ϵ was chosen arbitrarily so we may let $\epsilon \to 0$ and we get the inequality $|c| \|f\| \leq |cf(s)|$. Now take the supremum over $s \in S$ to get opposite inequality $|c| \|f\| \leq \|cf\|$ and it follows that $|c| \|f\| = \|cf\|$. The triangle inequality follows in a similar way. Given an f and g we see using the triangle inequality in X that for all $s \in S$ we have $|f(s) + g(s)| \leq |f(s)| + |g(s)| \leq \|f\| + \|g\|$; taking the supremum over $s \in S$ we get $\|f + g\| \leq \|f\| + \|g\|$. $\square$

Now we have the following extension result

LEMMA 15.112. *Let X be a normed vector space and let Y be a Banach space. Suppose that $V \subset X$ is a subspace and $A : V \to Y$ is a bounded linear map, then A has a unique extension $\overline{A} : \overline{V} \to Y$ from the closure of V into Y. Moreover if C is a bound on A the C is also a bound on $\overline{A}$.*

PROOF. TODO: $\square$

As is usual to compute with integrals it is imperative to connect the integration with differentiation. Since we are dealing with the Riemann integral we must use relatively strong hypotheses however these results will suffice for our applications and the proof are very simple. We start with the Fundamental Theorem of Calculus.

THEOREM 15.113 (Fundamental Theorem of Calculus). *Let X be a Banach space and let $f : [a,b] \to X$ be continuously differentiable then*

$$f(b) - f(a) = \int_a^b Df(t)dt$$

PROOF. First we let $g(t)$ be a regulated function from $[a,b]$ to X and consider the integral $\int_a^s g(t)\, dt$. Suppose that g is continuous at $c \in [a,b]$ and let $\epsilon > 0$ be given. By right continuity we may find $\delta > 0$ such that $|g(c+h) - g(c)| < \epsilon$ for all $|h| < \delta$. If we let $G(s) = \int_a^s g(t)\,dt$ then if $|h| < \delta$ we have

$$\begin{aligned}\left|\frac{G(c+h)-G(c)}{h} - g(c)\right| &= \left|\frac{1}{h}\int_c^{h+c} g(t)\,dt - \frac{1}{h}\int_c^{h+c} g(c)\,dt\right| \\ &= \left|\frac{1}{h}\int_c^{h+c} (g(t)-g(c))\,dt\right| \\ &\leq \frac{1}{|h|}\,|h| \sup_{c \leq s \leq c+h} |g(t)-g(c)| = \sup_{c \leq s \leq c+h} |g(t)-g(c)|\end{aligned}$$

By continuity TODO □

PROPOSITION 15.114. *Let X be a Banach space and let $f : [a,b] \to L(X,Y)$ be regulated then it follows that*

$$\left\|\int_a^b f(t)\,dt\right\| \leq \int_a^b \|f(t)\|\,dt$$

PROOF. TODO □

PROPOSITION 15.115. *Let X and Y be Banach spaces and let $f : [a,b] \to L(X,Y)$ be regulated then it follows that for every $x \in X$ we have*

$$\int_a^b f(t)x\,dt = \int_a^b f(t)\,dt \cdot x$$

PROOF. TODO □

DEFINITION 15.116. Let X and Y be Banach spaces then an *unbounded operator* is A a linear map of a subspace of X into Y. We let $\mathcal{D}(A)$ be the domain of A. We say that A is a *closed operator* if its graph is a closed linear subspace of $X \times Y$; that is to say if $v_n \to v$ in X and $Av_n \to w$ in Y then $v \in \mathcal{D}(A)$ and $y = Av_n$.

The domain of a linear operator is a crucial part of its definition and there is no small amount of pain in having to be careful about handling to be careful when dealing with these domains. In particular, if one is given two linear operators $A : X \to Y$ and $B : X \to Y$ then we can define $A + B : X \to Y$ where $\mathcal{D}(A+B) = \mathcal{D}(A) \cap \mathcal{D}(B)$. Given two linear operators $A : X \to Y$ and $B : Y \to Z$ then we can define $B \circ A : X \to Z$ where $\mathcal{D}(B \circ A) = \{v \in \mathcal{D}(A) \mid Av \in \mathcal{D}(B)\}$.

PROPOSITION 15.117. *Let X and Y be Banach spaces, let A be a closed linear operator from X to Y and let $f : [a,b] \to X$ be continuous such that*

(i) *$f(t) \in \mathcal{D}(A)$ for all $a \leq t \leq b$*
(ii) *$Af : [a,b] \to Y$ is continuous*

then it follows that $\int_a^b f(t)\,dt \in \mathcal{D}(A)$ and

$$A\int_a^b f(t)\,dt = \int_a^b Af(t)\,dt$$

PROOF. TODO □

11.2. Bochner Integrals. TODO: We develop Bochner integrals on σ-finite measure spaces; does it exist without that assumption? What if one uses nets for convergence of simple functions instead of sequences?

TODO: Preliminaries on norming subspaces.

TODO: The Bochner integral is the analogue of a vector valued integral using general measure theory. Though it can be developed in a bit more generality, we define it here for Banach space valued functions. The first thing to do is to observe that the defintion of simple functions extends trivially to this context.

DEFINITION 15.118. Given a set $(\Omega, \mathcal{A})$, a Banach space X sets $A_1, \dots, A_n \subset \Omega$ and $v_1, \dots, v_n \in X$ a linear combination $v_1 \mathbf{1}_{A_1} + \dots + v_n \mathbf{1}_{A_n}$ is called a *simple function*.

LEMMA 15.119. *A function $f : \Omega \to X$ is simple if and only it takes a finite number of values. A simple function is measurable if an only if $f^{-1}(v_j)$ is measurable for each of its distinct values $v_j \in X$.*

PROOF. The proof of Lemma 2.17 applies here with essentially no changes. $\square$

From this point on we will tacitly assume that all simple functions are measurable. Integrals of simple functions can be defined in the obvious way.

The natural way to proceed would be to observe that X can be given the Borel σ-algebra and thus we can talk of measurable functions. One might then try to show that all measurable functions can be approximated by simple functions and the integral can be extended by use of such approximations. In fact that is a little too much to hope for in a non-separable Banach space (not all measurable functions can be thus approximated) and we have to restrict ourselves to the class of functions that can be approximated by simple functions. We'll have to spend a bit of time understanding this class of functions.

DEFINITION 15.120. Let $(\Omega, \mathcal{A})$ be a measurable space and let X be a Banach space then a function $f : \Omega \to X$ is said to be *strongly measurable* if and only if there exist simple functions $f_n : \Omega \to X$ such that $f_n(\omega)$ converges to $f(\omega)$ for all $\omega \in \Omega$.

Intuitively strongly measurable functions must have some kind of countability restriction since they are a limit of a countable number of finite valued functions. This is indeed true and is at the heart of matter why can't approximate a general Borel measurable function with simple functions. To make this precise we need a couple of definitions before stating a theorem that provides a descriptive characterization of strongly measurable functions.

DEFINITION 15.121. Let Ω be a set, a function $f : \Omega \to X$ is *separably valued* if there exists a closed separable subspace $V \subset X$ such that $f(\Omega) \subset V$.

DEFINITION 15.122. Let $(\Omega, \mathcal{A})$ be a measurable space, a function $f : \Omega \to X$ is *weakly measurable* if for every $\lambda \in X^*$ the function $\lambda \circ f : \Omega \to \mathbb{R}$ is measurable.

THEOREM 15.123. *Let $(\Omega, \mathcal{A})$ be a measurable space and X be a Banach space then f is strongly measurable if and only if f is separably valued and weakly measurable. In fact it suffices to show that f is separably valued and $\lambda \circ f$ is measurable for λ in a norming subspace of E^*.*

PROOF. TODO $\square$

It is useful to note that one cannot approximate any more functions by using strongly measurable functions rather than just simple functions.

Corollary 15.124. *A pointwise limit of strongly measurable functions is strongly measurable.*

Proof. TODO □

We also have the following useful consequence that shows that a strongly measurable function is a separably valued Borel measurable function. In particular, all measurable functions with values in a separable Banach space are strongly measurable.

Now we introduce a σ-finte measure and consider the measure space $(\Omega, \mathcal{A}, \mu)$.

Definition 15.125. Given a set $(\Omega, \mathcal{A})$, a Banach space X, sets $A_1, \dots, A_n \in \mathcal{A}$ with $\mu(A_j) < \infty$ for $j = 1, \dots, n$ and $v_1, \dots, v_n \in X$ a linear combination $v_1 \mathbf{1}_{A_1} + \cdots + v_n \mathbf{1}_{A_n}$ is called a *μ-simple function*.

12. Differentiation in Banach Spaces

TODO:

- Absolute convergence of a infinite series in a Banach space
- Define space of linear maps with operator norm
- Show that Frechet deriviative is equal to Jacobian matrix on finite dimensional spaces

Proposition 15.126. *Let X be a Banach space then if $\sum_{j=0}^{\infty} a_j$ converges absolutely then $\sum_{j=0}^{\infty} a_j$ converges in X.*

Proof. By completeness of X it suffices to show that $S_n = \sum_{j=0}^{n} a_j$ is a Cauchy sequence. Let $\epsilon > 0$ be given and pick $n > 0$ such that $\sum_{j=n}^{\infty} \|a_j\| < \epsilon$. Then for all $m \geq n$ we have

$$\begin{aligned} \|S_m - S_n\| &\leq \left\| \sum_{j=n}^{m-1} a_j \right\| \\ &\leq \sum_{j=n}^{m-1} \|a_j\| \leq \sum_{j=n}^{infty} \|a_j\| < \epsilon \end{aligned}$$

and we are done. □

Proposition 15.127. *Let X be a Banach space. The set of invertible maps in $L(X)$ is open, moreover for any invertible map $A \in L(X)$ and any $\|A - B\| < \|A\|^{-1}$ we have*

$$B^{-1} = \sum_{n=0}^{\infty} A^{-n-1}(A-B)^n$$

In particular, the inversion map is continuously differentiable on its domain.

Proof. We first assume that $A = I$ is the identity map. If we let $\|B\| < 1$ then note that

$$\left\| \sum_{n=0}^{m} B^n \right\| \leq \sum_{n=0}^{m} \|B\|^n < \sum_{n=0}^{\infty} \|B\|^n = \frac{1}{1 - \|B\|} < \infty$$

which shows that $\sum_{n=0}^{\infty} B^n$ converges absolutely and is well defined in $L(X)$. Moreover we have

$$\left\|(1-B)\sum_{n=0}^{\infty} B^n\right\|$$

TODO: Finish.... □

We present some of the basic results on differentiation in Banach spaces.

DEFINITION 15.128. Let X and Y be Banach spaces, let $U \subset X$ be open and let $f : U \to Y$ be a map. We say that f is Frechet differentiable at $x \in U$ if there exists a bounded linear map $L : X \to Y$ such that

$$\lim_{h \to 0} \frac{f(x+h) - f(x) - Lh}{\|h\|} = 0$$

We call the linear map L the *Frechet derivative* of f at x and denote it $Df(x)$.

As it stand, we have been a little loose in defining *the* Frechet derivative as we have not ruled out the possibility that multiple linear maps may satisfy the defining property. The first task is to show that in fact the Frechet derivative is uniquely defined provided it exists.

PROPOSITION 15.129. *Suppose A and B are bounded linear maps satisfying the defining property of the Frechet derivative then $A = B$.*

PROOF. Let $\epsilon > 0$ be given and pick $\delta > 0$ so that we have $\|f(x+h) - f(x) - Ah\| < \epsilon \|h\|$ for $\|h\| < \delta$ and similarly for B. It then follows that

$$\|Ah - Bh\| \leq \|f(x+h) - f(x) - Ah\| + \|f(x+h) - f(x) - Bh\| < 2\epsilon \|h\|$$

so by linearity we see that $\|A - B\| < 2\epsilon$. Since $\epsilon > 0$ was arbitrary it follows that $\|A - B\| = 0$ and therefore $A = B$. □

There are weaker forms of derivative that one can consider. For the most part we shall be concerned with only the Frechet derivative but it can be helpful to be aware of the alternatives if for no other reason than to refine one's understanding of the nature of the Frechet derivative.

DEFINITION 15.130. Let X and Y be Banach spaces, let $U \subset X$ be open and let $f : U \to Y$ be a map. Let $v \in X$, the we say that f has a directional derivative at x in the direction of v if the limit

$$df(x, v) = \lim_{t \to 0} \frac{f(x+tv) - f(x)}{t}$$

exists. We say that f is *Gâteaux differentiable at x* if it has directional derivatives at all $v \in X$.

Note that some authors reserve the term Gâteaux differentiable for functions for which the directional derivatives are a linear form. Note that under any circumstances we have homogeneity. We will later show an example of a nonlinear Gâteaux derivative.

TODO: Example of nonlinear Gâteaux derivative

PROPOSITION 15.131. *Let X and Y be Banach spaces, let $U \subset X$ be open and let $f : U \to Y$ be a map such that f is Gâteaux differentiable at $x \in U$, then for all $c \in \mathbb{R}$ and $v \in X$, $df(x, cv) = c\,df(x, v)$.*

Proof. This follows by a simple change of variable in the limit

$$df(x,cv) = \lim_{t\to 0} \frac{f(x+tcv)-f(x)}{t} = \lim_{t\to 0} \frac{f(x+tv)-f(x)}{t/c} = c \lim_{t\to 0} \frac{f(x+tv)-f(x)}{t} = cdf(x,v)$$

□

Now we observe that Frechet differentiability implies Gâteaux differentiability.

Proposition 15.132. *Let X and Y be Banach spaces, let $U \subset X$ be open and let $f : U \to Y$ be a Fréchet differentiable at $x \in U$. Then f is Gâteaux differentiable at x and the directional derivative at v is equal to $Df(x)v$.*

Proof. Let $\epsilon > 0$ be given and pick $\delta > 0$ such that $\|f(x+h) - f(x) - Df(x)h\| \leq \epsilon \|h\|$ for all $\|h\| < \delta$. With $v \in X$ fixed and suppose that $\|v\| = 1$; we note that for all $|t| < \delta$ we have $\|f(x+tv) - f(x) - tDf(x)v\| \leq \epsilon |t|$ and thus

$$\left\| \frac{f(x+tv)-f(x)}{t} - Df(x)v \right\| < \epsilon$$

so that $df(x,v) = Df(x)v$. Now it is a simple matter to validate that $df(x,tv) = tdf(x,v) = Df(x) \cdot tv$ for all $t \in \mathbb{R}$. □

In general Gâteaux derivatives need not be linear (i.e. even though $df(x,tv) = tdf(x,v)$ it is not necessarily the case that $df(x,v+w) = df(x,v) + df(x,w)$) and even if linear need not be bounded. Somewhat more surprising is that even if the Gâteaux derivative exists and is bounded and linear the Fréchet derivative may not exist. What is necessary is that the limits $\lim_{t\to 0} \frac{f(x+tv)-f(x)}{t}$ converge uniformly for v in the unit sphere.

We calculate some trivial Frechet derivatives.

Example 15.133. Let $f : X \to Y$ be a constant map $f(x) = y$ for some fixed $y \in Y$, then f is differentiable at every point $x \in X$ and moreover $Df(x) = 0$.

Example 15.134. Let $A : X \to Y$ be a bounded linear map, then A is differentiable at every point $x \in X$ and moreover $Df(x) = A$.

The following example generalizes the product rule of calculus.

Example 15.135. Let $A : X_1 \times \cdots \times X_n \to Y$ be a bounded multilinear map, then A is differentiable at every point $x \in X_1 \times \cdots \times X_n$ and moreover

$$Df(x_1,\dots,x_n)(h_1,\dots,h_n) = A(h_1,x_2,\dots,x_n) + A(x_1,h_2,x_3,\dots,x_n) + \cdots + A(x_1,x_2,\dots,h_n)$$

Another important case is the behavior of deriviative when composing with a linear map.

Example 15.136. Let X, Y and Z be Banach spaces, let $U \subset X$ be open, let $f : U \to Y$ be differentiable and let $A : Y \to Z$ be a bounded linear map, then $D(A \circ f)(x) = A \circ Df(x)$.

Note that this would follow from the Chain Rule below (Proposition 15.137) but is worth showing this directly to get some practice with the definitions. Let $\epsilon > 0$ be given and pick $\delta > 0$ such that $\|f(x+h) - f(x) - Df(x)h\| \leq \frac{\epsilon}{\|A\|} \|h\|$ for all $\|h\| < \delta$. Note that

$$\|Af(x+h) - Af(x) - ADf(x)h\| \leq \|A\| \|f(x+h) - f(x) - Df(x)h\| \leq \epsilon \|h\|$$

for all $\|h\| < \delta$ which shows the result.

PROPOSITION 15.137. *Let X and Y be Banach spaces, let $U \subset X$ be open and let $f : U \to Y$ be differentiable at $x \in U$ then f is continuous at x.*

PROOF. Let $\epsilon > 0$ be given and define $0 < \delta < \frac{\epsilon}{1+\|Df(x)\|}$ small enough so that $\|f(x+h) - f(x) - Df(x)h\| \leq \|h\|$ for all $\|h\| < \delta$ then

$$\begin{aligned}\|f(x+h) - f(x)\| &\leq \|f(x+h) - f(x) - Df(x)h\| + \|Df\| \|h\| \\ &\leq \|h\| + \|Df\| \|h\| < \epsilon\end{aligned}$$

and continuity is proven. □

PROPOSITION 15.138 (Chain Rule). *Let X, Y and Z be Banach spaces, let $U \subset X$ and $f : U \to Y$ be differentiable at $x \in U$, let $V \subset Y$ with $f(U) \subset V$, $g : V \to Z$ be differentiable at $f(x)$ then $g \circ f : U \to Z$ is differentiable at x and moreover*

$$D(g \circ f)(x) = Dg(f(x)) \circ Df(x)$$

PROOF. Let ϵ be given. Let $\tilde{\delta} > 0$ be chosen so that $\|g(f(x)+h) - g(f(x)) - Dg(f(x))h\| < \frac{1}{2}\epsilon \|h\|$ for all $\|h\| < \tilde{\delta}$. By continuity of f at x we can choose $\delta_1 > 0$ such that $\|f(x+h) - f(x)\| < \tilde{\delta}$ for all $\|h\| < \delta_1$ and by differentiability we may choose $\delta_2 > 0$ such that $\|f(x+h) - f(x) - Df(x)h\| < \frac{\epsilon}{\epsilon + 2\|Dg(f(x))\|} \|h\|$ for all $\|h\| < \delta_2$. Let $\delta = \delta_1 \vee \delta_2$ and then for $\|h\| < \delta$ we compute

$$\begin{aligned}&\|g(f(x+h)) - g(f(x)) - Dg(f(x))Df(x)h\| \\ &\leq \|g(f(x+h)) - g(f(x)) - Dg(f(x))(f(x+h) - f(x))\| \\ &+ \|Dg(f(x))(f(x+h) - f(x)) - Dg(f(x))Df(x)h\| \\ &\leq \frac{1}{2}\epsilon \|f(x+h) - f(x)\| + \|Dg(f(x))\| \|f(x+h) - f(x) - Df(x)h\| \\ &\leq \frac{1}{2}\epsilon \|h\| + (\frac{\epsilon}{2} + \|Dg(f(x))\|) \frac{\epsilon}{\epsilon + 2\|Dg(f(x))\|} \|h\| = \epsilon \|h\|\end{aligned}$$

and we're done. □

12.1. Higher Order Derivatives and Taylor's Theorem.

THEOREM 15.139 (Mean Value Theorem). *Let X and Y be Banach spaces, $U \subset X$ be open and let $f : U \to Y$ be continuously differentiable. Suppose $x \in U$ and $y \in X$ such that $x + ty \in U$ for all $0 \leq t \leq 1$ then*

$$f(x+y) - f(x) = \int_0^1 Df(x+ty)y\,dt = \int_0^1 Df(x+ty)dt \cdot y$$

PROOF. Define $g(t) = f(x+ty)$. Then by the Chain Rule it follows that $g(t)$ is continuously differentiable and $Dg(t) = Df(x+ty)y$. Since $Dg(t)$ is continuous we may apply the Fundamental Theorem of Calculus (Theorem 15.112) to conclude that

$$f(x+y) - f(x) = g(1) - g(0) = \int_0^1 Df(x+ty)y\,dt = \int_0^1 Df(x+ty)dt \cdot y$$

where in the last inequality we have use Proposition 15.114. □

Higher order derivatives are defined by iterating Frechet derivatives. For example if we assume that the map $f : U \to Y$ differentiable on all of U then the second derivative is obtained by taking the derivative of the map $Df : U \to L(X, Y)$ whereever it exists. Thus the second derivative is a map $D^2 f : U \to L(X, L(X, Y))$.

EXAMPLE 15.140. Let $A : X \to Y$ be a bounded linear map then $D^2 A = 0$.

Based on the definition via induction we think of $D^n f$ as a map from U to $L(X, \cdots, L(X, Y) \cdots)$. The range here actually has a more convenient representation as the space of multilinear maps $X \times \cdots \times X \to Y$. For example given an element in $f \in L(X, L(X, Y))$ we may define $\tilde{f}(u, v) = f(u)v$ and note that

$$\tilde{f}(au + bv, w) = f(au + bv)w = af(u)w + bf(v)w = a\tilde{f}(u, w) + b\tilde{f}(v, w)$$

and

$$\tilde{f}(u, av + bw) = f(u)(av + bw) = af(u)v + bf(u)w = a\tilde{f}(u, v) + b\tilde{f}(u, w)$$

so that $\tilde{f}$ is indeed bilinear. It is easy to see that this is an isomorphism and that the construction extends to general n.

TODO: Do this in the required excruciating detail...

In the sequel, it will be convenient to view higher derivatives as maps from U to the space of multilinear maps. It turns out that higher derivatives are not arbitrary multilinear maps but also have the property of being symmetric.

PROPOSITION 15.141. *Let $U \subset X$ be open and $f : U \to Y$ be C^p then $D^p f(x)$ is multilinear and symmetric for every $x \in U$.*

PROOF. TODO: We first consider the case $p = 2$. Let $u, v \in X$ and consider

$$D^2 f(x)(u, v) =$$

$\square$

A more complicated but important example is the computation of the derivative of the inverse in a Banach algebra.

PROPOSITION 15.142. *The map $\phi(A) = A^{-1}$ on $L(X, X)$ is C^∞ on the open set of invertible maps. In fact we have*

$$D^n \phi(A)(h_1, \dots, h_n) = (-1)^n \sum_\sigma A^{-1} h_{\sigma_1} A^{-1} \cdots h_{\sigma_n} A^{-1}$$

where the summation is over all permutations of $\{1, \dots, n\}$.

PROOF. We first compute the first derivative of ϕ. Let A be invertible and observe that for $\|h\| < \left\|A^{-1}\right\|^{-1}$ we know that $I + A^{-1}h$ is invertible and moreover

$$(A + h)^{-1} = A^{-1}(I + hA^{-1})^{-1} = A^{-1} \sum_{n=0}^\infty (-1)^n h^n A^{-n}$$

and therefore using the absolute convergence of the series on the right we get

$$\begin{aligned}
\left\|(A + h)^{-1} - A^{-1} + A^{-1}hA^{-1}\right\| &\leq \sum_{n=2}^\infty \|h\|^n \left\|A^{-1}\right\|^n \\
&= \frac{\|h\|^2 \left\|A^{-1}\right\|^2}{1 - \|h\| \left\|A^{-1}\right\|} < \|h\|^2 \left\|A^{-1}\right\|^2
\end{aligned}$$

which shows us that $D\phi(A)h = -A^{-1}hA^{-1}$ (for $\epsilon > 0$ let $\delta < \epsilon \left\|A^{-1}\right\|^{-2}$).

Now to see that ϕ is in fact C^∞, we do an induction. TODO: Finish ∎

With the defintion of higher derivatives available we are now able to extend the Mean Value Theorem to Taylor's Theorem in Banach spaces.

THEOREM 15.143 (Taylor's Theorem). *Let X and Y be Banach spaces and let $U \subset X$ be open and of class C^p. Suppose that $x \in U$ and $y \in X$ such that $x+ty \in U$ for all $0 \leq t \leq 1$ then we have*

$$f(x+y) = f(x) + Df(x)y + \cdots + \frac{D^{p-1}f(x)y^{(p-1)}}{(p-1)!} + \int_0^1 \frac{(1-t)^{p-1}}{(p-1)!} D^p f(x+ty)y^{(p)} dt$$

where $y^{(k)} = (y, \dots, y) \in X^k$.

PROOF. TODO ∎

It is worth noting that in the case $Y = \mathbb{R}$ that Theorem 15.142 can be proven using the one dimensional version Theorem 1.20 and the chain rule Proposition 15.137. We'll show this in the proof of the Lagrange form of the remainder term below.

We've presented Taylor's Theorem in Banach spaces with the integral form of the remainder term. There are several different versions of the remainder and estimates derived therefrom that are useful to note. The first that we mention is applicable in the important case in which $Y = \mathbb{R}$; the Lagrange form of the remainder.

PROPOSITION 15.144. *There is a number $c \in (0,1)$ such that*

$$\int_0^1 \frac{(1-t)^{p-1}}{(p-1)!} D^p f(x+ty)y^{(p)} dt = \frac{D^p f(x+cy)y^{(p)}}{p!}$$

PROOF. We derive this from the one dimensional Taylor's Theorem. Note that $g(t) = f(x+ty)$ is C^p from $[0,1]$ to $\mathbb{R}$ and by the chain rule we have $g'(t) = Df(x+ty)y$. Now since evaluation $A \to Ay$ is a bounded linear map on $L(X,Y)$, an induction argument using either Example 15.135 or the chain rule shows that $g^{(k)}(t) = D^k f(x+ty)y^{(k)}$. Now apply Theorem 1.20 to see there is a $0 < c < 1$ such that

$$\begin{aligned} f(x+y) = g(1) &= g(0) + g'(0) + \cdots + \frac{g^{(p-1)}(0)}{(p-1)!} + \frac{g^{(p)}(c)}{p!} \\ &= f(x) + Df(x)y + \cdots + \frac{D^{p-1}f(x)y^{(p-1)}}{(p-1)!} + \frac{D^p f(x+cy)y^{(p)}}{p!} \end{aligned}$$

∎

TODO: Analytic functional calculus and beyond.

Frechet derivatives of functions that map into product spaces are easily computed.

PROPOSITION 15.145. *Let $f : U \to Y \times Z$ be a function then f is Frechet differentiable at $x \in U$ if and only if the coordinate functions $f_1 : U \to Y$ and $f_2 : U \to Z$ are Frechet differentiable at $x \in U$ and in this case $Df(x)v = (Df_1(x)v, Df_2(x)v)$.*

PROOF. Let π_1 and π_2 be the projections from $Y \times Z$ to Y and Z respectively. Suppose that f is Frechet differentiable at x then by the Chain Rule and the linearity of projections π_i we know that $f_i = \pi_i \circ f$ is Frechet differentiabe and $Df_i = \pi_i \circ Df$.

Suppose that $Df_1(x)$ and $Df_2(x)$ both exist. Then for any $\epsilon > 0$ there exists $\delta > 0$ such that $\|f_i(x+h) - f_i(x) - Df_i(x)h\| < \epsilon \|h\| / \sqrt{2}$ for any $\|h\| < \delta$ and $i = 1, 2$. The derivative of f follows from

$$\begin{aligned} \|f(x+h) - f(x) - (Df_1(x)h, Df_2(x)h)\| &\leq \Big(\|f_1(x+h) - f_1(x) - Df_1(x)h\|^2 + \|f_2(x+h) - f_2(x) - Df_2(x)h\|^2 \\ &\leq \epsilon \|h\|^{-1} \end{aligned}$$

□

There is an generalization of the notion of partial derivatives to the Frechet differentiation on Banach spaces. Recall that given Banach spaces $X_1, X_2, \ldots, X_n$ the cartesian product $X_1 \times \cdots \times X_n$ with norm $\|(v_1, \ldots, v_n)\| = \|v_1\| \vee \cdots \vee \|v_n\|$ is a Banach space.

DEFINITION 15.146. Let $Y, X_1, X_2, \ldots, X_n$ be Banach spaces and let $U_i \subset X_i$ be open for $i = 1, \ldots, n$ and $f : U_1 \times \cdots U_n \to Y$ be a map. Let $v = (v_1, \ldots, v_n) \in U_1 \times \cdots U_n$ then we say that f has a partial derivative with respect to i at v if there exists a bounded linear map $L_i : X_i \to Y$ such that

$$\lim_{h \to 0} \frac{f(v_1, \ldots, v_{i-1}, v_i + h, v_{i+1}, \ldots, v_n) - f(v_1, \ldots, v_n) - L_i h}{\|h\|} = 0$$

We say that L_i is the i^{th} partial derivative and is denoted $D_i f(v)$.

In general the existence of partial derivatives does not guarantee the existence of total Frechet deriviative however in the presence of continuity it does.

PROPOSITION 15.147. *Let $Y, X_1, X_2, \ldots, X_n$ be Banach spaces and let $U_i \subset X_i$ be open for $i = 1, \ldots, n$ and $f : U_1 \times \cdots \times U_n \to Y$ be a map then f is continuously differentiable on $U_1 \times \cdots \times U_n$ if and only if $D_i f : U_1 \times \cdots \times U_n \to L(X_i, Y)$ exists and is continuous for each $i = 1, \ldots, n$. In this case,*

$$Df(v)h = \sum_{i=1}^{n} D_i f(v) h_i$$

PROOF. Suppose that continuous partial derivatives exist, let $v = (v_1, \ldots, v_n) \in U_1 \times \cdots \times U_n$ and let $h = (h_1, \ldots, h_n) \in X_1 \times \cdots \times X_n$. For each $i = 0, \ldots, n$ let h^i be obtained by h by setting the first i coordinates to 0; in particular $h^0 = h$ and $h^n = 0$ and for $i = 1, \ldots, n$ $h^{i-1} - h^i$ is non-zero only in the i^{th} coordinate and that coordinate is h_i. Writing a telescoping sum and using the Mean Value Theorem 1

$$\begin{aligned} f(v+h)-f(v) &= \sum_{i=1}^{n}(f(v+h^{i-1})-f(v+h^i)) = \sum_{i=0}^{n-1}\int_0^1 D_if(v+h^i+t(h^{i-1}-h^i))h_i\,dt \\ &= \sum_{i=0}^{n-1}\int_0^1 D_if(v)h_i\,dt + \int_0^1\sum_{i=0}^{n-1}\int_0^1 (D_if(v+h^i+t(h^{i-1}-h^i))h_i - D_if(v)h_i)\,dt \\ &= \sum_{i=0}^{n-1} D_if(v)h_i + \int_0^1\sum_{i=0}^{n-1}\int_0^1 (D_if(v+h^i+t(h^{i-1}-h^i))h_i - D_if(v)h_i)\,dt \end{aligned}$$

It suffices to show that each term $\int_0^1 (D_if(v+h^i+t(h^{i-1}-h^i))h_i - D_if(v)h_i)\,dt$ is $o(h)$. To see this recall $\|h_i\| \leq \|h_1\| \vee \cdots \|v_n\| = \|v\|$ and thus we have the simple bound

$$\begin{aligned} &\frac{\left\|\int_0^1 (D_if(v+h^i+t(h^{i-1}-h^i))h_i - D_if(v)h_i)\,dt\right\|}{h} \\ &\leq \frac{\left\|\int_0^1 (D_if(v+h^i+t(h^{i-1}-h^i))h_i - D_if(v)h_i)\,dt\right\|}{h_i} \\ &\leq \frac{\int_0^1 \left\|D_if(v+h^i+t(h^{i-1}-h^i))h_i - D_if(v)h_i\right\|\,dt}{h_i} \\ &\leq \frac{\|h_i\| \sup_{0\leq t\leq 1}\left\|D_if(v+h^i+t(h^{i-1}-h^i)) - D_if(v)\right\|}{h_i} \\ &= \sup_{0\leq t\leq 1}\left\|D_if(v+h^i+t(h^{i-1}-h^i)) - D_if(v)\right\| \end{aligned}$$

Since D_if is continuous at v, for every $\epsilon > 0$ we may find a $\delta > 0$ such that $\|D_if(v+u) - D_if(v)\| < \epsilon$ for every $\|u\| < \delta$. Thus if $\|h\| < \delta$ then $\left\|h^i + t(h^{i-1}-h^i)\right\| = |t|\,\|h_i\| \leq \|h\| < \delta$ and therefore $\sup_{0\leq t\leq 1}\left\|D_if(v+h^i+t(h^{i-1}-h^i)) - D_if(v)\right\| \leq \epsilon$. Thus

$$\begin{aligned} &\lim_{h\to 0}\frac{\left\|\int_0^1 (D_if(v+h^i+t(h^{i-1}-h^i))h_i - D_if(v)h_i)\,dt\right\|}{h} \\ &\leq \lim_{h\to 0}\sup_{0\leq t\leq 1}\left\|D_if(v+h^i+t(h^{i-1}-h^i)) - D_if(v)\right\| = 0 \end{aligned}$$

To see the other direction suppose that Df exists and is continuous on $U_1 \times \cdots \times U_n$. Let $\epsilon > 0$ be given. Since $Df(v)$ exists we may find $\delta > 0$ so that $\|h\| < \delta$ we have $\|f(v+h) - f(v) - Df(v)h\| < \|h\|\,\epsilon$. Let $v = (v_1, \ldots, v_n) \in U_1 \times \cdots \times U_n$ and $h \in X_i$ and write h_i be vector in $U_1 \times \cdots \times U_n$ with h the i^{th} coordinate and 0 in the others. Note that $\|h\| = \|h_i\|$. Thus for all $\|h\| < \delta$

$$\begin{aligned} &\|f(v_1, \ldots, v_{i-1}, v_i + h, v_{i+1}, \ldots, v_n) - f(v_1, \ldots, v_n) - Df(v)h_i\| \\ &= \|f(v+h_i) - f(v) - Df(v)h_i\| \leq \|h_i\|\,\epsilon = \|h\|\,\epsilon \end{aligned}$$

which shows that $D_if(v)$ exists and $D_if(v)h = D_if(v)h_i$. Continuity of D_if follows from the continuity of Df and the continuity $\square$

12.2. Inverse and Implicit Function Theorems.

THEOREM 15.148. *Let X and Y be Banach spaces let $U \subset X$ be an open subset of X and suppose that $f : U \to Y$ is continuously differentiable and $Df(x)$ is invertible at $x \in U$. There is an open set $V \subset U$ containing x and an open set $W \subset Y$ containing $f(x)$ such that $f : V \to W$ is a bijection and f^{-1} is continuously differentialble on W.*

We'll be a bit redundant a provide both the general proof in Banach spaces as well as a proof for the finite dimensional case that is more verbose but is very elementary.

For the finite dimensional proof we use the following simple consequence of the mean value theorem that shows a continuously differentiable function is Lipschitz continuous on a bounded domain.

LEMMA 15.149. *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be differentiable on an open rectangle $R = (a_1, b_1) \times \cdots \times (a_n, b_n)$ such that*

$$\left| \frac{\partial f_i}{\partial x_j}(x) \right| \leq M$$

for all $1 \leq i, j \leq n$ and $x \in R$ then it follows that $\|f(y) - f(x)\| \leq M \cdot n^2 \cdot \|y - x\|$ for all $x, y \in R$.

PROOF. By expanding as a telescoping sum and the one dimensional mean value theorem we get for every $i = 1, \ldots, n$

$$\begin{aligned} f_i(y) - f_i(x) &= \sum_{j=1}^n f_i(y_1, \ldots, y_j, x_{j+1}, \ldots, x_n) - f_i(y_1, \ldots, y_{j-1}, x_j, \ldots, x_n) \\ &= \sum_{j=1}^n \frac{\partial f_i}{\partial x_j}(y_1, \ldots, y_{j-1}, y_j^*, x_{j+1}, \ldots, x_n)(y_j - x_j) \end{aligned}$$

where $a_j < y_j^* < b_j$ (in fact $x_j \leq y_j^* \leq y_j$ when $x_j \leq y_j$ and similarly when $y_j < x_j$). Now by the triangle inequality and the bound on partials of f we get

$$\begin{aligned} \|f(y) - f(x)\| &\leq \sum_{i=1}^n |f_i(y) - f_i(x)| \\ &= \sum_{i=1}^n \sum_{j=1}^n \left| \frac{\partial f_i}{\partial x_j}(y_1, \ldots, y_{j-1}, y_j^*, x_{j+1}, \ldots, x_n) \right| |y_j - x_j| \\ &\leq \sum_{i=1}^n \sum_{j=1}^n M \|y - x\| = M \cdot n^2 \cdot \|y - x\| \end{aligned}$$

□

Now we can proceed with the proof of the theorem in the finite dimensional case.

PROOF. We first make a reduction to the case in which $Df(x)$ is the identity. If the result is proven in that case then for general f we can define $Df(x)^{-1} \circ f : X \to X$ where from the Chain Rule it follows that $D(Df(x)^{-1} \circ f)(x)$ is the identity. Applying the inverse function theorem we see there exists open sets V

and $\tilde{W}$ containing x and $Df(x)^{-1}f(x)$ respectively such $Df(x)^{-1} \circ f$ is a bijection from V to $\tilde{W}$ with $(Df(x)^{-1} \circ f)^{-1}$ continuously differentiable. Now we define $W = Df(x)(\tilde{W})$ which is open by continuity of $Df(x)^{-1}$ and contains $f(x)$. Since $f^{-1} = Df(x) \circ Df(x)^{-1} \circ f$ it follows by the Chain Rule that f^{-1} is continuously differentiable on W.

CLAIM 15.149.1. There is an open ball $B(x, \delta) \subset U$ such that f is injective on the closure of $B(x, \delta)$, $Df(y)$ is invertible for all $y \in B(x, \delta)$ and

$$\left| \frac{\partial f_i}{\partial x}(y) - \frac{\partial f_i}{\partial x}(x) \right| < \frac{1}{2n^2} \text{ for all } 1 \leq i, j \leq n \text{ and } y \in B(x, \delta) \tag{27}$$

By the the openness of U, triangle inequality and the fact that $Df(x)$ is the identity we know that we can find $\delta > 0$ such that $B(x, \delta) \subset U$ and

$$\begin{aligned} \|f(x+h) - f(x)\| &= \|f(x+h) - f(x) - h + h\| \geq \|h\| - \|f(x+h) - f(x) - h\| \\ &\geq \frac{1}{2}\|h\| \end{aligned}$$

so injectivity on $B(x, \delta)$ follows; by continuity of f the bound and hence the injectivity extends to the closure of $B(x, \delta)$. Since the invertible linear maps are an open subset of $L(X, Y)$ and Df is continuous we may also assume that $\delta > 0$ is chosen so that Df is invertible on $B(x, \delta)$. Similarly continuity of Df implies the continuity of each partial derivative $\frac{\partial f_i}{\partial x}$ and therefore (15.148.1) follows for sufficiently small $\delta > 0$.

The next claim should be thought of as asserting that the inverse of f is Lipschitz. As it turns out the estimate is useful in showing that f^{-1} exists.

CLAIM 15.149.2. $\|y - z\| \leq 2\|f(y) - f(z)\|$ for all $y, z \in B(x, \delta)$.

Define $g(x) = f(x) - x$ on $B(x, \delta)$. Because $Df(x)$ is the identity we know that $\frac{\partial f_i}{\partial x}(x) = \delta_{ij}$ and therefore

$$\left| \frac{\partial g_i}{\partial x}(y) \right| = \left| \frac{\partial f_i}{\partial x}(y) - \frac{\partial f_i}{\partial x}(x) \right| \leq \frac{1}{2n^2}$$

Since y and z are contained in some open rectangle that is a subset of $B(x, \delta)$ we can apply Lemma 15.148 and the triangle inequality to conclude that

$$\begin{aligned} \frac{1}{2}\|y - z\| &\geq \|g(y) - g(z)\| \\ &= \|f(y) - y - g(z) + z\| \\ &\geq \|y - z\| - \|f(y) - g(z)\| \end{aligned}$$

and the claim follows by collecting terms.

The next step in the proof is to validate that the image of $B(x, \delta)$ under f contains an open set (on which we will then have a bijection). Consider the function $f(y) - f(x)$. It is continuous and by compactness of the boundary $\partial B(x, \delta)$ and the injectivity of f on the closed ball we know that there exists an $\epsilon > 0$ such that $g(y) \geq \epsilon$ on $\partial B(x, \delta)$. Define $W = B(f(x), \epsilon/2)$ and notice that by the choice of ϵ, the triangle inequality and the previous claim we have for all $z \in W$ and

$y \in \partial B(x, \delta)$

$$\|z - f(y)\| \geq \|f(x) - f(y)\| - \|f(x) - z\| \geq d - \frac{d}{2} \tag{28}$$

$$= \frac{d}{2} > \|z - f(x)\| \tag{29}$$

This estimate is used to construct an open set in the image of f.

CLAIM 15.149.3. For every $z \in W$ there is a unique $y \in B(x, \delta)$ such that $f(y) = z$.

To see existence we let $z \in W$ be given and we define the function $h(y) = \|f(y) - z\|^2 = \sum_{j=1}^{n}(f_j(y) - z_j)^2$. Differentiability of h follows from the differentiability of f and the chain rule. In particular, h is continuous and therefore by compactness of the closed ball $\overline{B}(x, \delta)$ it attains its minimum. By the estimate (12.2) we see that the minimum must occur in the interior of the ball. Therefore we know that the derivative of h must vanish at the minimum so by the Chain Rule we know that for all $v \in X$

$$0 = D\|f - z\|^2(y) \cdot v = 2\|Df(y) \cdot v\| f(y) - z$$

and by the invertibility of $Df(y)$ it follows that we must have $f(y) = z$ at the minimum.

The uniqueness of y follows from the injectivity of f.

Now we define $V = f^{-1}(W) \cap B(x, \delta)$ and it follows that f is a bijection from V to W. Now that f^{-1} is well defined we immediately get its continuity.

CLAIM 15.149.4. f^{-1} is continuous on W.

The second claim proved the bound $\|y - z\| \leq 2\|f(y) - f(z)\|$ on $B(x, \delta)$ which certainly shows that $\left\|f^{-1}(z) - f^{-1}(w)\right\| \leq 2\|z - w\|$ on W so that f^{-1} is Lipschitz in particular continuous.

It remains to show that f^{-1} is differentiable.

CLAIM 15.149.5. f^{-1} is continuously differentiable on W.

In fact we show (as would follow from the Chain Rule) that $Df^{-1}(z) = \left[Df(f^{-1}(z))\right]^{-1}$ for all $z \in W$. Note that this is well defined since Df is invertible on all of V. To clean up the notation a bit let $A = Df(f^{-1}(z))$. Let $\epsilon > 0$ be given. Using differentiability of f at $f^{-1}(z)$ we choose $\tilde{\eta} > 0$ such that

$$\left\|f(f^{-1}(z) + h) - f(f^{-1}(z)) - Ah\right\| < \frac{\epsilon}{2\|A^{-1}\|}\|h\| \text{ for all } \|h\| < \tilde{\eta}$$

By continuity of f^{-1} at z we choose $\eta > 0$ such that $\left\|f^{-1}(z + h) - f^{-1}(z)\right\| < \tilde{\eta}$ for all $\|h\| < \eta$. Pick $h \in Y$ with $\|h\| < \eta$ and compute using the Lipschitz continuity of f^{-1}

$$\begin{aligned}
&\left\|f^{-1}(z + h) - f^{-1}(z) - A^{-1}h\right\| \\
&= \left\|A^{-1}\left(f(f^{-1}(z + h)) - f(f^{-1}(z)) - A(f^{-1}(z + h) - f^{-1}(z))\right)\right\| \\
&\leq \left\|A^{-1}\right\|\left\|f(f^{-1}(z + h)) - f(f^{-1}(z)) - A(f^{-1}(z + h) - f^{-1}(z))\right\| \\
&\leq \frac{1}{2}\epsilon\left\|f^{-1}(z + h) - f^{-1}(z)\right\| \leq \epsilon\|h\|
\end{aligned}$$

which gives us $Df^{-1}(z) = \left[Df(f^{-1}(z))\right]^{-1}$. Continuity of Df^{-1} follows from the continuity of Df, continuity of f^{-1} and continuity of inversion of invertible maps in $L(X,Y)$. $\square$

The proof of the Inverse Function Theorem in general Banach spaces rests on a simple result that is of broad applicability. The result actually doesn't use the vector space structure and is valid in general complete metric spaces; it provides a very general mechanism for solving equations in such spaces.

PROPOSITION 15.150 (Contraction Mapping Principle). *Let (S,d) be a complete metric space space, let $F \subset S$ be a closed subset and let $g : F \to F$ be a mapping such that there exists a constant $0 < K < 1$ such that*

$$d(g(x), g(y)) \leq K d(x,y) \text{ for all } x, y \in F$$

then there exists a unique $x_0 \in F$ such that $g(x_0) = x_0$ and moreover given any $x \in F$ the sequence $\{g^n(x)\}$ is Cauchy and converges to x_0.

PROOF. First we prove uniqueness. Suppose there are two points x and y satisfying $g(x) = x$ and $g(y) = y$ then we know that

$$d(x,y) = d(g(x), g(y)) \leq K d(x,y)$$

and since $0 < K < 1$ this shows that $d(x,y) = 0$.

Now we let $x \in F$ be arbitrary and show that $g^n(x)$ is Cauchy. Suppose that $m > n$ and observe that by a simple induction

$$d(g^n(x), g^m(x)) \leq K^n d(x, g^{m-n}(x))$$

In particular, we have that $d(g^n(x), g^{n+1}(x)) \leq K^n d(x, g(x))$ and therefore by the triangle inequality

$$d(x, g^n(x)) \leq d(x, g(x)) + \cdots + d(g^{n-1}(x), g^n(x)) \leq (1 + \cdots + K^{n-1}) d(x, g(x)) < \frac{d(x, g(x))}{1-K}$$

Putting these two bounds together we see that $d(g^n(x), g^m(x)) \leq \frac{K^n d(x,g(x))}{1-K}$ hence $g^n(x)$ is Cauchy.

Since F is a closed subset of a complete metric space, it is complete and we know that $g^n(x)$ converges to some $x_0 \in F$; it remains to show that x_0 is a fixed point of g. Let $\epsilon > 0$ be given and chose $N > 0$ such that $d(x_0, g^n(x)) < \epsilon$ for all $n \geq N$. Then we know that

$$d(g(x_0), g^n(x)) \leq K d(x_0, g^{n-1}(x)) \leq K\epsilon < \epsilon$$

for all $n \geq N + 1$ which shows that $g^n(x)$ converges to $g(x_0)$. It follows that $x_0 = g(x_0)$. $\square$

PROOF. The first step in the proof of the Inverse Function Theorem is the construction of the inverse. In order to use the contraction mapping principle to build the inverse we have to formulate a fixed point problem. A little bit of experimentation witht the case $X = Y$ will suggest that if we let $y \in Y$ and define $g_y(x) = x + f(x) - y$ then finding a fixed point $g_y(x) = x$ is equivalent to $f(x) = y$. This idea is on target but doesn't yet work. Obviously we have the restriction of the assumption $X = Y$ to lift but deeper is the observation that g_y as defined above is not a contraction mapping in general. If we calculate the Frechet derivative we we get $Dg_y(x_0) = \mathrm{Id} + Df(x_0)$ which can be arbitrarily expansive (consider the example of $X = Y = \mathbb{R}$, $x_0 = y_0 = 0$ and $f(x) = ax$ for a large constant a).

The trick is to use the inverse of $Df(x_0)$ in the fixed point formulation in order to counteract the expansiveness of f itself. Doing this properly we also create a fixed point problem that makes sense in the case that $X \neq Y$. Specifically, for each $y \in Y$ define

$$g_y(x) = Df(x_0)^{-1}(Df(x_0) \cdot x + f(x) - y) = x + Df(x_0)^{-1} f(x) - Df(x_0)^{-1} y$$

then a simple calculation shows that $g_y(x) = x$ if and only if $f(x) = y$ (note that the use of $Df(x_0)^{-1}$ in this way is equivalent to redefining f as $Df(x_0)^{-1} \circ f$ and thereby reducing to the case in which $Df(x_0) = \mathrm{Id}$). As a quick check that we are on the right track now note that

$$g_{y_0}(x_0) = x_0 + Df(x_0)^{-1} f(x_0) - Df(x_0)^{-1} y_0 = x_0 + Df(x_0)^{-1} y_0 - Df(x_0)^{-1} y_0 = x_0$$

as we should expect.

Claim 15.150.1. There exists an $r > 0$ such that for all $y \in \overline{B}(y_0, r/2 \left\| Df(x_0)^{-1} \right\|)$ there exists a unique $x \in \overline{B}(x_0, r)$ such that $f(x) = y$.

Define g_y as above and we will show that for y in a neighborhood of y_0 the mapping g_y is a contraction on a neighborhood of x_0. First note that $Dg_y(x_0) = 0$ and apply continuity of Df to pick an $r > 0$ such that $\|Dg_y(x)\| \leq \frac{1}{2}$ for all $x \in \overline{B}(x_0, 2r)$. By the Mean Value Theorem 1 and Proposition 15.113 we know that for every $x \in \overline{B}(x_0, r)$ we have

$$\|g_y(x) - g_y(x_0)\| = \left\| \int_0^1 Dg_y(x_0 + t(x - x_0))(x - x_0)\, dt \right\| \leq \frac{1}{2} \|x - x_0\|$$

From this we compute for $y \in \overline{B}(y_0, r/2 \left\| Df(x_0)^{-1} \right\|)$ and $x \in \overline{B}(x_0, r)$

$$\|g_y(x) - x_0\| \leq \|g_y(x) - g_{y_0}(x)\| + \|g_{y_0}(x) - g_{y_0}(x_0)\| \quad = \left\| Df(x_0)^{-1} \right\| \|y - y_0\| + \frac{1}{2} \|x - x_0\| \leq r$$

which shows that g_y is a mapping from $\overline{B}(x_0, r)$ to itself. To see that g_y is a contraction mapping let $x_1, x_2 \in \overline{B}(x_0, r)$ and note that for all $0 \leq t \leq 1$,

$$\|x_1 + t(x_2 - x_1) - x_0\| = \|(1 - t)x_1 + tx_2 - (1 - t)x_0 - tx_0\| \leq (1 - t) \|x_1 - x_0\| + t \|x_2 - x_0\| \leq r$$

so another application of the Mean Value Theorem 1, Proposition 15.113 and the bounds on $Df(x)$ on $\overline{B}(x_0, r)$ yields

$$\begin{aligned} \|g_y(x_2) - g_y(x_1)\| &= \left\| \int_0^1 Dg_y(x_1 + t(x_2 - x_1))(x_2 - x_1)\, dt \right\| \\ &\leq \int_0^1 \|Dg_y(x_1 + t(x_2 - x_1))(x_2 - x_1)\|\, dt \leq \frac{1}{2} \|x_2 - x_1\| \end{aligned}$$

Since a closed subset of a Banach space is a complete metric space, we may apply the Contraction Mapping Principle Proposition 15.149 to finish the proof of the claim.

Pick $r > 0$ as in the claim and define $U_1 = B(x_0, r) \cap f^{-1}(B(y_0, r/2 \left\| Df(x_0)^{-1} \right\|)$. We know from continuity of f that U_1 is open. Let $V_1 = f(U_1)$. We know that f is an injective map from U_1 to V_1 and therefore way define the inverse map $\phi : V_1 \to U_1$. Note that the prior claim does not actually show that $V_1 = B(y_0, r/2 \left\| Df(x_0)^{-1} \right\|)$ since that claim leaves open the possibility that f may map a point on the boundary of $\overline{B}(x_0, r)$ to a point in the interior of $\overline{B}(y_0, r/2 \left\| Df(x_0)^{-1} \right\|)$;

hence we don't know that V_1 is an open set. We prove that and more in the next claim.

CLAIM 15.150.2. V_1 is an open set and ϕ is a continuous map.

To prove the claim, let $x_1 \in U_1$ and define $y_1 = f(x_1)$ so that $y_1 \in B(y_0, r/2 \left\|Df(x_0)^{-1}\right\|)$. Pick another $y_2 \in B(y_0, r/2 \left\|Df(x_0)^{-1}\right\|)$ and apply the previous claim to construct the unique $x_2 \in \overline{B}(x_0, r)$ such that $f(x_2) = y_2$. Using the triangle inequality

$$\begin{aligned}\|x_2 - x_1\| &\le \|f(x_2) - f(x_1)\| + \|x_2 - f(x_2) - (x_1 - f(x_1))\| \\ &= \|f(x_2) - f(x_1)\| + \|g_{y_0}(x_2) - g_{y_0}(x_1)\| \\ &\le \|f(x_2) - f(x_1)\| + \frac{1}{2}\|x_2 - x_1\|\end{aligned}$$

and therefore $\|x_2 - x_1\| \le 2\|f(x_2) - f(x_1)\|$. From this inequality we see that V_1 is open; given $y_1 \in V_1$ if we pick an $\epsilon > 0$ such that $B(\phi(y_1), \epsilon) \subset U_1$ then for any $\|y_2 - y_1\| < \epsilon/2$ we have $\|x_2 - x_1\| < \epsilon$ for the x_2 with $f(x_2) = y_2$ which implies $x_2 \in U_1$ and therefore $y_2 \in V_1$. The continuity of ϕ also follows from this inequality by essentially the same argument. Given $y, y_1, y_2, \ldots \in V_1$ with $\lim_{n\to\infty} y_n = y$ let $x = \phi(y)$ $(f(x) = y)$ and $x_n = \phi(y_n)$ $(f(x_n) = y_n)$. For any $\epsilon > 0$ small enough that $B(\phi(y), \epsilon) \subset U_1$ we pick $N > 0$ such that $\|y_n - y\| < \epsilon/2$ and therefore $\|phi(y_n) - \phi(y)\| = \|x_n - x\| \le 2\|y_n - y\| < \epsilon$ which shows $\lim_{n\to\infty} \phi(y_n) = \phi(y)$.

It remains to show that ϕ is continuously differentiable on a neighborhood of y_0.

CLAIM 15.150.3. ϕ is C^1 on a neighborhood of y_0.

Since f is continuously differentiable on a neighborhood of y_0, $Df(x_0)$ is invertible and the set of invertible linear maps is an open set of $L(X, Y)$ we may pick an $\epsilon > 0$ such that $B(Df(x_0), \epsilon)$ are invertible maps and the define $U_2 = U_1 \cap Df^{-1}(B(Df(x_0), \epsilon))$ which an open neighborhood of x_0 such that $Df(x)$ is invertible for all $x \in U_2$. Define $V_2 = \phi^{-1}(U_2)$ which is an open neighborhood of y_0 by continuity of ϕ and note that f is a continuous bijection of U_2 to V_2 with continuous inverse ϕ and we still have the inequality $\|x_2 - x_1\| \le 2\|f(x_2) - f(x_1)\|$ for all $x_1, x_2 \in U_2$. Given $x_1 \in U_2$ we pick ψ such that $f(x) - f(x_1) - Df(x_1)(x - x_1) = \|x - x_1\|\psi(x - x_1)$ and $\lim_{h\to 0}\psi(h) = 0$. Let $y_1 = f(x_1) \in V_2$ then

$$\begin{aligned}\left\|\phi(y) - \phi(y_1) - Df(x_1)^{-1}(y - y_1)\right\| &= \left\|x - x_1 - Df(x_1)^{-1}(f(x) - f(x_1))\right\| \\ &= \left\|x - x_1 - Df(x_1)^{-1}(Df(x_1)(x - x_1) + \|x - x_1\|\psi(x - x_1))\right\| \\ &= \left\|Df(x_1)^{-1}\|x - x_1\|\psi(x - x_1))\right\| \\ &\le \left\|Df(x_1)^{-1}\right\|\|x - x_1\|\|\psi(x - x_1)\| \\ &\le 2\left\|Df(x_1)^{-1}\right\|\|f(x) - f(x_1)\|\|\psi(\phi(y) - \phi(y_1))\|\end{aligned}$$

By continuity of ϕ and the fact that $\lim_{h\to 0}\psi(h) = 0$ we have $\lim_{y\to y_1} 2\left\|Df(x_1)^{-1}\right\|\|\psi(\phi(y) - \phi(y_1))\| = 0$ which shows that $D\phi(y_1) = Df(\phi(y_1))^{-1}$. The continuity of $D\phi$ thus follows from the continuity of ϕ, Df and Proposition 15.141. $\square$

The Inverse Function Theorem has the following equally important consequence that is known at the Implicit Function Theorem.

THEOREM 15.151 (Implicit Function Theorem). *Let X, Y and Z be a Banach spaces, let $U \subset X \times Y$ be an open set and let $f : U \to Z$ be C^p. Suppose $(x_0, y_0) \in U$, that $f(x_0, y_0) = 0$ and $Df(x_0, y_0)(0, v)$ defines an invertible map from $Y \to Z$, then there exists an open set $V \subset X$ such that $x_0 \in V$, an open set $W \subset Y$ such that $y_0 \in W$ and a function $g : V \to W$ such that g is continuously differentiable, $f(x, g(x)) = 0$ for all $x \in V$ and $f(x, y) = 0$ if and only if $y = g(x)$ for all $(x, y) \in V \times W$.*

PROOF. First define the map $g : U \to X \times Z$ by $\psi(x, y) = (x, f(x, y))$. We claim that g has a local inverse at (x_0, y_0). For this we apply Proposition 15.146 to calculate the derivate

$$D\psi(x_0, y_0)(u, v) = \begin{bmatrix} \mathrm{Id} & 0 \\ D_1 f(x_0, y_0) & D_2 f(x_0, y_0) \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = (u, D_1 f(x_0, y_0)u + D_2 f(x_0, y_0)v)$$

Since $D_2\psi(x_0, y_0)$ is assumed invertible it is easy to see that

$$D\psi(x_0, y_0)^{-1} = \begin{bmatrix} \mathrm{Id} & 0 \\ -D_2 f(x_0, y_0)^{-1} \circ D_1 f(x_0, y_0) & D_2 f(x_0, y_0)^{-1} \end{bmatrix}$$

and therefore we may apply the Inverse Function Theorem 15.147 to conclude that there exists an open set $W \subset X \times Z$ and a $\phi : W \to U$ with $(x_0, 0) \in W$, ϕ continuously differentiable on W and $\psi(\phi(x, z)) = (x, z)$ on W. It follows that $\phi(x, z) = (x, \phi_2(x, z))$ for a continuously differentiable ϕ_2. Pick an open ball V containing x_0 such that $V \times \{0\} \subset W$ and define $g(x) = \phi_2(x, 0)$ on V then $(x, 0) = \psi(\phi(x, 0)) = (x, f(x, g(x)))$; so $f(x, g(x)) = 0$ as required. Note that $g(x)$ is the unique point in $\phi(W) \cap \{x\} \times Y$ satisfying $f(x, g(x)) = 0$ for $x \in V$. This follows since ψ is injective on $\phi(W)$ and therefore if $f(x, a) = f(x, b)$ then $\psi(x, a) = (x, f(x, a)) = (x, f(x, b)) = \psi(x, b)$ which imples $a = b$. ☐

13. Linear Algebra

This is a little refresher on linear algebra with more of a focus on matrix factorizations.

DEFINITION 15.152. Let $\mathbb{F}$ be a field, then a *vector space* over $\mathbb{F}$ is an abelian group $(V, +, 0)$ together with a multiplication operator $\mathbb{F} \times V \to V$ such that $a(v + w) = av + aw$.

DEFINITION 15.153. If $W \subset V$ is a vector space then we say that W is a *subspace* of V. We say that a set of elements $v_1, \dots, v_n$ of V is *linearly independent* if and only if $a_1 v_1 + \cdots + a_n v_n = 0$ implies $a_1 = \cdots = a_n = 0$. If a set is not linearly independent then we say it is *linearly dependent.* The *dimension* of a vector space V is the supremum of cardinalities of linearly independent sets in V. Given elements v_α in V the *linear span* is the intersection of all subspaces of V containing the v_α.

PROPOSITION 15.154. *Let V be a vector space of $\mathbb{F}$.*

(i) *Let W_α be a collection of subspaces of V, then $W = \cap_\alpha W_\alpha$ is a subspace.*
(ii) *Let v_α be elements in V then the span of v_α is a subspace and moreover the span is precisely the set of finite linear combinations of elements of v_α.*
(iii) *The dimension of the span of $\{v_1, \dots, v_n\}$ is less than or equal to n. It is equal to n if and only if $v_1, \dots, v_n$ are linearly independent.*
(iv) $\dim\{v_1, \dots, v_{n+1}\} \leq \dim\{v_1, \dots, v_n\} + 1$.

PROOF. The fact that W is a subspace is simple and left to the read. The fact that a linear span is a subspaces follows from the first assertion as it was defined as an intersection of subspaces.

To see that the set of finite linear combinations is indeed the span, first note that it is clear all finite linear combinations belong to every subspace containing the v_α. It suffices to show that the set of finite linear combinations is a vector space. Given u and w which are finite linear combinations with index sets $A \subset \Lambda$ and $B \subset \Lambda$ respectively. By use of zero coefficients we express both u and w using the index set $A \cup B$. So to be concrete we write $u = \sum_{i=1}^n u_i v_{\alpha_i}$ and $w = \sum_{i=1}^n w_i v_{\alpha_i}$. Let $a, b \in \mathbb{F}$ and we compute $au + bw = \sum_{i=1}^n (au_i + bw_i) v_{\alpha_i}$.

To see (iii) we use induction. For $n = 1$ it is clear that $\dim\{v_1\} = 1$ if $v_1 \neq 0$ and 0 if $v_1 = 0$ and that $v_1 \neq 0$ if and only if $\{v_1\}$ is linearly independent. Suppose the result is true for $n-1$ and consider $v_1, \dots, v_n$. Suppose $w_1, \dots, w_m$ is a linearly independent set in the span of the v_i. Write $w_i = \sum_{j=1}^n a_{ij} v_j$ for $1 \leq i \leq m$. If $a_{in} = 0$ for all $1 \leq i \leq n$

TODO: Finish □

DEFINITION 15.155. Let V and W be vector spaces over $\mathbb{F}$ then a function $A : V \to W$ is said to be a *linear map* if $A(av + bw) = aAv + bAw$ for all $v, w \in V$ and $a, b \in \mathbb{F}$. In the special case that $W = \mathbb{F}$ we may say that A is a *linear functional*. The set of linear functionals on V is denoted V^* and is called the *dual space* to V. Given any linear map $A : V \to W$ we define the *dual map* $A^* : W^* \to V^*$ by $A^*(\lambda)(v) = \lambda(Av)$.

Note that the dual map is well defined since

$$A^*(\lambda)(av + bw) = \lambda(A(av + bw)) = \lambda(aAv + bAw) = a\lambda(Av) + b\lambda(Aw) = aA^*(\lambda)(v) + bA^*(\lambda)(w)$$

shows that $A^*(\lambda)$ is a linear functional. The dual space is easily seen to be a vector space and the dual map is easily shown to be a linear map.

PROPOSITION 15.156. *V^* is a vector space over $\mathbb{F}$ with addition and scalar multiplication defined pointwise as $(a\lambda + b\mu)(v) = a\lambda v + b\mu(v)$. With repsect to this vector space structure the dual map A^* is linear. If V is finite dimensional then V^* is finite dimensional and* $\dim V = \dim V^*$.

PROOF. The proof that V^* is a vector space is elementary and left to the reader. To see that A^* is a linear map we just compute using the definitions

$$A^*(a\lambda + b\mu)(v) = (a\lambda + b\mu)(Av) = a\lambda(Av) + b\mu(Av) = (aA^*(\lambda) + bA^*(\mu))(v) \text{ for all } v \in V$$

Now if V is finite dimensional we can select a basis $v_1, \dots, v_n$. Define $v_i^* \in V^*$ by $v_i^*(v_j) = \delta_{ij}$ for $1 \leq i, j \leq n$. We claim that v_i^* is a basis for V^*. Clearly v_i^* spans V^* since if we are given an arbitrary λ then by linearity

$$\left(\sum_{i=1}^n \lambda(v_i) v_i^* \right) v = \sum_{i=1}^n \lambda(v_i) v_i^* (\sum_{j=1}^n a_j v_j) = \sum_{i=1}^n a_i \lambda(v_i) = \lambda(v)$$

Moreover the v_i^* are seen to be linearly independent since if $\sum_{i=1}^n a_i v_i^* = 0$ then for each $1 \leq j \leq n$ we have $0 = (\sum_{i=1}^n a_i v_i^*) v_j = a_j$. Thus v_i^* is a basis hence $\dim V = \dim V^* = n$. □

The basis v_i^* constructed from the basis v_i in the above proof is referred to as the *dual basis*.

Definition 15.157. Let A be an $m \times n$ real matrix, then a triple comprising an $m \times m$ orthogonal matrix U, an $n \times n$ orthogonal matrix V and a $m \times n$ diagonal matrix Σ with $\Sigma_{11} \geq \cdots \geq \Sigma_{pp}$ where $p = m \wedge n$ such that $A = U\Sigma V^T$ is called a *singular value decomposition.*

Theorem 15.158. *Singular value decompositions exist.*

Proof. The result is trivially true if $A = 0$ (let $\Sigma = 0$, U and V be identity matrices), so assume that $A \neq 0$. Let σ_1 be the L^2 operator norm of A. By compactness of the unit sphere we can find a unit vector $x_1 \in \mathbb{R}^n$ such that $0 \neq \sigma_1 = \|Ax_1\|$. Define $y_1 = \sigma_1^{-1} A x_1$ so that y_1 is a unit vector in $\mathbb{R}^m$. We can now find an orthonormal basis $\{x_2, \ldots, x_n\}$ of $x_1^\perp$ and an orthonormal basis $\{y_2, \ldots, y_m\}$ of $y_1^\perp$. Define the orthogonal matrices $U_1 = [y_1, \ldots, y_m]$ and $V_1 = [x_1, \ldots, x_n]$. From the fact that $Ax_1 = \sigma_1 y_1$ we can write

$$U_1^T A V_1 = \begin{bmatrix} \sigma_1 & w^T \\ 0 & B \end{bmatrix}$$

where $w \in \mathbb{R}^{n-1}$ and B is an $(m-1) \times (n-1)$ matrix. Observe that

$$\begin{aligned} \left\|U_1^T A V_1\right\| = \sup_{v \neq 0} \frac{\left\|U_1^T A V_1 v\right\|}{\|v\|} &\geq \frac{\left\| U_1^T A V_1 \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} \right\|}{\left\| \begin{bmatrix} \sigma_1 \\ w \end{bmatrix} \right\|} = \frac{\left\| \begin{bmatrix} \sigma_1^2 + w^T w \\ Bw \end{bmatrix} \right\|}{\sqrt{\sigma_1^2 + w^T w}} \\ &= \frac{\sqrt{(\sigma_1^2 + w^T w)^2 + w^T B^T B w}}{\sqrt{\sigma_1^2 + w^T w}} \geq \sqrt{\sigma_1^2 + w^T w} \end{aligned}$$

On the other hand, by orthogonality of U_1 and V_1 we know that $\left\|U_1^T A V_1\right\| = \|A\| = \sigma_1$ and therefore we see that $w^T w = 0$ hence $w = 0$. Now use the induction hypothesis to conclude that there exist U_2 and V_2 such that $U_2^T B V_2 = \Sigma_2$ is diagonal with nonincreasing entries on the diagonal and note that if we define

$$U = U_1 \begin{bmatrix} 1 & 0 \\ 0 & U_2 \end{bmatrix}, \qquad V = V_1 \begin{bmatrix} 1 & 0 \\ 0 & V_2 \end{bmatrix}$$

Then

$$U^T A V = \begin{bmatrix} 1 & 0 \\ 0 & U_2^T \end{bmatrix} U_1^T A V_1 \begin{bmatrix} 1 & 0 \\ 0 & V_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & U_2^T \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 \\ 0 & B \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & V_2 \end{bmatrix} = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}$$

Lastly note that embedding $\mathbb{R}^{n-1}$ into $\mathbb{R}^n$ as the vectors of the form $\begin{bmatrix} 0 \\ v \end{bmatrix}$ we get

$$\begin{aligned} \sigma_1 = \|A\| = \left\|U_1^T A V_1\right\| &= \sup_{\substack{w \in \mathbb{R}^n \\ w \neq 0}} \frac{\left\| \begin{bmatrix} \sigma_1 & 0 \\ 0 & B \end{bmatrix} w \right\|}{\|w\|} \\ &\geq \sup_{\substack{v \in \mathbb{R}^{n-1} \\ v \neq 0}} \frac{\|Bv\|}{\|v\|} = \|B\| = (\Sigma_2)_{11} \end{aligned}$$

and therefore the diagonal entries of $\Sigma = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}$ are nonincreasing and we are done. ∎

The fact that we may perform matrix factorizations in linear algebra in a Borel measurable way is easy to verify with the following elegant method relying on the "principle of measurable choice".

THEOREM 15.159. *Let S and T be separable complete metric spaces and let $A \subset S \times T$ be closed and σ-compact. Then $\pi_1(A)$ is Borel and there exists a Borel measurable function $f : \pi_1(A) \to T$ such that the graph of f is contained in A.*

PROOF. We start with the following

CLAIM 15.159.1. Let F be a closed set in T then $\pi_1(A \cap S \times F)$ is Borel measurable.

Write $A = \cup_n K_n$ with K_n compact. Since F be a closed set it follows that $A \cap S \times F$ is closed $A \cap S \times F = \cup_n K_n \cap S \times F$ where each $K_n \cap S \times F$ is compact. Since π_1 is continuous each $\pi_1(K_n \cap S \times F)$ is compact in S and $\pi_1(A \cap S \times F) = \cup_n \pi_1(K_n \cap S \times F)$ is therefore Borel.

Applying the claim with $F = T$ we see that $\pi_1(A)$ is Borel.

Take a countable dense subset $\{y_n\}$ of T. We define f by an iterative approximation scheme. For $x \in \pi_1(A)$ define $f_1(x)$ to be the first y_n (in index order) such that $A \cap \{x\} \times \overline{B}(y_n, 1/2) \neq \emptyset$ where $\overline{B}(z, r)$ represents the closed ball of radius r centered at z. Clearly f_1 is well defined since there exists an $(x, y) \in A$ and by density of $\{y_n\}$ in T for any $r > 0$ there exists a y_n with $(x, y) \in \{x\} \times \overline{B}(y_n, r)$.

Next observe that $f_1(x)$ is Borel measurable. To see this, for each $r > 0$ and $n \in \mathbb{N}$ we define

$$\begin{aligned} C_{n,r} &= \pi_1(A \cap S \times \overline{B}(y_n, r)) \\ &= \{x \in S \mid A \cap \{x\} \times \overline{B}(y_n, r) \neq \emptyset\} \end{aligned}$$

which is Borel by the claim. Moreover it follows that $f_1^{-1}(y_n) = C^c_{1,1/2} \cap \cdots \cap C^c_{n-1,1/2} \cap C_{n,1/2}$ is Borel. Since f_1 is countably valued it follows that f_1 is Borel measurable.

Now define $f_k(x)$ be the first y_n such that $A \cap \{x\} \times \overline{B}(y_n, 1/2^k) \neq \emptyset$ and $d(f_{k-1}(x), y_n) \leq 1/2^{k-2}$; again density of the $\{y_n\}$ shows that $f_k(x)$ is well defined. Borel measurability of f_k follows by a simple induction as the Borel measurability of f_{k-1} and Lipschitz continuity of d imply that $D_{n,r} = \{x \mid d(f_{k-1}(x), y_n) \leq r\}$ is Borel measurable for every $n \in \mathbb{N}$ and $r > 0$. Since

$$f_k^{-1}(y_n) =$$
$$(C_{1,1/2^k} \cap D_{1,1/2^{k-2}})^c \cap \cdots \cap (C_{n-1,1/2^k} \cap D_{n-1,1/2^{k-2}})^c \cap C_{n,1/2^k} \cap D_{n,1/2^{k-2}}$$

and f_k is countably valued it follows that f_k is Borel measurable.

For a fixed x note that for $j > k$ we have the triangle inequality

$$d(f_k(x), f_j(x)) \leq \sum_{i=k}^{j-1} d(f_i(x), f_{i+1}(x)) \leq \sum_{i=k}^{j-1} 1/2^{i-1} \leq \sum_{i=k}^{\infty} 1/2^{i-1} = 1/2^{k-2}$$

which shows that $f_k(x)$ is Cauchy. By completeness of T, we can take the limit of f_k which is a Borel measurable function (Lemma 2.15). Since A is closed and $\lim_{k\to\infty} d(f_k(x), A) = 0$ it follows that $(x, f(x)) \in A$ and we are done.

TODO: Do we ever use Polishness of S? Note the reference Azoff "Borel Measurability in Linear Algebra" in the Proceedings of the AMS who in turn references Bourbaki. $\square$

We now illustrate how matrix factorizations (and canonical forms) may be shown to be Borel measurable by using the singular value decomposition as an example.

COROLLARY 15.160. *There exists a Borel measurable function $f(A) = (U, \Sigma, V)$ from $\mathbb{R}^{m\times n}$ to $\mathbb{R}^{m\times m} \times \mathbb{R}^{m\times n} \times \mathbb{R}^{n\times n}$ such that $A = U\Sigma V^T$ is a singular value decomposition.*

PROOF. Let

$$F = \{(A, U, \Sigma, V) \mid U, V \text{ are orthogonal, } \Sigma \text{ is diagonal and } A = U\Sigma V^T\}$$

and note that F is closed (because the spaces of orthogonal and diagonal matrices are closed and matrix multiplication is continuous). F is also σ-compact because the ambient space is (just write $F = \cup_n F \cap \overline{B}(0, n)$). Since singular value decompositions exist (Theorem 15.157) we know that $\pi_1(F) = \mathbb{R}^{m\times n}$ and therefore the result follows immediately from Theorem 15.158. □

Béla Sz Nagy also seems to prove measurability of eigenvalues and eigenvectors using the minimax criterion in "Harmonic Analysis of Operators on Hilbert Space". From this one can bootstrap up to the measurability of the Schur decomposition and then get to the measurability of the SVD by considering Schur decompositions of $A^T A$ and AA^T. There are also some more powerful section theorems that may have some relevance.

DEFINITION 15.161. Let A be a $d \times d$ matrix and let Σ_d be the set of permutations of $\{1, \dots, d\}$ then the *determinant* of A is the number $\det(A) = \sum_{\sigma\in\Sigma_d} \mathrm{sgn}(\sigma) A_{1\sigma(1)} \cdots A_{d\sigma(d)}$.

The formula we have used in the definition of the determinant is called Leibnitz's formula.

PROPOSITION 15.162. *Let A be a $d \times d$ matrix,*

(i) $\det(A) = \det(A^T)$
(ii) *let $\tau \in \Sigma_d$ and define $(A^\tau)_{ij} = A_{i\tau(j)}$ then* $\det(A^\tau) = \mathrm{sgn}(\tau)\det(A)$.
(iii) *let $\tau \in \Sigma_d$ and define $(A_\tau)_{ij} = A_{\tau(i)j}$ then* $\det(A_\tau) = \mathrm{sgn}(\tau)\det(A)$.
(iv) *If any two columns of A are identical or any two rows of A are identical then* $\det(A) = 0$

PROOF. To see (i)

$$\begin{aligned}
\det(A^T) &= \sum_{\sigma\in\Sigma_d} \mathrm{sgn}(\sigma) A^T_{1\sigma(1)} \cdots A^T_{d\sigma(d)} = \sum_{\sigma\in\Sigma_d} \mathrm{sgn}(\sigma) A_{\sigma(1)1} \cdots A_{\sigma(d)d} \\
&= \sum_{\sigma\in\Sigma_d} \mathrm{sgn}(\sigma) A_{\sigma(1)\sigma^{-1}(\sigma(1))} \cdots A_{\sigma(d)\sigma^{-1}(\sigma(d))} = \sum_{\sigma\in\Sigma_d} \mathrm{sgn}(\sigma^{-1}) A_{1\sigma^{-1}(1)} \cdots A_{d\sigma^{-1}(d)} \\
&= \det(A)
\end{aligned}$$

To see (ii) we note that by properties of permutations we have

$$\begin{aligned}
\det(A^\tau) &= \sum_{\sigma\in\Sigma_d} \operatorname{sgn}(\sigma) A^\tau_{1\sigma(1)}\cdots A^\tau_{d\sigma(d)} = \sum_{\sigma\in\Sigma_d} \operatorname{sgn}(\sigma) A_{1\tau(\sigma(1))}\cdots A_{d\tau(\sigma(d))} \\
&= \sum_{\sigma\in\Sigma_d} \operatorname{sgn}(\tau^{-1}\circ\tau\circ\sigma) A_{1\tau(\sigma(1))}\cdots A_{d\tau(\sigma(d))} \\
&= \sum_{\sigma\in\Sigma_d} \operatorname{sgn}(\tau^{-1}) \operatorname{sgn}(\tau\circ\sigma) A_{1\tau(\sigma(1))}\cdots A_{d\tau(\sigma(d))} \\
&= \operatorname{sgn}(\tau) \sum_{\sigma\in\Sigma_d} \operatorname{sgn}(\sigma) A_{1\sigma(1)}\cdots A_{d\sigma(d)} = \operatorname{sgn}(\tau)\det(A)
\end{aligned}$$

To see (iii) we simply note that $A_\tau = ((A^T)^\tau)^T$ and using (i) and (ii).

To see (iv) suppose $1 \le j < k \le d$ and that columns j and k of A are equal; if we let (jk) be the permutation that exchanges j and k then we restate the hypothesis as $A^{(jk)} = A$. Now apply (ii) and use the fact that $\operatorname{sgn}(jk) = -1$. The case of equal rows follow by transposition (or equivalently by using (iii) with the appropriate transposition). □

THEOREM 15.163 (Cauchy-Binet Theorem). *Let A be a $d \times m$ matrix and let B be an $m \times d$ matrix. Let Ψ be the set of all increasing injective functions from $\{1, \dots, d\}$ to $\{1, \dots, m\}$ and for every $\psi \in \Psi$ let*

$$A_\psi = A(:, [\psi(1)\cdots\psi(d)])$$

and let

$$B_\psi = B([\psi(1)\cdots\psi(d)], :)$$

then

$$\det(AB) = \sum_{\psi\in\Psi} \det(A_\psi)\det(B_\psi)$$

PROOF. By the formula for matrix multiplication note that the i^{th} row of AB is a linear combination of the m rows of B; specifically $(AB)(i,:) = \sum_{k=1}^m A_{ik} B(k,:)$. Iteratively using this fact together with the linearity of the determinant we get

$$\begin{aligned}
\det(AB) = \det\begin{bmatrix}(AB)(1,:)\\ \cdots \\ (AB)(d,:)\end{bmatrix} &= \sum_{k_1=1}^m A_{1k_1} \det\begin{bmatrix} B(k_1,:) \\ (AB)(2,:) \\ \cdots \\ (AB)(d,:)\end{bmatrix} \\
&= \sum_{k_1=1}^m \cdots \sum_{k_d=1}^m A_{1k_1}\cdots A_{dk_d} \det\begin{bmatrix} B(k_1,:)\\ \cdots \\ B(k_d,:)\end{bmatrix} \\
&= \sum_{(k_1,\dots,k_d)\in\{1,\dots,m\}^d} A_{1k_1}\cdots A_{dk_d} \det B([k_1\cdots k_d],:) \\
&= \sum_{\substack{(k_1,\dots,k_d)\in\{1,\dots,m\}^d \\ k_i \text{ distinct}}} A_{1k_1}\cdots A_{dk_d} \det B([k_1\cdots k_d],:)
\end{aligned}$$

where in the last line we have used the fact that for any set of indices $(k_1, \ldots, k_d)$ for which there exists $1 \leq i < j \leq d$ and $k_i = k_j$ we have $\det B([k_1 \cdots k_d], :) = 0$ (assertion (iv) of Proposition 15.161).

Given any d-tuple $(k_1, \ldots, k_d)$ with k_i distinct there is a unique permutation σ of $\{1, \ldots, d\}$ such that $1 \leq k_{\sigma(1)} < \cdots < k_{\sigma(d)} \leq m$. These increasing sequences $1 \leq k_{\sigma(1)} < \cdots < k_{\sigma(d)} \leq m$ are precisely the ranges of functions $\psi \in \Psi$. Thus if we let Σ_d be set of permutations of $\{1, \ldots, d\}$ then every d-tuple $(k_1, \ldots, k_d)$ with k_i distinct can be written uniquely as $(\psi(\sigma(1)), \ldots, \psi(\sigma(d)))$ where $\psi \in \Psi$ and $\sigma \in \Sigma_d$. Now we compute using Proposition 15.161

$$\begin{aligned}
\det(AB) &= \sum_{\psi \in \Psi} \sum_{\sigma \in \Sigma_d} A_{1\psi(\sigma(1))} \cdots A_{d\psi(\sigma(d))} \det B([\psi(\sigma(1)) \cdots \psi(\sigma(d))], :) \\
&= \sum_{\psi \in \Psi} \sum_{\sigma \in \Sigma_d} (A_\psi)_{1\sigma(1)} \cdots (A_\psi)_{d\sigma(d)} \det(B_\psi)([\sigma(1) \cdots \sigma(d)], :) \\
&= \sum_{\psi \in \Psi} \sum_{\sigma \in \Sigma_d} (A_\psi)_{1\sigma(1)} \cdots (A_\psi)_{d\sigma(d)} \operatorname{sgn}(\sigma) \det(B_\psi) \\
&= \sum_{\psi \in \Psi} \det(A_\psi) \det(B_\psi)
\end{aligned}$$

□

Corollary 15.164. *Let A and B be $d \times d$ matrices then* $\det(AB) = \det(A)\det(B)$.

Proof. Apply the Cauchy-Binet formula in the case $d = m$ (here Ψ comprises on the identity map). □

Corollary 15.165. *Let A be an orthogonal matrix then* $|\det(A)| = 1$.

Proof. By the previous Corollary and Proposition 15.161

$$1 = \det(\mathrm{Id}) = \det(A^T A) = \det(A^T)\det(A) = \det(A)^2$$

□

CHAPTER 16

Optimization in Banach Spaces

As one will undoubtedly remember from one's first calculus class the derivative is an extraordinarily useful tool for finding maxima and minima of functions of a real variable. Essentially all of that theory carries over to the case of general Banach space domains. One of the goals in this subsection is to develop that basic theory.

In a multivariate calculus course the reader almost certainly encounterd problems of constrained optimization as well: learning the tool of the Lagrange multiplier for solving such problems. This theory also carries over to the Banach space setting and we develop it here.

What one has learned up to this point is the theory of equality constrained optimization. Though it tends not to be taught in the introductory calculus curriculum, in applications it is equally important to be able to solve optimization with both equality and inequality constraints. Having the tools we have developed it is no harder to develop such theory in the general Banach space setting and we do so here.

We will discuss optimization problems in terms of minimization; as a general rule there is no loss of generality in doing so as maximization of a function f may be performed by minimizing the function $-f$.

First we distinguish the different kinds of minima that we may characterize. The primary distinction is the dichotomy between local and global minimization. There are subtler distinctions to be made between different type of local minima. The definitions make sense for arbitrary topological spaces.

Definition 16.1. Let X be a topological space and let $f : X \to \mathbb{R}$ be a function. We say that $x^* \in X$ is a *global minimizer* of f if $f(x^*) \leq f(x)$ for all $x \in X$. We say that $x^* \in X$ is a *local minimizer* if there exists an open set $U \subset X$ such that $x^* \in U$ and $f(x^*) \leq f(x)$ for all $x \in U$. We say that $x^* \in X$ is a *strict local minimizer* if there exists an open set $U \subset X$ such that $x^* \in U$ and $f(x^*) < f(x)$ for all $x \in U$ with $x^* \neq x$. We say that $x^* \in X$ is an *isolated local minimizer* if there exists and open set $U \subset X$ such that $x^* \in U$ and x^* is the only local minimizer in U.

Example 16.2. Let
$$f(x) = \begin{cases} x^4 \cos(1/x) + 2x^4 & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$
then 0 is a strict local minimizer that is not isolated. TODO: Show this

Note that minimizers are not guaranteed to exist since functions may be unbounded below. More subtely may have a lower bound but may never take the value of its greatest lower bound. We already know a case in which both of these problems are avoided: namely continuous images of compact sets are compact in $\mathbb{R}$ and this guarantees the exists of a miminum (Theorem 1.31). As it turns out this fact can be generalized a bit by relaxing the property of continuity.

Definition 16.3. Let X be a topological space and let $f : X \to \mathbb{R}$ be a function, we say that f is *lower semicontinuous* (resp. *upper semicontinuous*) if for every $\epsilon > 0$ there exists an open set U containing x such that $f(y) \geq f(x) - \epsilon$ (resp. $f(y) \leq f(x) + \epsilon$) for every $y \in U$. We say that f is *sequentially lower semicontinuous* (resp. *sequentially upper semicontinuous*) at x if for every sequence x_n such that $\lim_{n\to\infty} x_n = x$ we have $f(x) \leq \liminf_{n\to\infty} f(x_n)$ (resp. $f(x) \geq \limsup_{n\to\infty} f(x_n)$).

It is simple to see that f is lower semicontinuous (resp. sequentially lower semicontinuous) if and only if $-f$ is upper semicontinuous (resp. sequentially upper semicontinuous).

A function is lower (resp. upper) semicontinuous at x if its values near x are either close to $f(x)$ or larger (resp. smaller) than $f(x)$. In general sequential semicontinuity is a weaker property than semicontinuity (since sequences do not characterize convergence in general topological spaces); however in metric spaces the two concepts are equivalent.

Proposition 16.4. *Let X be a topological space and let f be a lower (resp. upper) semicontinuous at x, then f is sequentially lower (resp. upper) semicontinuous at x. If X is a metric space and f is sequentially lower (resp. upper) semicontinuous at x then f is lower (resp. upper) semicontinuous at x.*

Proof. If suffices to handle the cases of lower semicontinuity since the upper semicontinuity results follow by applying the lower semicontinuity case to $-f$.

If f is lower semicontinuous and $x_n \to x$. Let $\epsilon > 0$ be given an find an open neighborhood U of x such that $f(y) \geq f(x) - \epsilon$ for all $y \in U$. Since $x_n \to x$ we know that there exists $N > 0$ such that $x_n \in U$ for $n \geq N$ and thus $\inf_{m\geq n} f(x_m) \geq f(x) - \epsilon$ for every $n \geq N$. We take the limit at $n \to \infty$ to get $\liminf_{n\to\infty} f(x_n) \geq f(x) - \epsilon$. Since $\epsilon > 0$ was arbitrary we conclude that f is sequentially lower semicontinuous at x.

Now let X be a metric space and suppose that f is not lower semicontinuous at x. Then there exists an $\epsilon > 0$ such that for every $n \in \mathbb{N}$ there exists x_n with $d(x, x_n) < 1/n$ such that $f(x_n) < f(x) - \epsilon$. Clearly $x_n \to x$ and moreover $\liminf_{n\to\infty} f(x_n) \leq f(x) - \epsilon < f(x)$ which shows that f is not sequentially lower semicontinuous at x. □

Compactness and sequential lower semicontinuity suffice to show that a function has a global minimizer.

Theorem 16.5. *Let X be a topological space, let $f : X \to [-\infty, \infty]$ be a sequentially lower semicontinuous function and suppose that there exists $M \in \mathbb{R}$ such that $\{x \in X \mid f(x) \leq M\}$ is non-empty and compact then f has a global minimizer. Moreover if f is lower semicontinuous, the set of global minimizers is compact.*

Proof. We first show the existence of a global minimizer. Let $\alpha = \inf_{x\in X} f(x)$ and note that we know $\alpha \leq M < \infty$. If $\alpha = M$ then in fact $\{x \in X \mid f(x) \leq M\} = \{x \in X \mid f(x) = M\}$ and is assumed non-empty and we are done. Therefore we assume that $\alpha < M$. Let x_n be chosen so that $\lim_{n\to\infty} f(x_n) = \alpha$ (if $\alpha > -\infty$ then choose x_n such that $f(x_n) < \alpha + 1/n$ otherwise choose x_n so that $f(x_n) \le -n$). As $\alpha < M$ we know that there exists $N > 0$ such that $f(x_n) \le M$ for all $n \ge N$ and therefore by compactness there exists a convergent subsequence x_{n_j}. Let $x = \lim_{j\to\infty} x_{n_j}$ and note that by sequential lower semicontinuity at x

$$\alpha \le f(x) \le \liminf_{j\to\infty} f(x_{n_j}) = \alpha$$

which shows that $f(x) = \alpha$.

Let $G = \{x \in X \mid f(x) = \alpha\}$. Now we know that since $\alpha \le M$ that $G \subset \{x \in X \mid f(x) \le M\}$ hence it suffices to show that the set of global minimizers is closed (Corollary 1.30). Let x be in $\overline{G}$ and let $\epsilon > 0$ be given. Since f is lower semicontinuous we can find an open set U containing x such that $f(y) \ge f(x) - \epsilon$. However we know that $G \cap U \neq \emptyset$ therefore $\alpha \ge f(x) - \epsilon$. Since ϵ is arbitrary we conclude $\alpha \ge f(x)$ and thus $x \in G$. □

Given that derivatives are determined by the behavior of functions on arbitrarily small neighborhoods of a point it is clear that they have little to say about when a point is a global minimizer. On the other hand derivatives are rather informative about local minimizers and we turn our attention to this.

1. Unconstrained Optimization

The first thing to do is to note that there are *necessary* conditions for a point being a local minimizer that are described by derivatives. The first such is the vanishing of the first derivative.

THEOREM 16.6. *Let X be a Banach space, let $f : X \to \mathbb{R}$ be a function and let x^* be a local minimum. If f is C^1 on an open neighborhood of x^* then $Df(x^*) = 0$.*

PROOF. The proof is by contradiction. Suppose that $Df(x^*) \neq 0$. Thus there exists $y \in X$ such that $Df(x^*)y > 0$. Let f be C^1 on an open neighborhood U of x^*. By continuity of $Df(x)$ on U we may also find a $\delta > 0$ such that $Df(x)y > 0$ for all $x \in B(x^*, \delta) \subset U$. By multiplying y by an appropriate positive constant we may assume that $\|y\| < \delta$. Now we can apply the Mean Value Theorem to conclude that

$$f(x^* - y) = f(x^*) - \int_0^1 Df(x^* - ty)y \, dt < f(x^*)$$

which shows that x^* is not a local minimizer.

Here is an alternative proof that avoid appealing to the Mean Value Theorem. Find an open ball $\delta > 0$ such that $f(x^*) \le f(y)$ for all y with $\|y - x\| < \delta$. Pick an arbitrary $y \in X$ then for all $0 < t < \delta / \|y\|$ we have $f(x^* + ty) - f(x^*) \ge 0$ which implies

$$\lim_{t\to 0^+} \frac{f(x^* + ty) - f(x^*)}{t} \ge 0$$

Arguing similarly for $-\delta / \|y\| < t < 0$ we get

$$\lim_{t\to 0^-} \frac{f(x^* + ty) - f(x^*)}{t} \ge 0$$

On the other hand we have assumed that the Gâteaux derivative exists and thus

$$\begin{aligned}\lim_{t\to 0}\frac{f(x^*+ty)-f(x^*)}{t} &= \lim_{t\to 0^-}\frac{f(x^*+ty)-f(x^*)}{t} \leq 0 \\ &\leq \lim_{t\to 0^+}\frac{f(x^*+ty)-f(x^*)}{t} = \lim_{t\to 0}\frac{f(x^*+ty)-f(x^*)}{t}\end{aligned}$$

which shows that $df(x^*,y) = \lim_{t\to 0}\frac{f(x^*+ty)-f(x^*)}{t} = 0$.

If we assume that f is Frechet differentiable at x^* then by Proposition 15.131 we know that the Frechet derivative is equal to the directional derivative and thus $Df(x^*) = 0$.

TODO: This latter proof seems to suggest that the vanishing doesn't require C^1 or even Frechet differentiability at x^* just Gâteaux differentiability at x^*. Am I missing something? I don't think so; I should probably rephrase the result in terms of Gâteaux derivatives. □

When f has two derivatives then we can say even more.

THEOREM 16.7. *Let X be a Banach space, let $f : X \to \mathbb{R}$ be a function and let x^* be a local minimum. If f is C^2 on an open neighborhood of x^* then $Df(x^*) = 0$ and $D^2f(x^*)$ is positive semidefinite (i.e. $D^2f(x^*)(v,v) \geq 0$ for all $v \in X$).*

PROOF. Again we proceed by contradiction. Suppose that $D^2f(x^*)(v,v) < 0$. By continuity of D^2f we may find a $\delta > 0$ such that $D^2f(x)(v,v) < 0$ for all $x \in B(x^*,\delta) \subset U$. If necessary multiply v be a small positive constant to guarantee that $\|v\| < \delta$. By Theorem 16.6 we know that $Df(x^*) = 0$ so Taylor's Theorem says

$$f(x^*+v) = f(x^*) + \int_0^1 (1-t)D^2f(x^*+tv)(v,v)\,dt < f(x^*)$$

which is a contradiction. □

When f has two derivatives there also exists sufficient conditions that a point be a local minimizer.

THEOREM 16.8. *Let X be a Banach space, let $f : X \to \mathbb{R}$ be a function and suppose f is C^2 on an open neighborhood U of x^*. If $Df(x^*) = 0$ and $D^2f(x^*)$ is positive definite (i.e. there exists an $\alpha > 0$ such that $D^2f(x^*)(v,v) > \alpha\|v\|^2$ for all $v \in X$ with $v \neq 0$) then x^* is a strict local minimizer of f.*

PROOF. Using continuity of D^2f at x^* we may find a $\delta > 0$ such that $B(x^*,\delta) \subset U$ and $\left\|D^2f(x^*+y) - D^2f(x^*)\right\| < \frac{\alpha}{2}$ for all $\|y\| < \delta$. Note in particular that

$$(D^2f(x^*+y) - D^2f(x^*))(v,w) > -\frac{\alpha\|v\|\|w\|}{2} \text{ for all } \|y\| < \delta \text{ and } v,w \in X$$

By Taylor's Theorem, for all y with $\|y\| < \delta$

$$\begin{aligned}
f(x^* + y) - f(x) &= \int_0^t (1-t) D^2 f(x^* + ty)(y,y)\, dt \\
&= \frac{1}{2} D^2 f(x^*)(y,y) + \int_0^t (1-t)(D^2 f(x^* + ty) - D^2 f(x^*))(y,y)\, dt \\
&\geq \frac{\alpha \|y\|^2}{2} - \frac{\sup_{0 \leq t \leq 1} \left\| D^2 f(x^* + ty) - D^2 f(x^*) \right\| \|y\|^2}{2} \\
&\geq \frac{\alpha \|y\|^2}{4} > 0
\end{aligned}$$

which shows that x^* is a local minimizer. □

Note that in finite dimensions the condition of positive definiteness is equivalent to the apparently weaker condition $D^2 f(x^*)(v, v) > 0$ for all $v \neq 0$.

2. Constrained Optimization

We first consider an abstract version of constrained optimization. Let X be a Banach space and consider a function $f : X \to \mathbb{R}$ then given a closed set $F \subset X$ we can consider the problem of finding a minimizer of f restricted to F. Note that the meaning of finding a constrained minimizer is captured by using our existing definitions of minimizers on the space F with the relative topology.

In order to apply derivatives to the problem of characterizing minimizers on F we need to restrict them to directions that don't leave F; if we have a point $x \in F$ and f is decreasing at x in a direction that immediately takes one out of F then that alone won't means that x isn't a minimizer when restricted to F. This leads us to a definition of direction tangent to a closed set F. Note that if a direction is tangent to a set at a point then any positive multiple should be tangent (though if the set has corners then negative multiples may fail to be tangents; consider the behavior of $|x|$ at the origin). As a result of this observation we should be seeking to characterize a cone of tangent directions.

Definition 16.9. Let X be a Banach space, let F be a closed subset and let $x \in F$ then we say that $v \in X$ is a *tangent vector to F at x* if there is a sequence x_n such that $x_n \in F$ and $\lim_{n \to \infty} x_n = x$ and a sequence of positive real numbers t_n such that $\lim_{n \to \infty} t_n = 0$ that together satisfy

$$\lim_{n \to \infty} \frac{x_n - x}{t_n} = v$$

The set $T_F(x)$ of all tangent vectors to F at x is called the *tangent cone to F at x*.

We call out the fact that the tangent cone is in fact a cone.

Proposition 16.10. *The tangent cone $T_F(x)$ is a cone (i.e. for every $\alpha \geq 0$ and $v \in T_F(x)$ we have $\alpha v \in T_F(x)$).*

Proof. It is trivial to see that $0 \in T_F(x)$ since we can just pick the $x_n \equiv x$. Let $v \in T_F(x)$, $\alpha > 0$ and pick sequences x_n and t_n such that $x_n \to x$, $t_n \to 0$ and $\frac{x_n - x}{t_n} = v$. Then let $\tilde{t}_n = t_n / \alpha$ and note that $\tilde{t}_n \to 0$ and $\frac{x_n - x}{\tilde{t}_n} = \alpha v$. □

It is also useful to note the following equivalent form of the definition of a tangent vector.

PROPOSITION 16.11. *Let X be a Banach space, let F be a closed subset and let $x \in F$ then $= v \in X$ is a tangent vector to F at x if and only if*

$$\liminf_{t \to 0^+} \frac{d(F, x + tv)}{t} = 0$$

PROOF. Suppose $\liminf_{t\to 0^+} \frac{d(F,x+tv)}{t} = 0$. Since $\frac{d(F,x+tv)}{t}$ is non-negative it is equivalent that $\inf_{0<t<h} \frac{d(F,x+tv)}{t} = 0$ for every $h > 0$. Thus for every $n \in \mathbb{N}$ there exists $0 < t_n < 1/n$ and $x_n \in F$ such that

$$\frac{d(x_n, x + t_n v)}{t_n} = \left\| \frac{x_n - x}{t_n} - v \right\| < 1/n$$

By construction, $\lim_{n\to\infty} t_n = 0$ and by the triangle inequality,

$$\lim_{n\to\infty} \|x - x_n\| < \lim_{n\to\infty} (\|v\| \, t_n + t_n/n) = 0$$

thus v is a tangent vector.

If v is a tangent vector to F at x then pick $x_n \in F$ and t_n such that $\lim_{n\to\infty} x_n = x$ and $\lim_{n\to\infty}(x - x_n)/t_n = v$. Let $\epsilon > 0$ be given and pick an $N > 0$ such that $\|x - x_n - t_n v\| / t_n < \epsilon$ for all $n \geq N$. Then since $\lim_{n\to\infty} t_n = 0$ it follows that for any $h > 0$ there exist $0 < t_n < h$ such that $\|x - x_n - t_n v\| / t_n < \epsilon$ which implies $\inf_{0<t<h} \frac{d(F,x+tv)}{t} = 0$ for all $h > 0$. ∎

The first hint that we have the correct notion of tangent vector is the following necessary condition for a local minimizer to exist.

A few facts about Landau notation.

DEFINITION 16.12. Let X, Y and Z be Banach spaces. Let x_n be a sequence in X and let y_n be a sequence in Y we say that $x_n = o(y_n)$ as $n \to \infty$ if $\lim_{n\to\infty} \frac{x_n}{\|y_n\|} = 0$ (equivalently $\lim_{n\to\infty} \frac{\|x_n\|}{\|y_n\|} = 0$). We say that $x_n = O(y_n)$ if there exists $M > 0$ and $N \geq 0$ such that $\frac{\|x_n\|}{\|y_n\|} \leq M$ for all $n \geq N$. Given functions $f : X \to Y$, $g : X \to Z$ and $x_0 \in X$ we say that $f(x)$ is $o(g(x))$ as $x \to x_0$ if $\lim_{x\to x_0} \frac{f(x)}{\|g(x)\|} = 0$ (equivalently $\lim_{x\to x_0} \frac{\|f(x)\|}{\|g(x)\|} = 0$) and we say that $f(x)$ is $O(g(x))$ if there exists $M > 0$ and $\delta > 0$ such that $\frac{\|f(x)\|}{\|g(x)\|} \leq M$ for all $\|x - x_0\| < \delta$.

Because the definitions above really only depend on the norms of the sequences and functions in question, it is often useful to say that a sequence $x_n \in X$ is $o(\|y_n\|)$ or a function $f(x)$ is $o(\|g(x)\|)$. It is also worth pointing out that Landau notation is confusing for uninitiated in large part because of its abuse of the equality sign.

PROPOSITION 16.13. *The following are true:*

(i) $o(y_n) + o(y_n) = o(y_n)$.
(ii) *Suppose $z_n = O(y_n)$ then if $x_n = o(z_n)$ it follows that $x_n = o(y_n)$. In shorthand we say that $o(O(y_n)) = o(y_n)$.*

PROOF. (i) follows from linearity: if $x_n = o(y_n)$ and $z_n = o(y_n)$ then it follows that

$$\lim_{n\to\infty} \frac{x_n + z_n}{\|y_n\|} = \lim_{n\to\infty} \frac{x_n}{\|y_n\|} + \lim_{n\to\infty} \frac{z_n}{\|y_n\|} = 0$$

To see (ii), we know that $\lim_{n\to\infty} \frac{\|x_n\|}{\|z_n\|} = 0$ and there exist $M, N \geq 0$ such that $\|z_n\| \leq M \|y_n\|$ for all $n \geq N$, therefore

$$0 \leq \lim_{n\to\infty} \frac{\|x_n\|}{\|y_n\|} \leq M \lim_{n\to\infty} \frac{\|x_n\|}{\|z_n\|} = 0$$

□

THEOREM 16.14. *Let X be a Banach space, $F \subset X$ be closed and let $f : U \to \mathbb{R}$ be C^1 on an open set $U \supset F$. Then if x^* is a local minimizer of f on F we have $Df(x^*)v \geq 0$ for all $v \in T_F(x^*)$.*

PROOF. Suppose that we have $x_n \in F$ with $x_n \to x$, $t_n > 0$ with $t_n \to 0$ and $(x_n - x)/t_n \to v$. By Taylor's Theorem and the fact that x^* is a local minimizer we know that we can find a neighborhood $x^* \in V \subset U$ such that

$$f(y) - f(x^*) = Df(x^*)(y - x) + o(\|y - x^*\|) \geq 0$$

and for $y \in V$. From the fact that $x_n \to x$ we can find an $N \in \mathbb{N}$ such that $x_n \in V$ for all $n \geq N$. Since $(x_n - x)/t_n \to v$ we know that $\|x_n - x^*\|$ is $O(t_n)$ and therefore $o(\|x_n - x\|) = o(t_n)$ (Proposition 16.13) and since $x_n - x - t_n v = o(t_n)$ we have

$$t_n Df(x^*)v + o(t_n) = f(x_n) - f(x^*) \leq 0$$

which implies that $Df(x^*)v \geq 0$ (divide by $t_n > 0$ and let $n \to \infty$). □

TODO: Define the normal cone(s)...

DEFINITION 16.15. Let X be a Hilbert space, $F \subset X$ be closed and let $x \in F$ then we say that $v \in X$ is a *regular normal vector to F at x* if $\langle v, w \rangle \leq 0$ for all $w \in T_F(x)$. The set of all regular normal vectors to F at x is called the *regular normal cone* and is denoted $\widehat{N}_F(x)$. We say that $v \in X$ is a *limiting normal vector to F at x* or simply a *normal vector to F at x* if there are sequences x_n, v_n with v_n a regular normal vector to F at x_n and $\lim_{n\to\infty} x_n = x$ and $\lim_{n\to\infty} v_n = v$.

Facts:

The proximal normal cone is a convex cone but may not be closed The limiting normal cone is a closed cone but may not be convex. The limiting normal cone may be defined as the limit of proximal normal vectors as well as by using regular normal vectors.

DEFINITION 16.16. Let X be a Hilbert space, $F \subset X$ be closed and let $x \in F$ we say that F is *Clarke regular at x* if F is locally closed at x and $\widehat{N}_F(x) = N_F(x)$.

PROPOSITION 16.17. *$\widehat{N}_F(x)$ is a closed convex cone. Moreover, v is a regular normal to F at x if and only if $\langle v, y - x \rangle \leq o(y - x)$ for $y \in F$ and $x \neq y$. That is to say for every sequence $x_n \in F$ with $x_n \neq x$ and $\lim_{n\to\infty} x_n = x$, we have*

$$\limsup_{n\to\infty} \frac{\langle v, x_n - x \rangle}{\|x_n - x\|} \leq 0$$

PROOF. The fact that $\widehat{N}_F(x)$ is a closed convex cone follows from the fact that is is an intersection of closed halfspaces.

Suppose that there exists an sequence $x_n \in F$ with $x_n \neq x$, $x_n \to x$ and

$$\limsup_{n\to\infty} \frac{\langle v, x_n - x \rangle}{\|x_n - x\|} > 0$$

By passing to a subsequence we may assume that the lim sup may be replaced by a limit. Let $w_n = \frac{x_n - x}{\|x_n - x\|}$ and by compactness we may pass to a further subsequence and assume that w_n converges to a unit vector w and $\langle v, w \rangle = \lim_{n\to\infty} \langle v, w_n \rangle > 0$. On the other hand, w is seen to be a tangent vector to F at x because it is the limit $\frac{x_n - x}{\|x_n - x\|}$ and $\|x_n - x\| \to 0$.

Now suppose that $\langle v, y - x \rangle \leq o(y - x)$ and let $w \in T_F(x)$. Pick a defining sequence $x_n \to x$ with $x_n \in F$ and $t_n \downarrow 0$ such that $\lim_{n\to\infty} \frac{x_n - x}{t_n} = w$.

$$\begin{aligned}
\langle v, w \rangle &= \lim_{n\to\infty} \langle v, \frac{x_n - x}{t_n} \rangle \\
&\leq \limsup_{n\to\infty} \langle v, \frac{x_n - x}{\|x_n - x\|} \rangle \lim_{n\to\infty} \frac{\|x_n - x\|}{t_n} \\
&= \|w\| \limsup_{n\to\infty} \langle v, \frac{x_n - x}{\|x_n - x\|} \rangle \leq 0
\end{aligned}$$

□

THEOREM 16.18. *v is a regular normal to F at x if and only if there exists a function f differentiable at x such that f has a local minimum on F at x and $\nabla f(x) = v$. In fact, we can choose f be differentiable on all of $\mathbb{R}^n$ with a global minimum at x.*

PROOF. TODO: Let $v \in \widehat{N}_F(x)$ be given. The first step to building f is to define

$$\theta_0(r) = \sup\{\langle v, y - x \rangle \mid y \in F \text{ and } \|y - x\| \leq r\}$$

It is simple to see that θ_0 is a non-decreasing function of r, $\theta_0(0) = 0$ and $\theta_0(r) \leq \|v\| \, r$ so $\lim_{r\downarrow 0} \theta_0(r) = 0$. Moreover from Proposition 16.17 we get for any $\epsilon > 0$ there exists a $\delta > 0$ such that $\langle v, y - x \rangle \leq \epsilon \|y - x\|$ for $y \in F$ and $\|y - x\| \leq \delta$. Thus for $0 \leq r \leq \delta$,

$$\begin{aligned}
\theta_0(r) &= \sup\{\langle v, y - x \rangle \mid y \in F \text{ and } \|y - x\| \leq r\} \\
&\leq \sup\{\epsilon \|y - x\| \mid y \in F \text{ and } \|y - x\| \leq r\} \\
&\leq \epsilon r
\end{aligned}$$

and thus $\theta_0(r) = o(r)$. Now let

$$h_0(y) = \langle v, y - x \rangle - \theta_0(\|y - x\|)$$

and note since $\theta_0(r) = o(r)$,

$$\lim_{w\to 0} \frac{h_0(x + w) - h_0(x) - \langle v, w \rangle}{\|w\|} = \lim_{w\to 0} \frac{-\theta_0(\|w\|)}{\|w\|} = 0$$

which shows h_0 is differentiable at x and moreover $\nabla h_0(x) = v$. □

DEFINITION 16.19. Let X be a Hilbert space, $F \subset X$ be closed and let $x \in F$ then we say that $v \in X$ is a *proximal normal vector to F at x* if there exists an $M \geq 0$ such that $\langle v, y - x \rangle \leq M \|y - x\|^2$ for all $y \in F$.

Here is a useful way to think about the definition of a proximal normal vector. First note that the set of proximal normal vectors is a cone so it suffices to think about unit proximal normal vectors. By translating F by $-x$ we can also assume that $x = 0$. Now $v \in X$ and $y \in F$ and let θ be the angle between v and y then the definition says that we can find a universal $M > 0$ such that

$$\langle v, y \rangle = \|y\| \cos\theta \leq M \|y\|^2$$

which is to say

$$\cos\theta \leq M \|y\|$$

So intuitively if y is infinitesimally close to 0 in F then $\pi/2 \leq \theta \leq 3\pi/2$.

Here is another interpretation of a proximal normal vector. A vector v is a non-zero proximal vector to F at x if and only if there exists a $z \in X \setminus F$ such that $z - x$ points in the same direction as v and x is the closest point in F to z.

Proposition 16.20. *Let F be a closed set, $x \in F$ then v is a proximal normal vector to F at x if and only if there exists $z \in X$ and $r > 0$ such that $v = r(z - x)$ and $\|z - x\| = \min_{y \in F} \|z - y\|$.*

Proof. v is a proximal normal to F at x if and only if there exists $M > 0$ such that $\langle v, y-x \rangle \leq M \|y - x\|^2$ for all $y \in F$ or equivalently $0 \leq \frac{1}{M}\langle v, x-y \rangle + \|y - x\|^2$ for all $y \in F$. Adding $\|v\|^2 / 4M^2$ to both sides this is equivalent to

$$(\|v\| / 2M)^2 \leq (\|v\| / 2M)^2 + \frac{1}{M}\langle v, x - y \rangle + \|y - x\|^2 = \|x + v/2M - y\|^2$$

for all $y \in F$. Now define $z = x + v/2M$ and $r = 2M$. ∎

It is worth noting that it may not be possible to find z in the direction of v such that x is the *only* point in F with $\|z - x\| = \min_{y \in F} \|z - y\|$ (is this true; can we get an example if so????)

TODO: Show that the any limiting normal vector is a limit of proximal normals.

2.1. KKT. For computational purposes (and in particular for numerical optimization problems) we are given a constraint set in some concrete form rather than the abstract formulation we've used. In practice it is useful to formulate a constraint set using a combination of equalities and inequalities. For the moment we specialize to the case of finite dimensions. Let X be a finite dimensional Banach space (i.e. $\mathbb{R}^d$) and suppose we are given finite sets $\mathcal{E}$ (the *equality constraints*) and $\mathcal{I}$ (the *inequality constraints*) and for each $i \in \mathcal{E} \cup \mathcal{I}$ we have a C^1 function $c_i : X \to \mathbb{R}$. Let $f : X \to \mathbb{R}$ be a C^1 function and we consider the constrained minimization problem for f with constraint set

$$F = \{x \in X \mid c_i(x) = 0 \text{ for all } i \in \mathcal{E} \text{ and } c_i(x) \geq 0 \text{ for all } i \in \mathcal{I}\}$$

It is clear from the continuity of the $c_i(x)$ that F is closed and therefore we have the first order necessary condition of Theorem 16.14 for local minimizers of f restricted to F. What we seek are conditions in terms of f and the c_i that are implied by the conditions in Theorem 16.14; for that we need to understand how $T_F(x)$ might be expressed in terms of f and the c_i. To that end, we first have the following definitions.

DEFINITION 16.21. Given a Banach space X, disjoint sets $\mathcal{E}$ and $\mathcal{I}$, functions $c_i : X \to \mathbb{R}$ for each $i \in \mathcal{E} \cup \mathcal{I}$ and the set

$$F = \{x \in X \mid c_i(x) = 0 \text{ for all } i \in \mathcal{E} \text{ and } c_i(x) \geq 0 \text{ for all } i \in \mathcal{I}\}$$

we say that a constraint c_i is *active* at $x \in F$ if either $i \in \mathcal{E}$ or $i \in \mathcal{I}$ and $c_i(x) = 0$. For each $x \in F$ we let the *active constraint set* be

$$\mathcal{A}(x) = \{i \in \mathcal{E} \cup \mathcal{I} \mid i \text{ is active at } x\}$$

Assume that the c_i are continuously differentiable, then the set of *linearized feasible directions at* x is defined to be

$$\mathcal{F}(x) = \left\{ v \in X \mid \begin{array}{l} Dc_i(x)v = 0 \text{ for all } i \in \mathcal{E} \\ Dc_i(x)v \geq 0 \text{ for all } i \in \mathcal{A}(x) \cap \mathcal{I} \end{array} \right\}$$

Note that it is trivial to see that $\mathcal{F}(x)$ is a cone. The first thing is to note that every tangent vector is a linearized feasible direction.

PROPOSITION 16.22. $T_F(x) \subset \mathcal{F}(x)$.

PROOF. Let $v \in T_F(x)$ and pick a feasible sequence $x_n \to x$ and sequence of positive numbers $t_n \to 0$ such that $x - x_n = t_n v + o(t_n)$. Applying Taylor's Theorem we can conclude that

$$\begin{aligned} c_i(x_n) &= c_i(x) + Dc_i(x)(x_n - x) + o(\|x_n - x\|) \\ &= c_i(x) + t_n Dc_i(x)v + o(t_n) \end{aligned}$$

so if $i \in \mathcal{A}(x)$ we have $c_i(x_n) = t_n Dc_i(x)v + o(t_n)$. Dividing by t_n and taking the limit as $n \to \infty$ we get $Dc_i(x)v = \lim_{n\to\infty} \frac{c_i(x_n)}{t_n}$. Thus it follows that $i \in \mathcal{E}$ implies $Df(x)v = 0$ and $i \in \mathcal{A}(x) \cap \mathcal{I}$ implies $Df(x)v \geq 0$. $\square$

In general it is not true that $T_F(x) = \mathcal{F}(x)$ yet the result that we want to demonstrate requires that this equality holds. A set of conditions that we place on the c_i that guarantees such an equality is called a *constraint qualification*; more generally a constraint qualification may a bit weaker than that and simply imply that $T_F(x)$ and $\mathcal{F}(x)$ aren't too different. There are a variety of choices of constraint qualifications we state a conceptually straightforward and useful one.

DEFINITION 16.23. A set of constraints c_i satisfies the *linearly independent constraint qualification (LICQ)* at x if the set of derivatives $\{Dc_i(x)\}$ for $i \in \mathcal{A}(x)$ is linearly independent in X^*.

The LICQ is a sufficient criterion for the equality of the tangent cone and the linearized feasible set.

EXAMPLE 16.24. Consider the set $F \subset \mathbb{R}^2$ defined by the constraints $c_1(x, y) = 1 - x^2 - (y-1)^2 \geq 0$ and $c_2(x, y) = -y \geq 0$.

TODO: Show that $T_F(x)$ is a strict subset of $\mathcal{F}(x)$.

PROPOSITION 16.25. *Let F be defined by a set of constraints $c_i(x)$ which satisfy the LICQ at x then $T_F(x) = \mathcal{F}(x)$.*

PROOF. Let $v \in \mathcal{F}(x)$, we need to show that v is a tangent vector producing a sequence $x_n \in F$ and $t_n > 0$ such that $v = x_n + o(t_k)$. By assumption the set of derivatives $Dc_i(x)$ for $i \in \mathcal{A}(x)$ is linearly independent hence is a basis for the linear span $V = \{Dc_i(x)\}_{i \in \mathcal{A}(x)}$. Let m be the cardinality of $\mathcal{A}(x)$ and $c : X \to \mathbb{R}^m$

be defined by $c(y) = (c_{i_1}(y), \ldots, c_{i_m}(y))$ where $\{i_1, \ldots, i_m\} = \mathcal{A}(x)$. Take the orthogonal complement W of V in X^* and pick a basis w_j for W and let $w : X \to \mathbb{R}^{d-m}$ be defined by $w(y) = (w_1(y), \ldots, w_{d-m}(y))$. Now define $R : X \times \mathbb{R} \to \mathbb{R}^d$ by

$$R(y,t) = \begin{bmatrix} c(y) - tDc(x)v \\ w(y - x - tv) \end{bmatrix}$$

and note that $R(x,0) = 0$. Moreover

$$D_1R(x,0)(u,0) = \begin{bmatrix} Dc(x)u \\ w(u) \end{bmatrix}$$

which is invertible by construction of w. Now we can apply the Implicit Function Theorem 15.150 to conclude that there exists an $\epsilon > 0$ and a differentiable function $f : (-\epsilon, \epsilon) \to X$ such that $f(0) = x$, $R(f(t), t) = 0$ for all $-\epsilon < t < \epsilon$ and moreover $f(t)$ is the unique solution to the equation $R(x,t) = 0$. In addition note that since we have assumed that $v \in \mathcal{F}(x)$ we have from $R(f(t), t) = 0$,

$$\begin{aligned} &c_i(f(t)) = tDc_i(x)v = 0 \text{ for } i \in \mathcal{E} \\ &c_i(f(t)) = tDc_i(x)v >= 0 \text{ for } t > 0 \text{ and } i \in \mathcal{A}(x) \cap \mathcal{I} \end{aligned}$$

and moreover by continuity we have $c_i(f(t)) > 0$ for $i \in \mathcal{I} \setminus \mathcal{A}(x)$ (here we may need to shrink ϵ for this to be true). Thus we have $f(t) \in F$.

Now pick any sequence $0 < t_n < \epsilon$ with $\lim_{n\to\infty} t_n = 0$ and define $x_n = f(t_n)$; by continuity of f and and the fact that $f(0) = x$ we have $x_n \to x$. If we Taylor expand $R(y,t)$ around $(x,0)$ we get

$$0 = f(x_n, t_n) = \begin{bmatrix} Dc(x)(x_n - x - t_n v) \\ w(x_n - x - t_n v) \end{bmatrix} + o(\|(x_n, t_n) - (x, 0)\|)$$

Since $x_n = f(t_n)$ and f is differentiable it follows that $x_n - x = O(t_n)$ and therefore $o(\|(x_n, t_n) - (x, 0)\|) = o(t_n)$. Thus by considering the first component of the vector we have $Dc(x)(x_n - x - t_n v) = o(t_n)$ and since $Dc(x)$ is invertible we get that $x_n - x - t_n v = o(t_n)$ which shows that $v \in T_F(x)$. □

LEMMA 16.26 (Farkas' Lemma). *Let A be a $d \times m$ real matrix, B be a $d \times p$ real matrix and let*

$$K = \{Ay + Bw \mid y \in \mathbb{R}^m, y \geq 0 \text{ and } w \in \mathbb{R}^p\}$$

then for any $x \in \mathbb{R}^d$ either $x \in K$ or there exists $z \in \mathbb{R}^d$ such that

$$\langle z, x\rangle < 0 A^T z \geq 0 \qquad\qquad B^T z = 0$$

but not both.

PROOF. The proof begins by showing that the alternatives are mutually exclusive. Suppose that $x \in K$ and write $x = Ay + Bw$ for appropriate y and w. If z and in the Lemma exists as well then we can calculate

$$\langle z, x\rangle = y^T A^T z + w^T B^T z \geq 0$$

which contradicts $\langle z, x\rangle < 0$.

Now we suppose that $x \notin K$ and we show how to construct z.

CLAIM 16.26.1. K is closed.

First note that

$$K = \{Ay + Bw_+ - Bw_- \mid y \in \mathbb{R}^m, y \geq 0 \text{ and } w_\pm \in \mathbb{R}^p \text{ and } w_\pm \geq 0\}$$

so we can reduce ourselves to the case in which $B = 0$. For any subset $I \subset \{1, \dots, m\}$ we let A_I be the matrix with columns of index I. For any $x \in K$ we can find a representation $x = A_I y_I$ in which I has the smallest cardinality among such representations. Note that $y_I > 0$ and the columns of A_I must be linearly independent (if not select w_I such that $A_I w_I = 0$ and then pick a constant c to be the coefficient of y_I with the smallest magnitude; it follows that $x = A_I(y_I + cw_I)$ and $y_I + cw_I$ has coefficients equal to zero). Now select a convergent sequence x_k in with $x_k \in K$ and pick minimal representations $x_k = A_{I_k} y^k_{I_k}$. By passing to a subsequence we may assume that $I_k = I$ is constant. Since A_I has full column rank $A_I^T A_I$ is invertible and y^k_I is uniquely determined to be $y^k_I = (A^T_I A_I)^{-1} A^T_I x_k$. Let $x = \lim_{k \to \infty} x_k$ and let

$$y_I = (A^T_I A_I)^{-1} A^T_I x = \lim_{k \to \infty} (A^T_I A_I)^{-1} A^T_I x_k = \lim_{k \to \infty} y^k_I$$

It follows from $y^k_I > 0$ that $y_I \geq 0$ and by continuity of A_I, $A_I y_I = \lim_{k \to \infty} A_I y^k_I = \lim_{k \to \infty} x_k = x$. Thus $x \in K$.

Now given that K is closed we can pick an $s_0 \in K$ such that $\|s_0 - x\| = d(K, x) = \inf_{s \in K} \|s - x\| > 0$. Since K is a cone we know that $\lambda s_0 \in K$ for all $\lambda > 0$ and by choice of s_0 we know that $g(\lambda) = \|\lambda s_0 - x\|^2$ has a minimum at $\lambda = 1$. Now by Theorem 16.6 we know that $0 = \frac{dg}{d\lambda} \|\lambda s_0 - x\|^2 (1) = 2\langle s_0, s_0 \Re(-) 2\langle s_0, x\rangle$ and therefore we conclude that $\langle s_0, s_0 - x\rangle = 0$. Let $s \in K$ be arbitrary and since K is convex we know that $s_0 + \theta(s - s_0) \in K$ for all $0 \leq \theta \leq 1$ and by the defining property of s_0 and algebra we get

$$\begin{aligned} 0 &\leq \|s_0 + \theta(s - s_0) - x\|^2 - \|s_0 - x\|^2 \\ &= \langle s_0 - x + \theta(s - s_0), s_0 - x + \theta(s - s_0)\rangle - \langle s_0 - x, s_0 - x\rangle \\ &= 2\theta\langle s_0 - x, s - s_0\rangle + \theta^2 \|s - s_0\|^2 \end{aligned}$$

Dividing by θ, letting $\theta \to 0$ and using $\langle s_0, s_0 - x\rangle = 0$ we get

$$\langle s_0 - x, s\rangle = \langle s_0 - x, s - s_0\rangle = \lim_{\theta \to 0^+} \langle s_0 - x, s - s_0\rangle + \frac{1}{2}\theta \|s - s_0\|^2 \geq 0$$

Now define $z = s_0 - x$. As observed $z \neq 0$ and therefore

$$\langle z, x\rangle = \langle z, s_0 - z\rangle = \langle z, s_0\rangle - \|z\|^2 = \langle s_0 - x, s_0\rangle - \|z\|^2 = -\|z\|^2 < 0$$

Writing an arbitrary $s \in K$ as $s = Ay + Bw$ with $y \geq 0$ we know that $\langle z, Ay + Bw\rangle$. In particular if we let $y = 0$ and $w = -B^T z$ then we get

$$-\left\|B^T z\right\|^2 = -\langle B^T z, B^T z\rangle = -\langle z, BB^T z\rangle \geq 0$$

hence $B^T z = 0$. Choosing $w = 0$ and $y = e_j$ for $j = 1, \dots, m$ we get

$$(A^T z)_j = \langle A^T z, e_j\rangle = \langle z, Ae_j\rangle \geq 0$$

which shows that $A^T z \geq 0$. □

Geometrically, Farkas' Lemma says that given the cone K and a point x either the point $x \in K$ or there exists a hyperplane (defined by as the set $\{w \mid \langle z, w\rangle = 0\}$) that separates x and K. The reason why such a separating hyperplane condition is relevant to our discussion is made clear in the proof of the following extremely important result.

THEOREM 16.27 (Karush-Kuhn-Tucker Conditions). *If x^* is a local minimizer of f constrained to F then if we define the Lagrangian*

$$\mathcal{L}(x,\lambda) = f(x) - \lambda \sum_{i\in\mathcal{E}\cup\mathcal{I}} \lambda_i c_i(x)$$

there exists λ^* such that

(i) $\nabla_x \mathcal{L}(x^*, \lambda^*) = 0$
(ii) $\lambda_i^* \geq 0$ *for all* $i \in \mathcal{I}$
(iii) $\lambda_i^* c_i(x^*) = 0$ *for all* $i \in \mathcal{E}\cup\mathcal{I}$

PROOF. By Theorem 16.14 we know that $\langle \nabla_x f(x^*), v\rangle \geq 0$ for all $v \in T_F(x^*)$. By the LICQ we know from Proposition 16.25 that $T_F(x^*) = \mathcal{F}(x^*)$ and thus it follows that $\langle \nabla_x f(x^*), v\rangle \geq 0$ for all $v \in \mathcal{F}(x^*)$.

On the other hand, let A be the matrix with columns $\nabla c_i(x^*)$ for $i \in \mathcal{A}(x^*)\cap\mathcal{I}$ and B be the matrix with columns $\nabla c_i(x^*)$ for $i \in \mathcal{E}$. By Farkas' Lemma 16.26 applied to the cone $\{Ay + Bw \mid y \geq 0\}$ either there exist λ_i^* for $i \in \mathcal{A}(x^*)$ such that $\nabla_x f(x^*) = \sum_{i\in\mathcal{A}(x^*)} \lambda_i^* \nabla c_i(x^*)$ and $\lambda_i^* \geq 0$ for $i \in \mathcal{A}(x^*)\cap\mathcal{I}$ or there exists z such that $\langle z, \nabla_x f(x^*)\rangle < 0$, $A^T z \geq 0$ and $B^T z = 0$. Note however that the latter two conditions on z are precisely the conditions that determine $z \in T_F(x^*)$. Thus we can conclude that $\nabla_x f(x^*) = \sum_{i\in\mathcal{A}(x^*)} \lambda_i^* \nabla c_i(x^*)$ for λ_i^* as above. For $i \in \mathcal{I}\setminus\mathcal{A}(x^*)$ we define $\lambda_i^* = 0$.

Having defined λ^* we proceed to show that it satisfies the properties required by the theorem. To see (i) we simply take the gradient $\nabla_x \mathcal{L} = \nabla_x f - \lambda \sum_{i\in\mathcal{E}\cup\mathcal{I}} \lambda_i \nabla_x c_i$ and use the definition of λ^*. Property (ii) follows immediately since we had $\lambda_i^* \geq 0$ for $i \in \mathcal{A}(x^*)\cap\mathcal{I}$ by Farkas' Lemma and $\lambda_i^* = 0$ for $i \in \mathcal{I}\setminus\mathcal{A}(x^*)$ by definition. To see (iii) we have $\lambda_i^* c_i(x^*) = 0$ for $i \in \mathcal{I}\setminus\mathcal{A}(x^*)$ since $\lambda_i^* = 0$ and $\lambda_i^* c_i(x^*) = 0$ for $i \in \mathcal{A}(x^*)$ since for such indices $c_i(x^*) = 0$ by definition of $\mathcal{A}(x^*)$. □

LEMMA 16.28. *For $i = 1,\ldots,n$ let $P_i : \mathbb{R}^d \to \mathbb{R}$ be an affine-linear function (i.e. $P(x) = a_0 + \sum_{i=1}^d a_j x_j$ for $a_0,\ldots,a_d \in \mathbb{R}$). Then either there exists $x^* \in \mathbb{R}^d$ such that $P_i(x^*) \geq 0$ for all $i = 1,\ldots,n$ or there exist non-negative real numbers $q_1 \geq 0,\ldots,q_n \geq 0$ such that $\sum_{i=1}^n q_i P_i = -1$.*

PROOF. First, note that the result holds for a given set $P_1,\ldots,P_n$ if and only if it holds for the set $q_1P_1,\ldots,q_nP_n$ with $q_1 > 0,\ldots,q_n > 0$.

The proof is by induction on d. For the case $d = 1$ by multiplying each P_i by an appropriate positive constant we may assume that each P_i is either of the form $x - a_i$, $b_i - x$ or c_i with $a_i, b_i, c_i \in \mathbb{R}$. If any of the $c_i < 0$ then define $q_i = -c_i^{-1}$ and the rest of the q to be zero and we are done. Thus we may assume that all $c_i \geq 0$. In this case the c_i provide no obstruction to finding an x^* but neither to they present an obstruction to finding q_i such that $\sum_{i=1}^d q_i P_i = -1$ (just set the appropriate $q_i = 0$). Thus we are reduced to the case in which each P_i is either of the form $x - a_i$ or $b_i - x$. Suppose there is some $b_i < a_j$, then let $q_i = q_j = (a_j - b_i)^{-1}$ and set the rest of the q to zero. It follows that

$$\sum_{i=1}^n q_i P_i(x) = (a_j - b_i)^{-1}(P_i(x) + P_j(x)) = (a_j - b_i)^{-1}(b_i - x + x - a_j) = -1$$

On the other hand, any x^* that lies between the maximum of the a_i and the minimum of the b_i is a solution to $P_i(x^*) \geq 0$ for all $i = 1, \dots, n$.

For the induction step for $d \geq 2$, writing $x = (y, t) \in \mathbb{R}^d$ and arguing as in the case $d = 1$ we can assume that each $P_i(y, t)$ is of the form $t - a_i(y)$, $b_i(y) - t$ or $c_i(y)$. Assuming y^* is fixed, by the analysis in the case $d = 1$ we know that a t^* such that $P_i(y^*, t^*) \geq 0$ exists for each $i = 1, \dots, n$ if and only if $c_i(y^*) \geq 0$ and the maximum of the $a_i(y^*)$ is less than or equal to the minimum of the $b_i(y^*)$. So if these conditions hold for any $y^* \in \mathbb{R}^{d-1}$ we construct $x^* = (y^*, t^*)$ and we are done. Suppose on the other hand that for every $y \in \mathbb{R}^{d-1}$ either $c_i(y) < 0$ or some $b_j(y) - a_k(y) < 0$, the applying the induction hypothesis to the family c_i, $b_j - a_k$ we conclude that there exist non-negative constants r_i and $r_{j,k}$ such that

$$-1 = \sum_i r_i c_i(y) + \sum_{j,k} r_{j,k}(b_j(y) - a_k(y)) = \sum_i r_i P_i(x) + \sum_{j,k} r_{j,k}(P_j(x) + P_k(x)) = \sum_i r_i P_i(x) + \sum_j \sum_k r_{j,k} P_j($$

and if we define $q_i = r_i$, $q_j = \sum_k q_{j,k}$ and $q_k = \sum_j q_{j,k}$ we are done. $\square$

TODO: At this point Tao shows how to prove a separation lemma for convex polytopes by casting the problem of finding a separating hyperplane as solving a set of affine linear inequalities. I don't yet understand this problem formulation; seems like it might be clearer using the vertex description of the polytope but need the unbounded case so I think I need to see the translation in the hyperplane description.

Absent constraint qualifications the KKT conditions are not implied at a local minimum as the following example shows

Example 16.29. Consider the problem of minimizing $-x$ subject to $(x-1)^3 + y \geq 0$, $x \geq 0$ and $y \geq 0$. TODO:

3. Algorithms for Unconstrained Optimization

We have developed criteria for detecting minimizers (mostly local) however we have not yet addressed the issue of how we might find one. There are two basic paradigms to consider: line search and trust region methods. We first consider line search.

For motivation we give an interpretation of the Frechet derivative of a real valued function on a Hilbert space X. Since $Df(x)$ is a bounded linear functional on X, we know by Reisz representation that there is a unique element of X representing the functional.

Definition 16.30. Let X be a Hilbert space, $U \subset X$ be open and let $f : U \to \mathbb{R}$ be differentiable at $x \in U$. The *gradient of f at x* is the unique element $\nabla f(x)$ of X such that $\langle \nabla f(x), v \rangle = Df(x)v$ for all $v \in X$.

We now proceed to interpret the vector $-\nabla f(x)$ as the direction of steepest decrease of the function f. Suppose that we are at a point x for which $\nabla f(x) \neq 0$. To see this, let $v \in X$ be an arbitrary unit vector in X and consider the function of a single real variable $g_v(t) = f(x + tv)$. The question we ask is what is the direction v along which f is decreasing the fastest at x. By the Chain Rule, the definition of the gradient and Taylor's Theorem we can write

$$f(x + tv) = f(x) + t\langle \nabla f(x), v \rangle + o(t)$$

which implies that $g_v'(0) = \langle \nabla f(x), v \rangle$. So what we want is to find the unit vector v which minimizes the value of $g_v'(0)$. We can write $v = \alpha \nabla f(x) / \left\| \nabla f(x) \right\| + w$ where $\langle \nabla f(x), w \rangle = 0$. Note that on the one hand $\langle \nabla f(x), v \rangle = \alpha / \left\| \nabla f(x) \right\|$ and on the other hand from $\|v\| = 1$ we see that $-1 \leq \alpha \leq 1$. Therefore it is clear that the minimum of $g_v'(0)$ occurs for $\alpha = -1$ which implies that $v = -\nabla f(x) / \left\| \nabla f(x) \right\|$. It is colloquial to say that the direction of the gradient is the *direction of steepest descent of* f. Note that the computation above shows that $g_v'(0) < 0$ precisely when $\langle \nabla f(x), v \rangle < 0$ which motivates the following defintion

DEFINITION 16.31. Let X be a Hilbert space, $U \subset X$ be open and let $f : U \to \mathbb{R}$ be differentiable at $x \in U$ we say that $v \in X$ is a *descent direction for* f *at* x if $\langle \nabla f(x), v \rangle < 0$.

TODO: Discuss gradient flow in the Hilbert space and observe how the solutions of the differential equation have limit points equal to the stationary points of f.

Armed with the idea that when we are in possesion of derivatives we can find directions along which the values of a function decreases, we seek find an iterative algorithm for minimization. The obvious idea is that if at a given point x_k we can find a descent direction (e.g. the gradient $\nabla f(x_k)$ then we should move in that direction and thereby expect that the function decreases. There are three problems to address about such an algorithm. The first issue is that the descent direction is characterized by an infinitessimal condition and therefore there is no guarantee that a finite step in that direction will result in a decrease in the function value. The second issue is that if our step sizes in the descent direction are too small asymptotically we may never reach the minimum. The third issue is that if we choose a variable descent direction, the descent direction may get increasing close to being orthogonal to the gradient in which case function values may not decrease enough to converge (note this is a non-issue is we choose the steepest descent direction). We seek conditions on the choice of step sizes and descent directions that give us convergence to a stationary point of f.

CHAPTER 17

Skorohod Space

TODO: Currently going through this.

TODO: Show that for S complete, the sup norm makes $D([0,T];S)$ into a complete non-separable metric (Banach?) space. We actually use the completeness in showing that the metric d is complete.

Question 1: In the definition of the J_1 topology on $D([0,\infty);S)$ given a time shift $\lambda(t)$ we define $d(f,g,\lambda,u) = \sup_{t\geq 0} q(f(t\wedge u), g(\lambda(t)\wedge u))$ and take the distance given the time shift as $\int_0^\infty e^{-u} d(f,g,\lambda,u)\,du$. Why is d defined this way and not as $d(f,g,\lambda,u) = \sup_{0\leq t\leq u} q(f(t), g(\lambda(t)))$? Would the latter fail to define a metric or would it fail to be complete?

Question 2: Given a cadlag function $f : [0,1] \to S$, we know that f has only countably many jump discontinuities; is there some notion of uniform continuity that can be preserved? E.g. can we say that given $\epsilon > 0$ for all points of continuity x of f there exists a uniform $\delta > 0$ such that $|x-y| < \delta$ implies $q(f(x), f(y)) < \epsilon$? I think the modulus of continuity addresses this question.

TODO: It would be convenient to treat the case of $[0,\infty]$ below?

Definition 17.1. Let S be a topological space, then for every $0 < T < \infty$ we let $D([0,T];S)$ denote the set of functions $f : [0,T] \to S$ such that for every $0 \leq t < T$ we have $f(t) = \lim_{s\to t^+} f(s)$ and for every $0 < t \leq T$ the limit $\lim_{s\to t^-} f(s)$ exists and is finite. The space $D([0,\infty);S)$ is the set of functions $f : [0,\infty)$ such that for all $t \geq 0$ we have $f(t) = \lim_{s\to t^+} f(s)$ and for all $t > 0$ we have $\lim_{s\to t^-} f(s)$ exists and is finite.

In what follows we will often use the notation $f(t-)$ to denote the limit $\lim_{s\to t^-} f(s)$.

Lemma 17.2. *If $x \in D([0,T];S)$ or $x \in D([0,\infty);S)$ then x is continuous at all but a countable number of points.*

Proof. We begin by considering the case of $x \in D([0,T];S)$. Pick an $\epsilon > 0$ and define

$$A_\epsilon = \{0 \leq t \leq T \mid r(x(t-), x(t)) \geq \epsilon\}$$

Claim 17.2.1. A_ϵ is finite.

Suppose otherwise, then by compactness of $[0,T]$ there is an accumulation point t of A_ϵ. By passing to a further subsequence we can assume that we have a sequence t_n such that $t_n \in A_\epsilon$ and either $t_n \downarrow t$ or $t_n \uparrow t$. First consider the case $t_n \downarrow t$. For every n by the existence of the left limit $x(t_n-)$ we can find t'_n such that $t_{n+1} > t'_n > t_n$ and $r(x(t_n), x(t'_n)) > \epsilon/2$. Now by construction we know that $t'_n \downarrow t$ and by right continuity we get $\lim_{n\to\infty} x(t_n) = \lim_{n\to\infty} x(t_n) = x(t)$. However this is a contradiction since we can find $N > 0$ such that $r(x(t), x(t_N)) < \epsilon/4$ and $r(x(t), x(t'_N) < \epsilon/4$ which yields $r(x(t_N), x(t'_N)) < \epsilon/2$. If $t_n \uparrow t$ we argue similarly construction a sequence t'_n such that $t_{n-1} < t'_n < t_n$ and $r(x(t_n), x(t'_n)) > \epsilon/2$. By existence of left limits, we know that $\lim_{n\to\infty} x(t'_n) = \lim_{n\to\infty} x(t_n) = x(t-)$ and this gives a contradiction as before.

Now simply note that the set of discontinuities of x is $\cup_{n=1}^{\infty} A_{1/n}$ and is therefore countable. In a similar way we see that the set of discontinuities for $x \in D([0,\infty); S)$ is countable since it is equal to the union of the discontinuities of x restricted to $[0,n]$ for $n \in \mathbb{N}$. $\square$

Definition 17.3. Let (S,r) be a metric space, define Λ denote the set of all $\lambda : [0,T] \to [0,T]$ such that λ is continuous, strictly increasing and bijective. Then for each $\lambda \in \Lambda$ we define

$$\rho(x,y,\lambda) = \sup_{t\in[0,T]} |\lambda(t) - t| \vee \sup_{t\in[0,T]} r(x(t), y(\lambda(t)))$$

and define $\rho : D([0,T]; S) \times D([0,T]; S) \to \mathbb{R}$ by

$$\rho(x,y) = \inf_{\lambda\in\Lambda} \rho(x,y,\lambda) = \inf_{\lambda\in\Lambda} \sup_{t\in[0,T]} |\lambda(t) - t| \vee \sup_{t\in[0,T]} r(x(t), y(\lambda(t)))$$

Lemma 17.4. *ρ is a metric on $D([0,T]; S)$.*

Proof. It is clear that $\rho(x,y) \geq 0$, now suppose that $\rho(x,y) = 0$. By definition we can find a sequence $\lambda_n \in \Lambda$ such that $\sup_{t\in[0,T]} |\lambda_n(t) - t| < 1/n$ and $\sup_{t\in[0,T]} r(x(t), y(\lambda_n(t))) < 1/n$. From the former inequality we see that $\lim_{n\to\infty} \lambda_n(t) = t$ and the second inequality we see that $\lim_{n\to\infty} y(\lambda_n(t)) = x(t)$. The sequence $\lambda_n(t)$ has either a decreasing subsequence or a increasing subsequence therefore passing to that subsequence and using the fact that y is cadlag we see that either $x(t) = y(t)$ or $x(t) = y(t-)$. In particular, $x(t) = y(t)$ at all continuity points of $y(t)$ and since the set of discontinuity points of y is countable it follows that the set of continuity points is dense in $[0,T]$. Therefore for every $0 \leq t < T$ we can find a sequence t_n of continuity points of y such that $t_n \downarrow t$ and therefore by right continuity of y we conclude $x(t) = y(t)$. The fact that $x(T) = y(T)$ follows from the fact that $\lambda_n(T) = T$ for all $n \in \mathbb{N}$ thus

$$y(T) = \lim_{n\to\infty} y(\lambda_n(T)) = x(T)$$

To see symmetry of ρ we first note that $\lambda \in \Lambda$ implies $\lambda^{-1} \in \Lambda$. To see this, it is first off clear that λ^{-1} exists because λ is a bijection. The fact that λ^{-1} is strictly increasing follows because if $0 \leq t < s \leq T$ and $0 \leq \lambda^{-1}(s) \leq \lambda^{-1}(t) \leq T$ then strictly increasing and bijective nature of λ tells $s \leq t$ which is contradiction. To see that λ^{-1} is continuous, pick $0 < t < T$ and let $\epsilon > 0$ be given such that $0 < \lambda^{-1}(t) - \epsilon < \lambda^{-1}(t) < \lambda^{-1}(t) + \epsilon < T$. By strict increasingness and bijectivity of λ we know that $0 < \lambda(\lambda^{-1}(t) - \epsilon) < t < \lambda(\lambda^{-1}(t) + \epsilon) < T$. Let

$$\delta < (t - \lambda(\lambda^{-1}(t) - \epsilon)) \wedge (\lambda(\lambda^{-1}(t) + \epsilon) - t)$$

and note by the strict increasingness of λ^{-1} we have

$$0 < \lambda^{-1}(t) - \epsilon < \lambda^{-1}(t - \delta) < \lambda^{-1}(t) < \lambda^{-1}(t + \delta) < \lambda^{-1}(t) + \epsilon < T$$

Now by the bijectivity of λ we know that by a change of variables

$$\sup_{0\leq t\leq T} |\lambda(t) - t| = \sup_{0\leq s\leq T} \left|s - \lambda^{-1}(s)\right|$$

$$\sup_{0\leq t\leq T} r(x(t), y(\lambda(t))) = \sup_{0\leq s\leq T} r(x(\lambda^{-1}(s)), y(s)) = \sup_{0\leq s\leq T} r(y(s), x(\lambda^{-1}(s)))$$

and therefore $\rho(x, y, \lambda) = \rho(y, x, \lambda^{-1}$. Because inversion is a bijection on Λ we then get

$$\rho(x, y) = \inf_{\lambda\in\Lambda} \rho(x, y, \lambda) = \inf_{\lambda\in\Lambda} \rho(y, x, \lambda^{-1}) = \inf_{\lambda^{-1}\in\Lambda} \rho(y, x, \lambda^{-1}) = \rho(y, x)$$

□

The metric ρ defines the Skorohod J_1 topology on the space $D([0, T]; S)$. We emphasize here that we are actually interested in the underlying topology as much as the metric space structure itself since ρ is not a complete metric.

Example 17.5. Let $f_n = \mathbf{1}_{[1/2,1/2+1/(n+2))}$ for $n > 0$ be a sequence in $D([0, 1]; \mathbb{R})$. We show that f_n is a Cauchy sequence with respect to ρ but f_n does not converge in the J_1 topology. To see that f_n is Cauchy, let $n > 0$ be given and suppose $m \geq n$. Define

$$\lambda_{n,m}(t) = \begin{cases} t & \text{if } 0 \leq t \leq 1/2 \\ \frac{n+m+2}{n+2}(t - 1/2) + 1/2 & \text{if } 1/2 \leq t < 1/2 + 1/(n + m + 2) \\ \frac{\frac{1}{2}-\frac{1}{n}}{\frac{1}{2}-\frac{1}{n+m+2}}(t - \frac{1}{2} - \frac{1}{n+m+2}) + \frac{1}{2} + \frac{1}{n} & \text{if } 1/2 + 1/(n + m + 2) \leq t \leq 1 \end{cases}$$

so that $f_{n+m}(t) = f_n(\lambda_{m+n}(t))$ for all $t \in [0, 1]$ and $\sup_{0\leq t\leq 1} |\lambda_{m+n}(t) - t| = \frac{1}{n} - \frac{1}{n+m+2} < \frac{1}{n}$ which shows $\rho(f_n, f_{n+m}) < \frac{1}{n}$.

Claim 17.5.1. If f_n converges in then it must converge to 0.

Suppose that f_n converges to some $f \in D([0, 1]; \mathbb{R})$. Then there exist $\lambda_n \in \Lambda$ such that $\lim_{n\to\infty} \sup_{0\leq t\leq 1} |\lambda_n(t) - t| = 0$ and $\lim_{n\to\infty} \sup_{0\leq t\leq 1} |f_n(t) - f(\lambda_n(t))| = 0$. Therefore for each $0 \leq t \leq 1$ that is a point of continuity of f we have $\lim_{n\to\infty} f_n(t) = \lim_{n\to\infty} f(\lambda_n(t)) = f(t)$. By definition of $f_n(t)$ and Lemma 17.2 we see that $f(t) = 0$ for all but a countable number of $0 \leq t \leq 1$. Therefore by right continuity and the existence of left limits we conclude $f(t) = 0$ for all $0 \leq t \leq 1$. Since $f(\lambda(t))$ is identically zero for all $\lambda \in \Lambda$ we conclude that $\rho(f_n, 0) = 1$ hence f_n does not converge.

An extremely fundamental and useful fact is that a continuous function on is uniformly continuous on any compact interval; so in particular a continuous function on $[0, T]$ is uniformly continuous. For a cadlag function on $[0, T]$ it is reasonable to ask whether there is some analogue of uniform continuity that can be used to control the variation over small intervals. It is worth pondering a bit what kind of control we could have. Suppose we were given an $\epsilon > 0$, we know that there are only a finite number of jumps on $[0, T]$ that are of size bigger than ϵ. If we have two adjacent jumps at s and t we don't have a continuous function on $[s, t)$ but the deviations from continuity are bounded in some sense by ϵ. Furthermore if we restrict our function to $[s, t)$ and then extend to $[s, t)$ using the left limit at t then we have an "approximately" continuous function on a compact interval and we might how to extend that to some type of uniform approximate continuity. Indeed that can be done but the language for doing so is a bit indirect and uses a modulus of continuity that is appropriate for cadlag functions.

Definition 17.6. Given $f \in D([0,T];S)$ the function

$$w(f,\delta) = \inf_{\substack{0=t_0<t_1<\cdots<t_n=T \\ \min_{1\leq i\leq n}(t_i-t_{i-1})>\delta \\ n\in\mathbb{N}}} \max_{1\leq i\leq n} \sup_{t_{i-1}\leq s<t<t_i} r(f(s),f(t))$$

is called the modulus of continuity.

Lemma 17.7. *If $f \in D([0,T];S)$ then* $\lim_{\delta\to 0} w(f,\delta) = 0$.

Proof. First note that for fixed f the function $w(f,\delta)$ is a non-decreasing function of δ. This is simply because any candidate partition $0 = t_0 < t_1 < \cdots < t_n = T$ with $\min_{1\leq i\leq n}(t_i - t_{i-1}) > \delta$ is also a candidate for any smaller value of δ. Thus the set of candidate partitions gets larger as δ shrinks and the infimum over the set of candiates shrinks.

Let $\epsilon > 0$ be given. Define $t_0 = 0$ then so long as $t_{i-1} < T$ we inductively define $t_i = \inf\{t > t_{i-1} \mid r(f(t),f(t_{i-1})) > \epsilon\} \wedge T$. We claim that there exists n such $t_n = T$. First, note that the sequence t_i is strictly increasing while $t_i < T$ by the right continuity of f. If there are an infinite number of $t_i < T$ then by compactness of $[0,T]$ there is a limit point $0 \leq t \leq T$. However the existence of the left limit $f(t-)$ says exists $\delta > 0$ such that for all $0 < t-s < \delta$ we have $r(f(s),f(t-)) < \epsilon/3$. This is a contradiction since we can find an $n > 0$ such that for all $i \geq n$ we have $t - t_i < \delta$. By definition of the t_i for any $i \geq n+1$ we can pick $t_i \leq s < t$ such that $r(f(s),f(t_{i-1})) > \epsilon$ which provides us with $0 < t-s < \delta$ and

$$r(f(s),f(t-)) > r(f(s),f(t_{i-1})) - r(f(t_{i-1}),f(t-)) > \epsilon - \epsilon/2 = \epsilon/2$$

Thus we have constructed a sequence $0 = t_0 < t_1 < \cdots < t_n = T$ such that $\max_{1\leq i\leq n} \sup_{t_{i-1}<s<t<t_i} r(f(s),f(t)) < 2\epsilon$ so it we define $\delta = \frac{1}{2}\min_{1\leq i\leq n}(t_i - t_{i-1})$ we have shown $w(f,\delta) \leq 2\epsilon$. Since ϵ was arbitrary and $w(f,\delta)$ is a non-decreasing function of δ we are done. ∎

In order to understand compactness in $D([0,\infty);S)$ we need to have a notion of equicontinuity. The basic idea is that we express equicontinuity relative to bounded intervals $[0,T]$ and we must account for the existence jumps on such an interval and particularly at T itself.

Definition 17.8. Let (S,r) be a metric space and let $f \in D([0,\infty);S)$ be given then for each $\delta > 0$ and $T > 0$ we define the *modulus of continuity*

$$w'(f,\delta,T) = \inf_{\substack{0=t_0<t_1<\cdots<t_{n-1}<T\leq t_n \\ \min_{1\leq i\leq n}(t_i-t_{i-1})>\delta \\ n\in\mathbb{N}}} \max_{1\leq i\leq n} \sup_{t_{i-1}\leq s<t<t_i} r(f(s),f(t))$$

TODO: Unify the proofs here with the case of $D([0,T];S)$.

Note that we allow the right hand endpoint of the partition $\{t_i\}$ to extend beyond T in the definition above. TODO: Presumably this makes dealing with jumps at T easier; on the other hand we don't do this in the case of $D([0,T];S)$; is it necessary to define it this way?

Lemma 17.9. *Let $f,g \in D([0,\infty);S)$ then*

(i) *$w'(f,\delta,T)$ is a non-decreasing function of δ and T*

(ii) *For every $\delta > 0$ and $T > 0$*

$$w'(f, \delta, T) \leq w'(g, \delta, T) + 2 \sup_{0 \leq t \leq T+\delta} r(f(t), g(t))$$

(iii) *For each fixed $T > 0$, $w'(f, \delta, T)$ is a right continuous function of δ and*

$$\lim_{\delta \to 0^+} w'(f, \delta, T) = 0$$

(iv) *If $f_n \in D([0, \infty); S)$ and $\lim_{n\to\infty} d(f_n, g) = 0$ then for every $\delta > 0$, $T > 0$ and $\epsilon > 0$*

$$\limsup_{n\to\infty} w'(f_n, \delta, T) \leq w'(f, \delta, T + \epsilon)$$

(v) *For fixed $\delta > 0$ and $T > 0$, $w'(f, \delta, T)$ is a Borel measurable function of f.*

PROOF. (i) is immediate from the fact that the set of partitions over which the infimum is calculated is non-decreasing with respect to set inclusion as either δ decreases or T increases. Suppose $0 < \delta < \delta'$, then if we have a partition $t_0 < \cdots < t_{n-1} < T \leq t_n$ and $t_i - t_{i-1} > \delta'$ for all $i = 1, \ldots, n$ then it follows that $t_i - t_{i-1} > \delta$ for all $i = 1, \ldots, n$ as well. Thus

$$\begin{aligned}&\{(t_0, \ldots, t_n) \mid t_0 < \cdots < t_{n-1} < T \leq t_n \text{ and } t_i - t_{i-1} > \delta'\} \subset \\ &\{(t_0, \ldots, t_n) \mid t_0 < \cdots < t_{n-1} < T \leq t_n \text{ and } t_i - t_{i-1} > \delta\}\end{aligned}$$

and taking unions over $n \in \mathbb{N}$ and infimums it follows that $w(f, \delta, T) \leq w(f, \delta', T)$. Suppose that $0 < T < T'$, given a partition $t_0 < \cdots < t_{n-1} < T \leq t_n$ with $t_i - t_{i-1} > \delta$ for all $i = 1, \ldots, n$ we may extend it to a partition $t'_0 < \cdots < t'_{n'-1} < T' \leq t'_{n'}$ with $t'_i - t'_{i-1} > \delta$ for all $i = 1, \ldots, n'$ and $t'_i = t_i$ for all $0 \leq i \leq n$; moreover we have $\max_{1 \leq i \leq n} \sup_{t_{i-1} \leq s < t < t_i} r(f(s), f(t)) \leq \max_{1 \leq i \leq n'} \sup_{t'_{i-1} \leq s < t < t'_i} r(f(s), f(t))$ and therefore

$$\begin{aligned}
w'(f, \delta, T) &= \inf_{\substack{0 = t_0 < t_1 < \cdots < t_{n-1} < T \leq t_n \\ \min_{1 \leq i \leq n}(t_i - t_{i-1}) > \delta \\ n \in \mathbb{N}}} \max_{1 \leq i \leq n} \sup_{t_{i-1} \leq s < t < t_i} r(f(s), f(t)) \\
&\leq \inf_{\substack{0 = t_0 < t_1 < \cdots < t_{n-1} < T \leq t_n \\ \min_{1 \leq i \leq n}(t_i - t_{i-1}) > \delta \\ n \in \mathbb{N}}} \max_{1 \leq i \leq n'} \sup_{t'_{i-1} \leq s < t < t'_i} r(f(s), f(t)) \\
&\leq \inf_{\substack{0 = t_0 < t_1 < \cdots < t_{n-1} < T' \leq t_n \\ \min_{1 \leq i \leq n}(t_i - t_{i-1}) > \delta \\ n \in \mathbb{N}}} \max_{1 \leq i \leq n} \sup_{t_{i-1} \leq s < t < t_i} r(f(s), f(t)) \\
&= w(f, \delta, T')
\end{aligned}$$

(ii) follows from the fact that for fixed $\delta > 0$, $T > 0$ and any valid partition $t_1 < \cdots < t_{n-1} < T \leq t_n$ we have $0 \leq t_i \leq T + \delta$ for all $1 \leq i \leq n$ and therefore by the triangle inequality

$$\begin{aligned}
\sup_{t_{i-1} \leq s < t < t_i} r(f(s), f(t)) &\leq \sup_{t_{i-1} \leq s < t < t_i} r(f(s), g(s)) + \sup_{t_{i-1} \leq s < t < t_i} r(g(s), g(t)) + \sup_{t_{i-1} \leq s < t < t_i} r(g(t), f(t)) \\
&\leq \sup_{t_{i-1} \leq s < t < t_i} r(g(s), g(t)) + 2 \sup_{0 \leq t \leq T+\delta} r(g(t), f(t))
\end{aligned}$$

Now take the maximum over $1 \leq i \leq n$ and the infimum over all partitions.

To see the right continuity in (iii), let $T > 0$, $\delta > 0$ and $\epsilon > 0$ be given and pick a partition $t_0 < \cdots < t_{n-1} < T \leq t_n$ such that $\min_{1 \leq i \leq n}(t_i - t_{i-1}) > \delta$ and $\max_{1 \leq i \leq n} \sup_{t_{i-1} \leq s < t < t_i} r(f(s), f(t)) \leq w'(f, \delta, T) + \epsilon$. Using the fact that $w'(f, \delta, T)$ is a nondecreasing function of δ, the existence of this partition shows that $w'(f, \delta', T) < w'(f, \delta, T) + \epsilon$ for all $\delta' - \delta < \frac{1}{2} \min_{1 \leq i \leq n}(t_i - t_{i-1})$. The fact that $\lim_{\delta \to 0^+} w'(f, \delta, T) = 0$ follows by the same argument as in Lemma 17.7.

To see (iv) we know that there exist λ_n such that $\lim_{n \to \infty} \sup_{0 \leq t \leq T + \delta} |\lambda_n(t) - t| = 0$ and $\lim_{n \to \infty} \sup_{0 \leq t \leq T + \delta} r(f_n(t), f(\lambda_n(t))) = 0$. Define $\delta_n = \sup_{0 \leq t \leq T} |\lambda_n(t + \delta) - \lambda_n(t)|$ and note that $\delta_n \to 0$. If we let $t_0 < \cdots t_{m-1} < \lambda(T) \leq t_m$ be a partition with $t_i - t_{i-1} > \delta_n$ for all $1 \leq i \leq m$ then it follows that $\lambda_n^{-1}(t_0) < \cdots \lambda_n^{-1}(t_{m-1}) < T \leq \lambda_n^{-1}(t_m)$ and

$$\lambda_n^{-1}(t_i) - \lambda_n^{-1}(t_{i-1}) > \delta$$

for all $1 \leq i \leq m$. Since we also have $\sup_{\lambda_n^{-1}(t_{i-1}) \leq s < t < \lambda_n^{-1}(t_i)} r(f(s), f(t)) = \sup_{t_{i-1} \leq s < t < t_i} r(f(\lambda_n(s)), f(\lambda_n(t)))$ it follows that $w'(f \circ \lambda_n, \delta, T) \leq w'(f, \delta_n, \lambda_n(T))$. Using this fact, (i), (ii) and $\lim_{n \to \infty} \lambda_n(T) = T$ we get for any $\epsilon > 0$,

$$\begin{aligned}
\limsup_{n \to \infty} w'(f_n, \delta, T) &\leq \limsup_{n \to \infty} \left[w'(f \circ \lambda_n, \delta, T) + 2 \sup_{0 \leq t \leq T + \delta} r(f_n(t), f(\lambda_n(t))) \right] \\
&= \limsup_{n \to \infty} w'(f \circ \lambda_n, \delta, T) \\
&\leq \limsup_{n \to \infty} w'(f, \delta_n, \lambda_n(T)) \\
&\leq \limsup_{n \to \infty} w'(f, \delta \vee \delta_n, T + \epsilon) \\
&= w'(f, \delta, T + \epsilon)
\end{aligned}$$

To see (v) define $w'(f, \delta, T+) = \lim_{\epsilon \to 0^+} w'(f, \delta, T + \epsilon)$, then by the fact that $w'(f_n, \delta, T)$ is non-decreasing in T and (iv) we for every $\epsilon > 0$,

$$\limsup_{n \to \infty} w'(f_n, \delta, T+) \leq \limsup_{n \to \infty} w'(f_n, \delta, T + \epsilon) \leq w'(f, \delta, T + 2\epsilon)$$

Now let $\epsilon \to 0^+$ to conclude that $w'(f, \delta, T+)$ is an upper semicontinuous function of f, hence Borel measurable (TODO: where do we show this). Now we claim that $w'(f, \delta, T) = \lim_{n \to \infty} w'(f, \delta, (T - 1/n)+)$ which shows that $w'(f, \delta, T)$ is Borel measurable.

I believe that this last statement is supported by the following (which we should add to the statement of the Lemma):

Claim 17.9.1. $w'(f, \delta, T)$ is a left continuous function of T

Let $\epsilon > 0$ be given and using the existence of the left limit $\lim_{t \to T^-} f(t)$, pick $\rho > 0$ such that $\sup_{T - \rho \leq s < t < T} r(f(s), f(t)) < \epsilon/2$. Since $w'(f, \delta, T)$ is non-decreasing in T, we know that $\lim_{t \to T^-} w'(f, \delta, t)$ exists and $w'(f, \delta, s) \leq \lim_{t \to T^-} w'(f, \delta, t)$ for all $0 \leq s < T$. We can pick a partition $t_0 < \cdots < t_{m-1} < T - \rho \leq t_m$ with $\min_{1 \leq i \leq m}(t_i - t_{i-1}) > \delta$ and

$$\max_{1 \leq i \leq m} \sup_{t_{i-1} \leq s < t < t_i} r(f(s), f(t)) < w'(f, \delta, T - \rho) + \epsilon/2 \leq \lim_{t \to T^-} w'(f, \delta, t) + \epsilon/2$$

If we have $t_m \geq T$ then we can conclude

$$w'(f, \delta, T) \leq \max_{1 \leq i \leq m} \sup_{t_{i-1} \leq s < t < t_i} r(f(s), f(t)) \leq \lim_{t \to T^-} w'(f, \delta, t) + \epsilon/2$$

It that is not the case then we modify these chosen partition to make one with which we can bound $w'(f, \delta, T)$. Specifically supposing $t_m < T$, we define $\tilde{t}_j = t_j$ for $0 \leq j < m$ and define $\tilde{t}_m = T$. It is clear from the properties of the partition $\{t_i\}$ and the fact we have only moved the rightmost endpoint of the partition further to the right that $\tilde{t}_0 < \cdots < \tilde{t}_{m-1} < T = \tilde{t}_m$, that $\min_{1 \leq i \leq m}(\tilde{t}_i - \tilde{t}_{i-1}) \geq \min_{1 \leq i \leq m}(t_i - t_{i-1}) > \delta$ and for $1 \leq i < m$

$$\sup_{\tilde{t}_{i-1} \leq s < t < \tilde{t}_i} r(f(s), f(t)) = \sup_{t_{i-1} \leq s < t < t_i} r(f(s), f(t)) \leq \sup_{t_{i-1} \leq s < t < t_i} r(f(s), f(t)) + \epsilon/2$$

Moreover we may pick a point T^* such that

$$\tilde{t}_{m-1} = t_{m-1} < T - \rho < T^* < t_m < T = \tilde{t}_m$$

and use this to break down the supremum into cases:

$$\begin{aligned}
&\sup_{\tilde{t}_{m-1} \leq s < t < \tilde{t}_m} r(f(s), f(t)) \\
&= \sup_{\tilde{t}_{m-1} \leq s < t \leq T^*} r(f(s), f(t)) \vee \sup_{T^* \leq s < t < \tilde{t}_m} r(f(s), f(t)) \vee \sup_{\tilde{t}_{m-1} \leq s < T^* < t < \tilde{t}_m} r(f(s), f(t)) \\
&\leq \sup_{t_{m-1} \leq s < t < t_m} r(f(s), f(t)) \vee \sup_{T-\rho < s < t < T} r(f(s), f(t)) \vee \\
&\sup_{t_{m-1} \leq s < T^* < t < T} [r(f(s), f(T^*)) + r(f(T^*), f(t))] \\
&\leq \sup_{t_{m-1} \leq s < t < t_m} r(f(s), f(t)) \vee \sup_{T-\rho < s < t < T} r(f(s), f(t)) \vee \\
&\left(\sup_{t_{m-1} \leq s < t < t_m} r(f(s), f(t)) + \sup_{T-\rho < s < t \leq T} r(f(s), f(t)) \right) \\
&\leq \sup_{t_{m-1} \leq s < t < t_m} r(f(s), f(t)) + \epsilon/2
\end{aligned}$$

Thus we have

$$\begin{aligned}
w'(f, \delta, T) &\leq \max_{1 \leq i \leq m} \sup_{\tilde{t}_{i-1} \leq s < t < \tilde{t}_i} r(f(s), f(t)) \\
&\leq \max_{1 \leq i \leq m} \sup_{t_{i-1} \leq s < t < t_i} r(f(s), f(t)) + \epsilon/2 \\
&\leq \lim_{t \to T^-} w'(f, \delta, t) + \epsilon
\end{aligned}$$

Now since $\epsilon > 0$ was arbitrary we let $\epsilon \to 0$ and conclude $w'(f, \delta, T) \leq \lim_{t \to T^-} w'(f, \delta, t)$. On the other hand it is clear from the fact that $w'(f, \delta, T)$ is non-decreasing in T that $\lim_{t \to T^-} w'(f, \delta, t) \leq w'(f, \delta, T)$ and therefore $\lim_{t \to T^-} w'(f, \delta, t) = w'(f, \delta, T)$. ☐

Even though the metric ρ is not complete, the underlying topology is Polish because we can define a topologically equivalent metric that is complete. To repair the incompleteness of ρ we have to be a bit more strict about the types of time changes that are allowed; more specifically we have to prevent time changes are asymptotically flat (or by considering taking the inverse of a time change prevent time changes that are asymptotically vertical). The following is a way of quantifying such a requirement.

Definition 17.10. For every $\lambda \in \Lambda$ define

$$\gamma(\lambda) = \sup_{0 \leq s < t \leq T} \left| \log \frac{\lambda(t) - \lambda(s)}{t - s} \right|$$

For every $x, y \in D([0,T]; E)$ define

$$d(x,y) = \inf_{\substack{\lambda \in \Lambda \\ \gamma(\lambda) < \infty}} \gamma(\lambda) \vee \sup_{0 \leq t \leq T} r(x(t), y(\lambda(t)))$$

The main goal is to prove that d is a metric that is topologically equivalent to ρ. Before proving that we need a few facts about γ.

LEMMA 17.11. *$\gamma(\lambda) = \gamma(\lambda^{-1})$ and $\gamma(\lambda_1 \circ \lambda_2) \leq \gamma(\lambda_1) + \gamma(\lambda_2)$.*

PROOF. These both follow from reparameterizations using the fact that λ^{-1} is a strictly increasing bijection. For the first

$$\begin{aligned}
\gamma(\lambda) &= \sup_{0 \leq s < t \leq T} \left| \log \frac{\lambda(t) - \lambda(s)}{t - s} \right| \\
&= \sup_{0 \leq \lambda^{-1}(s) < \lambda^{-1}(t) \leq T} \left| \log \frac{\lambda(\lambda^{-1}(t)) - \lambda(\lambda^{-1}(s))}{\lambda^{-1}(t) - \lambda^{-1}(s)} \right| \\
&= \sup_{0 \leq s < t \leq T} \left| \log \frac{\lambda^{-1}(t) - \lambda^{-1}(s)}{t - s} \right|
\end{aligned}$$

and for the second

$$\begin{aligned}
\gamma(\lambda) &= \sup_{0 \leq s < t \leq T} \left| \log \frac{\lambda_2(\lambda_1(t)) - \lambda_2(\lambda_1(s))}{t - s} \right| \\
&\leq \sup_{0 \leq s < t \leq T} \left| \log \frac{\lambda_2(\lambda_1(t)) - \lambda_2(\lambda_1(s))}{\lambda_1(t) - \lambda_1(s)} \right| + \sup_{0 \leq s < t \leq T} \left| \log \frac{\lambda_1(t) - \lambda_1(s)}{t - s} \right| \\
&\leq \sup_{0 \leq s < t \leq T} \left| \log \frac{\lambda_2(t) - \lambda_2(s)}{t - s} \right| + \sup_{0 \leq s < t \leq T} \left| \log \frac{\lambda_1(t) - \lambda_1(s)}{t - s} \right| \\
&= \gamma(\lambda_2) + \gamma(\lambda_1)
\end{aligned}$$

∎

LEMMA 17.12. *For all $\lambda \in \Lambda$ such that $\gamma(\lambda) < 1/2$ we have $\sup_{0 \leq t \leq T} |\lambda(t) - t| \leq 2T\gamma(\lambda)$. For all $f, g \in D([0,T]; S)$ such that $d(f,g) < 1/2$ we have $\rho(f,g) \leq 2Td(f,g)$.*

PROOF. From the inequality $1 + x \leq e^x$ we have $\log(1 + 2x) \leq 2x$ for all $x > -1/2$ and therefore for $0 < x < 1/2$ we have $\log(1 - 2x) \leq -2x < -x < 0$. Similarly we have $\log(1 - 2x) \leq -2x$ for all $x < 1/2$ and therefore for $0 < x < 1/2$ we have $\log(1 - 2x) \leq -2x < -x < 0$ for $0 < x < 1/2$. On the other hand, we see that $\frac{d}{dx}(\log(1 + 2x) - x) = \frac{2}{1+2x} - 1$ is positive for $0 < x < 1/2$ and therefore we conclude

$$\log(1 - 2x) < -x < 0 < x < \log(1 + 2x) \text{ for } 0 < x < 1/2$$

Suppose $\gamma(\lambda) < 1/2$ and let $0 < t \leq T$. By definition and the fact that $\lambda(0) = 0$ we have

$$\left| \log \frac{\lambda(t)}{t} \right| \leq \sup_{0 \leq s < t \leq T} \left| \log \frac{\lambda(t) - \lambda(s)}{t - s} \right| = \gamma(\lambda)$$

and therefore we get

$$\log(1 - 2\gamma(\lambda)) < -\gamma(\lambda) < \log \frac{\lambda(t)}{t} < \gamma(\lambda) < \log(1 + 2\gamma(\lambda))$$

and exponentiating

$$1 - 2\gamma(\lambda) < \frac{\lambda(t)}{t} < 1 + 2\gamma(\lambda)$$

and therefore $|\lambda(t) - t| < 2T\gamma(\lambda)$ for $0 < t \leq T$. Since $\lambda(0) - 0 = 0$ it follows that $\sup_{0 \leq t \leq T} |\lambda(t) - t| \leq 2T\gamma(\lambda)$.

Now suppose we have $d(f, g) < 1/2$. Let $0 < \epsilon < 1/2 - d(f, g)$ be given and select $\lambda \in \Lambda$ such that $\gamma(\lambda) < d(f, g) + \epsilon$ and $\sup_{0 \leq t \leq T} r(f(t), g(\lambda(t))) < d(f, g) + \epsilon$. By what we have just shown, we get that $\sup_{0 \leq t \leq T} |\lambda(t) - t| < 2T\gamma(\lambda) < 2T(d(f, g) + \epsilon)$ and therefore $\rho(f, g) < 2T(d(f, g) + \epsilon)$. Now let $\epsilon \to 0$. □

LEMMA 17.13. *For any $\delta > 1/4$ if $\rho(f, g) < \delta^2$ then $d(f, g) \leq 4\delta + w(f, \delta)$.*

PROOF. Let $\delta > 0$ be given, by definition of $w(f, \delta)$ choose a partition $0 = t_0 < t_1 < \cdots < t_n = T$ such that $t_i - t_{i-1} > \delta$ and $\sup_{t_{i-1} \leq s \leq t < t_i} |f(t) - f(s)| < w(f, \delta) + \delta$ for all $i = 1, \ldots, n$. Using the fact that $\rho(f, g) < \delta^2$ to pick a $\mu \in \Lambda$ such that $\sup_{0 \leq t \leq T} r(f(t), g(\mu(t))) < \delta^2$ and $\sup_{0 \leq t \leq T} |\mu(t) - t| < \delta^2$. Note that by the properties of μ we also have $\sup_{0 \leq t \leq T} r(f(\mu^{-1}(t)), g(t)) < \delta^2$.

Now we construct an appropriate λ with which to bound $d(f, g)$. Define $\lambda(t_i) = \mu(t_i)$ for each $i = 0, \ldots, n$ and extend by linear interpolation

$$\lambda(t) = \frac{t - t_{i-1}}{t_i - t_{i-1}} \mu(t_i) + \frac{t - t_i}{t_{i-1} - t_i} \mu(t_{i-1}) \text{ for } t_{i-1} \leq t \leq t_i$$

From the fact that $\mu(t_{i-1}) < \mu(t_i)$ it follows that $\lambda(t)$ is strictly increasing, $\lambda(0) = \mu(0) = 0$, $\lambda(T) = \mu(T) = T$ and λ is piecewise linear hence continuous; thus $\lambda \in \Lambda$. Moreover by the increasingness of λ and μ (and μ^{-1}) we have $t_{i-1} \leq t \leq t_i$ is equivalent to $\lambda(t_{i-1}) \leq \lambda(t) \leq \lambda(t_i)$ which is in turn equivalent to $\mu^{-1}(\lambda((t_{i-1})) = t_{i-1} \leq \mu^{-1}(\lambda(t)) \leq t_i = \mu^{-1}(\lambda(t_i))$. Thus

$$\begin{aligned} r(f(t), g(\lambda(t))) &\leq r(f(t), f(\mu^{-1}(\lambda(t)))) + r(f(\mu^{-1}(\lambda(t))) - g(\lambda(t))) \\ &\leq w(f, \delta) + \sup_{0 \leq t \leq T} f(\mu^{-1}(\lambda(t))) \leq w(f, \delta) + \delta^2 \end{aligned}$$

As for bounding $\gamma(\lambda)$ we have

$$\begin{aligned} |\lambda(t_i) - \lambda(t_{i-1}) - (t_i - t_{i-1})| &= |\mu(t_i) - \mu(t_{i-1}) - (t_i - t_{i-1})| \\ &\leq |\mu(t_i) - t_i| + |\mu(t_{i-1}) - t_{i-1}| \leq 2\delta^2 < 2\delta(t_i - t_{i-1}) \end{aligned}$$

Recalling that λ is linear on each interval $[t_{i-1}, t_i]$ we note that this inequality simply says that the slope of λ on the linear segment $[t_{i-1}, t_i]$ is in the interval $(1 - 2\delta, 1 - 2\delta)$. Thus the inequality trivially extends to any $t_{i-1} \leq s \leq t \leq t_i$. For other $s < t$ pick $i < j$ with $t_{i-1} \leq s \leq t_i$ and $t_{j-1} \leq t \leq t_j$. Then we have

$$\begin{aligned} |\lambda(t) - \lambda(s) - (t - s)| &\leq |\lambda(t) - \lambda(t_{j-1}) - (t - t_{j-1})| + \sum_{k=i}^{j-2} |\lambda(t_{k+1}) - \lambda(t_k) - (t_{k+1} - t_k)| + \\ &\quad |\lambda(t_i) - \lambda(s) - (t_i - s)| \\ &\leq 2\delta(t - t_{j-1}) + 2\delta \sum_{k=i}^{j-2} (t_{k+1} - t_k) + 2\delta(t_i - s) \\ &= 2\delta(t - s) \end{aligned}$$

and therefore

$$\log(1-2\delta) \leq \log\left(\frac{\lambda(t)-\lambda(s)}{t-s}\right) \leq \log(1+2\delta)$$

For arbitrary δ we have $\log(1+2\delta) \leq 2\delta < 4\delta$ (Theorem C.1) and by Taylor's Theorem (Lemma 1.21) we have for any $0 < \delta < 1/4$ there is a $0 < c < \delta < 1/4$ such that

$$\log(1-2\delta) = \frac{-2\delta}{1-2c} > -4\delta$$

and therefore $\gamma(\lambda) < 4\delta$. □

Now we are ready to show that d is a metric and generates the same topology as ρ.

THEOREM 17.14. *d is a metric on $D([0,T];S)$ that is topologically equivalent to ρ.*

PROOF. The fact that $d(f,g) \geq 0$ is immediate. Suppose $d(f,g)$ and pick λ_n such that $\lim_{n\to\infty} \gamma(\lambda_n) = 0$ and $\lim_{n\to\infty} \sup_{0\leq t\leq T} r(f(t), g(\lambda_n(t))) = 0$. By Lemma 17.12 we know that $\lim_{n\to\infty} \sup_{0\leq t\leq T} |\lambda_n(t) - t| = 0$ as well and therefore we can repeat the argument of Lemma 17.4 to conclude $f = g$.

To see symmetry just note that by reparametrizing and Lemma 17.11

$$\begin{aligned} d(f,g) &= \inf_{\substack{\lambda\in\Lambda \\ \gamma(\lambda)<\infty}} \gamma(\lambda) \vee \sup_{0\leq t\leq T} r(f(t), g(\lambda(t))) \\ &= \inf_{\substack{\lambda\in\Lambda \\ \gamma(\lambda)<\infty}} \gamma(\lambda^{-1}) \vee \sup_{0\leq \lambda^{-1}(t)\leq T} r(g(t), f(\lambda^{-1}(t))) = d(g,f) \end{aligned}$$

and similarly with the triangle inequality.

TODO: Write out the triangle inequality part.

To see that ρ and d define the same topology we first show that every ball in the ρ metric contains a ball in the d metric and vice versa. For general notation let $B_\rho(f,\epsilon)$ and $B_d(f,\epsilon)$ denote balls of radius ϵ centered at f in the ρ and d metric respectively. Let f and $r > 0$ be given. Define $\delta < \frac{\epsilon}{2T} \wedge \frac{1}{4}$ and then apply Lemma 17.12 to conclude that $d(f,g) \leq \delta$ implies $\rho(f,g) \leq 2T d(f,g) < \epsilon$; thus $B_d(f,\delta) \subset B_\rho(f,\epsilon)$. On the other hand for a given $\epsilon > 0$ because $\lim_{\delta\to 0} 4\delta + w(f,\delta) = 0$ we can find $0 < \delta < 1/4$ such that $4\delta + w(f,\delta) < \epsilon$ and therefore by Lemma 17.13 we have $B_\rho(f,\delta^2) \subset B_d(f,\epsilon)$.

Now let U be an open set in the topology defined by d. For every $f \in U$, by openness of U we find $\epsilon_f > 0$ such that $B_d(f,\epsilon_f) \subset U$. By the above argument we may find $\delta_f > 0$ such that $B_\rho(f,\delta_f) \subset B_d(f,\epsilon_f) \subset U$. Therefore we can write $U = \cup_{f\in U} B_\rho(f,\delta_f)$ which shows that U is an open set in the topology defined by ρ as well. It is clear that the argument is symmetric in the role of ρ and d and therefore we have shown that d and ρ are topologically equivalent metrics. □

The goal in introducing d was to provide a complete metric; a useful thing to check first is that d fixes the example which showed ρ was not a complete metric.

EXAMPLE 17.15. Here we continue the Example 17.5 by showing directly that f_n is not Cauchy in the metric d. Because f_n are indicator functions it follows that $\sup_{0\leq t\leq 1} r(f_{n+m}(t), f_n(\lambda(t)))$ is either 0 or 1. Therefore if f_n is Cauchy then we can find $\lambda_{nm}(t)$ such that $\sup_{0\leq t\leq 1} r(f_{n+m}(t), f_n(\lambda_{nm}(t))) = 0$. By definition this tells us that $\lambda_{nm}([1/2, 1/2+1/n+m+2]) = [1/2, 1/2+1/n]$ (of course $\lambda_{nm}([0,1/2]) = [0,1/2]$ and $\lambda_{nm}([1/2+1/n+m+2, 1]) = [1/2+1/n, 1]$ as well). From this fact we see that $\gamma(\lambda_{mn}) \geq \frac{n+m+2}{n+2} > 1$ which shows that $d(f_n, f_{n+m}) \geq 1$ so f_n is not Cauchy with respect to d.

Now we show that the metric d is indeed complete. In addition we show that the J_1 topology is separable which shows us that it defines a Polish space. This will allow us to apply our theory of weak convergence.

THEOREM 17.16. *Let S be a complete metric space, then the metric d on $D([0,T];S)$ is complete. Morever if S is separable then $D([0,T];S)$ is separable in the J_1 topology.*

PROOF. Let f_n be a Cauchy sequence in $D([0,T];S)$ with the metric d. As a general principle of metric spaces it suffices to show that f_n has a convergent subsequence f_{n_j}. Suppose that such a subsequence exists and converges to f. Then we may find n_j such that $d(f, f_{n_j}) < \epsilon/2$ and $d(f_m, f_{n_j}) < \epsilon/2$ for all $m \geq n_j$; it follows that in fact $d(f, f_m) < \epsilon$ for all $m \geq n_j$ and thus $f_n \to f$.

Using the Cauchy property we can find a subsequence n_j such that $d(fn_j, f_{n_{j+1}}) < 2^{-j}$. Therefore we have λ_j such that $\gamma\lambda_j < 2^{-j}$ and $\sup_{0\leq t\leq T} r(f_{n_j}(t), f_{n_{j+1}}(\lambda_j(t))) < 2^{-j}$. Moreover from Lemma 17.12 we have $\sup_{0\leq t\leq T} |\lambda_j(t) - t| < T2^{-j+1}$ for all $j \in \mathbb{N}$. Therefore for every $m, n \in \mathbb{N}$,

$$\begin{aligned}
&\sup_{0\leq t\leq T} |\lambda_{m+n+1} \circ \lambda_{m+n} \circ \cdots \circ \lambda_n(t) - \lambda_{m+n} \circ \cdots \circ \lambda_n(t)| \\
&= \sup_{0\leq t\leq T} |\lambda_{m+n+1}(t) - t| \leq T2^{-(m+n)}
\end{aligned}$$

which shows us that for fixed $n \in \mathbb{N}$ the sequence $\lambda_{m+n} \circ \ldots \circ \lambda_n$ is Cauchy in $C([0,T], \mathbb{R})$ with the sup norm. Since this latter space is complete we know that there is a limit ν_n.

CLAIM 17.16.1. $\nu_n \in \Lambda$ and $\gamma(\nu_n) \leq 2^{-n+1}$.

It is clear that ν_n is continuous as the uniform limit of continuous functions. Moreover $\nu_n(0) = \lim_{m\to\infty} \lambda_{m+n} \circ \cdots \lambda_n(0) = 0$ and similarly $\nu_n(T) = T$. As a uniform limit of strictly increasing functions we also know that ν_n is non-decreasing. To estimate $\gamma(\nu_n)$ we compute

$$\begin{aligned}
&\left|\log \frac{\lambda_{m+n} \circ \cdots \circ \lambda_n(t) - \lambda_{m+n} \circ \cdots \circ \lambda_n(s)}{t-s}\right| \\
&\leq \gamma(\lambda_{m+n} \circ \cdots \circ \lambda_n) \\
&\leq \gamma(\lambda_{m+n}) + \cdots + \gamma(\lambda_n) < 2^{-(m+n)} + \cdots + 2^{-n} < 2^{-n+1}
\end{aligned}$$

and therefore taking the limit as $m \to \infty$ we conclude

$$\left|\log \frac{\nu_n(t) - \nu_n(s)}{t-s}\right| \leq 2^{-n+1}$$

which show both that ν_n is strictly increasing (hence $\nu_n \in \Lambda$) and moreover that $\gamma(\nu_n) < 2^{-n+1}$ and the claim is shown.

Now note that $\nu_n = \nu_{n+1} \circ \lambda_n$. From this it follows that

$$\sup_{0\leq t\leq T} r(f_{n_j}(\nu_j^{-1}(t)), f_{n_{j+1}}(\nu_{j+1}^{-1}(t))) = \sup_{0\leq t\leq T} r(f_{n_j}(t), f_{n_{j+1}}(\lambda_j(t))) < 2^{-j}$$

and therefore for $j, m \in \mathbb{N}$ we have

$$\sup_{0 \leq t \leq T} r(f_{n_j}(\nu_j^{-1}(t)), f_{n_{j+m}}(\nu_{j+m}^{-1}(t))) \leq 2^{-j} + \cdots + 2^{-j-m+1} < 2^{-j+1}$$

which shows that $f_{n_j} \circ \nu_j^{-1}$ is a Cauchy sequence in $D([0,T];S)$ with respect to the sup norm. Recalling that the sup norm makes $D([0,T];S)$ into a complete (but not separable) space we can find a limit $f \in D([0,T];S)$. Thus $\sup_{0 \leq t \leq T} r(f_{n_j}(\nu_j^{-1}(t)), f(t)) \to 0$ and moreover $\gamma(\nu_j) \to 0$ which shows us that $d(f_{n_j}, f) \to 0$.

TODO: Show separability □

1. Compactness and Tightness in Skorohod Space

The entire point behind constructing the J_1 topology on $D([0,T];S)$ was to make it a Polish space so that Prohorov's Theorem can be applied to understand weak convergence of cadlag stochastic processes. The other pieces of the puzzle in applying Prohorov's Theorem are being able to prove tightness and being able to characterize limits (e.g. by understanding the finite dimensional distributions). The latter problem has little to do with topological aspects of path space (and to be honest in many cases is an impossibly difficult nut to crack). The former problem is deeply tied into the nature of compactness in path space and we now turn to the consideration of such matters.

The first thing to do is to prove an analogue of the Arzela-Ascoli theorem that characterizes relatively compact sets in $D([0,T];S)$. The proof of such a theorem ultimately rests on approximation by step functions so we first prove that certain collections of step functions are compact.

Lemma 17.17. *Let (S,r) be a metric space and let $K \subset S$. Let $\delta > 0$ and define $A(K,\delta) \subset D([0,T];S)$ be the set of functions f for which there is a partition $0 = t_0 < t_1 < \cdots < t_n = T$ with $t_i - t_{i-1} > \delta$ for $i = 1, \ldots, n$, points $x_1, \ldots, x_n \in K$ with $x_j \neq x_{j-1}$ for $j = 2, \ldots, n$ such that $f(t) = x_i$ for $t_{i-1} \leq t < t_i$. If K is compact in S then $A(K,\delta)$ is relatively compact in the J_1 topology.*

Proof. It suffices to show that every sequence in $A(K,\delta)$ has a convergent subsequence. For each f_n we by the definition of $A(K,\delta)$, let $0 = t_{0,n} < \cdots < t_{m_n,n} = T$ be a partition with $t_{j,n} - t_{j-1,n} > \delta$ for $j = 1, \ldots, m_n$ and $x_{1,n}, \ldots, x_{m_n,n}$ be elements of K such that $f_n(t) = x_{j,n}$ for $t_{j-1,n} \leq t < t_{j,n}$. Let $m = \liminf_{n\to\infty} m_n$; that is to say m is the smallest m_n for which there are an infinite number of f_n whose partitions have length m. Note that $m \leq T/\delta < \infty$. Since we are only looking for a convergence subsequence we can pass to the subsequence of f_n for which $m_n = m$ and therefore we assume that all partitions have length m.

Consider the sequence $t_{1,n}$ and the sequence $x_{1,n}$. Let $t_1 = \limsup_{n\to\infty} t_{1,n}$. We know that $\delta < t_{1,n} \leq T$ thus $t_1 \in [\delta, T]$; pick a subsequence N^1 such that $t_{1,n} \to t_1$ along N^1. We also have $x_{1,n} \in K$ along N^1 so by compactness of K there is $x_1 \in K$ and a subsequence $N^2 \subset N^1$ such that $t_{1,n} \to t_1$ and $x_{1,n} \to x_1$ along N^2. In fact we can ask for another property of the sequence of $t_{1,n}$. The function $g(t) = \left|\frac{t_1}{t}\right|$ is continuous at t_1 and equals $g(t_1) = 0$ thus $g(t_{1,n}) \to 0$ along N^2 and we may pass to a further subsequence $N^3 \subset N^2$ (which we now denote by n_j) to arrange that

(i) $\lim_{j\to\infty} t_{1,n_j} = t_1$ with $\delta \leq t_1 \leq T$
(ii) $\left|\frac{t_1}{t_{1,n_j}}\right| < \frac{1}{j}$ for $j \in \mathbb{N}$

(iii) $\lim_{j\to\infty} x_{1,n_j} = x_1$ with $x_1 \in K$

TODO: Get rid of the $k = 1$ step since the induction step is clear once we define the trivial $k = 0$ step.

If $m = 1$ we stop here (and note that we must have $t_1 = T$), otherwise we iterate to find further subsequences. To see the induction step we suppose that we have run the procedure $k < m$ times so that we have a subsequence N^k and for each $1 \leq i \leq k$ we have $i\delta \leq t_i < T$, $x_i \in K$ such that

(i) $\lim_{j\to\infty} t_{i,n_j} = t_1$
(ii) $\left|\frac{t_i - t_{i-1}}{t_{i,n_j} - t_{i-1,n_j}}\right| < \frac{1}{j}$ for $j \in \mathbb{N}$
(iii) $\lim_{j\to\infty} x_{i,n_j} = x_1$

We now let $t_{k+1} = \limsup_{j\to\infty} t_{k+1,n_j}$. We know that $(k+1)\delta \leq t_{k+1} \leq T-(m-k)\delta$ and in fact $t_{k+1} = T$ if and only if $k+1 = m$. Now we replay the argument we used for $k = 1$. We extract a subsequence $N^{1,k} \subset N^k$ such that $t_{k+1,n_j} \to t_k$ along $N^{1,k}$. Note also that $m_{n_j} \geq k+1$ along this subsequence. We use compactness of K to the get a further subsequence $N^{2,k}$ such that there is $x_{k+1} \in K$ with $x_{k+1,n_j} \to x_{k+1}$ along $N^{2,k}$. The we use continuity of $\left|\frac{t_{k+1}-t_k}{t - t_{k,n_j}}\right|$ at t_k to arrange for a final subsequence $N^{3,k}$ so that if we redefine n_j to be the subsequence $N^{3,k}$ we have $\left|\frac{t_{k+1}-t_k}{t_{k+1,n_j} - t_{k,n_j}}\right| < \frac{1}{j}$ for all $j \in \mathbb{N}$.

Define $f(t)$ to be equal to x_k on $t_{k-1} \leq t < t_k$ for $k = 1, \dots, m$ (clearly $f \in D([0,T]; S)$). We claim $f_n \to f$ in the J_1 topology along the subsequence N^m. To see this, let $\lambda_j(t_{k,n_j}) = t_k$ for $k = 1, \dots, m$ and extend by linear interpolation. This is well defined because of the fact that $t_m = T$. Moreover from the property (ii) we have $\gamma(\lambda_j) \leq \frac{1}{j}$ so that $\lim_{j\to\infty} \gamma(\lambda_j) = 0$. We also have $r(f_{n_j}(t), f(\lambda_j(t))) = r(x_{n_j,k}, x_k)$ for $t_{k-1,n_j} \leq t < t_{k,n_j}$. Thus because there are only finitely many x_k we can conclude that $\lim_{j\to\infty} \sup_{0\leq t\leq T} r(f_{n_j}(t), f(\lambda_j(t))) = 0$ and therefore $\lim_{j\to\infty} d(f_{n_j}, f) = 0$.

TODO: Can the proof be simplified if we use the metric ρ to demostrate convergence? ☐

Theorem 17.18. *Let (S, r) be a complete metric space. A set $A \subset D([0,T]; S)$ is relatively compact in the J_1 topology if and only if*

(i) *for each rational number $t \in [0,T] \cap \mathbb{Q}$ there exists a compact set $K_t \subset S$ such that $\cup_{f\in A} f(t) \subset K_t$*
(ii) $\lim_{\delta\to 0} \sup_{f\in A} w(f, \delta) = 0$

In fact when $A \subset D([0,T]; S)$ is relatively compact in the J_1 topology then there is a compact set $K \subset S$ such that $\cup_{f\in A} \cup_{0\leq t\leq T} f(t) \subset K$.

Proof. We consider $D([0,T]; S)$ as a metric space with the complete metric d. We first suppose that A is a set satisfying (i) and (ii). Since a set A is totally bounded if and only if its closure is totally bounded and a closed set of a complete metric space is complete, it suffices to show that A is totally bounded with respect to d (Theorem 1.29). By (ii) every $k \in \mathbb{N}$ we pick $0 < \delta_k < 1$ such that $\sup_{f\in A} w(f, \delta_k) < \frac{1}{k}$. Pick $m_k \in \mathbb{N}$ with $\frac{1}{m_k} < \delta_k$ and define

$$K^{(k)} = \cup_{i=1}^{Tm_k} K_{i/m_k}$$

so that $K^{(k)}$ is compact and $A_k = A(K^{(k)}, \delta_k)$ is relatively compact by Lemma 17.17. Pick $f \in A$ and since $w(f, \delta_k) < \frac{1}{k}$ we may choose a partition $0 = t_0 < t_1 < \cdots < t_n = T$ such that for $i = 1, \dots, n$ we have $t_i - t_{i-1} > \delta_k$ and $r(f(s), f(t)) < \frac{1}{k}$ for all $s, t \in [t_{i-1}, t_i)$. Since we have chose $\frac{1}{m_k} < \delta_k < t_i - t_{i-1}$ note that every interval $[\frac{j-1}{m_k}, \frac{j}{m_k}]$ has either 0 or 1 element of t_i in it. Define $g \in D([0,T]; S)$ by

$$g(t) = f(\frac{\lfloor m_k t_{i-1} \rfloor + 1}{m_k}) \text{ for } t_{i-1} \leq t < t_i \text{ and } i = 1, \dots, n$$

(this says that we define $g(t)$ on $[t_{i-1}, t_i)$ to be the value $f(j/m_k)$ where j is the smallest integer such that $t_{i-1} < j/m_k$). It is clear that $g \in A_k$ since g is a step function on the partition $\{t_j\}$ and takes values in K_{j/m_k} for appropriate $0 \leq j \leq T m_k$. Furthermore we have for $t_{i-1} \leq t < t_i$ we have

$$r(f(t), g(t)) \leq r(f(t), f(\frac{\lfloor m_k t_{i-1} \rfloor + 1}{m_k})) < \frac{1}{k}$$

which shows that $d(f,g) < \frac{1}{k}$.

Now let $\epsilon > 0$ be given and pick $\frac{1}{k} < \epsilon/2$. Since A_k is relatively compact it is totally bounded and thus there exist $g_1, \dots, g_n$ such that the balls $B(g_i, \epsilon/2)$ cover A_k. By the above argument for any $f \in A$ we can find $g \in A_k$ such that $d(f,g) < \frac{1}{k} < \epsilon/2$ and therefore by the triangle inequality the balls $B(g_i, \epsilon)$ cover A and A is totally bounded.

Now assume that A is compact in $D([0,T]; S)$. To show (i) we actually show the stronger criteria that there exists a compact set K such that $\cup_{f \in A} \cup_{0 \leq t \leq T} f(t) \subset K$. It suffices to show that $\cup_{f \in A} \cup_{0 \leq t \leq T} f(t)$ is totally bounded. Let $\delta > 0$ given and use the fact that A is totally bounded to get a set $f_1, \dots, f_n \in A$ such that $B(f_j, \delta/2)$ cover A. For each f_j we pick a partition $0 = t_0^j < t_1^j < \cdots < t_{n_j}^j = T$ with $\max_{1 \leq i \leq n_j} \sup_{t_{i-1}^j \leq s < t < t_i^j} r(f_j(s), f_j(t)) < \delta/2$. We claim that the ball $B(f_j(t_i^j); \delta)$ for $0 \leq i \leq n_j$ cover $\cup_{f \in A} \cup_{0 \leq t \leq T} f(t)$. Now let $f \in A$ and $0 \leq t \leq T$ be given and pick $1 \leq j \leq n$ such that $f \in B(f_j, \delta/2)$ and then pick $\lambda \in \Lambda$ such that $\sup_{0 \leq s \leq T} r(f(s), f_j(\lambda(s))) < \delta/2$; in particular $r(f(t), f_j(\lambda(t))) < \delta/2$. Pick $1 \leq i \leq n_j$ such that $t_{i-1}^j \leq \lambda(t) < t_i^j$ (TODO: What if $t = T$?) and then it follows that

$$r(f(t), f_j(t_i^j)) \leq r(f(t), f_j(\lambda(t))) + r(f_j(\lambda(t)), f_j(t_{i-1}^j)) < \delta$$

To see (ii) we argue by contradiction. Suppose that there exists ϵ such that $\sup_{f \in A} w(f, \delta) \geq \epsilon$ for all $\delta > 0$. In particular we can find a sequence $f_n \in A$ such that $w(f_n, 1/n) \geq \epsilon$. By relative compactness of A we can pass to a convergent subsequence $f_{n_k} \to f$ with $f \in D([0,T]; S)$. But then we have from Lemma 17.7 (actually we have only proven the relevant part in Lemma 17.9)

$$\epsilon \leq \limsup_{k \to \infty} w(f_{n_k}, \delta) \leq w(f, \delta)$$

for all $\delta > 0$ which is a contradicts $\lim_{\delta \to 0} w(f, \delta) = 0$.

Here is another attempt at an argument to show (i) that I didn't finish. It suffices to show that $\cup_{f \in A} \cup_{0 \leq t \leq T} f(t) \subset K$ is relatively compact so let $f_n(t_n)$ be a sequence with $f_n \in A$ and $0 \leq t_n \leq T$. We first make some simple reductions. By compactness of $[0,T]$ we know there is a subsequence such that t_n converges to a value $0 \leq t \leq T$ and by passing to a further subsequence we may assume that $t_n < t$ or $t_n \geq t$. By passing to a third subsequence using the compactness of A

we may assume that f_n converges to some $f \in D([0,T];S)$ in the J_1 topology. For notational cleanliness we assume that $f_n(t_n)$ represents this final subsequence. By the fact that $f_n \to f$ in the J_1 topology we may assume that there exist $\lambda_n \in \Lambda$ such that $\lim_{n\to\infty} \sup_{0\leq s\leq T} |\lambda(s) - s| = 0$. In particular, $\lim_{n\to\infty} \lambda_n(t) = t$ and by passing to a subsequence we may assume that $\lambda_n(t) < t$ or $\lambda_n(t) \geq t$ for all n along the subsequence. TODO: How to finish this off; it is clear how this shows that $\cup_{f\in A} f(t)$ is relatively compact for a fixed t???? ☐

LEMMA 17.19. *For every $t \in [0,T]$ let $\pi_t : D([0,T];S) \to S$ be the evaluation map $\pi_t(f) = f(t)$. The Borel σ-algebra on $D([0,T];S)$ is equal to $\sigma(\{\pi_t \mid t \in [0,T]\})$ and therefore $\mathcal{B}(D([0,T];S)) = D([0,T];S) \cap \mathcal{B}(\mathbb{R})^{\otimes[0,\infty)}$.*

PROOF. TODO: ☐

2. The space $D([0,\infty);S)$

For applications it is equally or perhaps more important to deal with the space $D([0,\infty);S)$ of cadlag functions on the half infinite interval. Essentially all of the results we have proven for $D([0,T];S)$ hold but there is some subtlety in getting the definitions right so that the topology on $D([0,\infty);S)$ gives the right behavior to the restriction maps $D([0,\infty);S) \to D([0,T];S)$. The problem that has to be dealt with is best illustrated with an example. Consider the function

$$f(x) = \begin{cases} 0 & \text{if } 0 \leq x < 1 \\ 1 & \text{if } 1 \leq x \end{cases}$$

and the approximating sequence

$$f_n(x) = \begin{cases} 0 & \text{if } 0 \leq x < 1 + 1/n \\ 1 & \text{if } 1 + 1/n \leq x \end{cases}$$

Our intuition is the J_1 topology on $D([0,\infty);\mathbb{R})$ should make functions close if there is a small time change that makes them uniformly close; thus we should expect that f_n converges to f. However consider the restriction of these functions to $D([0,1];\mathbb{R})$. The restriction of f has a jump discontinuity of size 1 at the endpoint $x = 1$ while the restrictions of f_n are all identically zero. Because the endpoint of the domain $[0,1]$ cannot be moved by a time change it is easy to see that f_n does not converge to f in the J_1 topology on $D([0,1];\mathbb{R})$. Another way of looking at the example is to observe that is shows the restriction map $D([0,T];S)$ to $D([0,1];S)$ for $T > 1$ is not continuous in the J_1 topology. As a side effect of this one cannot simply glue the spaces $D([0,n];S)$ for $n \in \mathbb{N}$ together (formally to create a projective limit) to get a topology on $D([0,\infty);S)$ in the same way that one can do so with $C([0,T];S)$ and $C([0,\infty);S)$.

The good news is that the example we have given illustrates the only problem that has to be dealt with: namely restricting to a point of discontinuity of an element $f \in D([0,\infty),S)$. For a given f we have already seen that there are only a countable number of discontinuities of f so in particular the set of discontinuities has Lebesgue measure zero. The trick is that if we integrate the restrictions to $D([0,T];S)$ over $0 \leq T < \infty$ then these discontinuities won't matter and we can create a metric on $D([0,\infty);S)$. While one can proceed in this fashion by using the existing metric on $D([0,T];S)$ to create a metric on $D([0,\infty);S)$ it turns out to be about as much work to just start from scratch. The advantage in doing so is that the development can be formally independent of the case of a finite interval (though we do refer to some proofs for details that are left out).

DEFINITION 17.20. Let (S, r) be a metric space, define Λ_∞ denote the set of all $\lambda : [0, \infty) \to [0, \infty)$ such that λ is continuous, strictly increasing and bijective. For $\lambda \in \Lambda_\infty$ define

$$\gamma(\lambda) = \sup_{0 \leq s < t \leq T} \left| \log \frac{\lambda(t) - \lambda(s)}{t - s} \right|$$

Then for each $\lambda \in \Lambda_\infty$ we define

$$\psi(x, y, \lambda, T) = \sup_{0 \leq t < \infty} r(x(t \wedge T), y(\lambda(t) \wedge T)) \wedge 1$$

and define $\rho : D([0, \infty); S) \times D([0, \infty); S) \to \mathbb{R}$ by

$$\rho(x, y) = \inf_{\lambda \in \Lambda_\infty} \int_0^\infty \gamma(\lambda) \vee e^{-T} \left[\psi(x, y, \lambda, T) \vee (\sup_{0 \leq u \leq T} |\lambda(t) - t| \wedge 1) \right] dT$$

$d : D([0, \infty); S) \times D([0, infty); S) \to \mathbb{R}$ by

$$d(x, y) = \inf_{\substack{\lambda \in \Lambda_\infty \\ \gamma(\lambda) < \infty}} \left[\gamma(\lambda) \vee \int_0^\infty e^{-T} \psi(x, y, \lambda, T) \, dT \right]$$

So note with this definition for each T we are restricting each pair of cadlag functions x, y to the interval $[0, T]$ but also implicitly thinking of $D([0, T])$ as being embedded in $D([0, \infty))$ by extending as a constant function.

TODO: Note the measurability of $\psi(x, y, \lambda, T)$ as a function of T.

Note that $\psi(x, y, \lambda, T)$ is not a continuous function of T for fixed x, y, λ (however I believe it is cadlag?)

EXAMPLE 17.21. Let $f = \mathbf{1}_{[1,\infty)}$ and $g = \mathbf{1}_{[2,\infty)}$ and $\lambda(1) = 2$, then $\psi(f, g, \lambda, t) = \mathbf{1}_{[0,1) \cup [2,\infty)}$.

In general when thinking of convergence relative to d (which we have not yet shown is a metric of course) we usually use the following rendering of the definition. We warn the reader that this result is used so frequently that we will rarely make explicit note of it.

LEMMA 17.22. *Let f_n and g_n be sequences in $D([0, \infty); S)$ then $\lim_{n \to \infty} d(f_n, g_n) = 0$ if and only if there exist $\lambda_n \in \Lambda_\infty$ such that $\gamma(\lambda_n) < \infty$ for all $n \in \mathbb{N}$, $\lim_{n \to \infty} \gamma(\lambda_n) = 0$ and*

$$\lim_{n \to \infty} m\{t \in [0, T] \mid \sup_{0 \leq s < \infty} r(f_n(s \wedge t), g_n(\lambda_n(s) \wedge t)) > \epsilon\} = 0$$

for all $0 < T < \infty$ and $0 < \epsilon < 1$. Moreover if $\lim_{n \to \infty} \gamma(\lambda_n) = 0$ then we also have $\lim_{n \to \infty} \sup_{0 \leq t \leq T} |\lambda_n(t) - t| = 0$ for every $T > 0$.

PROOF. Suppose that such λ_n exist then for every $0 < T < \infty$ and $0 < \epsilon < 1$, then it follows from the definition of ψ that $\sup_{0 \leq s < \infty} r(f_n(s \wedge t), g_n(\lambda_n(s) \wedge t)) > \epsilon$

if and only if $\psi(f_n, g_n, \lambda_n, t) > \epsilon$

$$\begin{aligned}\lim_{n\to\infty} \int_0^\infty e^{-t}\psi(f_n, g_n, \lambda_n, t)\, dt &\leq \lim_{n\to\infty} \int_0^T e^{-t}\psi(f_n, g_n, \lambda_n, t)\, dt + \int_T^\infty e^{-t} dt \\ &\leq \epsilon \int_0^T e^{-t}\, dt + \lim_{n\to\infty} m\{t \in [0,T] \mid \psi(f_n, g_n, \lambda_n, t) > \epsilon\} + e^{-T} \\ &\leq \epsilon + e^{-T}\end{aligned}$$

now let $T \to \infty$ and $\epsilon \to 0$ to see that $\lim_{n\to\infty} \int_0^\infty e^{-t}\psi(f_n, g_n, \lambda_n, t)\, dt = 0$.

On the other hand, suppose that $\lim_{n\to\infty} d(f_n, g_n) = 0$ and pick λ_n such that $\lim_{n\to\infty} \gamma(\lambda_n) = 0$ and $\lim_{n\to\infty} \int_0^\infty e^{-t}\psi(f_n, g_n, \lambda_n, t)\, dt = 0$. By a Markov bound we know that $\psi(f_n, g_n, \lambda_n, t) \xrightarrow{P} 0$ with respect to the exponential distribution with rate 1 on $[0,\infty)$. Thus we have for every $T > 0$ and $\epsilon > 0$,

$$\begin{aligned}m\{t \in [0,T] \mid \psi(f_n, g_n, \lambda_n, t) > \epsilon\} &= \int_0^T \mathbf{1}_{\psi(f_n,g_n,\lambda_n,t)>\epsilon}\, dt \\ &\leq e^T \int_0^\infty e^{-t}\mathbf{1}_{\psi(f_n,g_n,\lambda_n,t)>\epsilon}\, dt = e^T \mathbf{P}\{\psi(f_n, g_n, \lambda_n, t) > \epsilon\}\end{aligned}$$

and let $n \to \infty$.

By the proof of Lemma 17.12 we know that for every $T > 0$ and for $n \in \mathbb{N}$ such that $\gamma(\lambda_n) < 1/2$ we have $\sup_{0\leq t\leq T} |\lambda_n(t) - t| \leq 2T\gamma(\lambda_n)$ and therefore $\lim_{n\to\infty} \sup_{0\leq t\leq T} |\lambda_n(t) - t| = 0$ for every $T > 0$. □

Proposition 17.23. *d is a metric on $D([0,\infty); S)$.*

Proof. It is immediate from the definition that $d(f,g) \geq 0$. Suppose $d(f,g) = 0$. Pick λ_n as per Lemma 17.22 applied to constant sequences $f_n \equiv f$ and $g_n \equiv g$. Thus, $\lim_{n\to\infty} \sup_{0\leq t\leq T} |\lambda_n(t) - t| = 0$ for all $0 < T < \infty$ and $\lim_{n\to\infty} m\{0 \leq t \leq T \mid \sup_{0\leq s<\infty} r(f(s\wedge t), g(\lambda_n(s)\wedge t)) > \epsilon\} = 0$ for all $0 < T < \infty$ and $0 < \epsilon < 1$. So for any fixed $0 \leq u < \infty$ we let $0 < \epsilon < 1$ be arbitrary and pick $T > u+\epsilon$. Therefore $m\{0 \leq t \leq T \mid \sup_{0\leq s<\infty} r(f(s\wedge t), g(\lambda_n(s)\wedge t)) > \epsilon\} < T - u - \epsilon$ for sufficiently large n thus $(u+\epsilon, T) \cap \{0 \leq t \leq T \mid \sup_{0\leq s<\infty} r(f(s\wedge t), g(\lambda_n(s)\wedge t)) \leq \epsilon\} \neq \emptyset$ for sufficiently large n and therefore for sufficiently large n we may pick $u+\epsilon < w_n < T$ such that

$$r(f(u), g(\lambda_n(u)\wedge w_n)) \leq \sup_{0\leq s<\infty} r(f(s\wedge w_n), g(\lambda_n(s)\wedge w_n)) \leq \epsilon$$

We also know that $\lim_{n\to\infty} \lambda_n(u) = u$ and by passing to a subsequence N, we may assume that $\lambda_n(u) \geq u$ or $\lambda_n(u) \leq u$ along N. In particular, $\lambda_n(u) < w_n$ for large n and thus $\lim_{n\to\infty} r(f(u), g(\lambda_n(u)\wedge w_n)) = r(f(u), g(u))$ or $\lim_{n\to\infty} r(f(u), g(\lambda_n(u)\wedge w_n)) = r(f(u), g(u-))$ where the limits are taken along the subsequence N. Now we argue as in Lemma 17.4, $f(u) = g(u)$ at all continuity points of g but the discontinuity points are a countable set thus $f(u) = g(u)$ everywhere by right continuity.

Symmetry of d follows just as for the $D([0,T])$ case Theorem 17.14 by using the facts that $\gamma(\lambda) = \gamma(\lambda^{-1})$ and $\psi(f, g, \lambda, t) = \psi(g, f, \lambda^{-1}, t)$ for all $f, g \in D([0,\infty), S)$ and $0 \leq t < \infty$.

TODO: Finish the triangle inequality part... □

Because the form of the Skorohod metric is a bit opaque, some effort will go into developing different criteria for convergence in the topology. A first such result follows; note that this result is very close to saying that implies that $\lim_{n\to\infty} d(f_n, f) = 0$ precisely when the restrictions to $[0,T]$ converge in $D([0,T];S)$ whenever T is a continuity point of f. TODO: What extra work has to be done to massage the λ_n so that the equivalence is proven???

Proposition 17.24. *Let $f, f_n \in D([0,\infty);S)$ then $\lim_{n\to\infty} d(f_n, f) = 0$ if and only if there exists $\lambda_n \in \Lambda_\infty$ such that $\lim_{n\to\infty} \gamma(\lambda_n) = 0$ and*

$$\lim_{n\to\infty} \psi(f_n, f, \lambda_n, t) = 0 \text{ for all continuity points of } f$$

If $\lim_{n\to\infty} d(f_n, f) = 0$ then $\lim_{n\to\infty} f_n(t) = f(t)$ when t is a continuity point of f.

Proof. Suppose that we have λ_n as in the hypotheses. Then as the the number of discontinuity points of f is countable, $\lim_{n\to\infty} \psi(f_n, f, \lambda_n, t) = 0$ for almost every $0 \leq t < \infty$. Since $0 \leq e^{-t}\psi(f_n, f, \lambda_n, t) \leq 1$ we may apply Dominated Convergence to conclude that $\lim_{n\to\infty} \int_0^\infty e^{-t}\psi(f_n, f, \lambda_n, t)\,dt = 0$.

On the other hand suppose $\lim_{n\to\infty} d(f_n, f) = 0$ and let $0 \leq T < \infty$ be fixed.

Claim 17.24.1. There exists λ_n and $T < T_n < \infty$ such that $\lim_{n\to\infty} T_n = \infty$, $\lim_{n\to\infty} \gamma(\lambda_n) = 0$ and

$$\lim_{n\to\infty} \sup_{0\leq t<\infty} r(f_n(t \wedge T_n), f(\lambda_n(t) \wedge T_n)) = 0$$

Pick λ_n as per Lemma 17.22 applied to sequences f_n and $g_n \equiv f$. So $\lim_{n\to\infty} \gamma(\lambda_n) = 0$ and for each $k \in \mathbb{N}$ we may find a N_k such that

$$m\{0 \leq t \leq T + k \mid \sup_{0\leq s<\infty} r(f_n(s \wedge t), f(\lambda_n(s) \wedge t)) > 1/k\} < 1/k \text{ for all } n \geq N_k$$

It follows that $(T+k-1, T+k) \not\subset \{0 \leq t \leq T+k \mid \sup_{0\leq s<\infty} r(f_n(s\wedge t), f(\lambda_n(s) \wedge t)) > 1/k\}$ for every $n \geq N_k$. Now working inductively on k, for $0 \leq n < N_1$ pick T_n arbitrarily and for $k \in \mathbb{N}$ and for each $N_k \leq n < N_{k+1}$, pick a T_n with $T \leq T + k - 1 < T_n < T + k$ such that

$$\sup_{0\leq s<\infty} r(f_n(s \wedge T_n), f(\lambda_n(s) \wedge T_n)) \leq 1/k$$

By construction, $\lim_{n\to\infty} T_n = \infty$ and $\lim_{n\to\infty} \sup_{0\leq s<\infty} r(f_n(s \wedge T_n), f(\lambda_n(s) \wedge T_n)) = 0$.

Let λ_n and T_n be chosen as in the claim, and by the triangle inequality

$$\begin{aligned}\sup_{0\leq t<\infty} r(f_n(t\wedge T), f(\lambda_n(t) \wedge T)) \leq &\sup_{0\leq t<\infty} r(f_n(t \wedge T), f(\lambda_n(t\wedge T) \wedge T_n)) + \\ &\sup_{0\leq t<\infty} r(f(\lambda_n(t\wedge T) \wedge T_n), f(\lambda_n(t) \wedge T))\end{aligned}$$

We look at each of the terms on the right hand side in detail. For the first term, since $T < T_n$,

$$\begin{aligned}\sup_{0\leq t<\infty} r(f_n(t \wedge T), f(\lambda_n(t\wedge T) \wedge T_n)) &= \sup_{0\leq t\leq T} r(f_n(t), f(\lambda_n(t) \wedge T_n)) \\ &= \sup_{0\leq t\leq T} r(f_n(t\wedge T_n), f(\lambda_n(t) \wedge T_n)) \\ &\leq \sup_{0\leq t} r(f_n(t\wedge T_n), f(\lambda_n(t) \wedge T_n))\end{aligned}$$

and for the second term,

$$\begin{aligned}
&\sup_{0\leq t<\infty} r(f(\lambda_n(t\wedge T)\wedge T_n), f(\lambda_n(t)\wedge T)) \\
&= \sup_{0\leq t\leq T} r(f(\lambda_n(t)\wedge T_n), f(\lambda_n(t)\wedge T)) \vee \sup_{T\leq t<\infty} r(f(\lambda_n(T)\wedge T_n), f(\lambda_n(t)\wedge T)) \\
&= \sup_{T\leq t\leq (\lambda_n(T)\wedge T_n)\vee T} r(f(t), f(T)) \vee \sup_{\lambda_n(T)\wedge T\leq t\leq T} r(f(\lambda_n(T)\wedge T_n), f(t)) \\
&\leq \sup_{T\leq t\leq \lambda_n(T)\vee T} r(f(t), f(T)) \vee \sup_{\lambda_n(T)\wedge T\leq t\leq T} r(f(T), f(t)) + r(f(\lambda_n(T)\wedge T_n), f(T))) \\
&\leq \sup_{\lambda_n(T)\wedge T\leq t\leq \lambda_n(T)\vee T} r(f(t), f(T)) + r(f(\lambda_n(T)\wedge T_n), f(T)))
\end{aligned}$$

Thus we have

$$\begin{aligned}
&\lim_{n\to\infty}\sup_{0\leq t<\infty} r(f_n(t\wedge T), f(\lambda_n(t)\wedge T)) \leq \lim_{n\to\infty}\sup_{0\leq t<\infty} r(f_n(t\wedge T_n), f(\lambda_n(t)\wedge T_n)) + \\
&\lim_{n\to\infty}\sup_{\lambda_n(T)\wedge T\leq t\leq \lambda_n(T)\vee T} r(f(t), f(T)) + \lim_{n\to\infty} r(f(\lambda_n(T)\wedge T_n), f(T))) \\
&= 0
\end{aligned}$$

where the first limit on the right hand side is zero by the claim and the second two limits are zero because $\lim_{n\to\infty}\lambda_n(T) = T$, $T < T_n$ and f is continuous at T. $\square$

TODO: The main part of the following proposition is (iii) implies (ii). I think there is a lot of redundancy between the argument provided and the one showing the modulus of continuity converges to zero. We should probably introduce the modulus of continuity first and then pick a partition to assist with the proof. Can we phrase the argument in terms of open balls as done in the $D([0,T])$ case that I culled from Bass?

PROPOSITION 17.25. *Let $f, f_n \in D([0,\infty);S)$ then the following are equivalent*

(i) $\lim_{n\to\infty} d(f_n, f) = 0$

(ii) *there exists* $\lambda_n \in \Lambda_\infty$ *with* $\lim_{n\to\infty}\gamma(\lambda_n) = 0$ *and*
$$\lim_{n\to\infty}\sup_{0\leq t\leq T} r(f_n(t), f(\lambda_n(t))) = 0 \text{ for all } T > 0$$

(iii) *for every* $T > 0$ *there exists* $\lambda_n \in \Lambda_\infty$ *with* $\lim_{n\to\infty}\sup_{0\leq t\leq T}|\lambda_n(t) - t| = 0$ *and*
$$\lim_{n\to\infty}\sup_{0\leq t\leq T} r(f_n(t), f(\lambda_n(t))) = 0$$

(iv) *there exists* $\lambda_n \in \Lambda_\infty$ *with* $\lim_{n\to\infty}\gamma(\lambda_n) = 0$ *and*
$$\lim_{n\to\infty}\sup_{0\leq t\leq T} r(f_n(\lambda_n(t)), f(t) = 0 \text{ for all } T > 0$$

(v) *for every* $T > 0$ *there exists* $\lambda_n \in \Lambda_\infty$ *with* $\lim_{n\to\infty}\sup_{0\leq t\leq T}|\lambda_n(t) - t| = 0$ *and*
$$\lim_{n\to\infty}\sup_{0\leq t\leq T} r(f_n(\lambda_n(t)), f(t) = 0$$

PROOF. TODO: Finish

(ii) $\implies$ (iii) is immediate since any single sequence λ_n that satisfies (ii) also satisfies (iii) for all $T > 0$ by Lemma 17.22.

(iii) $\implies$ (ii) For each $N \in \mathbb{N}$, pick $\lambda_n^N \in \Lambda_\infty$ such that $\lim_{n\to\infty} \sup_{0\leq t\leq N} \left|\lambda_n^N(t) - t\right| = 0$ and $\lim_{n\to\infty} \sup_{0\leq t\leq N} r(f_n(t), f(\lambda_n^N(t))) = 0$. Since the relevant conditions are independent of the values of $\lambda_n^N(t)$ for $t > N$ we may also assume that $\lambda_n^N(t) = \lambda_n^N(N) + t - N$ for $t > N$. Now define $\tau_0^N = 0$ and inductively define

$$\tau_k^N = \inf\{t > \tau_{k-1}^N \mid r(f(t), f(\tau_{k-1}^N)) > 1/N\}$$

if $\tau_{k-1}^N < \infty$ and define $\tau_k^N = \infty$ otherwise.

Claim 17.25.1. If $\tau_k^N < \infty$ then $\tau_k^N < \tau_{k+1}^N$.

By right continuity there exists $\delta > 0$ such that $r(f(\tau_k^N), f(t)) < 1/N$ for all $\tau_k^N < t < \tau_k^N + \delta$ and therefore $\tau_{k+1}^N \geq \tau_k^N + \delta$.

Claim 17.25.2. If $\lim_{k\to\infty} \tau_k^N = \infty$.

If $\lim_{k\to\infty} \tau_k^N < \infty$ then the previous claim and by the existence of left limits we know $f(\lim_{k\to\infty} \tau_k^N) = \lim_{k\to\infty} f(\tau_k^N) < \infty$. On the other hand we know that $r(f(\tau_k^N), f(\tau_{k+1}^N)) \geq 1/N$ which shows that such a limit cannot exist.

For each $n \in \mathbb{N}$ and $k \in \mathbb{Z}_+$ define

$$u_{k,n}^N = (\lambda_n^N)^{-1}(\tau_k^N)$$

when $\tau_k^N < \infty$ and $u_{k,n}^N = \infty$ if $\tau_k^N = \infty$. Define

$$\mu_n^N(t) = \begin{cases} \tau_k^N + \left(\frac{\tau_{k+1}^N - \tau_k^N}{u_{k+1,n}^N - u_{k,n}^N}\right)(t - u_{k,n}^N) & \text{for } t \in [u_{k,n}^N, u_{k+1,n}^N) \cap [0, N] \\ \mu_n^N(N) + t - N & \text{for } t > N \end{cases}$$

where we use the convention that $\frac{\infty}{\infty} = 1$. Note that $\mu_n^N(u_{k,n}^N) = \tau_k^N = \lambda_n^N(u_{k,n}^N)$. Note also that for $u_{k,n}^N \leq t < u_{k+1,n}^N$ we have $\tau_k^N \leq \lambda_t^N(t) < \tau_{k+1}^N$ which implies $r(f(\tau_k^N), f(\lambda_n^N(t))) < 1/N$ and similarly $r(f(\tau_k^N), f(\mu_n^N(t))) < 1/N$. Thus $\sup_{0\leq t\leq T} r(f(\lambda_n^N(t)), f(\mu_n^N(t))) < 2/N$.

Since we have defined μ_n^N to be piecewise linear it follows that

$$\gamma(\mu_n^N) = \max_{0\leq k\leq ?} \left|\log \frac{\tau_{k+1}^N - \tau_k^N}{u_{k+1,n}^N - u_{k,n}^N}\right| < \infty$$

(TODO: Note that μ_n^N only has a finite number of different slopes; but that exact number doesn't seem to have a simple formula) and moreover

$$\begin{aligned} \sup_{0\leq t\leq T} r(f_n(t), f(\mu_n^N(t))) &\leq \sup_{0\leq t\leq T} r(f_n(t), f(\lambda_n^N(t))) + \sup_{0\leq t\leq T} r(f_n(\lambda_n^N(t)), f(\mu_n^N(t))) \\ &\leq \sup_{0\leq t\leq T} r(f_n(t), f(\lambda_n^N(t))) + 2/N \end{aligned}$$

Claim 17.25.3. $\lim_{n\to\infty} \gamma(\mu_n^N) = 0$.

Since $\lim_{n\to\infty} \sup_{0\leq t\leq T} |\lambda_n(t) - t| = 0$ it follows that $\lim_{n\to\infty} u_{k,n}^N = \lim_{n\to\infty} (\lambda_n^N)^{-1}(\tau_k^N) = \tau_k^N$ hence

$$\lim_{n\to\infty} \gamma(\mu_n^N) = \max_{0\leq k\leq ?} \lim_{n\to\infty} \left|\log \frac{\tau_{k+1}^N - \tau_k^N}{u_{k+1,n}^N - u_{k,n}^N}\right| = 0$$

Let $n_0 = 0$ and for each $n \in \mathbb{N}$ pick $n_N > n_{N-1}$ such that

(i) $\gamma(\mu_n^N) < 1/N$ for all $n \geq n_N$

(ii) $\sup_{0\leq t\leq N} r(f_n(t), f(\lambda_n^N(t))) < 1/N$ for all $n \geq n_N$

For $1 \leq n < n_1$ define $\hat{\lambda}_n \in \Lambda_\infty$ satisfy $\gamma(\hat{\lambda}_n) < \infty$ but otherwise be arbitrary and for $n_N \leq n < n_{N+1}$ let $\hat{\lambda}_n = \mu_n^N$. Then for any $T > 0$, $N > \lceil T \rceil$ and $n_N \leq n < n_{N+1}$ we have

$$\begin{aligned}\sup_{0\leq t\leq T} r(f_n(t), f(\hat{\lambda}_n(t))) &\leq \sup_{0\leq t\leq N} r(f_n(t), f(\hat{\lambda}_n(t))) \\ &= \sup_{0\leq t\leq N} r(f_n(t), f(\mu_n^N(t))) \leq \sup_{0\leq t\leq N} r(f_n(t), f(\lambda_n^N(t))) + 2/N < 3/N\end{aligned}$$

which shows that $\lim_{n\to\infty} \sup_{0\leq t\leq T} r(f_n(t), f(\hat{\lambda}_n(t)))$.

We now show that (iv) is equivalent to (ii). Suppose that (ii) holds then pick λ_n with $\lim_{n\to\infty} \gamma(\lambda_n) = 0$ and $\lim_{n\to\infty} \sup_{0\leq t\leq T} r(f_n(t), f(\lambda_n(t))) = 0$. Define $\tilde{\lambda}_n = \lambda_n^{-1}$ and $\lim_{n\to\infty} \gamma(\tilde{\lambda}_n) = \lim_{n\to\infty} \gamma(\lambda_n) = 0$ by Lemma 17.11. Furthermore for any $T > 0$ we have

$$\lim_{n\to\infty} \sup_{0\leq t\leq T} r(f_n(\tilde{\lambda}_n(t)), f(t)) = \lim_{n\to\infty} \sup_{0\leq t\leq \tilde{\lambda}_n(T)} r(f_n(t), f(\lambda_n(t))) = 0$$

The same argument shows that (iv) implies (ii).

We now show that (v) is equivalent to (iii). Suppose (iii) holds. Let $T > 0$ be given. Pick λ_n such that $\lim_{n\to\infty} \sup_{0\leq t\leq T+1} |\lambda_n(t) - t| = 0$ and $\lim_{n\to\infty} r(f_n(t), f(\lambda_n(t))) = 0$. Define $\tilde{\lambda}_n = \lambda_n^{-1}$. Then for any $0 < \epsilon \leq 1$ we pick $N > 0$ such that $\sup_{0\leq t\leq T+1} |\lambda_n(t) - t| < \epsilon$ (in particular $\lambda_n(T) < T+\epsilon \leq T+1$) and $r(f_n(t), f(\lambda_n(t))) < \epsilon$ for all $n \geq N$. Then for all $n \geq N$,

$$\sup_{0\leq t\leq T} \left|\tilde{\lambda}_n(t) - t\right| = \sup_{0\leq t\leq \lambda_n(T)} |t - \lambda_n(t)| \leq \sup_{0\leq t\leq T+1} |\lambda_n(t) - t| < \epsilon$$

and

$$\sup_{0\leq t\leq T} \left|f_n(\tilde{\lambda}_n(t)) - f(t)\right| = \sup_{0\leq t\leq \lambda_n(T)} |f_n(t) - f(\lambda_n(t))| \leq \sup_{0\leq t\leq T+1} |f_n(t) - f(\lambda_n(t))| < \epsilon$$

which shows (v). The same argument shows that (v) implies (iii). □

Thus far, we have several different characterizations of convergence in the J_1 topology but all of them involve uniform convergence on compact sets. There are occasions in which it is more convenient to have a characterization in terms of sequential convergence.

TODO: Back in the $C([0,\infty);S)$ section state and prove the corresponding result.

Proposition 17.26. *Let (S,r) be a metric space, $f, f_1, f_2, \ldots \in D([0,\infty);S)$ then $\lim_{n\to\infty} f_n = f$ in the J_1 topology if and only if*

(i) *For every sequence $t, t_1, t_2, \ldots \in [0,\infty)$ such that $\lim_{n\to\infty} t_n = t$ we have $\lim_{n\to\infty} r(f_n(t_n), f(t)) \wedge r(f_n(t_n), f(t-)) = 0$.*

(ii) *If we have a sequence $t, t_1, t_2, \ldots \in [0,\infty)$ such that $\lim_{n\to\infty} t_n = t$ and $\lim_{n\to\infty} r(f_n(t_n), f(t)) = 0$ then for every sequence s_n such that $t_n \leq s_n \leq T$ and $\lim_{n\to\infty} s_n = t$ we have $\lim_{n\to\infty} r(f_n(s_n), f(t)) = 0$*

(ii) *If we have a sequence $t, t_1, t_2, \ldots \in [0,\infty)$ such that $\lim_{n\to\infty} t_n = t$ and $\lim_{n\to\infty} r(f_n(t_n), f(t-)) = 0$ then for every sequence s_n such that $0 \leq s_n \leq t_n$ and $\lim_{n\to\infty} s_n = t$ we have $\lim_{n\to\infty} r(f_n(s_n), f(t-)) = 0$*

Proof. First suppose that $f_n \to f$ in the J_1 topology and we are given $t, t_1, t_2, \ldots$ with $\lim_{n\to\infty} t_n = t$. Pick a $T > 0$ such that $t_n \leq T$ for all $n \in \mathbb{N}$ and therefore $t \leq T$ as well. Then there exists λ_n such that $\lim_{n\to\infty} \sup_{0\leq t\leq T} |\lambda_n(t) - t| = 0$ and $\lim_{n\to\infty} \sup_{0\leq t\leq T} r(f_n(t), f(\lambda_n(t))) = 0$. We have by the triangle inequality

$$\begin{aligned}
&r(f_n(t_n), f(t)) \wedge r(f_n(t_n), f(t-)) \\
&\leq (r(f_n(t_n), f(\lambda_n(t_n))) + r(f(\lambda_n(t_n)), f(t))) \wedge (r(f_n(t_n), f(\lambda_n(t_n))) + r(f(\lambda_n(t_n)), f(t-))) \\
&\leq \sup_{0\leq s\leq T} r(f_n(s), f(\lambda_n(s))) + r(f(\lambda_n(t_n)), f(t)) \wedge r(f(\lambda_n(t_n)), f(t-))
\end{aligned}$$

Now take the limit at $n \to \infty$. (Note that $\lim_{n\to\infty} r(f(\lambda_n(t_n)), f(t)) \wedge r(f(\lambda_n(t_n)), f(t-)) = 0$ since $\lambda(t_n) \to t$ and for any sequence $s_n \to s$ we have $r(f(s_n), f(s+)) \wedge r(f(s_n), f(s-)) = r(f(s_n), f(s)) \wedge r(f(s_n), f(s-)) \to 0$).

To see that (ii) holds, suppose that $s_n \to t$ and $t_n \leq s_n \leq T$. The triangle inequality shows

$$\begin{aligned}
r(f(\lambda_n(t_n)), f(t)) &\leq r(f(\lambda_n(t_n)), f_n(t_n)) + r(f_n(t_n), f(t)) \\
&\leq \sup_{0\leq s\leq T} r(f_n(s), f(\lambda_n(s))) + r(f_n(t_n), f(t))
\end{aligned}$$

Thus if we assume that $\lim_{n\to\infty} r(f_n(t_n), f(t)) = 0$ then we also have $\lim_{n\to\infty} r(f(\lambda_n(t_n)), f(t)) = 0$. By monotonicity of λ_n we know that $\lambda_n(s_n) \geq \lambda_n(t_n)$ and $\lim_{n\to\infty} \lambda_n(s_n) = t$ (let $\epsilon > 0$ be given and pick N such that $\sup_{0\leq s\leq T} |\lambda(s) - s| < \epsilon/2$ and $|s_n - t| < \epsilon/2$ for $n \geq N$ then $|\lambda_n(s_n) - t| \leq |\lambda_n(s_n) - s_n| + |s_n - t| \leq \sup_{0\leq s\leq T} |\lambda_n(s) - s| + |s_n - t| < \epsilon$) and therefore $\lim_{n\to\infty} r(f(\lambda_n(s_n)), f(t)) = 0$ (one way to see this is break down into two cases; either f is continuous at t or not. In the former case $\lim_{n\to\infty} f(\lambda_n(s_n)) = f(t)$ and in the later case we know that there must exist an $N > 0$ such that $\lambda_n(t_n) \geq t$ so the same is true of $\lambda(s_n)$ and $\lim_{n\to\infty} f(\lambda_n(s_n)) = f(t+) = f(t)$ by right continuity of f; TODO: Is there a simpler argument that avoids the case breakdown?). Another application of the triangle inequality shows that

$$\begin{aligned}
r(f_n(s_n), f(t)) &\leq r(f_n(s_n), f(\lambda_n(s_n))) + r(f(\lambda_n(s_n)), f(t)) \\
&\leq \sup_{0\leq s\leq T} r(f_n(s), f(\lambda_n(s))) + r(f(\lambda_n(s_n)), f(t))
\end{aligned}$$

and therefore we can conclude that $\lim_{n\to\infty} r(f_n(s_n), f(t)) = 0$.

TODO: Show (iii) by the same type of argument.

Now we assume that (i), (ii) and (iii) hold. We are going to need a couple of observations.

Claim 17.26.1. Suppose f is continuous at s then

$$\lim_{n\to\infty} f_n(s) = \lim_{n\to\infty} f_n(s-) = f(s)$$

By (i) for all sequences s_n such that $\lim_{n\to\infty} s_n = s$ we have $\lim_{n\to\infty} f_n(s_n) = f(s)$. In particular if we pick $s_n \equiv s$ then we get $\lim_{n\to\infty} f_n(s) = f(s)$. Furthermore if we let $\epsilon > 0$ be given then for each $n \in \mathbb{N}$ we may pick $s - 1/n < s_n < s$ such that $r(f_n(s_n), f_n(s-)) < \epsilon$; but then $\lim_{n\to\infty} s_n = s$ and so by the triangle inequality and (i) we get

$$\lim_{n\to\infty} r(f_n(s-), f(s)) \leq \lim_{n\to\infty} r(f_n(s-), f_n(s_n)) + \lim_{n\to\infty} r(f_n(s_n), f(s)) \leq \epsilon$$

Since $\epsilon > 0$ was arbitrary we get $\lim_{n\to\infty} r(f_n(s-), f(s)) = 0$ as well.

Claim 17.26.2. Let $0 < t < \infty$, there exist sequences a_n and b_n with $a_n < t < b_n$, $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = t$ and a_n, b_n continuity points of f such that

$$\lim_{n\to\infty} f_n(a_n) = \lim_{n\to\infty} f_n(a_n-) = \lim_{n\to\infty} f(a_n) = f(t-)$$

and

$$\lim_{n\to\infty} f_n(b_n) = \lim_{n\to\infty} f_n(b_n-) = \lim_{n\to\infty} f(b_n) = f(t)$$

We first construct a_n. By Lemma 17.2 we know that the set of continuity points of f is the complement of a countable set hence is dense in $[0,\infty)$ (Lemma 1.17). Thus we may find a sequence α_n of continuity points such that $\alpha_m < t$ and $\lim_{n\to\infty} \alpha_n = t$. By the previous claim we know that

$$\lim_{n\to\infty} f_n(\alpha_m) = \lim_{n\to\infty} f_n(\alpha_m-) = f(\alpha_m)$$

and clearly $\lim_{m\to\infty} f(\alpha_m) = f(t-)$. By a simple modification of Proposition 1.18 we can find l_n such that $\lim_{n\to\infty} l_n = \infty$ and $\lim_{n\to\infty} f_n(\alpha_{l_n}) = \lim_{n\to\infty} f_n(\alpha_{l_n}-) = f(t-)$. Define $a_n = \alpha_{l_n}$. Since $l_n \to \infty$ it follows that $\lim_{n\to\infty} a_n = t$ and therefore $\lim_{n\to\infty} f(a_n) = f(t-)$. The construction of b_n follows by essentially the same argument.

Fix a $T > 0$. Two challenges are how to construct the time changes λ_n to align jumps in the f_n and f and then how to lift ourselves up the closeness along sequences to uniform closeness over $[0,T]$. If we pick a $n \in \mathbb{N}$ and a $0 \leq t \leq T$ and then think about which $0 \leq s < \infty$ we to to map to t under $lambda_n$ we know that we want both $\lim_{u\to s^-} r(f_n(u), f(\lambda_n(u))) = r(f_n(s-), f(t-))$ and $\lim_{u\to s^+} r(f_n(u), f(\lambda_n(u))) = r(f_n(s), f(t))$ to be small. Moreover we want to pick s to be as close to t as possible (to give ourselves a hope that λ_n converges uniformly to the identity on $[0,T]$). With this motivation in hand for fixed $0 \leq t \leq T$, $\epsilon > 0$ and $n \in \mathbb{N}$ we define the set of candidate preimages of t under our to-be-constructed λ_n:

$$\Gamma(t,n,\epsilon) = \{s \in (t-\epsilon, t+\epsilon) \cap [0,\infty) \mid r(f_n(s), f(s)) < \epsilon \text{ and } r(f_n(s-), f(s-)) < \epsilon\}$$

For each $n \in \mathbb{N}$ we want to be able to pick a candidate for every $0 \leq t \leq T$ that is as close as possible and we want that closeness to be uniform over $[0,T]$ thus we define

$$\epsilon_n = 2\inf\{\epsilon > 0 \mid \Gamma(t,n,\epsilon) \neq \emptyset \text{ for all } 0 \leq t \leq T\}$$

(the point behind multiplying by 2 is simply that we want to define ϵ_n in such a way that $\Gamma(t,n,\epsilon_n) \neq \emptyset$ and ϵ_n is within a constant factor of being optimal).

A critical point is that optimal candidate distance converges to zero (of course this still leaves us with problem of picking candidates to make a continuous λ_n but given prior constructions the reader shouldn't be too surprised that a piecewise linear λ_n will suffice provided we take the modulus of continuity of f into consideration).

Claim 17.26.3. $\lim_{n\to\infty} \epsilon_n = 0$.

The proof of the claim is by contradiction. Suppose that ϵ_n does not converge to 0. Then there exists an ϵ such that $\epsilon_{n_j} \geq \epsilon$ for some subsequence $\lim_{j\to\infty} n_j = \infty$. From the definition of ϵ_n we see that this implies there exists $0 \leq t_j \leq T$ and such that $\Gamma(t_j, n_j, \epsilon) = \emptyset$ for all $j \in \mathbb{N}$. By passing to a further subsequence we may assume that there exists $0 \leq t \leq T$ such that $\lim_{j\to\infty} t_j = t$ and passing yet another subsequence we may assume that either $t_j < t$ for all $j \in \mathbb{N}$, $t_j > t$ for all $j \in \mathbb{N}$ or $t_j = t$ for all $j \in \mathbb{N}$. Let a_n and b_n be chosen as in Claim 17.26.2. We examine each case in turn.

Case 1 : If $t_j < t$ for all $j \in \mathbb{N}$ then for each $\epsilon > 0$ we pick $\delta > 0$ such that $t - \delta < s < t$ implies $r(f(s), f(t-)) < \epsilon$. For $N > 0$ such that $t - \delta < t_j < t$ for all $j \geq N$ we also have $r(f(t_j), f(t-)) < \epsilon$ which shows $\lim_{j\to\infty} f(t_j) = f(t-)$. Moreover for all such $j \geq N$ we may also pick an $t - \delta < s_j < t_j$ such that $r(f(s_j), f(t_j-)) < \epsilon$ and therefore

$$\lim_{j\to\infty} r(f(t_j-), f(t-)) \leq \lim_{j\to\infty} r(f(t_j-), f(s_j)) + \lim_{j\to\infty} r(f(s_j), f(t-)) < 2\epsilon$$

which shows that $\lim_{j\to\infty} f(t_j-) = f(t-)$ as well. From this and Claim 17.26.2 the following facts

- $t_j - \epsilon < a_{n_j} < t_j + \epsilon$ for all j sufficiently large since for j sufficiently large we can arrange for $t - \epsilon < t_j < t$ and $t - \epsilon < a_{n_j} < t$.
- $r(f_{n_j}(a_{n_j}), f(t_j)) < \epsilon$ for j sufficiently large follows since $\lim_{j\to\infty} f_{n_j}(a_{n_j}) = \lim_{j\to\infty} f(t_j) = f(t-)$
- $r(f_{n_j}(a_{n_j}-), f(t_j-)) < \epsilon$ for k sufficiently large follows since $\lim_{j\to\infty} f_{n_j}(a_{n_j}-) = \lim_{j\to\infty} f(t_j-) = f(t-)$

which show that $a_{n_j} \in \Gamma(t_j, n_j, \epsilon)$ for j sufficiently large which is a contradiction.

Case 2 : This follows by contradiction as in Case 1 by showing $\lim_{j\to\infty} f(t_j) = \lim_{j\to\infty} f(t_j) = f(t)$ and using $\lim_{j\to\infty} f_{n_j}(b_{n_j}) = \lim_{j\to\infty} f_{n_j}(a_{n_j}-) = f(t)$ to conclude $b_{n_j} \in \Gamma(t_j, n_j, \epsilon)$ for j sufficiently large.

Case 3 : We assume that $t_j = t$ for all $j \in \mathbb{N}$. If t is a continuity point of f then as we have seen in Claim 17.26.1 that $\lim_{n\to\infty} f_n(t) = \lim_{n\to\infty} f_n(t-) = f(t)$ so $t \in \Gamma(t, n, \epsilon)$ for sufficiently large n; in particular this is also true for all n_j for j sufficiently large which is contradiction. Therefore we assume that $r(f(t), f(t-)) = \delta > 0$. Since $\lim_{n\to\infty} f_n(a_n) = f(t-)$ and $\lim_{n\to\infty} f_n(b_n) = f(t)$ we may find an $N_1 > 0$ such that

$$r(f_n(a_n), f(t-)) \vee r(f_n(b_n), f(t)) < \frac{\delta \wedge \epsilon}{2} \tag{30}$$

for all $n \geq N_1$. We can find $N_2 > 0$ such that

$$\sup_{a_n \leq s \leq b_n} r(f_n(s), f(t-)) \wedge r(f_n(s), f(t)) < \frac{\delta \wedge \epsilon}{2} \tag{31}$$

for all $n \geq N_2$ since otherwise we could find a subsequence n_k and $a_{n_k} \leq s_k \leq b_{n_k}$ such that $r(f_{n_k}(s_k), f(t-)) \wedge r(f_{n_k}(s_k), f(t)) \geq \frac{\delta\wedge\epsilon}{2}$. Since $\lim_{k\to\infty} a_{n_k} = \lim_{k\to\infty} b_{n_k} = t$ it follows that $\lim_{k\to\infty} s_k = t$ and this contradicts (i). Since $\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = t$ we can find $N_3 > 0$ such that $t-\epsilon < a_n < t < b_n < t+\epsilon$ for all $n \geq N_3$. Let $N = N_1 \vee N_2 \vee N_3$ suppose $n \geq N$ and define

$$s_n = \inf\{s > a_n \mid r(f_n(s), f(t)) < \frac{\delta \wedge \epsilon}{2}\}$$

From (30) we know that $s_n \leq b_n$. By right continuity of f_n we have $r(f_n(s_n), f(t)) \leq \frac{\delta\wedge\epsilon}{2}$. This last fact also shows that $a_n < s_n$ since otherwise

$$\begin{aligned} r(f(t), f(t-)) &\leq r(f(t), f_n(a_n)) + r(f_n(a_n), f(t-)) \\ &\leq \frac{\delta \wedge \epsilon}{2} + r(f_n(a_n), f(t-)) \vee r(f_n(b_n), f(t)) < \delta \end{aligned}$$

which is a contradiction. We have

$$t - \epsilon < a_n < s_n \leq b_n < t + \epsilon$$

Since $a_n < s_n$ we know from the definition of s_n that for all $a_n < s < s_n$ we have $r(f_n(s), f(t)) \geq \frac{\delta\wedge\epsilon}{2}$. and therefore taking the limit $r(f_n(s_n-), f(t)) \geq \frac{\delta\wedge\epsilon}{2}$. Since (31) tells that either $r(f_n(s_n-), f(t-)) < \frac{\delta\wedge\epsilon}{2}$ or $r(f_n(s_n-), f(t)) < \frac{\delta\wedge\epsilon}{2}$ we see that in fact we must have $r(f_n(s_n-), f(t-)) < \frac{\delta\wedge\epsilon}{2}$. All of these facts together show that $s_n \in \Gamma(t, n, \epsilon)$ for $n \geq N$ which is a contradiction.

Now we turn to the construction of λ_n. For each $n \in \mathbb{N}$ pick an arbitrary partition $0 = t_0^n < t_1^n < \cdots < t_{m_n-1}^n < T \leq t_{m_n}^n$ with $\min_{1\leq i\leq m_n}(t_i^n - t_{i-1}^n) > 3\epsilon_n$ and

$$\max_{1\leq i\leq m_n} \sup_{t_{i-1}^n \leq s<t<t_i^n} r(f(s), f(t)) < w'(f, 3\epsilon_n, T) + \epsilon_n$$

and define $m_n^* = m_n - 1$ if $t_{m_n}^n > T$ and $m_n^* = m_n$ if $t_{m_n}^n = T$. Define λ_n by defining $\lambda_n(0) = 0$, picking an arbitrary $s_i^n \in \Gamma(t_i^n, n, \epsilon_n)$ and defining $\lambda_n(t_i^n) = s_i^n$ for each $1 \leq i \leq m_n^*$, extending λ_n by linearity on each $[t_{i-1}^n, t_i^n]$ for $1 \leq i \leq m_n^*$ and defining $\lambda_n(t) = t_{m_n^*}^n + (t - s_{m_n^*}^n)$ for $t > s_{m_n^*}^n$. By definition, $|\lambda_n(t_i^n) - t_i^n| < \epsilon_n$ and it follows that

$$\sup_{0\leq t\leq T} |\lambda_n(t) - t| \leq \sup_{0\leq t<\infty} |\lambda_n(t) - t| \qquad \leq \max_{0\leq i\leq m_n^*} |\lambda_n(t_i^n) - t_i^n| \leq \epsilon_n$$

and therefore $\lim_{n\to\infty} \sup_{0\leq t\leq T} |\lambda_n(t) - t| = 0$. Monotonicity of λ_n follows from

$$\lambda_n(t_i^n) - \lambda(t_{i-1}^n) > (t_i^n - \epsilon_n) - (t_{i-1}^n + \epsilon_n) > \epsilon_n$$

and the fact that λ_n is linear on each $[t_{i-1}^n, t_i^n]$.

By Proposition 17.25 the result is proven by the following claim.

CLAIM 17.26.4. $\lim_{n\to\infty} \sup_{0\leq t\leq T} r(f_n(\lambda_n(t)), f(t)) = 0$.

To see the claim it suffices to show that for each $t, t_1, t_2, \cdots \in [0, T]$ such that $\lim_{n\to\infty} t_n = t$ we have $\lim_{n\to\infty} r(f_n(\lambda_n(t_n)), f(t_n)) = 0$.

TODO: Put this in a separate lemma back in the $C([0, T]; S)$ section. If $f_n, f \in C([0, T]; S)$ are such that $\lim_{n\to\infty} \sup_{0\leq t\leq T} r(f_n(t), f(t)) \neq 0$ then there exists an $\epsilon > 0$ and a sequence n_j and $0 \leq t_j \leq T$ such that $r(f_{n_j}(t_j), f(t_j)) \geq \epsilon$. By compactness of $[0, T]$ and passing to a further subsequence we can actually assume that there exists $0 \leq t \leq T$ such that $\lim_{j\to\infty} t_j = t$. Now by continuity of f we can find $N > 0$ such that $r(f(t), f(t_j)) < \epsilon/2$ for all $j \geq N$ and therefore $r(f_{n_j}(t_j), f(t)) \geq r(f_{n_j}(t_j), f(t_j)) - r(f(t_j), f(t)) \geq \epsilon/2$ for all $j \geq N$ such shows that $\limsup_{n\to\infty} r(f_n(\lambda_n(t_n)), f(t_n)) \geq \epsilon/2$.

If f is continuous at t then by the triangle inequality

$$\begin{aligned} r(f_n(\lambda_n(t_n)), f(t_n)) &\leq r(f_n(\lambda_n(t_n)), f(t)) + r(f(t), f(t_n)) \\ &= r(f_n(\lambda_n(t_n)), f(t)) \wedge r(f_n(\lambda_n(t_n)), f(t-)) + r(f(t), f(t_n)) \end{aligned}$$

and the claim follows in this case by taking the limit as $n \to \infty$ and applying (i) and the continuity of f at t.

Now if f is not continuous at t then since both $\epsilon_n \to 0$ and $w'(f, 3\epsilon_n, T) \to 0$ there exists $N > 0$ such that $\max_{1 \leq i \leq m_n} \sup_{t^n_{i-1} \leq s < t < t^n_i} r(f(s), f(t)) < r(f(t), f(t-))$ for all $n \geq N$; in particular for every $n \geq N$ there exists $0 \leq i_n \leq m^*_n$ such that $t^n_{i_n} = t$. It suffices to consider separately cases in which $t \leq t_n \leq T$ for all $n \in \mathbb{N}$ and $0 \leq t_n < t$ for all $n \in \mathbb{N}$.

If $t \leq t_n \leq T$ for all $n \in \mathbb{N}$ then since by definition of $\lambda_n(t^n_{i_n}) \in \Gamma(t, n, \epsilon_n)$ we have $r(f_n(\lambda_n(t^n_{i_n})), f(t)) < \epsilon_n$ and therefore $\lim_{n\to\infty} r(f_n(\lambda_n(t^n_{i_n})), f(t)) = 0$. However $t^n_{i_n} = t \leq t_n$, therefore by monotonicity of λ_n we have $\lambda(t^n_{i_n}) \leq \lambda_n(t_n)$. By continuity of λ_n and the fact that $\lim_{n\to\infty} t_n = t$ we know that $\lim_{n\to\infty} \lambda_n(t_n) = t$ and therefore we may apply (ii) to conclude that $\lim_{n\to\infty} r(f_n(\lambda_n(t_n)), f(t)) = 0$. By right continuity of f we get $\lim_{n\to\infty} r(f_n(\lambda_n(t_n)), f(t_n)) = 0$.

If $0 \leq t_n < t$ then since by definition of $\lambda_n(t^n_{i_n}) \in \Gamma(t, n, \epsilon_n)$ we have $r(f_n(\lambda_n(t^n_{i_n})-), f(t-)) < \epsilon_n$. Also we know that $t_n < t = t^n_{i_n}$ and therefore by monotonicity of λ_n we get $\lambda_n(t_n) < \lambda(t^n_{i_n})$; we may therefore pick $\lambda_n(t_n) < s_n \leq \lambda(t^n_{i_n})$ such that $r(f_n(s_n), f(t-)) < \epsilon_n$. By (iii) we conclude that $\lim_{n\to\infty} r(f_n(\lambda_n(t_n)), f(t-)) = 0$ and therefore since $0 \leq t_n < t$ we have $\lim_{n\to\infty} r(f_n(\lambda_n(t_n)), f(t_n)) = 0$. □

LEMMA 17.27. *For every $t \in [0, \infty)$ let $\pi_t : D([0, \infty); S) \to S$ be the evaluation map $\pi_t(f) = f(t)$.*

(i) *For arbitrary S every π_t is Borel measurable.*
(ii) *$\sigma(\pi_t; 0 \leq t < \infty) = \sigma(\pi_t; t \in D)$ for any dense subset $D \subset [0, \infty)$.*
(iii) *If S is separable then the Borel σ-algebra on $D([0, \infty); S)$ is equal to $\sigma(\{\pi_t \mid t \in [0, \infty)\})$ and therefore $\mathcal{B}(D([0, \infty); S)) = D([0, \infty); S) \cap \mathcal{B}(S)^{\otimes [0,\infty)}$.*

PROOF. Let $\psi : S \to \mathbb{R}$ be a bounded continuous function and suppose $\epsilon > 0$ and $0 \leq t < \infty$ are fixed. Define

$$\psi^\epsilon_t(f) = \frac{1}{\epsilon} \int_t^{t+\epsilon} \psi(\pi_s(f))\, ds = \frac{1}{\epsilon} \int_t^{t+\epsilon} \psi(f(s))\, ds$$

CLAIM 17.27.1. $\psi^\epsilon_t : D([0, \infty); S) \to \mathbb{R}$ is a continuous function

Suppose that $f_n \to f$ in the J_1 topology. By Proposition 17.24 we know that $f_n(t) \to f(t)$ at every continuity point of f. Since the set of discontinuity points of f is countable it has measure zero and it follows that $f_n \to f$ almost everywhere. Thus by continuity of ψ we have $\psi \circ f_n \to \psi \circ f$ almost everywhere and thus by Dominated Convergence we conclude that $\psi^\epsilon_t(f_n) \to \psi^\epsilon_t(f)$.

By the Fundamental Theorem of Calculus we have $\lim_{\epsilon\to 0^+} \psi^\epsilon_t(f) = \psi(\pi_t(f))$ which shows that $\psi \circ \pi_t$ is Borel measurable for every bounded continuous $\psi : S \to \mathbb{R}$. Since every bounded Borel measurable function is a limit of bounded continuous functions (TODO:) if follows that $\psi \circ \pi_t$ is Borel measurable for every bounded measurable $\psi : S \to \mathbb{R}$. In particular for $A \in \mathcal{B}(S)$ then $\pi_t^{-1}(A) = (\mathbf{1}_A \circ \pi_t)^{-1}(\{1\})$ is Borel measurable.

To see (ii) let $0 \leq t < \infty$ be given and pick a sequence $t_1, t_2, \dots$ with $t_n \in D \cap [t, \infty)$ such that $\lim_{n\to\infty} t_n = t$. By right continuity of the elements of $D([0, \infty); S)$ we see that $\pi_t = \lim_{n\to\infty} \pi_{t_n}$.

To see (iii) we first set about showing that open balls are in $\sigma(\pi_t; 0 \leq t < \infty)$. To show that is equivalent to showing that for fixed $g \in D([0, \infty); S)$ the function $d(\cdot, g) : D([0, \infty); S) \to \mathbb{R}$ is $\sigma(\pi_t; 0 \leq t < \infty)$-measurable, so we set to work on that. To create approximations of $d(\cdot, g)$, we assume that a partition $0 = t_0 < t_1 < \cdots < t_n < t_{n+1} = \infty$ is given and define $\eta : S^{n+1} \to D([0,\infty);S)$ by

$$\eta(x_0, \dots, x_n)(t) = \sum_{j=0}^{n} x_j \mathbf{1}_{[t_j, t_{j+1})}(t)$$

By considering the $\lambda(t) = t$ we see that $d(\eta(x_0, \dots, x_n), \eta(y_0, \dots, y_n)) \leq \max_{0 \leq j \leq n} r(x_j, y_j)$ which shows that η is continuous. Since S is separable we know that $\mathcal{B}(S^{n+1}) = \mathcal{B}(S)^{\otimes n+1}$ and therefore η is Borel measurable as well. Now let $g \in D([0,\infty);S)$ be fixed and consider $d(\eta \circ (\pi_{t_0}, \dots, \pi_{t_n}), g) : D([0,\infty);S) \to \mathbb{R}$ which is therefore $\sigma(\pi_t; 0 \leq t < \infty)$-measurable. Now apply this construction to the sequence of partitions $t^m_j = j/m$ for $j = 0, \dots, m^2$ and $m \in \mathbb{N}$, letting η_m be the m^{th} constructed embedding.

Claim 17.27.2. $\lim_{m\to\infty} d(\eta_m \circ (\pi_{t^m_0}, \dots, \pi_{t^m_{m^2}}), g) = d(\cdot, g)$.

Suppose $\epsilon > 0$ and $T > 0$ be given. By Lemma 17.9 we may find a partition $0 = t_0 < t_1 < \cdots < t_{m-1} < T \leq t_m$ such that $\max_{1\leq i\leq m} \sup_{t_{i-1}\leq s<t<t_i} r(f(s), f(t)) < \epsilon$. Pick $N \in \mathbb{N}$ such that $1/N < \min_{1\leq i\leq m}(t_i - t_{i-1}) \wedge 1/t_m$ and for every $n \geq N$ and $1 \leq i \leq m$ we let $j(i,n)$ be the integer such that $\frac{j(i,n)-1}{n} < t_i \leq \frac{j(i,n)}{n}$. Define $\lambda_n(j(i,n)/n) = t_i$ for $1 \leq i \leq m$, linearly interpolate in between and $\lambda_n(t) = t_m + (t - j(m,n)/n)$ for $t > j(m,n)/n$. For any $0 \leq t \leq T$, we pick $k \in \mathbb{N}$ such that $k - 1/n \leq t < k/n$ and we have by definition $f_n(t) = f((k-1)/n)$. Furthermore, since $0 \leq t \leq T$, we may pick $1 \leq i \leq m$ such that $j(i-1,n) \leq k-1 < j(i,n)$ and it follows that $j(i-1,n)/n \leq t < j(i,n)/n$ hence $t_{i-1} \leq \lambda_n(t)) < t_i$ and also

$$t_{i-1} \leq \frac{j(i-1,n)}{n} \leq \frac{k-1}{n} \leq \frac{j(i,n)-1}{n} < t_i$$

Therefore we conclude $r(f_n(t), f(\lambda_n(t))) \leq \sup_{t_{i-1}\leq s<t<t_i} r(f(s), f(t)) < \epsilon$ for all $n \geq N$ and we have shown that $\lim_{n\to\infty} \sup_{0\leq t\leq T} r(f_n(t), f(\lambda_n(t)) = 0$. Since λ_n is defined to be piecewise linear it follows that $\sup_{0\leq t\leq T} |\lambda_n(t) - t| = \max_{0\leq m} |\lambda_n(j(i,n)/n) - j(i,n)/n| = \max_{0\leq m} |t_i - j(i,n)/n|$. It is clear by the definition of $j(i,n)$ that $\lim_{n\to\infty} j(i,n)/n = t_i$ and thus it follows that $\sup_{0\leq t\leq T} |\lambda_n(t) - t| \to 0$. Proposition 17.25 shows this is sufficient to show convergence (recall that we may pick λ_n depending on T according to that Lemma).

By the claim, we see that for every $g \in D([0,\infty);S)$ the function $d(\cdot, g)$ is $\sigma(\pi_t; 0 \leq t < \infty)$-measurable from which it follows that every open ball $B(g,r) \in \sigma(\pi_t; 0 \leq t < \infty)$. Since S is separable, every open set is a countable union of open balls and therefore every open set belongs to $\sigma(\pi_t; 0 \leq t < \infty)$ and we are done. □

Theorem 17.28. *Let (S,r) be a complete metric space. A set $A \subset D([0,\infty);S)$ is relatively compact in the J_1 topology if and only if*

(i) *for each rational number $t \in [0,\infty) \cap \mathbb{Q}$ there exists a compact set $K_t \subset S$ such that $\cup_{f\in A} f(t) \subset K_t$*
(ii) *For all $T > 0$, $\lim_{\delta\to 0} \sup_{f\in A} w'(f, \delta, T) = 0$*

Proof. TODO: Not significantly different than the case of $D([0,T];S)$. □

Theorem 17.29. *A set of Borel probability measures μ_α on $D([0,\infty);S)$ is tight if and only if*

(i) *For every $\epsilon > 0$ and $t \in [0,\infty) \cap \mathbb{Q}$ there exists a compact set $K_{\epsilon,t} \subset S$ such that $\sup_\alpha \mu_\alpha(f(t) \in K_{\epsilon,t}) > 1 - \epsilon$.*

(ii) *For every* $\lambda > 0$ *and* $T > 0$

$$\lim_{\delta \to 0} \sup_\alpha \mu_\alpha(w'(f, \delta, T) \geq \lambda) = 0$$

Proof. Let μ_α be tight. Let $\epsilon > 0$ be given and pick $K \subset D^\infty([0,\infty); S)$ compact with $\mu_\alpha(K) > 1 - \epsilon/2$ for all α. Then by Theorem 17.28 we know that for every $t \in [0,\infty) \cap \mathbb{Q}$ there exists a compact set $K_{\epsilon,t} \subset S$ such that $f(t) \in K_{\epsilon,t}$ for every $f \in K$ and therefore by a union bound

$$\sup_\alpha \mu_\alpha(f(t) \in K_{\epsilon,t}) \geq \sup_\alpha \mu_\alpha(K) \geq 1 - \epsilon/2 > 1 - \epsilon$$

Similarly applying Theorem 17.28 we know that for every $T > 0$ and $\lambda > 0$ there exists $\delta > 0$ such that $\sup_{f \in K} w'(f, \delta, T) < \lambda$. Therefore $\{f \mid w'(f,\delta,T) \geq \lambda\} \subset K^c$ and by a union bound applied for every α we have $\sup_\alpha \mu_\alpha(w'(f,\delta,T) \geq \lambda) \leq \mu_\alpha(K^c) < \epsilon$. Since $w'(f,\delta,T)$ is a non-decreasing function of δ this it follows that for all $0 < \rho \leq \delta$,

$$\sup_\alpha \mu_\alpha(w'(f,\rho,T) \geq \lambda) \leq \sup_\alpha \mu_\alpha(w'(f,\delta,T) \geq \lambda) < \epsilon$$

and we have shown (ii).

Now assume that (i) and (ii) hold and suppose that $\epsilon > 0$ is given. Let $q_1, q_2, \dots$ be an enumeration of $t \in [0,\infty) \cap \mathbb{Q}$. By (i) for every q_M there exists compact $K_{\epsilon,M} \subset S$ such that $\sup_\alpha \mu_\alpha(f(q_M) \notin K_{\epsilon,M}) < \epsilon/2^{M+1}$. By (ii) for every $N, k \in \mathbb{N}$, there exists a $\delta_{N,k}$ such that $\sup_\alpha \mu_\alpha(w'(f,\delta_{N,k},N) \geq 1/k) < \epsilon/2^{N+k+1}$. If we define

$$A_N = \{f \mid w'(f,\delta_{N,k},N) < 1/k \text{ for all } k \geq 1\}$$

so that $A_N^c \subset \cup_{k=1}^\infty \{f \mid w'(f,\delta_{N,k},N) \geq 1/k\}$ then by a union bound

$$\begin{aligned}
\sup_\alpha \mu_\alpha(A_N) &= \sup_\alpha (1 - \mu_\alpha(A_N^c)) \\
&\geq \sup_\alpha \left(1 - \sum_{k=1}^\infty \mu_\alpha(w'(f,\delta_{N,k},N) \geq 1/k)\right) \\
&\geq 1 - \epsilon/2^{T+1}
\end{aligned}$$

If we define $K = \cap_{M=1}^\infty \{f(q_M) \in K_{\epsilon,M}\} \cap \cap_{N=1}^\infty A_N$ then another union bound shows

$$\begin{aligned}
\sup_\alpha \mu_\alpha(K^c) &= \sup_\alpha \mu_\alpha(\cup_{M=1}^\infty \{f(q_M) \notin K_{\epsilon,M}\} \cup \cup_{N=1}^\infty A_N^c) \\
&\leq \sup_\alpha \left[\sum_{M=1}^\infty \mu_\alpha(f(q_M) \notin K_{\epsilon,M}) + \sum_{N=1}^\infty \mu_\alpha(A_N^c)\right] \\
&\leq \sum_{M=1}^\infty \epsilon/2^{M+1} + \sum_{N=1}^\infty \epsilon/2^{T+1} = \epsilon
\end{aligned}$$

and by construction the set K satisfies the conditions of Theorem 17.28 so is proven compact. □

TODO: Understand the Ethier and Kurtz construction and its relationship with Lindvall's. One proposal is that the only difference is the manner in which we are embedding $D([0,\infty))$ in D_0^∞ (accounting for the lack of linearity). Here we define for each $n \in \mathbb{N}$ the map $c_n : D([0,\infty)) \to D([0,\infty])$ by $c_n(f)(t) = f(t \wedge n)$

and then map f to $\Psi(f) = (c_1(f), c_2(f), \dots)$. The key things that we need to have be true are that

(i) Ψ is an injection
(ii) $\Psi(D([0,\infty)))$ is closed in $D_0[0,\infty]^\infty$.

It is obvious that Ψ is linear and furthermore if $f \neq 0$ the we pick $0 \leq t < \infty$ such that $f(t) \neq 0$ and then $n > t$ and it follows that $c_n(f)(t) = f(t) \neq 0$ thus $\Psi(f) \neq 0$.

Lemma 17.30. *For each $m \in \mathbb{N}$, c_m is continuous on the set*

$$D_0 = \{f \in D([0,\infty);S) \mid \lim_{t\to\infty} f(t) \text{ exists and is finite}\}$$

Furthermore, $\Psi(D)$ is closed.

Proof. Suppose $f_n \to f$ in D_0. Then there exist λ_n such that $\lim_{n\to\infty} \sup_{0\leq t<\infty} r(f_n(\lambda_n(t)), f(t)) = 0$ and $\lim_{n\to\infty} \sup_{0\leq t<\infty} |\lambda_n(t) - t| = 0$. We apply the sequence $c_m(f_n)(t) = f_n(t \wedge m)$.

$$\begin{aligned}\sup_{0\leq t<\infty} r(c_m(f_n)(\lambda_n(t)), c_m(f)(t)) &= \sup_{0\leq t<\infty} r(f_n(\lambda_n(t)\wedge m), f(t\wedge m)) \\ &= \sup_{0\leq t<\infty} r(f_n(\lambda_n(t)\wedge m), f_n(\lambda_n(t\wedge m))) + r(f_n(\lambda_n(t\wedge m)), f(t\wedge m))\end{aligned}$$

Note that

$$\sup_{0\leq t<\infty} r(f_n(\lambda_n(t\wedge m)), f(t\wedge m)) = \sup_{0\leq t\leq m} r(f_n(\lambda_n(t)), f(t)) \leq \sup_{0\leq t<\infty} r(f_n(\lambda_n(t)), f(t)) \to 0$$

Oops. It is not true that $\sup_{0\leq t<\infty} r(f_n(\lambda_n(t)\wedge m), f_n(\lambda_n(t\wedge m))) \to 0$ for a counterexample let $0 < x < \infty$ and define

$$\lambda_n(t) = \begin{cases} (1 - 1/nx)t & 0 \leq t \leq x \\ 2t - x - 1/n & x < t < x + 1/n \\ t & t \geq x + 1/n \end{cases}$$

so that $f_n(\lambda_n(x) \wedge x) = f_n((x - 1/n) \wedge x) = f_n(x - 1/n)$ and $f_n(\lambda_n(x \wedge x) = f_n(x)$. If f_n all have a jump of the same size $\epsilon > 0$ at $t = x$ it follows that $\sup_{0\leq t<\infty} r(f_n(\lambda_n(t)\wedge m), f_n(\lambda_n(t\wedge m))) \geq \epsilon > 0$.

Note that this doesn't imply that $f \to f(\cdot \wedge n)$ isn't a continuous map necessarily we might just need to be more creative in finding the sequence λ_n. In fact the example $f = \mathbf{1}_{[x,\infty)}$ and $f_n = \mathbf{1}_{[x+1/n)}$ provide an example showing that c_x is not continuous. Can it still be true that $\Psi(D)$ is closed? □

TODO: Lindvall's approach only works with S a linear space since we use multiplication to tamp things down. Maybe there are some worthwhile exercises to be culled from that. For each $n \in \mathbb{N}$ we define

$$g_n(t) = \begin{cases} 1 & 0 \leq t \leq n \\ 1 - t - n & n < t < n + 1 \\ 0 & t \geq n + 1 \end{cases}$$

and for every $f \in D([0,\infty)$.

3. The Aldous Criterion

TODO: How to tie this in with what Ethier and Kurtz claim to be the Aldous criterion. I am pretty sure that the EK development provides a necessary and sufficient condition for tightness that makes it more or less clear that the Aldous criterion is a sufficient condition. Here is how Aldous describes the criterion: "if you try to predict a jump you are right with probability close to 0".

LEMMA 17.31. *Let $\xi_1, \dots, \xi_n$ be nonnegative random variables, $S_n = \sum_{k=1}^n \xi_k$ and $t, c > 0$ then*

$$\boldsymbol{E}\left[e^{-S_n}; S_n < t\right] \leq e^{-nc} + \max_{1 \leq k \leq n} \boldsymbol{P}\{\xi_k < c; S_n < t\}$$

PROOF. By the multivariate Hölder Inequality (the one you get from the AMGM inequality) and a Markov bound

$$\begin{aligned}
\mathbf{E}\left[e^{-S_n}; S_n < t\right] &= \mathbf{E}\left[\prod_{k=1}^n e^{-\xi_k}; S_n < t\right] \leq \prod_{k=1}^n \left(\mathbf{E}\left[e^{-n\xi_k}; S_n < t\right]\right)^{1/n} \\
&= \prod_{k=1}^n \left(\mathbf{E}\left[e^{-n\xi_k}; S_n < t; \xi_k < c\right] + \mathbf{E}\left[e^{-n\xi_k}; S_n < t; \xi_k \geq c\right]\right)^{1/n} \\
&\leq \prod_{k=1}^n \left(e^{-nc} + \mathbf{P}\{S_n < t; \xi_k \geq c\}\right)^{1/n} \\
&\leq \left(\left(e^{-nc} + \max_{1 \leq k \leq n} \mathbf{P}\{S_n < t; \xi_k \geq c\}\right)^{1/n}\right)^n \\
&= e^{-nc} + \max_{1 \leq k \leq n} \mathbf{P}\{S_n < t; \xi_k \geq c\}
\end{aligned}$$

□

TODO: Update the following statement and proof to use arbitrary familiies of processes (and replace the $\limsup_{n\to\infty}$ with $\sup_{\alpha \in A}$)

THEOREM 17.32. *Let X^n be a sequence of processes in $D([0,T]; S)$ with (S, r) a metric space and suppose for every $\epsilon > 0$ and $t \in [0,T] \cap \mathbb{Q}$ there exists a compact set $K_{\epsilon,t}$ such that $\boldsymbol{P}\{X^n_t \in K_{\epsilon,t}\} > 1 - \epsilon$, then for every $\lambda > 0$ X^n satisfies $\lim_{\delta\to 0} \sup_n \boldsymbol{P}\{w(X^n, \delta) \geq \lambda\} = 0$ if and only if for any bounded sequence of $\mathcal{F}^{X^n}$-optional times τ_n and any sequence $\delta_n > 0$ with $\lim_{n\to\infty} \delta_n = 0$ we have*

$$r(X^n_{\tau_n}, X^n_{\tau_n + \delta_n}) \xrightarrow{P} 0$$

PROOF. In the proof we need to take suprema over a few different families of optional times, so let's first setup some notation for them. For $\delta > 0$, $t > 0$ and $n \in \mathbb{N}$ let

$$\begin{aligned}
\mathfrak{S}_{n,t} &= \{\tau \mid \tau \text{ is an } \mathcal{F}^{X^n}\text{-optional time with } \tau \leq t\} \\
\mathfrak{S}_{\delta,n,t} &= \{(\sigma, \tau) \mid \sigma, \tau \in \mathfrak{S}_{n,t} \text{ and } \sigma \leq \tau \leq \sigma + \delta\}
\end{aligned}$$

The first task is a couple of claims to establish equivalent stronger forms of the Aldous criterion.

CLAIM 17.32.1. The Aldous criterion is equivalent to the statement that

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\sup_{\tau\in\mathfrak{S}_{n,t}}\sup_{0\leq h\leq\delta}\mathbf{E}\left[r(X^n_\tau, X^n_{\tau+h})\wedge 1\right] = 0 \text{ for every } t>0$$

Suppose that the condition in the claim does not hold then there exists $\epsilon > 0$, $t > 0$, a sequence $\delta_m > 0$ such that $\lim_{m\to\infty}\delta_m = 0$ and

$$\limsup_{n\to\infty}\sup_{\tau\in\mathfrak{S}_{n,t}}\sup_{0\leq h\leq\delta_m}\mathbf{E}\left[r(X^n_\tau, X^n_{\tau+h})\wedge 1\right] > \epsilon \text{ for every } m\in\mathbb{N}$$

Thus the inequality holds for infinitely many n for each fixed m and therefore we can pick a subsequence n_m of $\mathbb{N}$, $\tau_m \in \mathfrak{S}_{n_m,t}$ and $0 \leq h_m \leq \delta_m$ such that

$$\mathbf{E}\left[r(X^{n_m}_{\tau_m}, X^{n_m}_{\tau_m+h_m})\wedge 1\right] \geq \epsilon \text{ for all } m\in\mathbb{N}$$

Since $\lim_{m\to\infty}\delta_m = 0$ it follows that $\lim_{m\to\infty} h_m = 0$ and thus the Aldous criterion fails as well by Lemma 5.9. On the other hand, if the Aldous criterion fails then by Lemma 5.9 and taking an appropriate subsequence there exists a $t > 0$, $\epsilon > 0$, a sequence $\delta_m > 0$ with $\lim_{m\to\infty}\delta_m = 0$ and a subsequence n_m of $\mathbb{N}$ such that

$$\mathbf{E}\left[r(X^{n_m}_{\tau_m}, X^{n_m}_{\tau_m+\delta_m})\wedge 1\right] > \epsilon \text{ for all } m\in\mathbb{N}$$

Since $\lim_{m\to\infty}\delta_m = 0$, for any fixed $m \in \mathbb{N}$ there are infinitely many $j \geq m$ such that $\delta_j \leq \delta_m$ and therefore

$$\limsup_{n\to\infty}\sup_{\tau\in\mathfrak{S}_{n,t}}\sup_{0\leq h\leq\delta_m}\mathbf{E}\left[r(X^n_\tau, X^n_{\tau+h})\wedge 1\right] > \epsilon$$

and the condition in the claim fails.

CLAIM 17.32.2. The Aldous criterion is equivalent to the statement that

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\sup_{(\sigma,\tau)\in\mathfrak{S}_{\delta,n,t}}\mathbf{E}\left[r(X^n_\sigma, X^n_\tau)\wedge 1\right] = 0 \text{ for every } t>0$$

That this condition implies the Aldous criterion follows from the prior claim and the fact that for every $\tau \in \mathfrak{S}_{n,t}$ and $0 \leq h \leq \delta$, we have $(\tau, \tau+h) \in \mathfrak{S}_{\delta,n,t}$. Let $(\sigma, \tau) \in \mathfrak{S}_{\delta,n,t}$ so that $\sigma \leq \tau \leq \sigma+\delta$. From this it follows that $[\tau, \tau+\delta] \subset [\sigma, \sigma+2\delta]$ and therefore

$$\begin{aligned}
r(X^n_\sigma, X^n_\tau)\wedge 1 &= \frac{1}{\delta}\int_0^\delta r(X^n_\sigma, X^n_\tau)\wedge 1\, dh \\
&\leq \frac{1}{\delta}\int_0^\delta (r(X^n_\sigma, X^n_{\tau+h}) + r(X^n_\tau, X^n_{\tau+h})\wedge 1\, dh \\
&\leq \frac{1}{\delta}\left[\int_0^\delta r(X^n_\sigma, X^n_{\tau+h})\wedge 1\, dh + \int_0^\delta r(X^n_\tau, X^n_{\tau+h}\wedge 1\, dh\right] \\
&\leq \frac{1}{\delta}\left[\int_0^{2\delta} r(X^n_\sigma, X^n_{\sigma+h})\wedge 1\, dh + \int_0^\delta r(X^n_\tau, X^n_{\tau+h}\wedge 1\, dh\right]
\end{aligned}$$

Taking expectations using we get by Tonelli's Theorem and the fact that $\sigma, \tau \leq \mathfrak{S}_{n,t}$

$$\begin{aligned}
\mathbf{E}\left[r(X^n_\sigma, X^n_\tau)\wedge 1\right] &\leq \frac{1}{\delta}\left[\int_0^{2\delta}\mathbf{E}\left[r(X^n_\sigma, X^n_{\sigma+h})\wedge 1\right]\, dh + \int_0^\delta \mathbf{E}\left[r(X^n_\tau, X^n_{\tau+h}\wedge 1\right]\, dh\right] \\
&\leq 3\sup_{\tau\in\mathfrak{S}_{n,t}}\sup_{0\leq h\leq 2\delta}\mathbf{E}\left[r(X^n_\tau, X^n_{\tau+h}\wedge 1\right]
\end{aligned}$$

and the claim follows from the prior claim by taking the supremum over $(\sigma,\tau) \in \mathfrak{S}_{\delta,n,t}$ and the limits over n and δ.

Claim 17.32.3. The Aldous criterion is equivalent to the statement that

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \sup_{(\sigma,\tau) \in \mathfrak{S}^+_{\delta,n,t}} \mathbf{E}\left[r(X^n_\sigma, X^n_\tau) \wedge 1\right] = 0 \text{ for every } t > 0$$

If this is not true then there exists $t > 0$, $\epsilon > 0$, $\delta_k \to 0$, $n_{k,m}$ with $\lim_{m\to\infty} n_{k,m} = \infty$ for each $k \in \mathbb{N}$ and $(\sigma_{k,m}, \tau_{k,m}) \in \mathfrak{S}^+_{\delta_k, n_{k,m}, t}$ such that

$$\mathbf{E}\left[r(X^{n_{k,m}}_{\sigma_{k,m}}, X^{n_{k,m}}_{\tau_{k,m}}) \wedge 1\right] > \epsilon \text{ for all } k, m \in \mathbb{N}$$

We know that we have $(\sigma_{k,m} + h, \tau_{k,m} + h) \in \mathfrak{S}_{\delta_k, n_{k,m}, t+h}$ for all $h > 0$. By right continuity and Dominated Convergence Theorem we know that

$$\lim_{h \to 0^+} \mathbf{E}\left[r(X^{n_{k,m}}_{\sigma_{k,m}+h}, X^{n_{k,m}}_{\tau_{k,m}+h}) \wedge 1\right] = \mathbf{E}\left[r(X^{n_{k,m}}_{\sigma_{k,m}}, X^{n_{k,m}}_{\tau_{k,m}}). \wedge 1\right] > \epsilon$$

thus in particular we know that

$$\mathbf{E}\left[r(X^{n_{k,m}}_{\sigma_{k,m}+h}, X^{n_{k,m}}_{\tau_{k,m}+h}) \wedge 1\right] > \epsilon$$

for all $k, m \in \mathbb{N}$ and sufficiently small $h > 0$ (where sufficiently small depends on k and m). This shows

$$\sup_{(\sigma,\tau) \in \mathfrak{S}_{\delta_k, n_{k,m}, t+1}} \mathbf{E}\left[r(X^{n_{k,m}}_\sigma, X^{n_{k,m}}_\tau) \wedge 1\right] > \epsilon \text{ for all } k, m \in \mathbb{N}$$

and therefore $\lim_{\delta\to 0} \limsup_{n\to\infty} \sup_{(\sigma,\tau)\in\mathfrak{S}_{\delta,n,t+1}} \mathbf{E}\left[r(X^n_\sigma, X^n_\tau) \wedge 1\right] \geq \epsilon$ which contradicts the prior claim.

Now for every $n \in \mathbb{N}$ and $\epsilon > 0$ we define a sequence of weakly $\mathcal{F}^{X^n}$-optional times via recursion

$$\sigma^n_{k+1} = \inf\{s > \sigma^n_k \mid r(X^n_{\sigma^n_k}, X^n_s) > \epsilon\}$$

By this definition, provided $\sigma^n_k < \infty$ for all $\sigma^n_k \leq s < t < \sigma^n_{k+1}$ we have via the triangle inequality

$$r(X^n_s, X^n_t) \leq r(X^n_{\sigma^n_k}, X^n_s) + r(X^n_{\sigma^n_k}, X^n_t) \leq 2\epsilon$$

Thus we can use the σ^n_k to give an upper bound on the modulus of continuity provided they form a valid partition. Our task is to bound the probability that they do not form a valid partition. In formulating that bound we need the following simple fact.

Claim 17.32.4. Given $T > 0$ and $0 < \delta < T$ and numbers $0 = t_0 < t_1 < \cdots < t_m$ with $t_m \geq T$ then if there is no $1 \leq k \leq m$ such that $0 = t_0 < t_1 < \cdots < T \leq t_k$ and $t_j - t_{j-1} > \delta$ for all $j = 1, \ldots, k$ then there exists $0 \leq k \leq m-1$ such that $t_k < T$ and $t_{k+1} - t_k \leq \delta$.

This follows easily by contradiction. Suppose that for every $0 \leq k \leq m-1$ such that $t_k \geq T$ or $t_{k+1} - t_k > \delta$. If there is $0 \leq k \leq m-1$ such that $t_k \geq T$ then let k be the smallest such (note such a k is nonzero) and it follows that $0 = t_0 < \cdots < t_{k-1} < T \leq t_k$ and we are guaranteed $t_j - t_{j-1} > \delta$ for all $1 \leq j \leq k$. It $t_k \geq T$ for all $0 \leq k \leq m-1$ then it follows that $0 = t_0 < \cdots < t_{m-1} < T \leq t_m$ and we are guaranteed $t_j - t_{j-1} > \delta$ for all $1 \leq j \leq m$.

Now we apply a few union bounds using the above claim

$$\begin{aligned}
&\mathbf{E}\left[w'(X^n,\delta,T)\wedge 1\right]\\
&\mathbf{E}\left[w'(X^n,\delta,T)\wedge 1; T\leq \sigma^n_m\right] + \mathbf{E}\left[w'(X^n,\delta,T)\wedge 1; \sigma^n_m < T\right]\\
&= \mathbf{E}\left[w'(X^n,\delta,T)\wedge 1; \cup_{k=1}^n\{0=\sigma^n_0<\sigma^n_1<\cdots<\sigma^n_{k-1}<T\leq\sigma^n_k;\sigma^n_1-\sigma^n_0>\delta;\cdots;\sigma^n_k-\sigma^n_{k-1}>\delta\}\right]\\
&+ \mathbf{E}\left[w'(X^n,\delta,T)\wedge 1; \cup_{k=0}^{m-1}\{\sigma^n_{k+1}\leq\sigma^n_k+\delta;\sigma^n_k<T\}\right]\\
&+ \mathbf{E}\left[w'(X^n,\delta,T)\wedge 1; \sigma^n_m < T\right]\\
&\leq 2\epsilon + \sum_{k=0}^{m-1}\mathbf{P}\{\sigma^n_{k+1}\leq\sigma^n_k+\delta;\sigma^n_k<T\} + \mathbf{P}\{\sigma^n_m<T\}
\end{aligned}$$

Now note that for any $\epsilon < 1$ and $\delta < 1$,

$$\begin{aligned}
\mathbf{P}\{\sigma^n_{k+1}<\sigma^n_k+\delta;\sigma^n_k<T\} &= \mathbf{P}\{r(X^n_{\sigma^n_k},X^n_{\sigma^n_{k+1}})\geq\epsilon;\sigma^n_{k+1}<\sigma^n_k+\delta;\sigma^n_k<T\}\\
&= \mathbf{P}\{r(X^n_{\sigma^n_k\wedge(T+1)},X^n_{\sigma^n_{k+1}\wedge(T+1)})\geq\epsilon;\sigma^n_{k+1}<\sigma^n_k+\delta;\sigma^n_k<T\}\\
&\leq \mathbf{P}\{(r(X^n_{\sigma^n_k\wedge(T+1)},X^n_{\sigma^n_{k+1}\wedge(T+1)})\wedge 1)\geq\epsilon\}\\
&\leq \epsilon^{-1}\mathbf{E}\left[r(X^n_{\sigma^n_k\wedge(T+1)},X^n_{\sigma^n_{k+1}\wedge(T+1)})\wedge 1\right]\\
&\leq \epsilon^{-1}\sup_{(\sigma,\tau)\in\mathfrak{S}^+_{\delta,n,T+1}}\mathbf{E}\left[r(X^n_\sigma,X^n_\tau)\wedge 1\right]
\end{aligned}$$

We also get for any $0<c<1$

$$\begin{aligned}
\mathbf{P}\{\sigma^n_m<T\} < e^T\mathbf{E}\left[e^{-\sigma^n_m};\sigma^n_m<T\right] &= e^T\mathbf{E}\left[e^{-\sum_{k=1}^m(\sigma^n_k-\sigma^n_{k-1})};\sigma^n_m<T\right]\\
&\leq e^T(e^{-mc}+\max_{1\leq k\leq m}\mathbf{P}\{\sigma^n_k-\sigma^n_{k-1}<c;\sigma^n_m<T\})\\
&\leq e^T(e^{-mc}+\epsilon^{-1}\sup_{(\sigma,\tau)\in\mathfrak{S}^+_{c,n,T+1}}\mathbf{E}\left[r(X^n_\sigma,X^n_\tau)\wedge 1\right])
\end{aligned}$$

Putting the estimates together and using Lemma

$$\begin{aligned}
&\lim_{\delta\to 0}\limsup_{n\to\infty}\mathbf{E}\left[w'(X^n,\delta,T)\wedge 1\right]\\
&\leq \lim_{\delta\to 0}\limsup_{n\to\infty}\left[2\epsilon+\epsilon^{-1}\sup_{(\sigma,\tau)\in\mathfrak{S}^+_{\delta,n,T+1}}\mathbf{E}\left[r(X^n_\sigma,X^n_\tau)\wedge 1\right]+e^T(e^{-mc}+\epsilon^{-1}\sup_{(\sigma,\tau)\in\mathfrak{S}^+_{c,n,T+1}}\mathbf{E}\left[r(X^n_\sigma,X^n_\tau)\wedge 1\right])\right]\\
&= 2\epsilon + e^Te^{-mc}
\end{aligned}$$

Now let $m\to\infty$, $c\to 0$ and $\epsilon\to 0$. ☐

In practice it turns out that it is difficult to apply the criterion $\lim_{\delta\to 0}\sup_\alpha\mu_\alpha(w'(f,\delta,T)\geq\lambda)=0$ to show tightness of a family of measures μ_α on Skorohod space. In some sense this is not surprising as, being an infimum over a set of partitions, the modulus of continuity is a rather complicated object. What we need to develop are tools for estimating $w'(f,\delta,T)$ that are strong enough to imply the tightness condition. One technique for finding upper bounds of $w'(f,\delta,T)$ is clear; one simply needs to find a particular partition $0=t_0<\cdots<t_{n-1}<T\leq t_n$ for which we can calculate (or upper bound) each term $\sup_{t_{i-1}\leq s<t<t_i} r(f(s),f(t))$.

Though it makes sense that one should be able come up with a construction of partitions that allows to provide upper bounds for $w'(f,\delta,T)$ and therefore get sufficient conditions for compactness one can do better. There is a related construction of a family of optional times that provides a necessary and sufficient condition for compactness. In order to come up with necessary conditions we need to be able to use $w'(f, \delta, T)$ to bound f. TODO: Motivate the appearance of terms of the form $(r(x+u, x) \wedge 1)(r(x, x-v) \wedge 1)$ in the following. Suppose we have a cadlag function f and we are given the value of $w'(f, \delta, T)$. We ask a simple question: what kinds of expressions involving f can be bounded by $w'(f, \delta, T)$? The case of a continuous function and the corresponding modulus of continuity provides bounds on terms $r(f(t+u), f(t))$ for any $0 \leq t \leq T - \delta$ and $-\delta < u < \delta$. In the cadlag case we allow for jump discontinuities to occur (manifested by the partitions over which we take the infimum) and only provide bounds in between adjacent jumps; thus we cannot bound a term $r(f(t+u), f(t))$ because we can always place a node of the partition between t and $t+u$. What comes the rescue is that the definition of $w'(f, \delta, T)$ restricts the distance between adjacent nodes of the partition to be bigger than δ. Therefore if we place a node of the partition between t and $t+u$ then we know that there cannot be a node of the partition between t and $t-v$ provided $|u-v| \leq \delta$. This discussion motivates the following simple proposition whose purpose is to set the stage for the types of expressions we will be using in the ensuing discussion.

Proposition 17.33. *Let (S, r) be a metric space and let $f \in D([0, \infty); S)$, then for all $T > 0$, $\delta > 0$, $0 \leq u \leq \delta$ and $0 \leq v \leq \delta \wedge t$*

$$(r(f(t+u), f(t)) \wedge 1)(r(f(t), f(t-v)) \wedge 1) \leq (r(f(t+u), f(t)) \wedge 1) \wedge (r(f(t), f(t-v)) \wedge 1) \leq w'(f, 2\delta, T+\delta)$$

Proof. TODO: ☐

Our first step is to work pointwise in $D([0, \infty); S)$ and show how create a useful partition for an arbitrary cadlag function f. For the construction we assume that $\epsilon > 0$ and $f \in D([0, \infty); S)$ are both given. First define inductively $\tau_0 = 0$ and $n \in \mathbb{N}$

$$\tau_n = \begin{cases} \inf\{t > \tau_{n-1} \mid r(f(t), f(\tau_{n-1})) > \epsilon/2\} & \text{if } \tau_{n-1} < \infty \\ \infty & \text{if } \tau_{n-1} = \infty \end{cases}$$

and then define for $n \in \mathbb{Z}_+$

$$\sigma_n = \begin{cases} \sup\{t \leq \tau_n \mid r(f(t), f(\tau_n)) \vee r(f(t-), f(\tau_n)) \geq \epsilon/2\} & \text{if } \tau_n < \infty \\ \infty & \text{if } \tau_n = \infty \end{cases}$$

Note that by right continuity of f we have $r(f(\tau_n), f(\tau_{n-1})) \geq \epsilon/2$ whenever $\tau_n < \infty$ (in particular $\tau_{n-1} < \tau_n$ whenever $\tau_{n-1} < \infty$).

Claim 17.33.1. Let $\delta > 0, T > 0$ be given. If $w'(f, \delta, T) < \epsilon/2$ then $\min\{\tau_{n+1} - \sigma_n \mid \tau_n < T\} > \delta$.

The claim is verified by contradiction, so suppose that we have $\tau_n < T$ and $\tau_{n+1} - \sigma_n \leq \delta$ for some $n \in \mathbb{Z}_+$. If we are given a partition $0 = t_0 < \cdots t_{m-1} < T \leq t_m$ with $\min_{1 \leq i \leq m}(t_i - t_{i-1}) > \delta$ it follows that there is some $1 \leq i \leq m$ such that $t_{i-1} \leq \tau_n < \tau_i$. If $\sigma_n \leq t_{i-1} < t_i \leq \tau_{n+1}$ then $\tau_{n+1} - \sigma_n \geq t_i - t_{i-1} > \delta$ which is a contradiction therefore either $t_{i-1} < \sigma_n \leq \tau_n < t_i$ or $t_{i-1} \leq \tau_n < \tau_{n+1} < t_i$ or both. In the first case by definition of σ_n we can find a $t_{i-1} < u \leq \sigma_n$ such that $r(f(u), f(\tau_n)) \geq \epsilon/2$ and in the second case we have already observed $r(f(\tau_n), f(\tau_{n-1})) \geq \epsilon/2$; thus $\max_{1 \leq i \leq m} \sup_{t_{i-1} \leq s < t < t_i} r(f(s), f(t)) \geq \sup_{t_{i-1} \leq s < t < t_i} r(f(s), f(t)) \geq \epsilon/2$. Now take the infimum over all partitions.

Next we define the partition that will generate our upper bound on the modulus of continuity.

$$s_n = \frac{\sigma_n + \tau_n}{2}$$

and note that

$$\sigma_n \leq s_n \leq \tau_n \leq \sigma_{n+1} \leq s_{n+1} \leq \tau_{n+1}$$

and

$$s_{n+1} - s_n = \frac{\sigma_{n+1} + \tau_{n+1}}{2} - \frac{\sigma_n + \tau_n}{2} \geq \frac{\tau_n + \tau_{n+1}}{2} - \frac{\sigma_n + \tau_n}{2} = \frac{\tau_{n+1} - \sigma_n}{2}$$

Claim 17.33.2. For every $\delta > 0, T > 0$ if $\min\{\tau_{n+1} - \sigma_n \mid \tau_n < T + \delta/2\} > \delta$ then $\min\{s_{n+1} - s_n \mid s_n < T\} > \delta/2$

We argue by contradiction, so suppose that $s_n < T$ and $s_{n+1} - s_n \leq \delta/2$. Then $\tau_n \leq s_{n+1} \leq s_n + \delta/2 < T + \delta/2$ and

$$\tau_{n+1} - \sigma_n \leq 2(s_{n+1} - s_n) \leq \delta$$

Claim 17.33.3. For every $\delta > 0, T > 0$ if $\min\{s_{n+1} - s_n \mid s_n < T\} > \delta/2$ then $w'(f, \delta/2, T) \leq \epsilon$

The claim follows if we can show $\sup_{s_n \leq s < t < s_{n+1}} r(f(s), f(t)) \leq \epsilon$ for then $0 = s_0 < \cdots < s_n < T \leq s_{n+1}$ is a partition which shows that $w'(f, \delta/2, T) \leq \epsilon$ (recall that $s_n \to \infty$ so there are only finitely many $s_n < T$). The property $\sup_{s_n \leq s < t < s_{n+1}} r(f(s), f(t)) \leq \epsilon$ follows from the triangle inequality if we can show that for any $s_n \leq s < s_{n+1}$ we have $r(f(s), f(\tau_n)) < \epsilon/2$. To see this last fact we consider two cases. First assume $s_n \leq \tau_n \leq s < s_{n+1} \leq \tau_{n+1}$ then by the definition of τ_{n+1} we know that $r(f(s), f(\tau_n)) < \epsilon/2$. If on the other hand, $s_n \leq s < \tau_n$ then this implies that $\sigma_n < \tau_n$ hence $\sigma_n < s_n$ and therefore either $\sigma_n < s < \tau_n$ By the definition of σ_n we know that $r(f(s), f(\tau_n)) < \epsilon/2$.

TODO: What are the steps and examples that lead one to considering the definitions of σ_n and s_n (the case for the definition of τ_n is clear).

Claim 17.33.4. If S is separable the τ_n, σ_n and s_n are Borel measurable functions on $D([0, \infty); S)$ to $[0, \infty]$.

TODO:

Lemma 17.34. *Let (S, r) be a separable metric space, let A be an arbitrary index set and let X^α for $\alpha \in A$ be a family of stochastic processes with values in $D([0, \infty); S)$. Define $\tau_n^{\alpha,\epsilon}, \sigma_n^{\alpha,\epsilon}$ and $s_n^{\alpha,\epsilon}$ as above then the following are equivalent*

(i) $\lim_{\delta \to 0} \inf_{\alpha \in A} \boldsymbol{P}\{w'(X^\alpha, \delta, T) < \epsilon\} = 1$ *for all $\epsilon > 0$ and $T > 0$*
(ii) $\lim_{\delta \to 0} \inf_{\alpha \in A} \boldsymbol{P}\{\min\{\tau_{n+1}^{\alpha,\epsilon} - \sigma_n^{\alpha,\epsilon} \mid \tau_n^{\alpha,\epsilon} < T\} \geq \delta\} = 1$ *for all $\epsilon > 0$ and $T > 0$*
(iii) $\lim_{\delta \to 0} \inf_{\alpha \in A} \boldsymbol{P}\{\min\{s_{n+1}^{\alpha,\epsilon} - s_n^{\alpha,\epsilon} \mid s_n^{\alpha,\epsilon} < T\} \geq \delta\} = 1$ *for all $\epsilon > 0$ and $T > 0$*

Proof. Let $\epsilon > 0$ and $T > 0$ be given then by (i) we know that $\lim_{\delta \to \infty} \inf_{\alpha \in A} \mathbf{P}\{w'(X^\alpha, \delta, T) < \epsilon/2\} = 1$. On the other hand, for all $\alpha \in A$ we know that

$$\begin{aligned} \mathbf{P}\{w'(X^\alpha, \delta, T) < \epsilon/2\} &\leq \mathbf{P}\{\min\{\tau_{n+1}^{\alpha,\epsilon} - \sigma_n^{\alpha,\epsilon} > \delta \mid \tau_n^{\alpha,\epsilon} < T\} > \delta\} \\ &\leq \mathbf{P}\{\min\{\tau_{n+1}^{\alpha,\epsilon} - \sigma_n^{\alpha,\epsilon} > \delta \mid \tau_n^{\alpha,\epsilon} < T\} \geq \delta\} \end{aligned}$$

Now take the infimum over $\alpha \in A$ and let $\delta \to 0$ to conclude (ii).

Assume (ii) holds. Let $\epsilon > 0$ and $T > 0$ be given. Now pick $T' > T$, let $\eta > 0$ then by (ii) there exists $0 < \delta < 2(T' - T)$ such that

$$\begin{aligned}
1 - \eta &< \inf_{\alpha \in A} \mathbf{P}\{\min\{\tau^{\alpha,\epsilon}_{n+1} - \sigma^{\alpha,\epsilon}_n \mid \tau^{\alpha,\epsilon}_n < T'\} \geq \delta\} \\
&\leq \inf_{\alpha \in A} \mathbf{P}\{\min\{\tau^{\alpha,\epsilon}_{n+1} - \sigma^{\alpha,\epsilon}_n \mid \tau^{\alpha,\epsilon}_n < T'\} > \rho\} \text{ for all } 0 < \rho < \delta \\
&\leq \inf_{\alpha \in A} \mathbf{P}\{\min\{s^{\alpha,\epsilon}_{n+1} - s^{\alpha,\epsilon}_n \mid s^{\alpha,\epsilon}_n < T' - \rho/2\} > \rho/2\} \text{ for all } 0 < \rho < \delta \\
&\leq \inf_{\alpha \in A} \mathbf{P}\{\min\{s^{\alpha,\epsilon}_{n+1} - s^{\alpha,\epsilon}_n \mid s^{\alpha,\epsilon}_n < T\} \geq \rho/2\} \text{ for all } 0 < \rho < \delta
\end{aligned}$$

which shows (iii).

Assume (iii) holds, let $T > 0$ and $\epsilon > 0$ be given. Let $\eta > 0$ be arbitrary and by (iii) pick a $\delta > 0$ such that

$$\begin{aligned}
1 - \eta &< \inf_{\alpha \in A} \mathbf{P}\{\min\{\tau^{\alpha,\epsilon}_{n+1} - \sigma^{\alpha,\epsilon}_n \mid \tau^{\alpha,\epsilon}_n < T\} \geq \delta\} \\
&\leq \inf_{\alpha \in A} \mathbf{P}\{\min\{\tau^{\alpha,\epsilon}_{n+1} - \sigma^{\alpha,\epsilon}_n \mid \tau^{\alpha,\epsilon}_n < T\} > \rho\} \text{ for all } 0 < \rho < \delta \\
&\leq \inf_{\alpha \in A} \mathbf{P}\{w'(X^\alpha, \rho, T) \leq \epsilon\} \text{ for all } 0 < \rho < \delta
\end{aligned}$$

which shows $\lim_{\delta \to 0} \inf_{\alpha \in A} \mathbf{P}\{w'(X', \rho, T) \leq \epsilon\} = 1$ for all $T > 0$ and $\epsilon > 0$. This is equivalent to (i) since for any $\alpha \in A$, $T > 0$ and $0 < \epsilon$ we have

$$\mathbf{P}\{w'(X^\alpha, \rho, T) \leq \epsilon/2\} \leq \mathbf{P}\{w'(X^\alpha, \rho, T) < \epsilon\} \leq \mathbf{P}\{w'(X^\alpha, \rho, T) \leq 2\epsilon\}$$

$\square$

LEMMA 17.35. *Let A be an arbitrary index set and for every $\alpha \in A$ let $0 = s^\alpha_0 < s^\alpha_1 < \cdots$ be a sequence of random variables such that $\lim_{n \to \infty} s^\alpha_n = \infty$. Let $T > 0$ be arbitrary and define $K(\alpha, T) = \max\{n \in \mathbb{N} \mid s^\alpha_n < T\}$ and $F : [0, \infty) \to [0, 1]$ by $F(t) = \sup_{\alpha \in A} \sup_{n \geq 0} \boldsymbol{P}\{(s^\alpha_{n+1} - s^\alpha_n) < t, s^\alpha_n < T\}$. Then for all $\delta > 0$ and $L \in \mathbb{Z}_+$*

$$F(\delta) \leq \sup_{\alpha \in A} \boldsymbol{P}\{\min_{0 \leq n \leq K(\alpha,T)} (s^\alpha_{n+1} - s^\alpha_n) < \delta\} \quad \leq LF(\delta) + e^T \int_0^\infty Le^{-Lt} F(t)\, dt$$

Therefore

$$\lim_{\delta \to 0} \sup_{\alpha \in A} \boldsymbol{P}\{\min_{0 \leq n \leq K(\alpha,T)} (s^\alpha_{n+1} - s^\alpha_n)\} = 0$$

in and only if $F(0+) = 0$.

PROOF. For each fixed $\alpha \in A$ and every $n \in \mathbb{N}$,

$$\{(s^\alpha_{n+1} - s^\alpha_n) < \delta, s^\alpha_n < T\} \subset \{\min_{0 \leq n \leq K(\alpha,T)} (s^\alpha_{n+1} - s^\alpha_n) < \delta\}$$

and therefore a union bound, taking the supremum over $n \in \mathbb{N}$ and $\alpha \in A$ yields the first inequality.

To see the second inequality,

$$
\begin{aligned}
&\mathbf{P}\{\min_{0 \le n \le K(\alpha,T)} (s^\alpha_{n+1} - s^\alpha_n) < \delta\} \\
&= \mathbf{P}\{\min_{0 \le n \le K(\alpha,T)} (s^\alpha_{n+1} - s^\alpha_n) < \delta, K(\alpha,T) \le L-1\} + \mathbf{P}\{\min_{0 \le n \le K(\alpha,T)} (s^\alpha_{n+1} - s^\alpha_n) < \delta, K(\alpha,T) \ge L\} \\
&\le \sum_{n=0}^{L-1} \mathbf{P}\{(s^\alpha_{n+1} - s^\alpha_n) < \delta, s^\alpha_n < T\} + \mathbf{P}\{K(\alpha,T) \ge L\} \\
&\le LF(\delta) + \mathbf{E}\left[e^{T - \sum_{n=0}^{L-1}(s^\alpha_{n+1} - s^\alpha_n)}; K(\alpha,T) \ge L\right] \\
&\le LF(\delta) + e^T \Pi_{n=0}^{L-1} \mathbf{E}\left[e^{-L(s^\alpha_{n+1} - s^\alpha_n)}; K(\alpha,T) \ge L\right]^{1/L} \\
&\le LF(\delta) + e^T \Pi_{n=0}^{L-1} \mathbf{E}\left[e^{-L(s^\alpha_{n+1} - s^\alpha_n)}; s^\alpha_k < T\right]^{1/L}
\end{aligned}
$$

Here there is a multivariate generalization of Cauchy Schwartz used I think ($\mathbf{E}\left[\Pi_{k=1}^L f_k\right]^L \le \Pi_{k=1}^L \mathbf{E}\left[f_k^L\right]$ for $f_k \ge 0$; TODO: prove this, instead of Young's inequality use the AMGM inequality). ☐

The last two lemmas can be combined to yield the following additional equivalent criterion for equicontinuity of a family of stochastic processes.

PROPOSITION 17.36. *Let (S,r) be a separable metric space, let A be an arbitrary index set and let X^α for $\alpha \in A$ be a family of stochastic processes with values in $D([0,\infty);S)$. Define $\tau^{\alpha,\epsilon}_n, \sigma^{\alpha,\epsilon}_n$ and $s^{\alpha,\epsilon}_n$ as above then the following are equivalent*

(i) $\lim_{\delta \to 0} \inf_{\alpha \in A} \boldsymbol{P}\{w'(X^\alpha, \delta, T) < \epsilon\} = 1$ *for all* $\epsilon > 0$ *and* $T > 0$
(ii) $\lim_{\delta \to 0} \sup_{\alpha \in A} \sup_{n \ge 0} \boldsymbol{P}\{\tau^{\alpha,\epsilon}_{n+1} - \sigma^{\alpha,\epsilon}_n < \delta, \tau^{\alpha,\epsilon}_n < T\} = 0$ *for all* $\epsilon > 0$ *and* $T > 0$

PROOF.

CLAIM 17.36.1. For all $\epsilon > 0$, $T > 0$, $\delta > 0$ and $\alpha \in A$,

$$
\mathbf{P}\{s^{\alpha,\epsilon}_{n+1} - s^{\alpha,\epsilon}_n < \delta/2, s^{\alpha,\epsilon}_n < T\} \le \mathbf{P}\{\tau^{\alpha,\epsilon}_{n+1} - \sigma^{\alpha,\epsilon}_n < \delta, \tau^{\alpha,\epsilon}_n < T + \delta\} \le \mathbf{P}\{s^{\alpha,\epsilon}_{n+1} - s^{\alpha,\epsilon}_n < \delta, s^{\alpha,\epsilon}_n < T + \delta\}
$$

Recalling the defintion $s^{\alpha,\epsilon}_n = (\tau^{\alpha,\epsilon}_n + \sigma^{\alpha,\epsilon}_n)/2$ we see that if $s^{\alpha,\epsilon}_{n+1} - s^{\alpha,\epsilon}_n < \delta/2$ and $s^{\alpha,\epsilon}_n < T$ then it follows from $\tau_n \le \sigma_{n+1}$ that

$$
\begin{aligned}
\tau^{\alpha,\epsilon}_{n+1} - \sigma^{\alpha,\epsilon}_n &\le \tau^{\alpha,\epsilon}_{n+1} - \sigma^{\alpha,\epsilon}_n + \sigma^{\alpha,\epsilon}_{n+1} - \tau^{\alpha,\epsilon}_n \\
&= 2(s^{\alpha,\epsilon}_n - \sigma^{\alpha,\epsilon}_n) < \delta
\end{aligned}
$$

and from $\tau^{\alpha,\epsilon}_n \le s^{\alpha,\epsilon}_{n+1}$,

$$
\tau^{\alpha,\epsilon}_n < s^{\alpha,\epsilon}_n + \delta/2 < T + \delta
$$

Further if $\tau^{\alpha,\epsilon}_{n+1} - \sigma^{\alpha,\epsilon}_n < \delta$ and $\tau^{\alpha,\epsilon}_n < T + \delta$ then we have

$$
s^{\alpha,\epsilon}_{n+1} - s^{\alpha,\epsilon}_n \le \tau^{\alpha,\epsilon}_{n+1} - \sigma^{\alpha,\epsilon}_n < \delta
$$

and $s^{\alpha,\epsilon}_n \le \tau^{\alpha,\epsilon}_n < T + \delta$. The claim follows from the union bound implied by these set inclusions.

Now if we assume (i) then ϵ and $T > 0$ be given. By Lemma 17.34 we know that

$$\begin{aligned} 1 &= \lim_{\delta \to 0} \inf_{\alpha \in A} \mathbf{P}\{\min\{s^{\alpha,\epsilon}_{n+1} - s^{\alpha,\epsilon}_n \mid s^{\alpha,\epsilon}_n < T\} \geq \delta\} \\ &= 1 - \lim_{\delta \to 0} \sup_{\alpha \in A} \mathbf{P}\{\min\{s^{\alpha,\epsilon}_{n+1} - s^{\alpha,\epsilon}_n \mid s^{\alpha,\epsilon}_n < T\} < \delta\} \end{aligned}$$

thus by the claim and Lemma 17.35 we know that

$$\lim_{\delta \to 0} \sup_{\alpha \in A} \sup_{n \geq 0} \mathbf{P}\{(\tau^{\alpha,\epsilon}_{n+1} - \sigma^{\alpha,\epsilon}_n) < \delta, \tau^{\alpha,\epsilon}_n < T\} \leq \lim_{\delta \to 0} \sup_{\alpha \in A} \sup_{n \geq 0} \mathbf{P}\{(s^{\alpha,\epsilon}_{n+1} - s^{\alpha,\epsilon}_n) < \delta, s^{\alpha,\epsilon}_n < T\} = 0$$

If we have (ii) then by the claim,

$$\begin{aligned} \lim_{\delta \to 0} \sup_{\alpha \in A} \sup_{n \geq 0} \mathbf{P}\{s^{\alpha,\epsilon}_{n+1} - s^{\alpha,\epsilon}_n < \delta, s^{\alpha,\epsilon}_n < T\} &\leq \lim_{\delta \to 0} \sup_{\alpha \in A} \sup_{n \geq 0} \mathbf{P}\{s^{\alpha,\epsilon}_{n+1} - s^{\alpha,\epsilon}_n < 2\delta, s^{\alpha,\epsilon}_n < T + 2\delta\} \\ &\leq \lim_{\delta \to 0} \sup_{\alpha \in A} \sup_{n \geq 0} \mathbf{P}\{s^{\alpha,\epsilon}_{n+1} - s^{\alpha,\epsilon}_n < \delta, s^{\alpha,\epsilon}_n < T + 1\} = 0 \end{aligned}$$

and then applying Lemma 17.35 and Lemma 17.34 we get (i). □

Having arrived at an alternative criterion for equicontinuity we want to develop some techniques for estimating the probabilities therein.

LEMMA 17.37. *Let (S, r) be a separable metric space, let X be a process with paths in $D([0, \infty); S)$ and let $T > 0$ and $\beta > 0$ be given. Then for each $\delta > 0$, $\lambda > 0$ and weakly $\mathcal{F}^X$-optional time τ with $\tau \leq T$*

$$\boldsymbol{P}\{\sup_{0 \leq u \leq \delta} r(X_{\tau+u}, X_\tau) \wedge 1 \geq \lambda, \sup_{0 \leq v \leq \delta \wedge \tau} r(X_\tau, X_{\tau - v}) \wedge 1 \geq \lambda\} \leq \lambda^{-2\beta}(a_\beta + 2a_\beta^2(a_\beta + 4a_\beta^2))C(\delta)$$

and

$$\boldsymbol{P}\{\sup_{0 \leq u \leq \delta} r(X_u, X_0) \wedge 1 \geq \lambda\} \leq \lambda^{-2\beta} \left[(a_\beta(a_\beta + 4a_\beta^2))C(\delta) + a_\beta \boldsymbol{E}\left[r^\beta(X_\delta, X_0) \wedge 1\right]\right]$$

where a_β is any number that satisfies $(x + y)^\beta \leq a_\beta(x^\beta + y^\beta)$ for all $x, y \geq 0$ (e.g. $a_\beta = 2^{(\beta-1)\vee 0}$) and

$$C(\delta) = \sup_{\tau \in \mathfrak{S}(T+2\delta)} \sup_{0 \leq u \leq 2\delta} \boldsymbol{E}\left[\sup_{0 \leq v \leq 3\delta \wedge \tau} (r^\beta(X_{\tau+u}, X_\tau) \wedge 1)(r^\beta(X_\tau, X_{\tau-v}) \wedge 1)\right]$$

PROOF. We start off with Markov bound. For that we use an optional time to transform one of the suprema to a more managable form. Suppose we are given $0 < \eta < \lambda$ and a weakly $\mathcal{F}^X$-optional time τ and define

$$\Delta = \inf\{t > 0 \mid X_{\tau+t} - X_\tau > \lambda - \eta\}$$

so that Δ is a weakly $\mathcal{F}^X$-optional by right continuity of X openness of $(\lambda - \eta, \infty)$ and Lemma 9.70.

CLAIM 17.37.1. For all $0 < \eta < \lambda$, l

$$\begin{aligned} &\mathbf{P}\{\sup_{0 \leq u \leq \delta} r(X_{\tau+u}, X_\tau) \wedge 1 \geq \lambda, \sup_{0 \leq v \leq \delta \wedge \tau} r(X_\tau, X_{\tau-v}) \wedge 1 \geq \lambda\} \\ &\quad \leq \frac{1}{(\lambda - \eta)^\beta \lambda^\beta} \mathbf{E}\left[\sup_{0 \leq v \leq \delta \wedge \tau} (r^\beta(X_{\tau+\Delta\wedge\delta}, X_\tau) \wedge 1)(r^\beta(X_\tau, X_{\tau-v}) \wedge 1)\right] \end{aligned}$$

If $\sup_{0 \leq u \leq \delta} r(X_{\tau+u}, X_\tau) \wedge 1 \geq \lambda$ then for all $0 < \eta < \lambda$ there exists some $0 \leq u \leq \delta$ such that $r(X_{\tau+u}, X_\tau) \wedge 1 > \lambda - \eta$. This implies $\Delta \leq u \leq \delta$ which in turn by right continuity of X implies that

$$r(X_{\tau+\Delta\wedge\delta}, X_\tau) \wedge 1 = r(X_{\tau+\Delta}, X_\tau) \wedge 1 \geq \lambda - \eta$$

Therefore the claim follows from a Markov bound

$$\begin{aligned}
&\mathbf{P}\{\sup_{0 \leq u \leq \delta} r(X_{\tau+u}, X_\tau) \wedge 1 \geq \lambda, \sup_{0 \leq v \leq \delta\wedge\tau} r(X_\tau, X_{\tau-v}) \wedge 1 \geq \lambda\} \\
&\leq \mathbf{P}\{r(X_{\tau+\Delta\wedge\delta}, X_\tau) \wedge 1 \geq \lambda - \eta, \sup_{0 \leq v \leq \delta\wedge\tau} r(X_\tau, X_{\tau-v}) \wedge 1 \geq \lambda\} \\
&\leq \mathbf{P}\{\sup_{0 \leq v \leq \delta\wedge\tau} (r^\beta(X_{\tau+\Delta\wedge\delta}, X_\tau) \wedge 1)(r^\beta(X_\tau, X_{\tau-v}) \wedge 1) \geq (\lambda-\eta)^\beta \lambda^\beta\} \\
&\leq \frac{1}{(\lambda-\eta)^\beta\lambda^\beta} \mathbf{E}\left[\sup_{0 \leq v \leq \delta\wedge\tau} (r^\beta(X_{\tau+\Delta\wedge\delta}, X_\tau) \wedge 1)(r^\beta(X_\tau, X_{\tau-v}) \wedge 1)\right]
\end{aligned}$$

We turn our attention to bounding the right hand side of the claim. Let $0 \leq v \leq \delta \wedge \tau$ and use two applications of the triangle inequality to see

$$\begin{aligned}
&(r^\beta(X_{\tau+\Delta\wedge\delta}, X_\tau) \wedge 1)(r^\beta(X_\tau, X_{\tau-v}) \wedge 1) \\
&\leq a_\beta(r^\beta(X_{\tau+\Delta\wedge\delta}, X_{\tau+\delta}) \wedge 1)(r^\beta(X_\tau, X_{\tau-v}) \wedge 1) + \\
&a_\beta(r^\beta(X_{\tau+\delta}, X_\tau) \wedge 1)(r^\beta(X_\tau, X_{\tau-v}) \wedge 1) \\
&\leq a_\beta^2(r^\beta(X_{\tau+\Delta\wedge\delta}, X_{\tau+\delta}) \wedge 1)(r^\beta(X_\tau, X_{\tau+\Delta\wedge\delta}) \wedge 1) + \\
&a_\beta^2(r^\beta(X_{\tau+\Delta\wedge\delta}, X_{\tau+\delta}) \wedge 1)(r^\beta(X_{\tau+\Delta\wedge\delta}, X_{\tau-v}) \wedge 1) + \\
&a_\beta(r^\beta(X_{\tau+\delta}, X_\tau) \wedge 1)(r^\beta(X_\tau, X_{\tau-v}) \wedge 1)
\end{aligned}$$

taking the supremum and expectation we get

$$\begin{aligned}
&\mathbf{E}\left[\sup_{0 \leq v \leq \delta\wedge\tau} (r^\beta(X_{\tau+\Delta\wedge\delta}, X_\tau) \wedge 1)(r^\beta(X_\tau, X_{\tau-v}) \wedge 1)\right] \\
&\leq 2a_\beta^2 \mathbf{E}\left[\sup_{0 \leq v \leq \delta\wedge\tau} (r^\beta(X_{\tau+\Delta\wedge\delta}, X_{\tau+\delta}) \wedge 1)(r^\beta(X_{\tau+\Delta\wedge\delta}, X_{\tau-v}) \wedge 1)\right] + \\
&a_\beta \mathbf{E}\left[\sup_{0 \leq v \leq \delta\wedge\tau} (r^\beta(X_{\tau+\delta}, X_\tau) \wedge 1)(r^\beta(X_\tau, X_{\tau-v}) \wedge 1)\right]
\end{aligned}$$

As for the second term on the right hand side the situation is pretty straightforward, using the fact that $\tau \leq T < T + 2\delta$ we get the bound

$$a_\beta \mathbf{E}\left[\sup_{0 \leq v \leq \delta\wedge\tau} (r^\beta(X_{\tau+\delta}, X_\tau) \wedge 1)(r^\beta(X_\tau, X_{\tau-v}) \wedge 1)\right] \leq a_\beta \mathbf{E}\left[\sup_{0 \leq v \leq 2\delta\wedge\tau} (r^\beta(X_{\tau+\delta}, X_\tau) \wedge 1)(r^\beta(X_\tau, X_{\tau-v}) \wedge 1)\right]$$

As for the first term on the right hand side, we have to work harder. We prove the following

Claim 17.37.2. Let $\delta > 0$, $T > 0$, τ be weakly $\mathcal{F}^X$-optional time such that $\tau \leq T + \delta$ and let ξ be a random variable such that $0 \leq \xi \leq \delta$ then

$$\mathbf{E}\left[\sup_{0 \leq u \leq 2\delta\wedge\tau} (r^\beta(X_{\tau+\xi}, X_\tau) \wedge 1)(r^\beta(X_\tau, X_{\tau-u}) \wedge 1)\right] \leq (a_\beta + 4a_\beta)C(\delta)$$

First note that by the triangle inequality, a change of integration variables and the fact that $0 \leq \xi \leq \delta$ we have

$$
\begin{aligned}
r^\beta(X_{\tau+\xi}, X_\tau) \wedge 1 = \delta^{-1} \int_\delta^{2\delta} & r^\beta(X_{\tau+\xi}, X_\tau) \wedge 1 \, d\theta \\
& \leq \delta^{-1} a_\beta \int_\delta^{2\delta} (r^\beta(X_{\tau+\xi}, X_{\tau+\theta}) \wedge 1 + r^\beta(X_{\tau+\theta}, X_\tau) \wedge 1) \, d\theta \\
& \leq \delta^{-1} a_\beta \int_{\delta-\xi}^{2\delta-\xi} r^\beta(X_{\tau+\xi}, X_{\tau+\xi+\theta}) \wedge 1 \, d\theta + \\
& \delta^{-1} a_\beta \int_\delta^{2\delta} r^\beta(X_{\tau+\theta}, X_\tau) \wedge 1 \, d\theta \\
& \leq \delta^{-1} a_\beta \int_0^{2\delta} r^\beta(X_{\tau+\xi}, X_{\tau+\xi+\theta}) \wedge 1 \, d\theta + \\
& \delta^{-1} a_\beta \int_\delta^{2\delta} r^\beta(X_{\tau+\theta}, X_\tau) \wedge 1 \, d\theta
\end{aligned}
$$

From the previous inequality, the triangle inequality and the fact that $[(\tau - 2\delta) \vee 0, \tau] \subset [(\tau + \xi - 3\delta) \vee 0, \tau + \xi]$

$$
\begin{aligned}
& \sup_{0 \leq u \leq 2\delta \wedge \tau} (r^\beta(X_{\tau+\xi}, X_\tau) \wedge 1)(r(X_\tau, X_{\tau-u}) \wedge 1) \\
& \leq \delta^{-1} a_\beta \int_0^{2\delta} \sup_{0 \leq u \leq 2\delta \wedge \tau} (r^\beta(X_{\tau+\xi}, X_{\tau+\xi+\theta}) \wedge 1)(r(X_\tau, X_{\tau-u}) \wedge 1) \, d\theta + \\
& \delta^{-1} a_\beta \int_\delta^{2\delta} \sup_{0 \leq u \leq 2\delta \wedge \tau} (r^\beta(X_{\tau+\theta}, X_\tau) \wedge 1)(r(X_\tau, X_{\tau-u}) \wedge 1) \, d\theta \\
& \leq \delta^{-1} a_\beta^2 \int_0^{2\delta} (r^\beta(X_{\tau+\xi}, X_{\tau+\xi+\theta}) \wedge 1)(r(X_\tau, X_{\tau+\xi}) \wedge 1) \, d\theta + \\
& \delta^{-1} a_\beta^2 \int_0^{2\delta} \sup_{0 \leq u \leq 2\delta \wedge \tau} (r^\beta(X_{\tau+\xi}, X_{\tau+\xi+\theta}) \wedge 1)(r(X_{\tau+\xi}, X_{\tau-u}) \wedge 1) \, d\theta + \\
& \delta^{-1} a_\beta \int_\delta^{2\delta} \sup_{0 \leq u \leq 2\delta \wedge \tau} (r^\beta(X_{\tau+\theta}, X_\tau) \wedge 1)(r(X_\tau, X_{\tau-u}) \wedge 1) \, d\theta \\
& \leq 2\delta^{-1} a_\beta^2 \int_0^{2\delta} \sup_{0 \leq u \leq 3\delta \wedge (\tau+\xi)} (r^\beta(X_{\tau+\xi}, X_{\tau+\xi+\theta}) \wedge 1)(r(X_{\tau+\xi}, X_{\tau+\xi-u}) \wedge 1) \, d\theta + \\
& \delta^{-1} a_\beta \int_\delta^{2\delta} \sup_{0 \leq u \leq 2\delta \wedge \tau} (r^\beta(X_{\tau+\theta}, X_\tau) \wedge 1)(r(X_\tau, X_{\tau-u}) \wedge 1) \, d\theta \\
& \leq 2\delta^{-1} a_\beta^2 \int_0^{2\delta} \sup_{0 \leq u \leq 3\delta \wedge (\tau+\xi)} (r^\beta(X_{\tau+\xi}, X_{\tau+\xi+\theta}) \wedge 1)(r(X_{\tau+\xi}, X_{\tau+\xi-u}) \wedge 1) \, d\theta + \\
& \delta^{-1} a_\beta \int_\delta^{2\delta} \sup_{0 \leq u \leq 3\delta \wedge \tau} (r^\beta(X_{\tau+\theta}, X_\tau) \wedge 1)(r(X_\tau, X_{\tau-u}) \wedge 1) \, d\theta
\end{aligned}
$$

Now note that $\tau \leq T + \delta$ implies $\tau + \xi \leq T + 2\delta$ (it is still weakly optional since ξ is nonnegative) so by taking expectations, using Tonelli's Theorem and the definition of $C(\delta)$ we get

$$\mathbf{E}\left[\sup_{0\leq u\leq 2\delta\wedge\tau}(r^\beta(X_{\tau+\xi},X_\tau)\wedge 1)(r(X_\tau,X_{\tau-u})\wedge 1)\right]$$
$$\leq 2\delta^{-1}a_\beta^2\int_0^{2\delta}\mathbf{E}\left[\sup_{0\leq u\leq 3\delta\wedge(\tau+\xi)}(r^\beta(X_{\tau+\xi},X_{\tau+\xi+\theta})\wedge 1)(r(X_{\tau+\xi},X_{\tau+\xi-u})\wedge 1)\right]\,d\theta+$$
$$\delta^{-1}a_\beta\int_\delta^{2\delta}\mathbf{E}\left[\sup_{0\leq u\leq 3\delta\wedge\tau}(r^\beta(X_{\tau+\theta},X_\tau)\wedge 1)(r(X_\tau,X_{\tau-u})\wedge 1)\right]\,d\theta$$
$$\leq 2\delta^{-1}a_\beta^2\int_0^{2\delta}C(\delta)\,d\theta+\delta^{-1}a_\beta\int_\delta^{2\delta}C(\delta)\,d\theta=(4a_\beta^2+a_\beta)C(\delta)$$

and the claim is proven. (TODO: Make explicit where we use the $\mathcal{F}^X_{\tau+}$-measurability of ξ).

Now we return to the first term on the right hand side, $\tau\leq T$ implies $\tau+\Delta\wedge\delta\leq T+\delta$ and since $\Delta+\delta>0$ we know $\tau+\Delta\wedge\delta\in\mathfrak{S}(T+\delta)$. Also $0\leq\delta-\Delta\wedge\delta\leq\delta$ and $\delta-\Delta\wedge\delta$ is $\mathcal{F}^X_{\tau+\Delta\wedge\delta+}$ measurable (TODO: Show this). Lastly we have $0\leq v\leq\delta\wedge\tau$ implies $0\leq\Delta\wedge\delta+v\leq\Delta\wedge\delta+\delta\wedge\tau\leq 2\delta$ and $\Delta\wedge\delta+v\leq\tau+\Delta\wedge\delta$. Thus we apply Claim with weakly optional time $\tau+\Delta\wedge\delta$ and $\xi=\delta-\Delta\wedge\delta$,

$$2a_\beta^2\mathbf{E}\left[\sup_{0\leq v\leq\delta\wedge\tau}(r^\beta(X_{\tau+\Delta\wedge\delta},X_{\tau+\delta})\wedge 1)(r^\beta(X_{\tau+\Delta\wedge\delta},X_{\tau-v})\wedge 1)\right]$$
$$=2a_\beta^2\mathbf{E}\left[\sup_{0\leq v\leq\delta\wedge\tau}(r^\beta(X_{\tau+\Delta\wedge\delta},X_{\tau+\Delta\wedge\delta+(\delta-\Delta\wedge\delta)})\wedge 1)(r^\beta(X_{\tau+\Delta\wedge\delta},X_{\tau+\Delta\wedge\delta-(v+\Delta\wedge\delta)})\wedge 1)\right]$$
$$=2a_\beta^2\mathbf{E}\left[\sup_{0\leq v\leq 2\delta\wedge(\tau+\Delta\wedge\delta)}(r^\beta(X_{\tau+\Delta\wedge\delta},X_{\tau+\Delta\wedge\delta+(\delta-\Delta\wedge\delta)})\wedge 1)(r^\beta(X_{\tau+\Delta\wedge\delta},X_{\tau+\Delta\wedge\delta-v})\wedge 1)\right]$$
$$\leq 2a_\beta^2(a_\beta+2a_\beta^2)C(\delta)$$

Putting everything together we see that for all $0<\eta<\lambda$,

$$\mathbf{P}\{\sup_{0\leq u\leq\delta}r(X_{\tau+u},X_\tau)\wedge 1\geq\lambda,\ \sup_{0\leq v\leq\delta\wedge\tau}r(X_\tau,X_{\tau-v})\wedge 1\geq\lambda\}$$
$$\leq\frac{1}{(\lambda-\eta)^\beta\lambda^\beta}(a_\beta+2a_\beta^2(a_\beta+2a_\beta^2))C(\delta)$$

and we simply let $\eta\to 0$.

Now define

$$\Delta=\inf\{t>0\mid X_t-X_0>\lambda-\eta\}$$

and we get

$$r^{2\beta}(X_{\Delta\wedge\delta},X_0)\wedge 1\leq a_\beta(r^\beta(X_{\Delta\wedge\delta},X_\delta)\wedge 1)(r^\beta(X_{\Delta\wedge\delta},X_\delta)\wedge 1)+$$
$$a_\beta(r^\beta(X_0,X_\delta)\wedge 1)(r^\beta(X_{\Delta\wedge\delta},X_\delta)\wedge 1)$$

TODO: Finish □

Theorem 17.38. *Let (S,r) be a complete, separable metric space, let A be an arbitrary index set and let X^α be a stochastic process with values in $D([0,\infty);S)$ for all $\alpha\in A$. Suppose that for every $\epsilon>0$ and $t\in\mathbb{Q}_+$ there exists a compact set $K_{\epsilon,t}$ such that $\sup_{\alpha\in A}\boldsymbol{P}\{X^\alpha_t\in K_{\epsilon,t}\}>1-\epsilon$ then the following are equivalent*

(i) $\{X^\alpha\}$ *is relatively compact*
(ii) *For every* $T > 0$ *there exists* $\beta > 0$ *and non-negative random variables* $\gamma_{\alpha,\delta}$ *for* $\alpha \in A$, $0 < \delta < 1$ *such that*
 a) $\lim_{\delta \to 0} \sup_{\alpha \in A} \boldsymbol{E}[\gamma_{\alpha,\delta}] = 0$
 b) *For all* $0 \leq t \leq T$, $0 \leq u \leq \delta$ *and* $0 \leq v \leq \delta \wedge t$

$$\boldsymbol{E}\left[r^\beta(X^\alpha_{t+u}, X^\alpha_t) \wedge 1 \mid \mathcal{F}^{X^\alpha}_t\right](r^\beta(X^\alpha_t, X^\alpha_{t-v}) \wedge 1) \leq \boldsymbol{E}\left[\gamma_{\alpha,\delta} \mid \mathcal{F}^{X^\alpha}_t\right] \quad a.s.$$

 c) $\lim_{\delta \to 0} \sup_{\alpha \in A} \boldsymbol{E}\left[r^\beta(X^\alpha_\delta, X^\alpha_0) \wedge 1\right] = 0$
(iii) *For every* $T > 0$ *there exists* $\beta > 0$ *such that*

$$C(\alpha, \delta) = \sup_{\tau \in \mathfrak{S}^\alpha_0(T)} \sup_{0 \leq u \leq \delta} \boldsymbol{E}\left[\sup_{0 \leq v \leq \delta \wedge \tau} (r^\beta(X^\alpha_{\tau+u}, X^\alpha_\tau) \wedge 1)(r^\beta(X^\alpha_\tau, X^\alpha_{\tau-v}) \wedge 1)\right]$$

is defined for all $\alpha \in A$ *and* $0 < \delta < 1$ *and*
 a) $\lim_{\delta \to 0} \sup_{\alpha \in A} C(\alpha, \delta) = 0$
 b) $\lim_{\delta \to 0} \sup_{\alpha \in A} \boldsymbol{E}\left[r^\beta(X^\alpha_\delta, X^\alpha_0) \wedge 1\right] = 0$
($\mathfrak{S}^\alpha_0(T)$ *is the set of countably valued* $\mathcal{F}^{X^\alpha}$*-optional times* τ *such that* $\tau \leq T$*).*

PROOF. (i) $\implies$ (ii). Suppose $0 < \delta < 1$ and $T > 0$ are given and we select $0 \leq t \leq T$, $0 \leq u \leq \delta$ and $0 \leq v \leq \delta \wedge t$. If we let $0 = t_0 < t_1 < \cdots < t_{n-1} < T \leq t_n$ be any partition such that $\max_{1 \leq i \leq n}(t_i - t_{i-1}) > 2\delta$ then since $(t+u) - (t-v) \leq 2\delta$ it follows that there is an $1 \leq i \leq n$ such that either $t_{i-1} \leq t - v < t < t_i$ or $t_{i-1} \leq t < t + u < t_i$ or both. Thus we have

$$\begin{aligned}(r(X^\alpha_{t+u}, X^\alpha_t) \wedge 1)(r(X^\alpha_t, X^\alpha_{t-v}) \wedge 1) &\leq (r(X^\alpha_{t+u}, X^\alpha_t) \vee r(X^\alpha_t, X^\alpha_{t-v})) \wedge 1 \\ &\leq \max_{1 \leq i \leq n} \sup_{t_{i-1} \leq x < y < t_i} r(X^\alpha_x, X^\alpha_y) \wedge 1\end{aligned}$$

Taking the infimum over all partitions we get $(r(X^\alpha_{t+u}, X^\alpha_t) \wedge 1)(r(X^\alpha_t, X^\alpha_{t-v}) \wedge 1) \leq w'(X^\alpha, 2\delta, T) \wedge 1$ and by monotonicity of conditional expectation (iib) follows with $\gamma_{\alpha,\delta} = w'(X^\alpha, 2\delta, T) \wedge 1$. Property (iia) follows from Theorem 17.29; for any $0 < \lambda < 1$,

$$\begin{aligned}\mathbf{E}\left[w'(X^\alpha, 2\delta, T) \wedge 1\right] &= \mathbf{E}\left[w'(X^\alpha, 2\delta, T) \wedge 1; w'(X^\alpha, 2\delta, T) \geq \lambda\right] + \\ &\quad \mathbf{E}\left[w'(X^\alpha, 2\delta, T) \wedge 1; w'(X^\alpha, 2\delta, T) < \lambda\right] \\ &\leq \mathbf{P}\{w'(X^\alpha, 2\delta, T) \geq \lambda\} + \lambda \\ &\leq \sup_{\alpha \in A} \mathbf{P}\{w'(X^\alpha, 2\delta, T) \geq \lambda\} + \lambda\end{aligned}$$

and let $\delta \to 0$ and $\lambda \to 0$. By similar reasoning, we see that $r^\beta(X^\alpha_\delta, X^\alpha_0) \wedge 1 \leq w'(X^\alpha, \delta, 1) \wedge 1$ and therefore (iic) follows.

(ii) $\implies$ (iii). First let $\tau \in \mathfrak{S}^\alpha_0(T)$ with range $\{t_n\}_{n=1}^\infty$. Then by Lemma 9.31 $\mathcal{F}^{X^\alpha}_\tau$ agrees with $\mathcal{F}^{X^\alpha}_{t_n}$ on $\{\tau = t_n\}$ thus we may apply localization of conditional expectation Lemma 8.14 and (ii) to conclude for all $0 < \delta < 1$, $0 \leq u \leq \delta$ and

$0 \leq v \leq \delta \wedge \tau$,

$$\begin{aligned}
&\mathbf{E}\left[r^\beta(X^\alpha_{\tau+u}, X^\alpha_\tau) \wedge 1 \mid \mathcal{F}^{X^\alpha}_\tau\right](r^\beta(X^\alpha_\tau, X^\alpha_{\tau-v}) \wedge 1) \\
&= \sum_{n=1}^\infty \mathbf{E}\left[r^\beta(X^\alpha_{t_n+u}, X^\alpha_{t_n}) \wedge 1 \mid \mathcal{F}^{X^\alpha}_{t_n}\right](r^\beta(X^\alpha_{t_n}, X^\alpha_{t_n-v}) \wedge 1)\mathbf{1}_{\tau=t_n} \\
&\leq \sum_{n=1}^\infty \mathbf{E}\left[\gamma_{\alpha,\delta} \mid \mathcal{F}^{X^\alpha}_{t_n}\right]\mathbf{1}_{\tau=t_n} \\
&= \mathbf{E}\left[\gamma_{\alpha,\delta} \mid \mathcal{F}^{X^\alpha}_\tau\right]
\end{aligned}$$

where the equalities are all almost sure (taking a countable intersection of the almost sure events for each t_n). Now taking another countable intersection of almost sure events, we may take the supremum over $v \in [0, \delta \wedge \tau] \cap \mathbb{Q}$ to get

$$\mathbf{E}\left[r^\beta(X^\alpha_{\tau+u}, X^\alpha_\tau) \wedge 1 \mid \mathcal{F}^{X^\alpha}_\tau\right] \sup_{v \in [0,\delta\wedge\tau]\cap\mathbb{Q}} (r^\beta(X^\alpha_\tau, X^\alpha_{\tau-v}) \wedge 1) \leq \mathbf{E}\left[\gamma_{\alpha,\delta} \mid \mathcal{F}^{X^\alpha}_\tau\right] \text{ a.s.}$$

and by right continuity of X^α we may extend this to the supremum over all $0 \leq v \leq \delta \wedge \tau$. Now take the expectation and then the supremum over $0 \leq u \leq \delta$ and $\tau \in \mathfrak{S}^\alpha_0(T)$ to conclude $C(\alpha, \delta) \leq \mathbf{E}[\gamma_{\alpha,\delta}]$; (iiia) follows directly from (iia).

(iii) $\implies$ (i).

Claim 17.38.1.

$$C(\alpha, \delta) = \sup_{\tau \in \mathfrak{S}^\alpha_0(T)} \sup_{0 \leq u \leq \delta} \mathbf{E}\left[\sup_{0 \leq v \leq \delta\wedge\tau} (r^\beta(X^\alpha_{\tau+u}, X^\alpha_\tau) \wedge 1)(r^\beta(X^\alpha_\tau, X^\alpha_{\tau-v}) \wedge 1)\right]$$

TODO: Finish □

4. Test Function Methods

Is this really the same as what Kushner calls the perturbed test function method?

Theorem 17.39. *Let (S, r) be a complete, separable metric space, let A be an arbitrary index set and let X^α be a process with values in $D([0,\infty); S)$ for every $\alpha \in A$. Suppose for every $\epsilon > 0$ and $T > 0$ there exists a compact set $K_{\epsilon,T} \subset S$ such that*

$$\inf_{\alpha \in A} \boldsymbol{P}\{X^\alpha_t \in K_{\epsilon,T} \text{ for all } 0 \leq t \leq T\} \geq 1 - \epsilon$$

Suppose that H is a dense subset of $C_b(S; \mathbb{R})$ (topology of uniform convergence on compacts). Then $\{X^\alpha\}$ is relatively compact if and only if $\{h \circ X^\alpha\}$ is relatively compact for every $h \in H$.

Proof. If we assume that $\{X^\alpha\}$ is relatively compact then since $X \mapsto g \circ X$ is a continuous map from $D([0,\infty); S)$ to $D([0,\infty); \mathbb{R})$ for every continuous $g : S \to \mathbb{R}$ (Exercise 60) we know that $\{g \circ X^\alpha\}$ is relatively compact by the Continuous Mapping Theorem; a fortiori for every $h \in H$.

Now suppose H is as above and $\{h \circ X^\alpha\}$ is relatively compact for every $h \in H$. Let $g : S \to \mathbb{R}$ be an arbitrary continuous function, $\epsilon > 0$ and $T > 0$ and note that

$g(K_{\epsilon,T})$ is compact and

$$\inf_{\alpha\in A}\mathbf{P}\{g(X^\alpha_t)\in g(K_{\epsilon,T}) \text{ for all } 0\le t\le T\}\ge \inf_{\alpha\in A}\mathbf{P}\{X^\alpha_t\in K_{\epsilon,T} \text{ for all } 0\le t\le T\}\ge 1-\epsilon$$

thus the compact containment condition holds for every $h\in H$ and $g\in C_b(S;\mathbb{R})$.

We now observe that the relative compactness of $h\circ X^\alpha$ for $h\in H$ extends to the closure of H.

Claim 17.39.1. $\{g\circ X^\alpha\}$ is relatively compact for every $g\in C_b(S;\mathbb{R})$.

Let $g\in C_b(S;\mathbb{R})$, $T>0$, $\lambda>0$ and $0<\epsilon<\lambda$ be given. Select $h\in H$ such that

$$\sup_{x\in K_{\epsilon,T+1}} r(h(x),g(x))<\epsilon/2$$

and $\alpha\in A$. Using the fact that

$$w'(g\circ X^\alpha,\delta,T)\le w'(h\circ X^\alpha,\delta,T)+2\sup_{0\le t\le T+\delta} r(g(X^\alpha_t),h(X^\alpha_t))$$

we get for all $0<\delta<1$

$$\begin{aligned}\mathbf{P}\{w'(g\circ X^\alpha,\delta,T)\ge\lambda\}&\le \mathbf{P}\{w'(g\circ X^\alpha,\delta,T)\ge\lambda; X^\alpha_t\in K_{\epsilon,T+1} \text{ for all } 0\le t\le T+1\}+\epsilon\\&\le \mathbf{P}\{w'(h\circ X^\alpha,\delta,T)\ge\lambda-\epsilon\}+\epsilon\end{aligned}$$

Take the supremum over α and the limit $\delta\to 0$ and use Theorem 17.29

$$\lim_{\delta\to 0}\sup_{\alpha\in A}\mathbf{P}\{w'(g\circ X^\alpha,\delta,T)\ge\lambda\}\le\epsilon$$

Letting $\epsilon\to 0$ and using Theorem 17.29 we conclude that $\{g\circ X^\alpha\}$ is relatively compact.

Pick $\epsilon>0$, $T>0$ and $\eta>0$. By compactness of $K_{\epsilon,T}$ we may find a finite set $z_1,\dots,z_N$ such that the open balls $B(z_i,\eta)$ cover $K_{\epsilon,T}$. By the claim we know that $(r(\circ,z_i)\wedge 1)\circ X^\alpha$ is relatively compact for each $i=1,\dots,N$.

Claim 17.39.2. For every $y\in K_{\epsilon,T}$ and $x\in S$ then

$$(r(x,y)\wedge 1)\le\max_{1\le i\le N}|(r(x,z_i)\wedge 1)-(r(z_i,y)\wedge 1)|+2\eta$$

Pick $1\le i\le N$ such that $r(y,z_i)<\eta$. By the triangle inequality

$$(r(x,y)\wedge 1)\le(r(x,z_i)\wedge 1)+(r(z_i,y)\wedge 1)$$

If $r(x,z_i)\le r(z_i,y)<\eta$ then

$$\begin{aligned}(r(x,z_i)\wedge 1)+(r(z_i,y)\wedge 1)&=(r(z_i,y)\wedge 1)-(r(x,z_i)\wedge 1)+2(r(x,z_i)\wedge 1)\\&\le|(r(x,z_i)\wedge 1)-(r(z_i,y)\wedge 1)|+2\eta\end{aligned}$$

and if $r(z_i,y)<r(x,z_i)$ then

$$\begin{aligned}(r(x,z_i)\wedge 1)+(r(z_i,y)\wedge 1)&=(r(x,z_i)\wedge 1)-(r(z_i,y)\wedge 1)+2(r(z_i,y)\wedge 1)\\&\le|(r(x,z_i)\wedge 1)-(r(z_i,y)\wedge 1)|+2\eta\end{aligned}$$

From the claim and Proposition 17.33, for all $0 \leq t \leq T$, $0 \leq u \leq \delta$ and $0 \leq v \leq \delta \wedge t$

$$
\begin{aligned}
&(r(X^\alpha_{t+u}, X^\alpha_t) \wedge 1)(r(X^\alpha_t, X^\alpha_{t-v}) \wedge 1) \\
&\leq (r(X^\alpha_{t+u}, X^\alpha_t) \wedge 1)(r(X^\alpha_t, X^\alpha_{t-v}) \wedge 1)\mathbf{1}_{X^\alpha_t \in K_{\epsilon,T} \text{ for all } 0\leq t\leq T} + \mathbf{1}_{X^\alpha_t \notin K_{\epsilon,T} \text{ for some } 0\leq t\leq T} \\
&\leq \max_{1\leq i\leq N} (\left|(r(X^\alpha_{t+u}, z_i) \wedge 1) - (r(z_i, X^\alpha_t) \wedge 1)\right| + 2\eta)(\left|(r(X^\alpha_t, z_i) \wedge 1) - (r(z_i, X^\alpha_{t-v}) \wedge 1)\right| + 2\eta)\mathbf{1}_{X^\alpha_t \in K_{\epsilon,T} \text{ for all } 0\leq t} \\
&+ \mathbf{1}_{X^\alpha_t \notin K_{\epsilon,T} \text{ for some } 0\leq t\leq T} \\
&\leq \max_{1\leq i\leq N} \left|(r(X^\alpha_{t+u}, z_i) \wedge 1) - (r(z_i, X^\alpha_t) \wedge 1)\right| \left|(r(X^\alpha_t, z_i) \wedge 1) - (r(z_i, X^\alpha_{t-v}) \wedge 1)\right| + 4(\eta + \eta^2) \\
&+ \mathbf{1}_{X^\alpha_t \notin K_{\epsilon,T} \text{ for some } 0\leq t\leq T} \\
&\leq \max_{1\leq i\leq N} w'((r(\cdot, z_i) \wedge 1) \circ X^\alpha, 2\delta, T + \delta) \wedge 1 + 4(\eta + \eta^2) + \mathbf{1}_{X^\alpha_t \notin K_{\epsilon,T} \text{ for some } 0\leq t\leq T} \\
&\leq \sum_{i=1}^N w'((r(\cdot, z_i) \wedge 1) \circ X^\alpha, 2\delta, T + \delta) \wedge 1 + 4(\eta + \eta^2) + \mathbf{1}_{X^\alpha_t \notin K_{\epsilon,T} \text{ for some } 0\leq t\leq T}
\end{aligned}
$$

This estimate is just about what we need in order to apply Theorem 17.38. The right hand side of the estimate has quite a few parameters (e.g. ϵ, η, δ) and we just need to make some choices in order get upper bounds $\gamma_{\alpha,\delta}$ such that $\lim_{\delta\to 0} \sup_{\alpha\in A} \mathbf{E}\left[\gamma_{\alpha,\delta}\right] = 0$. For each $n \in \mathbb{N}$, we pick $K_{1/3n,T}$ and $\eta_n > 0$ such that $4(\eta_n + \eta_n^2) < 1/3n$. Now choose $z_1, \dots, z_N$ and use the fact that for each $1 \leq i \leq N$

$$\lim_{\delta\to 0} \sup_{\alpha\in A} \mathbf{E}\left[w'((r(\cdot, z_i) \wedge 1) \circ X^\alpha, 2\delta, T + \delta) \wedge 1\right] = 0$$

(relative compactness of $\{(r(\cdot, z_i) \wedge 1) \circ X^\alpha\}$ and Theorem 17.29) to pick $\delta_n < \delta_{n-1}$ such that

$$\sum_{i=1}^N \sup_{\alpha\in A} \mathbf{E}\left[w'((r(\cdot, z_i) \wedge 1) \circ X^\alpha, 2\delta, T + \delta) \wedge 1\right] < 1/3n \text{ for all } \delta \leq \delta_n$$

Lastly define for $\delta_{n+1} < \delta \leq \delta_n$,

$$\gamma_{\alpha,\delta} = \sum_{i=1}^N w'((r(\cdot, z_i) \wedge 1) \circ X^\alpha, 2\delta, T + \delta) \wedge 1 + 1/3n + \mathbf{1}_{X^\alpha_t \notin K_{1/3n,T} \text{ for some } 0\leq t\leq T}$$

and we have

$$\sup_{\substack{0\leq t\leq T \\ 0\leq u\leq\delta \\ 0\leq v\leq\delta\wedge t}} (r(X^\alpha_{t+u}, X^\alpha_t) \wedge 1)(r(X^\alpha_t, X^\alpha_{t-v}) \wedge 1) \leq \gamma_{\alpha,\delta}$$

and

$$\lim_{\delta\to 0} \sup_{\alpha\in A} \mathbf{E}\left[\gamma_{\alpha,\delta}\right] = 0$$

which imply conditions (iia) and (iib) of Theorem 17.38.

By a similar but easier argument, we see that with $\epsilon > 0$ $z_1, \ldots, z_N$ and $\eta > 0$ chosen as above we get

$$\begin{aligned}\lim_{\delta \to 0} \sup_{\alpha \in A} \mathbf{E}\left[r(X^\alpha_\delta, X^\alpha_0) \wedge 1\right] &\leq \lim_{\delta \to 0} \sup_{\alpha \in A} \mathbf{E}\left[\max_{1 \leq i \leq N} \left|(r(X^\alpha_\delta, z_i) \wedge 1) - (r(X^\alpha_0, z_i) \wedge 1)\right| + 2\eta + \mathbf{1}_{X^\alpha_0 \notin K_{\epsilon,T}}\right] \\ &\leq \sum_{i=1}^N \lim_{\delta \to 0} \sup_{\alpha \in A} \mathbf{E}\left[\left|(r(X^\alpha_\delta, z_i) \wedge 1) - (r(X^\alpha_0, z_i) \wedge 1)\right|\right] + 2\eta + \mathbf{P}\{X^\alpha_0 \notin K_{\epsilon,T}\}\end{aligned}$$

where we have applied Theorem 17.38 to the relatively compact families $\{(r(\cdot, z_i) \wedge 1) \circ X^\alpha\}$. Now we may let $\eta \to 0$ and then $\epsilon \to 0$ to conclude $\lim_{\delta \to 0} \sup_{\alpha \in A} \mathbf{E}\left[r(X^\alpha_\delta, X^\alpha_0) \wedge 1\right] = 0$. Now apply Theorem 17.38 to conclude that $\{X^\alpha\}$ is relatively compact. $\square$

THEOREM 17.40. *Let (S, r) be a metric space, let A be an arbitrary set and for each $\alpha \in A$ let $(\Omega^\alpha, \mathcal{A}^\alpha, P^\alpha)$ be a probability space with filtration $\mathcal{F}^\alpha$ and let X^α be an $\mathcal{F}^\alpha$-adapted stochastic process with values in $D([0, \infty); S)$. Let C be a subalgebra of $U_b(S; \mathbb{R})$ where for every $f \in C$ and $\epsilon > 0$, $T > 0$ and $\alpha \in A$ there exist $\mathcal{F}^\alpha$-progressive processes Y^α and Z^α satisfying*

(i) $\sup_{0 \leq t < \infty} \boldsymbol{E}\left[|Y^\alpha_t|\right] < \infty$
(ii) $\sup_{0 \leq t < \infty} \boldsymbol{E}\left[|Z^\alpha_t|\right] < \infty$
(iii) $Y^\alpha_t - \int_0^t Z^\alpha_s \, ds$ *is an* $\mathcal{F}^\alpha$*-martingale*
(iv) $\sup_{\alpha \in A} \boldsymbol{E}\left[\sup_{t \in [0,T] \cap \mathbb{Q}} |Y^\alpha_t - f(X^\alpha_t)|\right] < \epsilon$
(v) $\sup_{\alpha \in A} \|Z^\alpha\|_{p,T} < \infty$ *for some* $1 < p \leq \infty$ *where* $\|Z\|_{p,T} = \boldsymbol{E}\left[(\infty_0^T |Z_s|^p \, ds)^{1/p}\right]$ *for* $1 < p < \infty$ *and* $\|Z\|_{\infty,T} = \boldsymbol{E}\left[\operatorname{ess\,sup}_{0 \leq t \leq T} |Z_t|\right]$

Then $\{f \circ X^\alpha\}$ is relatively compact for every $f \in C$ and in fact the family $\{(f_1, \ldots, f_k) \circ X^\alpha\}$ is relatively compact for every $f_1, \ldots, f_k \in C$.

PROOF. Let $f \in C$, $\epsilon > 0$ and $T > 0$ be given. Pick Y^α and Z^α be selected satisfying the hypotheses for $f \in C$ and $T+1$ and let $\tilde{Y}^\alpha$ and $\tilde{Z}^\alpha$ be selected satisfying the hypotheses of the theorem for $f^2 \in C$ and $T+1$. Thus we have

$$\sup_{\alpha \in A} \mathbf{E}\left[\sup_{t \in [0,T+1] \cap \mathbb{Q}} |Y^\alpha_t - f(X^\alpha_t)|\right] < \epsilon$$
$$\sup_{\alpha \in A} \mathbf{E}\left[\sup_{t \in [0,T+1] \cap \mathbb{Q}} \left|\tilde{Y}^\alpha_t - f^2(X^\alpha_t)\right|\right] < \epsilon$$
$$\sup_{\alpha \in A} \|Z^\alpha\|_{p,T+1} < \infty \text{ for some } 1 < p \leq \infty$$
$$\sup_{\alpha \in A} \left\|\tilde{Z}^\alpha\right\|_{\tilde{p},T+1} < \infty \text{ for some } 1 < \tilde{p} \leq \infty$$

Let $0 < \delta < 1$ be given and choose $t \in [0, T] \cap \mathbb{Q}$ and $u \in [0, \delta] \cap \mathbb{Q}$ then

$$
\begin{aligned}
&\mathbf{E}\left[(f(X^\alpha_{t+u}) - f(X^\alpha_t))^2 \mid \mathcal{F}^\alpha_t\right] \\
&= \mathbf{E}\left[f^2(X^\alpha_{t+u}) - 2f(X^\alpha_t)f(X^\alpha_{t+u}) + f^2(X^\alpha_t) \mid \mathcal{F}^\alpha_t\right] \\
&= \mathbf{E}\left[f^2(X^\alpha_{t+u}) - f^2(X^\alpha_t) - 2f(X^\alpha_t)(f(X^\alpha_{t+u}) - f(X^\alpha_t)) \mid \mathcal{F}^\alpha_t\right] \\
&= \mathbf{E}\left[f^2(X^\alpha_{t+u}) - \tilde{Y}^\alpha_{t+u} \mid \mathcal{F}^\alpha_t\right] + \mathbf{E}\left[\tilde{Y}^\alpha_t - f^2(X^\alpha_t) \mid \mathcal{F}^\alpha_t\right] + \mathbf{E}\left[\tilde{Y}^\alpha_{t+u} - \tilde{Y}^\alpha_t \mid \mathcal{F}^\alpha_t\right] - \\
&2f(X^\alpha_t)\{\mathbf{E}\left[f(X^\alpha_{t+u}) - Y^\alpha_t \mid \mathcal{F}^\alpha_t\right] + \mathbf{E}\left[Y^\alpha_{t+u} - f(X^\alpha_t) \mid \mathcal{F}^\alpha_t\right] + \mathbf{E}\left[Y^\alpha_{t+u} - Y^\alpha_t \mid \mathcal{F}^\alpha_t\right]\} \\
&\leq 2\mathbf{E}\left[\sup_{s \in [0,T+1]\cap\mathbb{Q}} \left|\tilde{Y}^\alpha_s - f^2(X^\alpha_s)\right| \mid \mathcal{F}^\alpha_t\right] + \mathbf{E}\left[\sup_{0 \leq r \leq T} \int_r^{r+\delta} \left|\tilde{Z}^\alpha_s\right| \, ds \mid \mathcal{F}^\alpha_t\right] + \\
&4\left\|f\right\|_\infty \mathbf{E}\left[\sup_{s \in [0,T+1]\cap\mathbb{Q}} \left|Y^\alpha_s - f(X^\alpha_s)\right| \mid \mathcal{F}^\alpha_t\right] + 2\left\|f\right\|_\infty \mathbf{E}\left[\sup_{0 \leq r \leq T} \int_r^{r+\delta} \left|Z^\alpha_s\right| \, ds \mid \mathcal{F}^\alpha_t\right]
\end{aligned}
$$

Take the conjugate pairs q and $\tilde{q}$ so that $\frac{1}{p} + \frac{1}{q} = 1$ and $\frac{1}{\tilde{p}} + \frac{1}{\tilde{q}} = 1$. By Hölder's Inequality we have

$$
\begin{aligned}
\int_r^{r+\delta} \left|Z^\alpha_s\right| \, ds &\leq \left(\int_r^{r+\delta} \left|Z^\alpha_s\right|^p \, ds\right)^{1/p} \left(\int_r^{r+\delta} ds\right)^{1/q} \\
&\leq \delta^{1/q} \left(\int_0^{T+1} \left|Z^\alpha_s\right|^p \, ds\right)^{1/p}
\end{aligned}
$$

and thus if we define

$$
\begin{aligned}
\gamma_{\alpha,\delta} = 2 \sup_{s \in [0,T+1]\cap\mathbb{Q}} &\left|\tilde{Y}^\alpha_s - f^2(X^\alpha_s)\right| + \delta^{1/\tilde{q}} \left(\int_0^{T+1} \left|\tilde{Z}^\alpha_s\right|^{\tilde{p}} \, ds\right)^{1/\tilde{p}} \\
4\left\|f\right\|_\infty \sup_{s \in [0,T+1]\cap\mathbb{Q}} &\left|Y^\alpha_s - f(X^\alpha_s)\right| + 2\delta^{1/q} \left(\int_0^{T+1} \left|Z^\alpha_s\right|^p \, ds\right)^{1/p}
\end{aligned}
$$

we have $\mathbf{E}\left[(f(X^\alpha_{t+u}) - f(X^\alpha_t))^2 \mid \mathcal{F}^\alpha_t\right] \leq \mathbf{E}\left[\gamma_{\alpha,\delta} \mid \mathcal{F}^\alpha_t\right]$ for all $t \in [0, T] \cap \mathbb{Q}$. Note that this extends to all $0 \leq t \leq T$ and all $0 \leq u \leq \delta$ by right continuity of $f \circ X^\alpha$, monotone convergence for conditional expectations and the tower property of conditional expectations. For example to extend to arbitrary t,

$$
\begin{aligned}
\mathbf{E}\left[(f(X^\alpha_{t+u}) - f(X^\alpha_t))^2 \mid \mathcal{F}^\alpha_t\right] &= \lim_{\substack{r \to t^+ \\ r \in \mathbb{Q}}} \mathbf{E}\left[(f(X^\alpha_{r+u}) - f(X^\alpha_r))^2 \mid \mathcal{F}^\alpha_t\right] \\
&= \lim_{\substack{r \to t^+ \\ r \in \mathbb{Q}}} \mathbf{E}\left[\mathbf{E}\left[(f(X^\alpha_{r+u}) - f(X^\alpha_r))^2 \mid \mathcal{F}^\alpha_r\right] \mid \mathcal{F}^\alpha_t\right] \\
&\leq \lim_{\substack{r \to t^+ \\ r \in \mathbb{Q}}} \mathbf{E}\left[\mathbf{E}\left[\gamma_{\alpha,\delta} \mid \mathcal{F}^\alpha_r\right] \mid \mathcal{F}^\alpha_t\right] \\
&= \mathbf{E}\left[\gamma_{\alpha,\delta} \mid \mathcal{F}^\alpha_t\right]
\end{aligned}
$$

and the argument for extending to arbitrary u is even easier.

We also have the bound

$$\begin{aligned}\sup_{\alpha \in A} \mathbf{E}\left[\gamma_{\alpha,\delta}\right] \leq 2\epsilon + \delta^{1/\tilde{q}} \sup_{\alpha \in A} \left\| \tilde{Z}^{\alpha} \right\|_{\tilde{p},T} \\ + 4 \left\| f \right\|_{\infty} \epsilon + 2\delta^{1/q} \sup_{\alpha \in A} \left\| Z^{\alpha} \right\|_{p,T}\end{aligned}$$

This bound implies conditions (iia), (iib) and (iic) of Theorem 17.38 from which we conclude that $\{f \circ X^{\alpha}\}$ is relatively compact. (TODO: Write down the song and dance about selecting ϵ decreasing to make this work as in the proof of Theorem 17.39)

TODO: Finish the case of $(f_1, \ldots, f_k)$. $\square$

CHAPTER 18

Feller Processes

We now specialize to the case of time homogeneous Markov processes and develop an approach that allow one to bring powerful tools of functional analysis to bear on the theory of Markov processes and ultimately elucidates a deep connection between Markov processes and partial differential equations. Any treatment of this topic must make several pedagogical decisions. The functional analysis tools we will use are part of the theory of operator semigroups and one could simply assume the reader has been exposed to them and quote the results with appropriate references. We deem such an approach an undue burden on the reader as any treatment of semigroups is likely deeply embedded in a textbook in which the core results and difficult to extract efficiently (much of semigroup theory is motivated by differential equations and not probability theory). Thus we have the choice of how to present the required functional analysis. One choice is to present the results in a separate chapter or appendix and the other is to present the results on an as needed basis. While we have relegated the basic theory of Banach spaces to an appendix, we have chosen the second path for the theory of operator semigroups hoping that the probabilistic development can provide motivation for the functional analysis and make it easier to digest. The disadvantage in doing things this way is that functional analysis results become spread out thinly through the text and that a reader looking for a particular result cannot find it without knowing or guessing the places that it might be used. We accept this disadvantage in hopes that the spirit of the interaction between the fields can be better appreciated.

TODO:

- Chapman Kolmogorov relation is equivalent to semigroup property
- Feller process defined in terms of Feller semigroup properties
- Feller semigroup generators
- Feller semigroup strongly continuous (seen via Yosida approximation)
- Kolmogorov forward/backward equations follow from strong continuity
- Generators of strongly continuous semigroups and Feller semigroups characterized (Hille-Yosida)
- Convergence/Approximation of semigroups in terms of generators
- Convergence of Markov processes in terms of convergence of semigroups/generators
- Every Feller semigroup has an associated cadlag Feller process (approximation by pure jump-type processes or by Kinney regularization)
- Approximation of Feller semigroup by Markov chains
- Continuous sample paths and elliptic generators

1. Semigroups and Generators

The first step is to change the point of view on transition kernels slightly. In the case of a time homogeneous Markov process, the family of transition kernels is a single parameter family of kernels μ_t. Note that in the discrete time case it is clear that the entire family of kernels is generated by the single time unit kernel $\mu = \mu_1$ via kernel multiplication $\mu_n = \mu^n$ (in the case of discrete time Markov chains this is just matrix multiplication). The first question that we will pursue is whether there is an analogy in the continuous time case. The Chapman Kolmogorov relation gives us a hint on how to proceed. In the time homogeneous case the Chapman Kolmogorov relation says that $\mu_s \mu_t = \mu_{s+t}$ which is the *semigroup property* and suggests that we may be able to write μ_s as $exp(sA)$ for some appropriately defined A. With some additional assumptions this may be done, but first we want to recast the transition kernels in a different light in which these questions may be more naturally resolved. Let f be a measurable function on S that is either non-negative or bounded. For any probability kernel $\mu : S \to \mathcal{P}(S)$, by Lemma 8.29 we know that $\int f(t)\, \mu(s, dt)$ is a itself a measurable function of s that is non-negative or bounded when f is. Thus if we are given the transition kernels of a time homogeneous Markov process we may define an operator

$$T_t f(s) = \int f(u)\, \mu_t(s, du)$$

on an appropriate space of measurable functions to itself (say the space of bounded measurable functions). The first thing to observe is that the Chapman-Kolmogorov relations are equivalent to the semigroup property for these operators.

Proposition 18.1. *Let μ_t for $t \geq 0$ be a family of probability kernels on a measurable space $(S, \mathcal{S})$ and define $T_t f(x) = \int f(s)\, \mu_t(x, ds)$ for all bounded measurable function $f : S \to \mathbb{R}$, then μ_t satisfies the Chapman-Kolmogorov relations if and only if $T_t T_s = T_{t+s}$ for all $t, s \geq 0$.*

Proof. Let $A \in \mathcal{S}$ then we have $T_{t+s}\mathbf{1}_B(x) = \mu_{t+s}(x, B)$ and

$$\begin{aligned} T_t T_s \mathbf{1}_B(x) &= \int T_s \mathbf{1}_B(y)\, \mu_t(x, dy) = \int \mu_s(y, B)\, \mu_t(x, dy) \\ &= \mu_t \mu_s(B) \end{aligned}$$

therefore the Chapman-Kolomogorov relations are equivalent to $T_t T_s \mathbf{1}_B = T_{t+s}\mathbf{1}_B$ for all $B \in \mathcal{S}$. Therefore the semigroup property implies the Chapman-Kologorov relations and if the the Chapman-Kolmogorov relations hold then the semigroup property holds for simple functions by linearity. If f is positive, bounded and measurable then find simple functions $f_n \uparrow f$ then by Monotone Convergence

$$T_t f(x) = \int f(s)\, \mu_t(x, ds) = \lim_{n\to\infty} \int f_n(s)\, \mu_t(x, ds) = \lim_{n\to\infty} T_t f_n(x)$$

Moreover $T_t f_n(x)$ is increasing by positivity of integral and therefore another application of Monotone Convergence shows

$$T_t T_s f(x) = \lim_{n\to\infty} T_t T_s f_n(x) = \lim_{n\to\infty} T_{t+s} f_n(x) = T_{t+s} f(x)$$

The semigroup property extends to arbitrary bounded measurable f by writing $f = f_+ - f_-$ with $f_\pm \geq 0$ and using linearity of T_t. $\square$

In the case that the kernels μ_t are the transition kernels of a Markov process we will call the semigroup T_t the transition semigroup of the Markov process. It is worth recording for future use the expression for the transition semigroup in terms of the underlying Markov process.

PROPOSITION 18.2. *Let X be a homogeneous Markov process with time scale $\mathbb{R}_+$ and state space S then for every bounded measurable function $f : S \to \mathbb{R}$ and $s \geq 0$,*

$$T_t f(X_s) = \boldsymbol{E}[f(X_{t+s}) \mid X_s] = \boldsymbol{E}[f(X_{t+s}) \mid \mathcal{F}_s]$$

in particular, if X_t is a Markov family (e.g. X is canonical) then

$$T_t f(x) = \boldsymbol{E}_x\left[f(X_t)\right]$$

PROOF. By definition we know that $\mathbf{P}\{X_{t+s} \in \cdot \mid X_s\}(\omega) = \mu_t(X_s(\omega), \cdot)$ and therefore by distintigration (Theorem 8.35) we have

$$\mathbf{E}\left[f(X_{t+s}) \mid X_s\right](\omega) = \int f(s)\mu_t(X_s(\omega), ds) = T_t f(X_s)$$

If X is a Markov family then for $x \in S$ under the measure P_x we have by the Markov property (Theorem 13.29) and the first part of this result

$$\mathbf{E}_x\left[f(X_t)\right] = \mathbf{E}\left[f(X_t) \mid X_0\right] = T_t f(X_0) = T_t f(x)$$

□

PROPOSITION 18.3. *Let X be a pure jump-type Markov process with state space S and a bounded rate kernel α. For every bounded measurable function $f : S \to \mathbb{R}$ define*

$$Af(x) = \int \left(f(y) - f(x)\right) \alpha(x, dy)$$

then A is a bounded linear operator from $B(S) \to B(S)$ and $T_t = e^{tA}$.

PROOF. Linearity of A is immediate from the linearity of the integral. Write $\alpha(x,\cdot) = c(x)\mu(x,\cdot)$ so that $Af(x) = c(x)\int (f(y) - f(x))\, \mu(x, dy) = c(x)(\int f(y)\, \mu(x, dy) - f(x))$ and the boundedness of α means that $\|c\|_\infty < \infty$. Now given $f \in B(S)$, the fact that α is a kernel (Lemma 13.77) and Lemma 8.29 imply that Af is a measurable function on S. In addition, we have $|Af(x)| \leq |c(x)| \int |f(y) - f(x)|\, \mu(x, dy) \leq 2\|c\|_\infty \|f\|_\infty$ which shows that $Af \in B(S)$ and moreover shows upon taking the supremum over $x \in S$ that $\|Af\| \leq 2\|c\|_\infty \|f\|_\infty$ so that A is a bounded operator with $\|A\| \leq 2\|c\|_\infty$.

TODO: Finish. Kallenberg's proof is very elegant as it reduces to the case in which $c(x)$ is constant and therefore we have a pseudo-Poisson process. I haven't proven that result as it is tied up a bit in my confusion about how to sort out Markov families. I should try to come up with a proof that avoids the reduction and just uses the Strong Markov property. □

Given a general operator semigroup T_t we can think a bit about how find an operator A such that $T_t = e^{tA}$. There are a couple of ways to proceed, but perhaps the simplest is to observe that looking at the equality pointwise in $B(S)$ and formally differentiating at $t = 0$ we'd get $\lim_{t\to 0} \frac{T_t f - f}{t} = Af$. The trick in making this formal is to not assume that the resulting A will be defined for all $B(S)$.

Definition 18.4. Let T_t be an operator semigroup on a Banach space X then the *generator* is the operator A defined by $Av = \lim_{t\to 0} \frac{T_t v - v}{t}$ for all $v \in X$ for which the limit on the right exists.

We reiterate that A is not necessarily defined everywhere and that part of the definition of the generator is its domain of definition. Unless it is explicity noted otherwise we will use $\mathcal{D}(A)$ to denote the domain of a generator.

It is trivial to see the following.

Proposition 18.5. *The generator is a linear operator on its domain of definition.*

Proof. This follows directly from linearity of the T_t and linearity of limits. If $v, w \in \mathcal{D}(A)$ and $a \in \mathbb{R}$ then

$$\lim_{t\to 0} \frac{T_t av - av}{t} = a \lim_{t\to 0} \frac{T_t v - v}{t} = aAv$$

$$\lim_{t\to 0} \frac{T_t(v+w) - (v+w)}{t} = \lim_{t\to 0} \frac{T_t v - v}{t} + \lim_{t\to 0} \frac{T_t w - w}{t} = Av + Aw$$

□

2. Strongly Continuous and Feller Semigroups

We begin the process of seeing how the theory of operator semigroups and Markov processes interacts from the semigroup point of view. We shall enumerate some properties of semigroups that make them well behaved in important ways and then define an associated class of Markov processes as those whose semigroup posess these properties.

Definition 18.6. A *semigroup of operators* T_t on a Banach space X is one parameter family of bounded continuous operators $T_t : X \to X$ such that for all $0 \geq s, t < \infty$ we have $T_s \circ T_t = T_{s+t}$.

Definition 18.7. A semigroup of operators T_t on a Banach space X is said to be *strongly continuous* if $\lim_{t\to 0} T_t v = v$ for all $v \in X$.

Example 18.8. Let $A : X \to X$ be a bounded linear operator then $T_t = e^{tA}$ is a strongly continuous semigroup. Strong continuity follows from the much stronger property that e^{tA} is a continuous function $\mathbb{R}$ into $L(X)$.

In fact paths of strongly continuous semigroups are continuous; to see this we first need the following result.

Lemma 18.9. *Let T_t be a strongly continuous semigroup on a Banach space X then there exists constants $M \geq 1$ and $c > 0$ such that $\|T_t\| \leq Me^{ct}$.*

Proof.

Claim 18.9.1. There exists a $t_0 > 0$ and $M \geq 1$ such that $\|T_s\| \leq M$ for all $0 \leq s \leq t_0$

Suppose that the claim is not true then clearly we can find $0 \leq t_1 \leq 1$ such that $\|T_{t_1}\| > 1$; since $\|T_0\| = 1$ we actually know that $t_1 > 0$. If we have $0 \leq t_n \leq \cdots \leq t_1$ with $t_j \leq 1/j$ and $T_{t_j} \geq j$ for $j = 1, \ldots, n$ then there must also be $0 \leq t_{n+1} \leq t_n \wedge 1/(n+1)$ such that $\left\|T_{t_{n+1}}\right\| > n+1$; as before since $\|T_0\| = 1$ we know that $0 < t_{n+1}$. In this way, we find a sequence $t_n \downarrow 0$ for which $\|T_{t_n}\| \geq n$. Now by the Principle of Uniform Boundedness Theorem 15.47 we conclude that T_{t_n} are not pointwise bounded so there must exist a $v \in X$ for which $sup_n \|T_{t_n} v\| = \infty$. It then follows that $\lim_{n\to\infty} T_{t_n} v \neq v$ which is a contradiction.

Now let $c = t_0^{-1} \ln M$ and for an arbitrary $t \geq 0$ write $t = nt_0 + s$ for $0 \leq s < t_0$ then we have

$$\|T_t\| = \left\|T_s \circ T_{t_0}^k\right\| \leq MM^n \leq MM^{t/t_0} = Me^{\frac{t}{t_0}\ln M} = Me^{ct}$$

$\square$

PROPOSITION 18.10. *Let T_t be a strongly continuous semigroup on a Banach space X then for each $v \in X$ the function $T_t v : [0,\infty) \to X$ is a continuous function.*

PROOF. Let $v \in X$ and $t \geq 0$ then

$$\lim_{h\downarrow 0} \|T_{t+h}v - T_t v\| = \lim_{h\downarrow 0} \|T_t(T_h v - v)\| \leq Me^{ct} \lim_{h\downarrow 0} \|T_h v - v\| = 0$$

for $t > 0$ we have

$$\lim_{h\downarrow 0} \|T_{t-h}v - T_t v\| = \lim_{h\downarrow 0} \|T_{t-h}(v - T_h v)\| \leq Me^{c(t-h)} \lim_{h\downarrow 0} \|v - T_h v\| = 0$$

$\square$

Note that if A is a bounded generator then paths $e^{tA}v$ are actually differentiable and are solutions of the abstract Cauchy problem $\frac{d}{dt}f(t) = Af(t)$ with $f(0) = v$. As it turns out, for an arbitrary strongly continuous contraction semigroup, paths $T_t v$ are differentiable solutions of an abstract Cauchy problem so long as $v \in \mathcal{D}(A)$. In this context we may also refer to the differential equation $\frac{d}{dt}T_t v = AT_t v$ as the Kologorov forward equation. It also turns out that the Kolmogorov backward equation generalizes to arbitrary strongly continuous semigroups.

PROPOSITION 18.11. *Let T_t be a strongly continuous semigroup on a Banach space X with generator A then*

(i) *if $v \in X$ and $t \geq 0$ then $\int_0^t T_s v\, ds \in \mathcal{D}(A)$ and moreover $T_t v - v = A\int_0^t T_s v\, ds$*
(ii) *if $v \in \mathcal{D}(A)$ then $T_t v \in \mathcal{D}(A)$ for all $t \geq 0$ and moreover $\frac{d}{dt}T_t v = AT_t v = T_t Av$*
(iii) *for each $v \in \mathcal{D}(A)$ and $t \geq 0$ we have $T_t v - v = \int_0^t AT_s v\, ds = \int_0^t T_s Av\, ds$*

PROOF. To see (i), by strong continuity of T_t, Proposition 18.10 implies that $T_t v$ is continuous hence Riemann integrable on all finite intervals. The same follows for $T_u T_t v = T_{u+t} v$ thus we may calculate using Proposition 15.116, a change of integration variable and the Fundamental Theorem of Calculus 15.112

$$\begin{aligned}
A \int_0^t T_s v \, ds &= \lim_{h \to 0} \frac{T_h \int_0^t T_s v \, ds - \int_0^t T_s v \, ds}{h} \\
&= \lim_{h \to 0} \frac{\int_0^t T_{h+s} v \, ds - \int_0^t T_s v \, ds}{h} \\
&= \lim_{h \to 0} \frac{\int_h^{t+h} T_s v \, ds - \int_0^t T_s v \, ds}{h} \\
&= \lim_{h \to 0} \frac{\int_t^{t+h} T_s v \, ds - \int_0^h T_s v \, ds}{h} = T_t v - v
\end{aligned}$$

To see (ii), note that a trivial consequence of the semigroup property is that for all $s, t \geq 0$ we have $T_s \circ T_t = T_{s+t} = T_t \circ T_s$. Using this fact and the continuity of T_t we calculate for any $v \in \mathcal{D}(A)$

$$\begin{aligned}
AT_t v &= \lim_{h \to 0} \frac{T_h T_t v - T_t v}{h} = \lim_{h \to 0} \frac{T_t T_h v - T_t v}{h} \\
&= T_t \lim_{h \to 0} \frac{T_h v - v}{h} = T_t A v
\end{aligned}$$

It remains to show that for $t > 0$ we also have $\lim_{h \to 0} \frac{T_{t-h} v - T_t v}{-h} = T_t A v$. We compute

$$\begin{aligned}
\lim_{h \to 0} \frac{T_t v - T_{t-h} v}{h} &= \lim_{h \to 0} T_{t-h} \frac{T_h v - v}{h} = \lim_{h \to 0} T_{t-h} \left[\frac{T_h v - v}{h} - Av \right] + \lim_{h \to 0} T_{t-h} A v \\
&= \lim_{h \to 0} T_{t-h} \left[\frac{T_h v - v}{h} - Av \right] + T_t A v
\end{aligned}$$

To see that the first limit on the last line is zero we can use Lemma 18.9 and the fact that $v \in \mathcal{D}(A)$ to see for all $0 < h \leq t$,

$$\lim_{h \to 0} \left\| T_{t-h} \left[\frac{T_h v - v}{h} - Av \right] \right\| \leq M e^{ct} \lim_{h \to 0} \left\| \frac{T_h v - v}{h} - Av \right\| = 0$$

To see (iii) we apply the Fundamental Theorem of Calculus 15.112 and (ii) to see that

$$\int_0^t A T_s v \, ds = \int_0^t A T_s v \, ds = \int_0^t \frac{d}{ds} T_s v \, ds = T_t v - v$$

□

As a consequence of the fact $\frac{d}{dt} T_t v = A T_t v$ for $v \in \mathcal{D}(A)$ by analogy it is sometime helpful to think of general paths $T_t v$ as being a type of generalized or weak solution to the forward equation.

Corollary 18.12. *Let A be the generator of a strongly continuous semigroup on the Banach space X then A is a closed operator and $\mathcal{D}(A)$ is dense.*

Proof. If we let $v \in X$ then by Proposition 18.11 we know that $\int_0^t T_s v \, ds \in \mathcal{D}(A)$ for all $t \geq 0$ so it follows that $t^{-1} \int_0^t T_s v \, ds \in \mathcal{D}(A)$ for all $t > 0$. Now observe that $\lim_{t \to 0} t^{-1} \int_0^t T_s v \, ds = v$ and we see that $\mathcal{D}(A)$ is dense.

Now suppose that $v_n \in \mathcal{D}(A)$ and suppose that $\lim_{n\to\infty} v_n = v$ and $\lim_{n\to\infty} Av_n = w$. Applying continuity of T_t and Proposition 18.11 we see that

$$T_t v - v = \lim_{n\to\infty} (T_t v_n - v_n) = \lim_{n\to\infty} \int_0^t T_s A v_n \, ds = \int_0^t T_s w \, ds$$

TODO: Justify exchanging the limit and the integral...

From this is follows that

$$\lim_{t\to 0} \frac{T_t v - v}{t} = \lim_{t\to 0} t^{-1} \int_0^t T_s w \, ds = w$$

which shows that $v \in \mathcal{D}(A)$ and $Av = w$. □

2.1. The Hille-Yosida Theorem. Our next goal is to prove a significant theorem that characterizes completely the unbounded operators A that are generators of strongly continuous contraction semigroups. In particular, we need to be able to take a given operator A and construct from it the corresponding semigroup T_t. Due to the Kolmogorov Forward Equation in Proposition 18.11 one useful way of think about the problem (which is in fact the historical motivation for the theory we present) is that we are trying to construct solutions to the differential equation $\frac{d}{dt} f(t) = Af(t)$ given an unbounded operator A. If one chooses to stress the analogy with the case of a bounded operator A, another way of thinking about the task is that we are trying characterize those unbounded operators A for which we can define e^{tA} for $t \geq 0$. In fact the problem that we are posing is a bit more restricted than we've indicated. Since we are considering contraction semigroups that means that we are looking for *bounded* solutions to the aforementioned problems.

TODO : Motivate the study of the resolvent by showing what bad thing would happen if we have a singular value in $(0, \infty)$. Note that a bounded operator can arbitrary spectrum since the exponential function is entire however if there is positive real spectrum then the contraction property is not obeyed. We really don't need spectral theory in what follows; essentially we just need a proper definition of a value that *isn't* in the spectrum of an unbounded operator.

DEFINITION 18.13. Let $A : X \to X$ be a closed linear operator and the *resolvent set* $\rho(A)$ is the set of $\lambda \in \mathbb{R}$ such that

(i) $\lambda - A$ is injective on $\mathcal{D}(A)$.
(ii) $\Re(\lambda - A) = X$.
(iii) $(\lambda - A)^{-1}$ is a bounded linear operator

The operator $R_\lambda = (\lambda - A)^{-1}$ the *resolvent* of A.

Though we don't make use of the concept, the complement of the resolvent set (actually extended to the complex numbers case) is called the spectrum of A.

The resolvent operator of the generator of a strongly continuous contraction semigroup is also the Laplace transform of the semigroup.

PROPOSITION 18.14. *Let T_t be a strongly continuous contraction semigroup on X with generator A then $(0, \infty) \subset \rho(A)$ and for all $0 < \lambda < \infty$ and $v \in X$ we have*

$$R_\lambda v = (\lambda - A)^{-1} v = \int_0^\infty e^{-\lambda t} T_t v \, dt$$

and $\|R_\lambda\| \leq \lambda^{-1}$.

PROOF. Let $0 < \lambda < \infty$. We know that $e^{-\lambda t} T_t v$ is a continuous function of t since T_t is strongly continuous (Proposition 18.10). Moreover, $\int_0^\infty \left\| e^{-\lambda t} T_t v \right\| \, dt \leq \|v\| \int_0^\infty e^{-\lambda t} \, dt \leq \|v\| \, \lambda^{-1}$ and therefore $\int_0^\infty e^{-\lambda t} T_t v \, dt$ is well defined and in fact defines a bounded linear operator U_λ with $\|U_\lambda\| \leq \lambda^{-1}$.

CLAIM 18.14.1. For every $0 < \lambda < \infty$ and $v \in X$ we have $\int_0^\infty e^{-\lambda t} T_t v \, dt \in \mathcal{D}(A)$ and $(\lambda - A) \cdot \int_0^\infty e^{-\lambda t} T_t v \, dt = v$.

Applying Proposition 15.116, a change of integration variable, L'Hopital's Rule, the Fundamental Theorem of Calculus (Theorem 15.112) and strong continuity of T_t

$$
\begin{aligned}
&\lim_{t \to 0} \frac{T_t \cdot \int_0^\infty e^{-\lambda s} T_s v \, ds - \int_0^\infty e^{-\lambda s} T_s v \, ds}{t} \\
&= \lim_{t \to 0} \frac{\int_0^\infty e^{-\lambda s} T_{t+s} v \, ds - \int_0^\infty e^{-\lambda s} T_s v \, ds}{t} \\
&= \lim_{t \to 0} \frac{\int_t^\infty e^{-\lambda (s-t)} T_s v \, ds - \int_0^\infty e^{-\lambda s} T_s v \, ds}{t} \\
&= \lim_{t \to 0} \frac{e^{\lambda t} \int_0^\infty e^{-\lambda s} T_s v \, ds - \int_0^\infty e^{-\lambda s} T_s v \, ds - e^{\lambda t} \int_0^t e^{-\lambda s} T_s v \, ds}{t} \\
&= \lim_{t \to 0} \frac{e^{\lambda t} - 1}{t} \int_0^\infty e^{-\lambda s} T_s v \, ds - \lim_{t \to 0} \frac{e^{\lambda t}}{t} \int_0^t e^{-\lambda s} T_s v \, ds \\
&= \lambda \int_0^\infty e^{-\lambda s} T_s v \, ds - \lim_{t \to 0} \lambda e^{\lambda t} \int_0^t e^{-\lambda s} T_s v \, ds - \lim_{t \to 0} e^{\lambda t} e^{-\lambda t} T_t v \\
&= \lambda \int_0^\infty e^{-\lambda s} T_s v \, ds - v
\end{aligned}
$$

which shows the claim.

From the claim, it follows that $(\lambda - A)$ is surjective. To see that $(\lambda - A)$ is injective on $\mathcal{D}(A)$ we let $v \in \mathcal{D}(A)$ and the apply Proposition 18.11 and Proposition 15.116 to see

$$
\int_0^\infty e^{-\lambda s} T_s A v \, ds = \int_0^\infty A T_s v \, ds = A \int_0^\infty T_s v \, ds
$$

By this fact and the claim, if $(\lambda - A)v = 0$ then

$$
0 = \int_0^\infty e^{-\lambda s} T_s (\lambda - A) v \, ds = (\lambda - A) \int_0^\infty e^{-\lambda s} T_s v \, ds = v
$$

thus $(\lambda - A)$ is injective.

We know know that $(\lambda - A)^{-1}$ is a well defined linear operator and the claim shows that $(\lambda - A)^{-1} = U_\lambda$. The rest of the result follows from the properties already proven of U_λ. ☐

We now want to derive some simple properties of resolvents.

PROPOSITION 18.15. *Let A be a closed linear operator then*

(i) $(\mu - A)^{-1} (\lambda - A)^{-1} = (\lambda - A)^{-1} (\mu - A)^{-1} = (\lambda - \mu)^{-1} \left((\mu - A)^{-1} - (\lambda - A)^{-1} \right)$ *for all* $\mu, \lambda \in \rho(A)$.
(ii) $(\lambda - A)^{-1} A = A (\lambda - A)^{-1}$ *on* $\mathcal{D}(A)$ *for all* $\lambda \in \rho(A)$.

PROOF. To see (i) we first write $(\lambda - \mu) = ((\lambda - A) - (\mu - A))$ and note that the right hand size has domain $\mathcal{D}(A)$. Since $\Re((\lambda - A)^{-1}) = \mathcal{D}(A)$ for $\lambda \in \rho(A)$ we can compute

$$\begin{aligned}(\lambda - \mu)(\mu - A)^{-1}(\lambda - A)^{-1} &= (\mu - A)^{-1}((\lambda - A) - (\mu - A))(\lambda - A)^{-1} \\ &= (\mu - A)^{-1} - (\lambda - A)^{-1}\end{aligned}$$

To see (ii) note that $\Re((\lambda - A)^{-1}) = \mathcal{D}(A)$ so both sides of the equality are well defined operators with the left hand size having with domain $\mathcal{D}(A)$ and right hand size domain of X. Now for $v \in \mathcal{D}(A)$,

$$\begin{aligned}(\lambda - A)^{-1}Av &= -(\lambda - A)^{-1}(\lambda - A)v + \lambda(\lambda - A)^{-1}v \\ &= -v + \lambda(\lambda - A)^{-1}v \\ &= -(\lambda - A)(\lambda - A)^{-1}v + \lambda(\lambda - A)^{-1}v = A(\lambda - A)^{-1}v\end{aligned}$$

□

PROPOSITION 18.16. *Let T_t be a strongly continuous contraction semigroup on X with generator A then $\rho(A)$ is open. For any $\lambda \in \rho(A)$ and $|\mu - \lambda| < \|R_\lambda\|^{-1}$ we have $\mu \in \rho(A)$ and*

$$R_\mu = \sum_{n=0}^{\infty} (\lambda - \mu)^n R_\lambda^{n+1}$$

PROOF. First note that for $|\mu - \lambda| < \|R_\lambda\|^{-1}$ we have

$$\sum_{n=0}^{\infty} |\lambda - \mu|^n \|R_\lambda\|^{n+1} \leq \|R_\lambda\| \sum_{n=0}^{\infty} |\lambda - \mu|^n \|R_\lambda\|^n = \frac{\|R_\lambda\|}{1 - |\lambda - \mu| \|R_\lambda\|} < \infty$$

and therefore $\sum_{n=0}^{\infty}(\lambda - \mu)^n R_\lambda^{n+1}$ defines a bounded linear operator. Let $v \in X$ then for every $m \in \mathbb{N}$ $\sum_{n=0}^{m}(\lambda - \mu)^n R_\lambda^{n+1} v \in \mathcal{D}(A)$ therefore we can compute

$$(\lambda - A) \sum_{n=0}^{m} (\lambda - \mu)^n R_\lambda^{n+1} v = \sum_{n=0}^{m} (\lambda - \mu)^n R_\lambda^n v$$

and

$$\sum_{n=0}^{\infty} |\lambda - \mu|^n \|R_\lambda\|^n v = \frac{\|v\|}{1 - |\lambda - \mu| \|R_\lambda\|} < \infty$$

Thus $(\lambda - A) \sum_{n=0}^{m}(\lambda - \mu)^n R_\lambda^{n+1} v$ converges absolutely and since $\lambda - A$ is closed (Corollary 18.12) it follows that $\sum_{n=0}^{\infty}(\lambda - \mu)^n R_\lambda^{n+1} v \in \mathcal{D}(A)$ and

$$(\lambda - A) \sum_{n=0}^{\infty} (\lambda - \mu)^n R_\lambda^{n+1} v = \sum_{n=0}^{\infty} (\lambda - \mu)^n R_\lambda^n v$$

Now we compute for any $v \in X$,

$$\begin{aligned}(\mu - A) \sum_{n=0}^{\infty} (\lambda - \mu)^n R_\lambda^{n+1} v &= (\mu - \lambda) \sum_{n=0}^{\infty} (\lambda - \mu)^n R_\lambda^{n+1} v + \sum_{n=0}^{\infty} (\lambda - \mu)^n R_\lambda^n v \\ &= -\sum_{n=1}^{\infty} (\lambda - \mu)^n R_\lambda^n v + \sum_{n=0}^{\infty} (\lambda - \mu)^n R_\lambda^n v = v\end{aligned}$$

from which it follows that $\Re(\mu - A) = X$. To see that $\mu - A$ is injective on $\mathcal{D}(A)$, assume that $v \in \mathcal{D}(A)$ and $(\mu - A)v = 0$. We apply $\sum_{n=0}^{\infty}(\lambda - \mu)^n R_\lambda^{n+1}$ and use the fact that A and R_λ commute (Proposition 18.15)

$$0 = \sum_{n=0}^{\infty}(\lambda - \mu)^n R_\lambda^{n+1}(\mu - A)v = (\mu - A)\sum_{n=0}^{\infty}(\lambda - \mu)^n R_\lambda^{n+1} v = v$$

and this computation also shows that

$$R_\mu = (\mu - A)^{-1} = \sum_{n=0}^{\infty}(\lambda - \mu)^n R_\lambda^{n+1}$$

which we have already shown to be a bounded operator. Thus $\mu \in \rho(A)$. □

Definition 18.17. A linear operator $A : X \to Y$ is *dissipative* if $\|\lambda v - Av\| \geq \lambda \|v\|$ for every $v \in \mathcal{D}(A)$ and $\lambda > 0$.

The reason this definition is of interest is the following.

Example 18.18. The generator of any strongly continuous contraction semigroup is dissipative. This follows from Proposition 18.14 since by that result for any $v \in \mathcal{D}(A)$ and $\lambda > 0$

$$\|v\| = \left\|(\lambda - A)^{-1}(\lambda - A)v\right\| \leq \lambda^{-1}\|(\lambda - A)v\|$$

We also record the following example of a class of bounded dissipative operators.

Example 18.19. Let $A : X \to X$ be a bounded linear operator then for any $c \geq \|A\|$ the operator $A - c$ is dissipative. This follows from the triangle inequality

$$\|(\lambda - (A - c))v\| = \|(\lambda + c)v\| - \|Av\| \geq \|(\lambda + c)v\| - \|A\|\,\|v\| \geq \lambda \|v\|$$

It is also useful to note that the set of dissipative operators is a cone.

Example 18.20. If A is dissipative and $c > 0$ then cA is dissipative since

$$\|(\lambda - cA)v\| = \|(\lambda/c - A)cv\| \geq \lambda/c\,\|cv\| = \lambda \|v\|$$

Now we turn to less trivial matters. It turns out that the property of being dissipative is a key part of the characterization of generators of strongly continuous semigroups. As a step in the direction of proving such a result we first study the resolvent set properties of dissipative operators.

Proposition 18.21. *Let A be dissipative and $\lambda > 0$ then A is closed if and only if $\Re(\lambda - A)$ is closed. Thus if A is dissipative and $\Re(\lambda - A)$ is closed for a single value of $\lambda > 0$ it is closed for all $\lambda > 0$.*

Proof. Let A be closed. Suppose $\lambda > 0$ and $(\lambda - A)v_n$ converges, then it is Cauchy. Let $\epsilon > 0$ be given. We may find $N > 0$ such that $\|(\lambda - A)(v_n - v_m)\| \leq \epsilon$ for all $n, m \geq N$. Applying the fact that A is dissipative we have $\|v_n - v_m\| \leq \lambda^{-1}\epsilon$ which shows that v_n is Cauchy in X and therefore convergent; let v be the limit of v_n. Since λv_n converges to λv it follows that $Av_n = lambdav_n - (\lambda - A)v_n$ converges and since A is assumed closed we know that $v \in \mathcal{D}(A)$ and $\lim_{n\to\infty} Av_n = Av$. It follows that $\lim_{n\to\infty}(\lambda - A)v_n = (\lambda - A)v$ hence $\Re(\lambda - A)$ is closed.

Now assume $\Re(\lambda - A)$ is closed for all $\lambda > 0$. Suppose $\lim_{n\to\infty} v_n = v$ and $\lim_{n\to\infty} Av_n = w$. It follows that $\lim_{n\to\infty}(\lambda - A)v_n = \lambda v - w$. Since $\Re(\lambda - A)$ is closed there exists $u \in \mathcal{D}(A)$ such that $(\lambda - A)u = \lambda v - w$. Applying the fact that A is dissipative,

$$\lim_{n\to\infty} \|v_n - u\| \leq \lambda^{-1} \lim_{n\to\infty} \|(\lambda - A)(v_n - u)\| = \lambda^{-1} \lim_{n\to\infty} \|(\lambda - A)v_n - \lambda v + w\| = 0$$

Therefore $v = u$ hence $v \in \mathcal{D}(A)$ and

$$\lim_{n\to\infty} Av_n = \lim_{n\to\infty} \lambda v_n - \lim_{n\to\infty} (\lambda - A)v_n = \lambda v - (\lambda v - Av) = Av$$

hence A is closed.

If A is dissipative and $\Re(\lambda - A)$ for a single $\lambda > 0$ then A is closed and dissipative and it follows that $\Re(\lambda - A)$ is closed for all $\lambda > 0$. ☐

LEMMA 18.22. *Let A be a closed dissipative operator on X, if there exists $\lambda > 0$ such that $\lambda \in \rho(A)$ then $\lambda \in \rho(A)$ for all $\lambda > 0$.*

PROOF. We need a simple property of the topology of the real line.

CLAIM 18.22.1. If $A \subset (0,\infty)$ is non-empty, closed and open then $A = (0,\infty)$.

We know that all open sets of $(0,\infty)$ are countable unions of disjoint open intervals (Lemma 1.16). Suppose that there is a open interval $(a, b) \subset A$ with either $a \neq 0$ or $b \neq \infty$ then either a or b is a limit point of A that is not contained in A which contradicts the fact that A is closed.

By the claim, the fact that $(0,\infty) \cap \rho(A)$ is assumed to be non-empty and the fact that $(0,\infty) \cap \rho(A)$ is open (Proposition 18.16), it suffices to show that $(0,\infty) \cap \rho(A)$ is closed. Thus suppose that we have a sequence $\lambda_n \in (0,\infty) \cap \rho(A)$ such that $\lim_{n\to\infty} \lambda_n = \lambda > 0$. Let $v \in X$ and define $v_n = (\lambda - A)R_{\lambda_n} v$ which is well defined since $\Re(R_{\lambda_n}) = \mathcal{D}(A)$. Applying the norm bound $\|R_{\lambda_n}\| \leq \lambda_n^{-1}$ (Proposition 18.14),

$$\begin{aligned}\lim_{n\to\infty} \|v_n - v\| &= \lim_{n\to\infty} \|(\lambda - A)R_{\lambda_n} v - (\lambda_n - A)R_{\lambda_n} v\| \\ &= \lim_{n\to\infty} \|(\lambda - \lambda_n)R_{\lambda_n} v\| \leq \lim_{n\to\infty} \frac{|\lambda - \lambda_n|}{\lambda_n} \|v\| = 0\end{aligned}$$

Therefore $\lim_{n\to\infty} v_n = v$; in particular $\Re(\lambda - A)$ is dense in X. However, since A is dissipative and closed we can apply Proposition 18.21 to conclude that $\Re(\lambda - A)$ is closed so in fact we have $\Re(\lambda - A) = X$. To see that $\lambda - A$ is injective, suppose $(\lambda - A)v = 0$ then using the fact that A is dissipative, $\|v\| \leq \lambda^{-1} \|(\lambda - A)v\| = 0$. Similarly using the fact that A is dissipative we know that

$$\left\|(\lambda - A)^{-1} v\right\| \leq \lambda^{-1} \left\|(\lambda - A)(\lambda - A)^{-1} v\right\| = \lambda^{-1} \|v\|$$

which shows that $(\lambda - A)^{-1}$ is a bounded operator hence $\lambda \in \rho(A)$. ☐

DEFINITION 18.23. Let $A : X \to X$ be a closed linear operator and suppose $\lambda \in \rho(A)$ the we define the *Yosida approximation* be be

$$A_\lambda = \lambda A R_\lambda$$

To see why we refer to A_λ as an approximation consider the case in which A is bounded. In that case for $\lambda > \|A\|$ we can write

$$A_\lambda = A(1 - \lambda^{-1} A)^{-1} = A + \lambda^{-1} A^2 + \lambda^{-2} A^3 + \cdots$$

which shows that $\lim_{\lambda\to\infty} A_\lambda = A$. The next result shows how this idea can be applied with unbounded A : closed dissipative operators can be approximated by bounded operators using the Yosida approximation.

LEMMA 18.24 (Yosida approximation). *Let A be a closed dissipative operator with $\mathcal{D}(A)$ dense and $(0,\infty) \subset \rho(A)$, then A_λ satisfies*

(i) *A_λ is a a bounded linear operator and e^{tA_λ} is a strongly continuous contraction semigroup.*
(ii) *$A_\lambda A_\mu = A_\mu A_\lambda$ for all $\lambda, \mu > 0$*
(iii) *$\lim_{\lambda\to\infty} A_\lambda v = Av$ for all $v \in \mathcal{D}(A)$.*

PROOF. We begin by showing (i). Let $\lambda > 0$. Then since $\lambda \in \rho(A)$ we have R_λ is defined and $\Re(R_\lambda) = \mathcal{D}(A)$. Because A is dissipative we have for every $v \in X$,

$$\|R_\lambda v\| \leq \lambda^{-1} \|(\lambda - A)R_\lambda v\| = \lambda^{-1} \|v\| \tag{32}$$

hence $\|R_\lambda\| \leq \lambda^{-1}$.

Using $(\lambda - A)R_\lambda = \mathrm{Id}$ on X and $R_\lambda(\lambda - A) = \mathrm{Id}$ on $\mathcal{D}(A)$ we get

$$A_\lambda = \lambda A R_\lambda = \lambda(\lambda R_\lambda - \mathrm{Id}) = \lambda^2 R_\lambda - \lambda \text{ on } X \tag{33}$$

and it follows that A_λ is a bounded linear operator. Therefore e^{tA_λ} is a strongly continuous semigroup (Example 18.8). Furthermore by (32)

$$\left\|e^{tA_\lambda}\right\| = \left\|e^{t\lambda^2 R_\lambda} e^{-t\lambda}\right\| \leq e^{t\lambda^2 \|R_\lambda\|} e^{-t\lambda} \leq e^{t\lambda} e^{-t\lambda} = 1$$

so e^{tA_λ} is contractive. Thus (i) is shown.

Now (ii) follows from (33) and the commutativity of resolvents (Proposition 18.15)

$$\begin{aligned}
A_\lambda A_\mu &= (\lambda^2 R_\lambda - \lambda)(\mu^2 R_\mu - \mu) \\
&= \lambda^2 \mu^2 R_\lambda R_\mu - \lambda^2 \mu R_\lambda - \mu^2 \lambda R_\mu + \lambda\mu) \\
&= \lambda^2 \mu^2 R_\mu R_\lambda - \lambda^2 R_\lambda - \mu^2 R_\mu + \mu\lambda) \\
&= A_\mu A_\lambda
\end{aligned}$$

To see (iii) first we have

CLAIM 18.24.1. $\lim_{\lambda\to\infty} \lambda R_\lambda v = v$ for all $v \in X$.

For $v \in \mathcal{D}(A)$ we have using Proposition 18.15 and Proposition 18.14

$$\begin{aligned}
\|\lambda R_\lambda v - v\| &= \|\lambda R_\lambda v - (\lambda - A)R_\lambda v\| \\
&= \|A R_\lambda v\| = \|R_\lambda A v\| \leq \lambda^{-1} \|Av\|
\end{aligned}$$

Now take the limit $\lambda \to \infty$ to see the claim for $v \in \mathcal{D}(A)$. By assumption $\mathcal{D}(A)$ is dense in X so for general $v \in X$ let $v_n \in \mathcal{D}(A)$ such that $\lim_{n\to\infty} v_n = v$. Then for all $n \in \mathbb{N}$

$$\begin{aligned}
\|\lambda R_\lambda v - v\| &\leq \|(\lambda R_\lambda - \mathrm{Id})(v - v_n)\| + \|\lambda R_\lambda v_n - v_n\| \\
&\leq 2\|v - v_n\| + \|\lambda R_\lambda v_n - v_n\|
\end{aligned}$$

so take the limit as $\lambda \to \infty$ and then as $n \to \infty$.

To finish (iii) we use the claim and Proposition 18.15 to see that for all $v \in \mathcal{D}(A)$,

$$\lim_{\lambda\to\infty} A_\lambda v = \lim_{\lambda\to\infty} \lambda R_\lambda A v = Av$$

□

LEMMA 18.25. *Suppose A and B are communting bounded linear operators on a Banach space X such that $\left\|e^{tA}\right\| \leq 1$ and $\left\|e^{tB}\right\| \leq 1$ for all $t \geq 0$. It follows that*

$$\left\|e^{tA} - e^{tB}\right\| \leq t \left\|A - B\right\|$$

PROOF. Let $v \in X$ and $t \geq 0$ then by the Fundamental Theorem of Calculus (Theorem 15.112) and the fact that A and B commute we get

$$e^{tA}v - e^{tB}v = \int_0^t \frac{d}{ds} e^{sA} e^{(t-s)B} v\, ds = \int_0^t e^{sA} e^{(t-s)B} (A - B) v\, ds$$

Now if we take norms over both sides and use Proposition 15.113 and the hypotheses that $\left\|e^{sA}\right\|, \left\|e^{(t-s)B}\right\| \leq 1$ the result follows. □

THEOREM 18.26 (Hille-Yosida Theorem). *Let X be a Banach space, then a linear operator $A : X \to X$ is the generator of a strongly continuous contraction semigroup if and only if*

(i) *$\mathcal{D}(A)$ is dense in X*
(ii) *A is dissipative*
(iii) *$\Re(\lambda_0 - A) = X$ for some $\lambda_0 > 0$.*

PROOF. If A is the generator of a strongly continuous contraction semigroup then Corollary 18.12 implies (i), Example 18.18 implies (ii) and Proposition 18.14 implies that $\Re(\lambda - A) = X$ for all $\lambda > 0$, so in particular (iii) holds.

Let us assume that (i), (ii) and (iii) hold.

CLAIM 18.26.1. A is a closed operator and $\lambda_0 \in \rho(A)$.

By (iii) we know $\Re(\lambda_0 - A) = X$. This implies $\Re(\lambda_0 - A)$ is closed and since A is dissipative Proposition 18.21 implies that A is closed. The fact that A is dissipative implies that $\lambda_0 - A$ is injective (in fact for all $\mu > 0$, if $(\mu - A)v = 0$ then $\|v\| \leq \mu^{-1} \|(\mu - A)v\| = 0$). The fact that A is dissipative implies that for all $v \in X$, $\|R_{\lambda_0} v\| \leq \lambda_0^{-1} \|(\lambda_0 - A) R_{\lambda_0} v\| = \lambda_0^{-1} v$ hence R_{λ_0} is a bounded linear operator. Therefore we have shown that $\lambda_0 \in \rho(A)$.

From the claim and Lemma 18.22 we know that $(0, \infty) \subset \rho(A)$. This fact, the fact that A is closed and dissipative and (i) means that we can apply the Yosida approximation Lemma 18.24. For $\lambda > 0$ define $A_\lambda = \lambda A R_\lambda$ and $T^\lambda_t = e^{tA_\lambda}$ so that T^λ_t is a strongly continuous contraction semigroup. By application of Lemma 18.25 we get the inequality

$$\left\|T^\lambda_t v - T^\mu_t v\right\| \leq t \left\|A_\lambda v - A_\mu v\right\| \text{ for all } t \geq 0 \text{ and } v \in X \tag{34}$$

Lemma 18.24 also tells us that $\lim_{\lambda \to \infty} A_\lambda v = Av$ for all $v \in \mathcal{D}(A)$. Therefore the sequence $A_n v$ is Cauchy hence (34) tells us that $T^n_t v$ is uniformly Cauchy on every bounded interval of $[0, \infty)$ and therefore converges uniformly on bounded intervals to a continuous function $T_t v$ (recall that $T^n_t v$ is continuous by Proposition 18.10 and Lemma 15.22).

CLAIM 18.26.2. For every fixed $v \in X$, $T^n_t v$ converges in $C([0, \infty); X)$.

By (i) we can find $v_m \in \mathcal{D}(A)$ such that $\lim_{m \to \infty} v_m = v$. Now by the triangle inequality, (34) and the fact that T^λ_t is a contraction semigroup for all $\lambda > 0$, for all $m \in \mathbb{N}$

$$\begin{aligned}\left\|T^\lambda_t v - T^\mu_t v\right\| &\leq \left\|T^\lambda_t v - T^\lambda_t v_m\right\| + \left\|T^\lambda_t v_m - T^\mu_t v_m\right\| + \left\|T^\mu_t v_m - T^\mu_t v\right\| \\ &\leq 2\left\|v - v_m\right\| + t\left\|A_\lambda v_m - A_\mu v_m\right\|\end{aligned}$$

The fact that $v_m \in \mathcal{D}(A)$ and Lemma 18.24 tell us that $\lim_{\lambda\to\infty} A_\lambda v_m = A v_m$ for all $m \in \mathbb{N}$. Thus for every $T \geq 0$ and $\epsilon > 0$ we can find $m \in \mathbb{N}$ such that $\|v - v_m\| \leq \epsilon/4$ and a $N \geq 0$ such that $\|A_\lambda v_m - A_\mu v_m\| \leq \epsilon/2T$ for $\lambda, \mu \geq N$ which shows

$$\sup_{0\leq t\leq T} \left\|T^\lambda_t v - T^\mu_t v\right\| \leq 2\left\|v - v_m\right\| + T \sup_{0\leq t\leq T}\left\|A_\lambda v_m - A_\mu v_m\right\| \leq 2\left\|v - v_m\right\| \leq \epsilon$$

which shows that $T^\lambda_t v$ is Cauchy in $C([0,T];X)$ and therefore converges in $C([0,T];X)$. It follows from Lemma 15.22 that $T^\lambda_t v$ converges in $C([0,\infty);X)$.

Now by the claim, for every $v \in X$ we have a $T_t v \in C([0,\infty);X)$ such that $T^\lambda_t v \to T_t v$ in $C([0,\infty);X)$.

Claim 18.26.3. T_t is a strongly continuous contraction semigroup.

To see the semigroup property we let $v \in X$ we use the fact that $T^\lambda_t v \to T_t v$ for every $0 \leq t < \infty$ and the fact that $\left\|T^\lambda_t\right\| \leq 1$ to see

$$\begin{aligned}\left\|T_{s+t} v - T_s T_t v\right\| &\leq \lim_{\lambda\to\infty}\left[\left\|T_{s+t}v - T^\lambda_{s+t}v\right\| + \left\|T^\lambda_s T^\lambda_t v - T_s T_t v\right\|\right] \\ &\leq \lim_{\lambda\to\infty}\left[\left\|T_{s+t}v - T^\lambda_{s+t}v\right\| + \left\|T^\lambda_s(T^\lambda_t v - T_t v)\right\| + \left\|T^\lambda_s T_t v - T_s T_t v\right\|\right] \\ &\leq \lim_{\lambda\to\infty}\left[\left\|T_{s+t}v - T^\lambda_{s+t}v\right\| + \left\|T^\lambda_t v - T_t v\right\| + \left\|T^\lambda_s T_t v - T_s T_t v\right\|\right] \\ &= 0\end{aligned}$$

The contraction property of T_t follows easily from the contraction property of T^λ_t,

$$\left\|T_t v\right\| \leq \lim_{\lambda\to\infty}\left[\left\|T^\lambda_t v\right\| + \left\|T_t v - T^\lambda_t v\right\|\right] leq \left\|v\right\| + \lim_{\lambda\to\infty}\left\|T_t v - T^\lambda_t v\right\| = \left\|v\right\|$$

For strong continuity we need to use the full power of the fact that $T^\lambda_t v \to T_t v$ in $C([0,\infty);X)$:

$$\begin{aligned}\lim_{t\to 0}\left\|T_t v - v\right\| &\leq \lim_{\lambda\to\infty}\lim_{t\to 0}\left[\left\|T^\lambda_t v - v\right\| + \left\|T_t v - T^\lambda_t v\right\|\right] \\ &\leq \lim_{\lambda\to\infty}\lim_{t\to 0}\left\|T^\lambda_t v - v\right\| + \lim_{\lambda\to\infty}\sup_{0\leq t\leq 1}\left\|T_t v - T^\lambda_t v\right\| \\ &= 0\end{aligned}$$

Claim 18.26.4. A is the generator of T_t.

By the Kolomogorov backward equation (Proposition 18.11) and the fact that $\mathcal{D}(A_\lambda) = X$ we have for all $v \in X$,

$$T^\lambda_t v - v = \int_0^t T^\lambda_s A_\lambda v\, ds$$

If we assume that $v \in \mathcal{D}(A)$ and $T > 0$ then using the fact that $\|T_t^\lambda\| \leq 1$, the Yosida approximation and the definition of T_t,

$$\begin{aligned}\lim_{\lambda \to \infty} \sup_{0 \leq t \leq T} \|T_t^\lambda A_\lambda v - T_t A v\| &\leq \lim_{\lambda \to \infty} \sup_{0 \leq t \leq T} \left[\|T_t^\lambda A_\lambda v - T_t^\lambda A v\| + \|T_t^\lambda A v - T_t A v\|\right] \\ &\leq \lim_{\lambda \to \infty} \|A_\lambda v - Av\| + \lim_{\lambda \to \infty} \sup_{0 \leq t \leq T} \|T_t^\lambda A v - T_t A v\| \\ &= 0\end{aligned}$$

From the uniform convergence of $T_t^\lambda A_\lambda v$ on compacts we conclude

$$\lim_{\lambda \to \infty} \int_0^t T_s^\lambda A_\lambda v \, ds = \int_0^t T_s A v \, ds$$

and therefore the Kolmogorov backward equation is established for T_t and A:

$$T_t v - v = \int_0^t T_s A v \, ds \text{ for } v \in \mathcal{D}(A)$$

The fact that $\lim_{t \to 0} t^{-1}(T_t v - v) = Av$ for all $v \in \mathcal{D}(A)$ follows from the Fundamental Theorem of Calculus and the fact that $T_0 = \mathrm{Id}$. If B is the generator of T_t we know that A and B agree on $\mathcal{D}(A)$ thus it remains to show that $\mathcal{D}(A) = \mathcal{D}(B)$. Suppose there exists $v \in \mathcal{D}(B) \setminus \mathcal{D}(A)$. By (iii) we know that there is $\lambda > 0$ and $w \in \mathcal{D}(A)$ such that $(\lambda - A)w = (\lambda - B)v$. On the other hand $(\lambda - A)w = (\lambda - B)w$ and by (ii) we know that $\lambda - B$ is injective which is a contradiction. □

TODO: Show this prior to proving Hille-Yosida and incorporate the next corollary into the statement of Hille-Yosida. The next important fact that we want to demonstrate is that the generator uniquely identifies a strongly continuous contraction semigroup. That fact (as well as a few others) is a corollary of the following technical lemma.

Lemma 18.27. *Let A be a dissipative linear operator on X, let $u : [0, \infty) \to X$ be a continuous path in X such that $u(t) \in \mathcal{D}(A)$ for all $t > 0$, $Au : (0, \infty) \to X$ is continuous and*

$$u(t) = u(\epsilon) + \int_\epsilon^t Au(s) \, ds \text{ for all } 0 < \epsilon < t \tag{35}$$

Then $\|u(t)\| \leq \|u(0)\|$.

Proof. Since $Au(s)$ is uniformly continuous on $[\epsilon, t]$ given any $\delta > 0$ we may choose partition $0 < \epsilon = t_0 < t_1 < \cdots < t_n = t$ such that $\max_{1 \leq i \leq n} \sup_{t_{i-1} \leq s \leq t_i} \|Au(s) - A(t_i)\| \leq \delta$. Then writing a telescoping sum, using eqref 35, the dissipative property (in the form $\|v\| \le \left\|v - \lambda^{-1}Av\right\|$) and the triangle inequality

$$\begin{aligned}
\|u(t)\| &= \|u(\epsilon)\| + \sum_{i=1}^n \left[\|u(t_i)\| - \|u(t_{i-1})\|\right] \\
&= \|u(\epsilon)\| + \sum_{i=1}^n \left[\|u(t_i)\| - \left\|u(t_i) - \int_{t_{i-1}}^{t_i} Au(s)\,ds\right\|\right] \\
&\le \|u(\epsilon)\| + \sum_{i=1}^n \left[\|u(t_i) - (t_i - t_{i-1})Au(t_i)\| - \left\|u(t_i) - \int_{t_{i-1}}^{t_i} Au(s)\,ds\right\|\right] \\
&\le \|u(\epsilon)\| + \sum_{i=1}^n \left\|\int_{t_{i-1}}^{t_i} Au(s)\,ds - (t_i - t_{i-1})A(t_i)\right\| \\
&= \|u(\epsilon)\| + \sum_{i=1}^n \left\|\int_{t_{i-1}}^{t_i} (Au(s) - A(t_i))\,ds\right\| \\
&\le \|u(\epsilon)\| + \sum_{i=1}^n \int_{t_{i-1}}^{t_i} \|Au(s) - A(t_i)\|\,ds \\
&\le \|u(\epsilon)\| + (t-\epsilon) \max_{1\le i\le n} \sup_{t_{i-1}\le s\le t_i} \|Au(s) - A(t_i)\| \\
&\le \|u(\epsilon)\| + (t-\epsilon)\delta
\end{aligned}$$

As $\delta > 0$ was arbitrary we may take the limit as $\delta \to 0$ conclude that $\|u(t)\| \le \|u(\epsilon)\|$. We may also take the limit as $\epsilon \to 0$ and use the continuity of u at 0 to conclude $\|u(t)\| \le \|u(0)\|$. □

COROLLARY 18.28. *Let T_t and S_t be strongly continuous contraction semigroups both with generator A then it follows that $T_t = S_t$ for all $t \ge 0$.*

PROOF. By Example 18.18 we know that A is dissipative. Let $v \in \mathcal{D}(A)$. Since $\mathcal{D}(A)$ is dense (Corollary 18.12) and T_t and S_t are bounded operators it suffices to show that $T_t v = S_t v$ for all $t \ge 0$. By Proposition 18.10 we know that $T_t v$ and $S_t v$ are continuous. By Proposition 18.11 we know that $AT_t v = T_t Av$ and $AS_t v = S_t Av$ hence $AT_t v$ and $AS_t v$ are continuous, that $T_t v, S_t v \in \mathcal{D}(A)$ for all $t \ge 0$ and that $T_t v - S_t v = \int_0^t A(T_s v - S_s v)\,ds$ for all $t \ge 0$. Thus applying Lemma 18.27 to the path $T_t v - S_t v$ we see that $\|T_t v - S_t v\| \le \|T_0 v - S_o v\| = 0$ for all $t \ge 0$. □

Now that we know that the generator determines the semigroup we know that a strongly continuous contraction semigroup is equal to the limit of the exponential of the Yosida approximations of its generators. Moreover

COROLLARY 18.29. *Let T_t be a strongly continuous contraction semigroup on X with generator A and let A_λ be the Yosida approximation of A, then for all $v \in \mathcal{D}(A)$ and $\lambda > 0$*

$$\left\|e^{tA_\lambda} - T_t v\right\| \le t\,\|A_\lambda v - Av\|$$

and therefore for all $v \in X$ and $0 \le t < \infty$, $\lim_{\lambda\to\infty} e^{tA_\lambda} = T_t v$ uniformly on bounded intervals.

PROOF. By Corollary 18.28 and the proof of the Hille-Yosida theorem we know that $T_t v = \lim_{\mu \to \infty} e^{tA_\mu} v$ for all $v \in X$ and $0 \leq t < \infty$ as noted in the proof that convergence is in $C([0,\infty); S)$ for fixed $v \in X$. From (34) in the proof of Hille-Yosida theorem we know that for all $t \geq 0$ and $v \in X$

$$\begin{aligned}
\left\|e^{\lambda t} v - T_t v\right\| &\leq \left\|e^{\lambda t} v - e^{\mu t} v\right\| + \left\|e^{\mu t} v - T_t v\right\| \\
&\leq t \left\|A_\lambda v - A_\mu v\right\| + \left\|e^{\mu t} v - T_t v\right\| \\
&\leq t \left\|A_\lambda v - A v\right\| + t \left\|A v - A_\mu v\right\| + \left\|e^{\mu t} v - T_t v\right\|
\end{aligned}$$

now let $\mu \to \infty$ and use the facts that $\lim_{\mu \to \infty} A_\mu v = Av$ and $\lim_{\mu \to \infty} e^{\mu t} v = T_t v$. ∎

The following corollary is a technical device that allows one to show that a semigroup preserves a subset by considering the behavior of its generator. In particular, we will be able to use this to show that a semigroup comprises positive operators on function spaces (i.e. the semigroup preserves the positive cone of functions $f \geq 0$) by showing positivity of the resolvents.

COROLLARY 18.30. *Let T_t be a strongly continuous contraction semigroup on X with generator A, let $Y \subset X$ and*

$$\Lambda_Y = \{\lambda > 0 \mid \lambda(\lambda - A)^{-1}(Y) \subset Y\}$$

If either

(i) *Y is a closed convex subset of X and Λ_Y is unbounded*
(ii) *Y is a closed subspace of X and Λ_Y is nonempty*

then $T_t(Y) \subset Y$ for all $0 \leq t < \infty$.

PROOF. TODO: ∎

Sometimes it is inconvenient to deal with the full domain of a generator. In particular, $\mathcal{D}(A)$ might be too big in the sense that one lacks a clean characterization of the elements it contains. In these situations it is often the case that there there is a subset of $\mathcal{D}(A)$ which is easy to identify and which is big enough to capture all of the information of A in the sense the pair of $(A, \mathcal{D}(A))$ is a limit of the restriction to the subset.

DEFINITION 18.31. A linear operator is said to be *closable* if it has a closed linear extension. The smallest closed linear extension of A is called the *closure* of A. Given a closable operator A the closure of A is denoted $\overline{A}$.

PROPOSITION 18.32. *Let A be a closable linear operator then the closure of the graph of A defines a single valued closed linear operator $\overline{A}$. Any closed extension of A is an extension of $\overline{A}$.*

PROOF. Suppose B is a closed linear extension of A. If we have a sequence (v_n, Av_n) with $v_n \in \mathcal{D}(A)$ that converges in $X \times X$ to (v, w) then since B is a closed extension of A we know that $v \in \mathcal{D}(B)$ and $w = Bv$. Thus if (v, w) and (v, u) are in the closure of the graph of A we have $w = u = Bv$; hence the closure of the graph of A defines a function $\overline{A} : X \to X$. The fact that $\overline{A}$ is linear follows from the linearity of limits and the linearity of A; pick $(v_n, Av_n) \to (v, Av)$ and $(w_n, Aw_n) \to (w, Aw)$ then

$$(av + bw, \overline{A}(av + bw)) = \lim_{n\to\infty} (av_n + bw_n, A(av_n + bw_n)) = \lim_{n\to\infty} (av_n + bw_n, aAv_n + bAw_n) = (av + bw, a\overline{A}v + b$$

If we let C be any closed extension of A and $v \in \mathcal{D}(\overline{A})$ then we may pick a sequence $v_n \in \mathcal{D}(A)$ such that $\lim_{n\to\infty}(v_n, Av_n) = (v, \overline{A}v)$; but since C is an extension of A, $(v_n, Av_n) = (v_n, Cv_n)$ converges in $X \times X$ and therefore since C is closed $(v, \overline{A}v) = (v, Cv)$ which shows that C is an extension of $\overline{A}$. $\square$

Our goal is to provide an alternative statement of the Hille-Yosida theorem in terms of closable operators rather than closed operators. We need to do a small bit of work to examine the interactions of some of our existing concepts with the new concept of the closure of a closable operator.

LEMMA 18.33. *Suppose A is a closable linear operator then A is dissipative if and only if $\overline{A}$ is dissipative.*

PROOF. First suppose $\overline{A}$ is dissipative. Then using the fact that $\overline{A}$ is an extension of A, for any $\lambda > 0$ and $v \in \mathcal{D}(A)$

$$\|(\lambda - A)v\| = \left\|(\lambda - \overline{A})v\right\| \geq \lambda \|v\|$$

On the other hand if A is dissipative then for any $v \in \mathcal{D}(\overline{A})$ we may pick a sequence $v_n \in \mathcal{D}(A)$ such that $\lim_{n\to\infty}(v_n, Av_n) = (v, \overline{A}v)$ and therefore for every $\lambda > 0$,

$$\left\|(\lambda - \overline{A})v\right\| = \lim_{n\to\infty} \|(\lambda - A)v_n\| \geq \lambda \lim_{n\to\infty} \|v_n\| = \lambda \|v\|$$

$\square$

PROPOSITION 18.34. *Let A be dissipative with $\mathcal{D}(A)$ dense then A is closable and $\overline{\Re(\lambda - A)} = \Re(\lambda - \overline{A})$ for all $\lambda > 0$.*

PROOF.

CLAIM 18.34.1. A is closable

Pick a sequence $v_n \in \mathcal{D}(A)$ such that $\lim_{n\to\infty} v_n = 0$ and $\lim_{n\to\infty} Av_n = w$. Since $\mathcal{D}(A)$ is dense we may pick a sequence $w_n \in \mathcal{D}(A)$ such that $\lim_{n\to\infty} w_n = w$. Therefore for all $m \in \mathbb{N}$ and $\lambda > 0$ using the continuity of norms and the fact that A is dissipative,

$$\begin{aligned}
\|(\lambda - A)w_m - \lambda w\| &= \lim_{n\to\infty} \|(\lambda - A)w_m - (\lambda - A)\lambda v_n\| \\
&\geq \lambda \lim_{n\to\infty} \|w_m - \lambda v_n\| = \lambda \|w_m\|
\end{aligned}$$

so after dividing by λ, and taking limits

$$\begin{aligned}
\|w\| = \lim_{m\to\infty} \|w_m\| &\leq \lim_{m\to\infty} \lim_{\lambda\to\infty} \left\|(\mathrm{Id} - \lambda^{-1} A)w_m - w\right\| \\
&\leq \lim_{m\to\infty} \lim_{\lambda\to\infty} \left[\|w_m - w\| + \lambda^{-1} \|Aw_m\|\right] = 0
\end{aligned}$$

and therefore $w = 0$.

Now to see that A is closable suppose that (v_n, Av_n) and (u_n, Au_n) are convergent sequences in $X \times X$ with $v_n, u_n \in \mathcal{D}(A)$ and $\lim_{n\to\infty} v_n = \lim_{n\to\infty} u_n$. Then if follows that $\lim_{n\to\infty}(v_n - u_n) = 0$ and therefore by the preceeding argument $\lim_{n\to\infty} A(v_n - u_n) = \lim_{n\to\infty} Av_n - \lim_{n\to\infty} Au_n = 0$; which is to say that A is closable.

Claim 18.34.2. $\Re(\lambda - \overline{A}) \subset \overline{\Re(\lambda - A)}$

Now suppose that $w \in \Re(\lambda - \overline{A})$ and pick $v \in \mathcal{D}(\overline{A})$ with $(\lambda - \overline{A})v = w$. By definition of $\overline{A}$ we may pick a sequence $v_n \in \mathcal{D}(A)$ such that $\lim_{n\to\infty} v_n$ and $\lim_{n\to\infty} Av_n = \overline{A}v = \lambda v - w$. Therefore $w = \lim_{n\to\infty}(\lambda v_n - Av_n)$ which shows that $w \in \overline{\Re(\lambda - A)}$.

Claim 18.34.3. $\overline{\Re(\lambda - A)} \subset \Re(\lambda - \overline{A})$

By Lemma 18.33 we know that $\overline{A}$ is dissipative and closed and by Proposition 18.21 we know this implies that $\Re(\lambda - \overline{A})$ is closed. Since $\overline{A}$ is an extension of A we know that $\Re(\lambda - A) \subset \Re(\lambda - \overline{A})$. The claim follows by combining these two observations and taking closures. ∎

We can have enough to state a prove an alternative formulation of the Hille-Yosida theorem in terms of closable operators.

Theorem 18.35 (Hille-Yosida Theorem for Closable Operators). *Let X be a Banach space, then a linear operator $A : X \to X$ is closable with $\overline{A}$ the generator of a strongly continuous contraction semigroup if and only if*

(i) *$\mathcal{D}(A)$ is dense in X*
(ii) *A is dissipative*
(iii) *$\Re(\lambda_0 - A)$ is dense in X for some $\lambda_0 > 0$.*

Proof. Suppose A is closable and $\overline{A}$ is generator of a strongly continuous contraction semigroup. Then by the Hille-Yosida Theorem 18.26 we know that $\mathcal{D}(\overline{A})$ is dense in X, $\overline{A}$ is dissipative and $\Re(\lambda_0 - \overline{A}) = X$ for some $\lambda_0 > 0$. By Lemma 18.33 we know that A is dissipative. To see that $\mathcal{D}(A)$ is dense, just note that for any $\epsilon > 0$ and $v \in X$ we may find a $w \in \mathcal{D}(\overline{A})$ such that $\|v - w\| \leq \epsilon/2$ but also that we may find a $u \in \mathcal{D}(A)$ such that $\|w - u\| \leq \epsilon/2$ (and $\left\|\overline{A}w - Au\right\| \leq \epsilon/2$ although we don't need this fact). Now that we know $\mathcal{D}(A)$ is dense it follow from Proposition 18.34 that $\overline{\Re(\lambda_0 - A)} = \Re(\lambda_0 - \overline{A}) = X$.

On the other hand, suppose that A satisfies (i), (ii) and (iii). By Proposition 18.34 we know that A is closable. Then by Lemma 18.33 we know that $\overline{A}$ is dissipative. It is immediate from (i) and the fact that $\mathcal{D}(A) \subset \mathcal{D}(\overline{A})$ that $\mathcal{D}(\overline{A})$ is dense. Lastly by (i), (ii) and (iii) we may apply Proposition 18.34 implies that $\Re(\lambda_0 - \overline{A}) = \overline{\Re(\lambda_0 - A)} = X$. Thus we may apply the Hille-Yosida Theorem 18.26 to conclude that $\overline{A}$ is the generator of a strongly continuous contraction semigroup. ∎

2.2. Feller Semigroups. TODO: Define Feller processes and use the Yosida approximation to show that they are strongly continuous.

Having presented the theory of semigroups in a generic Banach space setting we now specialize back to the case of interest for Markov processes. In fact we specialize the state space and assume that S is a locally compact separable metric space and consider the action on the Banach space of continuous functions vanishing at infinity (see Proposition 15.42). Recall that semigroups We want to apply the Hille-Yosida theorem in the special case and as a new property makes an appearance.

Definition 18.36. Let S be a topological space then a semigroup $T_t : C_0(S) \to C_0(S)$ is said to be *positive* if for every $f \geq 0$ we have $T_t f \geq 0$ for all $t \geq 0$. A positive strongly continuous contraction semigroup on $C_0(S)$ is called a *Feller* *semigroup*. An unbounded operator $A : C_0(S) \to C_0(S)$ is said to satisfy the *positive maximum principle* if given $f \in \mathcal{D}(A)$ and $x_0 \in S$ satisfying $\sup_{x \in S} f(x) = f(x_0) \geq 0$ we have $Af(x_0) \leq 0$.

LEMMA 18.37. *Let S be locally compact Hausdorff and suppose that A satisfies the positive maximum principle then A is dissipative.*

PROOF. Let $f \in \mathcal{D}(A)$ and suppose $\lambda > 0$. Using Corollary 15.44 pick $x_0 \in S$ such that $|f(x_0)| = \sup_{x \in S} |f(x)|$. First suppose that $f(x_0) \geq 0$. It follows that $f(x_0) \geq f(x)$ for all $x \in S$ and from the positive maximum principle that $Af(x_0) \leq 0$, thus

$$\|\lambda f - Af\| \geq |\lambda f(x_0) - Af(x_0)| = \lambda f(x_0) - Af(x_0) \geq \lambda f(x_0) = \lambda \|f\|$$

If $f(x_0) < 0$ then we can apply the same argument to $-f$. It follows that A is dissipative. ☐

THEOREM 18.38. *Let S be a locally compact separable metric space then $A : C_0(S) \to C_0(S)$ is closable and generates a Feller semigroup if and only if*

(i) *$\mathcal{D}(A)$ is dense in X*
(ii) *A satisfies the positive maximum principle*
(iii) *$\Re(\lambda_0 - A)$ is dense in X for some $\lambda_0 > 0$.*

PROOF. If $\overline{A}$ generates a Feller semigroup then we can apply the Hille-Yosida Theorem 18.35 to conclude (i) and (iii). Moreover we know that A is dissipative. Suppose $f \in \mathcal{D}(A)$ and $f(x_0) = \sup_{x \in S} f(x) \geq 0$. Using positivity and the contraction property we get for all $t \geq 0$,

$$T_t f(x_0) \leq T_t f_+(x_0) \leq \|f_+\| = f(x_0)$$

and therefore

$$Af(x_0) = \lim_{t \to 0} \frac{T_t f(x_0) - f(x_0)}{t} \leq 0$$

and it follows that A satisfies (ii).

On the other hand, suppose A satisfies (i), (ii) and (iii) by Lemma 18.37 we know that A is dissipative and therefore we can apply Hille-Yosida Theorem 18.35 to conclude that A is closable and $\overline{A}$ generates a strongly contractive semigroup T_t. It remains to prove that T_t is positive.

CLAIM 18.38.1. For all $f \in \mathcal{D}(\overline{A})$ and $\lambda > 0$ if $(\lambda - \overline{A})f \geq 0$ then $f \geq 0$.

We argue by contradiction. Pick $f \in \mathcal{D}(\overline{A})$, $\lambda > 0$ and suppose $\inf_{x \in S} f(x) < 0$. Pick $f_n \in \mathcal{D}(A)$ such that $\lim_{n \to \infty} f_n = f$ and $\lim_{n \to \infty} Af_n = \overline{A}f$. It follows that $\lim_{n \to \infty} (\lambda - A) f_n = (\lambda - \overline{A}) f$. Using Corollary 15.44 for each $n \in \mathbb{N}$ we select $x_n \in S$ such that $f_n(x_n) = \inf_{x \in S} f_n(x)$ and select $x_0 \in S$ such that $f(x_0) = \inf_{x \in S} f(x) < 0$.

We need to piece together some simple facts.

CLAIM 18.38.2. $\lim_{n \to \infty} f_n(x_n) = f(x_0)$

From the definition of x_n and $\lim_{n \to \infty} f_n = f$ we see that $\lim_{n \to \infty} f_n(x_n) \leq \lim_{n \to \infty} f_n(x_0) = f(x_0)$. On the other hand for every $\epsilon > 0$ there exists N such that $\sup_{x \in S} |f_n(x) - f(x)| < \epsilon$ for $n \geq N$ and therefore $f(x_0) - \epsilon \leq f(x_n) - \epsilon \leq f_n(x_n)$ for $n \geq N$. This implies $f(x_0) - \epsilon \leq \lim_{n \to \infty} f_n(x_n)$ and since $\epsilon > 0$ was arbitrary we get $f(x_0) \leq \lim_{n \to \infty} f_n(x_n)$.

Note that exactly the same argument shows that $\lim_{n\to\infty} \inf_{x\in S}(\lambda - A) f_n(x) = \inf_{x\in S}(\lambda - \overline{A}) f(x)$.

Claim 18.38.3. For n sufficiently large $A f_n(x_n) \geq 0$.

From the previous claim and the fact that $f(x_0) < 0$ we know that for sufficiently large n we must have $f_n(x_n) \leq 0$ and thus $-f_n(x_n) = \sup_{x\in S}(-f_n(x)) \geq 0$ for sufficiently large n; by the positive maximum principle applied to $-f_n$ we see that $A f_n(x_n) \geq 0$.

By the previous claim we get for every $n \in \mathbb{N}$,

$$\inf_{x\in S}(\lambda - A) f_n(x) \leq (\lambda - A) f_n(x_n) \leq \lambda f_n(x_n)$$

and therefore taking the limit as $n \to \infty$ we have

$$\inf_{x\in S}(\lambda - \overline{A}) f(x) = \lim_{n\to\infty} \inf_{x\in S}(\lambda - A) f_n(x) \leq \lim_{n\to\infty} \lambda f_n(x_n) = \lambda f(x_0) < 0$$

and Claim 18.38.1 is shown.

To finish, note that the positive cone $\{f \in C_0(S) \mid f \geq 0\}$ is closed and convex. Claim 18.38.1 shows that for all $\lambda > 0$ if $f \geq 0$ then $(\lambda - A)^{-1} f \geq 0$ so by Corollary 18.30 we conclude that T_t is positive for all $t \geq 0$. ☐

As it turns out the strong continuity required in the definition of a Feller semigroup may be derived from a weaker pointwise continuity property.

Proposition 18.39. *Let S be locally compact Hausdorff and suppose that $T_t : C_0(S) \to C_0(S)$ is a positive contraction semigroup such that*

$$\lim_{t\to 0} T_t f(x) = f(x) \text{ for all } f \in C_0(S) \text{ and } x \in S$$

then T_t is a Feller semigroup.

Proof. TODO: The proof follows by defining the resolvent for $\lambda > 0$ explicitly using the Laplace transform and then using the Yosida approximation. ☐

2.3. Cores.

Definition 18.40. Let A be a closed operator then $D \subset \mathcal{D}(A)$ is a *core* if $\overline{A\mid_D} = A$.

Proposition 18.41. *Let A be the generator of a strongly continuous contraction semigroup then D is a core for A if and only if $(\lambda_0 - A)(D)$ is dense for some $\lambda_0 > 0$. In either case $(\lambda - A)(D)$ is dense for all $\lambda > 0$.*

Proof. Suppose D is a core then let $v \in X$ then $R_\lambda v \in \mathcal{D}(A)$ and therefore we can find $w_n \in D$ such that $\lim_{n\to\infty} w_n = R_\lambda v$ and $\lim_{n\to\infty} A w_n = A R_\lambda v$. Therefore $\lim_{n\to\infty}(\lambda - A) w_n = (\lambda - A) R_\lambda v = v$.

Suppose $(\lambda_0 - A)(D)$ is dense for some $\lambda_0 > 0$. Let $v \in \mathcal{D}(A)$ and choose $v_n \in D$ such that $\lim_{n\to\infty}(\lambda_0 - A) v_n = (\lambda_0 - A) v$. From Proposition 18.14 we know that R_{λ_0} is bounded and therefore

$$\lim_{n\to\infty} v_n = R_{\lambda_0} \lim_{n\to\infty} (\lambda_0 - A) v_n = R_{\lambda_0}(\lambda_0 - A) v = v$$

from which we get

$$\lim_{n\to\infty} A v_n = \lim_{n\to\infty} \lambda_0 v_n - \lim_{n\to\infty} (\lambda_0 - A) v_n = \lambda_0 v - (\lambda_0 - A) v = A v$$

TODO: In EK they add the assumption that D is dense; do we need that or is it a consequence of $(\lambda_0 - A)(D)$ is dense? ☐

Proposition 18.42. *Let A be the generator of a strongly continuous contraction semigroup and let $D_0 \subset D \subset \mathcal{D}(A)$ be dense subspaces such that $T_t D_0 \subset D$ for all $t \geq 0$ then D is a core.*

Proof. Let $v \in D_0$, $\lambda > 0$ and define

$$v_n = \frac{1}{n} \sum_{k=0}^{n^2} e^{-\lambda k/n} T_{k/n} v \in D$$

for all $n \in \mathbb{N}$.

$$\begin{aligned}\lim_{n\to\infty} (\lambda - A) v_n &= \lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n^2} e^{-\lambda k/n} T_{k/n} (\lambda - A) v \\ &= \int_0^\infty e^{-\lambda t} T_t (\lambda - A) v \, dt = R_\lambda (\lambda - A) v = v\end{aligned}$$

which shows us that $D_0 \in \overline{(\lambda - A)(D)}$. Now by density of D_0 we see that $(\lambda - A)(D)$ is dense and by Proposition 18.41 we see that D is a core. □

3. Existence of Feller Processes

Our next goal is to show that there is a cadlag Feller process associated with any Feller semigroup. We begin with some motivational comments. It isn't too hard to come up with a rough sketch for how to proceed:

- since the semigroup is on the Banach space $C_0(S)$ for each $x \in S$ and $t \geq 0$ we get a positive linear functional $\lambda_{t,x} f = T_t f(x)$
- by the Riesz Representation Theorem we therefore get transition measures $\mu_t(s, \cdot)$ each of which is a finite Radon measure
- the semigroup property implies the Chapman Kolmogorov relations thus we can assert the existence of a Markov process with transition measures $\mu_t(x, \cdot)$.

A couple of things need to be sorted out. We have not come up with an obvious plan for how to prove that there is a cadlag modification of the constructed Markov process. Indeed this requires some real ingenuity and work. The second issue is much simpler; in the second step of our plan we mentioned that the Riesz Representation Theorem will only guarantee that we have a finite Radon measure but not necessarily a probability measure. Indeed the contraction property of T_t shows that the total mass of the measure is less than or equal to one (it is equal to $\|\lambda_{t,x}\|$); however, in our definition of a Feller semigroup we have not introduced the condition that is necessary to guarantee that the constructed transition measures are probability measures. From a Markov process point of view, we have not introduced the condition that prevents explosion. We do that now.

Definition 18.43. A Feller semigroup is *conservative* if and only if $\sup_{f \leq 1} T_t f(x) = 1$ for all $x \in S$ and all $t \geq 0$.

Morally we want the property of being conservative to simply say that $T_t 1 = 1$ for all $t \geq 0$ but we have chosen to define a Feller semigroup as being defined only on $C_0(S)$ and 1 does not vanish at infinity unless S is compact. As we will see it is in fact possible to extend every Feller semigroup to the entire space $B_b(S)$ in which case the property $T_t 1 = 1$ for all $t \geq 0$ is a valid (and equivalent) definition of conservativeness. We don't simply proceed by tacking on the hypothesis that T_t is conservative, rather we show that every Feller semigroup can naturally be extended to a conservative Feller semigroup. From a transition measure point of view we add a new state to S and put the missing mass there so that every transition measure becomes a probability measure. From a Markov process point of view we are adding an absorbing state to which explosions can transition.

We append a state Δ to the state space S. The manner in which this is done depends on whether S is compact or not. If S is compact then we make Δ an isolated point of $S^\Delta = S \cup \{\Delta\}$ otherwise we let Δ be the point at infinity in the one point compactification of S (Definition 15.52). In the latter case, recall from Theorem 15.53 that S^Δ is compact and from Proposition 15.43 that we can regard every element of $f \in C_0(S)$ isometrically as an element of $C(S^\Delta)$ by defining $f(\Delta) = 0$. In the former case we leave it to the reader to verify both properties (they follow immediately from the fact that $\{\Delta\}$ is an open set). It what follows we will freely identify $C_0(S)$ with the subspace of $f \in C(S^\Delta)$ for which $f(\Delta) = 0$.

Proposition 18.44. *Let S be locally compact separable metric space and let T_t be a Feller semigroup on S. Define T_t^Δ on $C(S^\Delta)$ by*

$$T_t^\Delta f = f(\Delta) + T_t(f - f(\Delta))$$

then T_t^Δ is a conservative Feller semigroup on S^Δ.

Proof. First note that for any $f \in C(S^\Delta)$, $T_t^\Delta f(\Delta) = f(\Delta)+T_t(f-f(\Delta))(\Delta) = f(\Delta)$ since $T_t(f - f(\Delta)) \in C_0(S)$.

Claim 18.44.1. T_t^Δ is a semigroup

To see the semigroup property of T_t^Δ use the semigroup property of T_t and the fact that $T_s^\Delta f(\Delta) = f(\Delta)$,

$$\begin{aligned} T_{t+s}^\Delta f &= f(\Delta) + T_{t+s}(f - f(\Delta)) = f(\Delta) + T_t T_s(f - f(\Delta)) \\ &= f(\Delta) + T_t(T_s^\Delta f - f(\Delta)) = T_s^\Delta f(\Delta) + T_t(T_s^\Delta f - T_s^\Delta f(\Delta))) = T_t^\Delta T_s^\Delta f \end{aligned}$$

and also

$$T_0^\Delta f = f(\Delta) + T_0(f - f(\Delta)) = f(\Delta) + f - f(\Delta) = f$$

Strong continuity of T_t^Δ follows easily from strong continuity of T_t,

$$\lim_{t\to 0} T_t^\Delta f = f(\Delta) + \lim_{t\to 0} T_t(f - f(\Delta)) = f(\Delta) + f - f(\Delta) = f$$

Claim 18.44.2. T_t^Δ is positive

Let $f \in C_0(S)$ and pick $\alpha \in \mathbb{R}$ such that $\alpha + f \geq 0$. Note that $\alpha + f \in C(S^\Delta)$ and by definition $T_t^\Delta(\alpha + f) = \alpha + T_t f$; also note that every $g \in C(S^\Delta)$ may be written in the form $\alpha + f$ for $f \in C_0(S)$. Write $f = f_+ - f_-$ with $f_\pm \geq 0$ so that $T_t f_\pm \geq 0$. Therefore $f_- + f = f_- + (f_+ - f_-) = f_+$ and positivity of T_t implies $-T_t f \leq T_t f_-$. Therefore

$$(T_t f)_- = (-T_t f \vee 0) \leq (T_t f_- \vee 0) = T_t f_-$$

Since T_t is a contraction, $\|T_t f_-\| \leq \|f_-\| \leq \alpha$ and therefore $(T_t f)_- \leq \alpha$ which imples $\alpha + T_t f \geq 0$.

To see that T_t^Δ is a contraction, note that since $\|f\| \pm f \geq 0$, by positivity $\left\|T_t^\Delta f\right\| \leq T_t^\Delta \|f\| = \|f\|$ and therefore $\|T_t\| \leq 1$.

To see that T_t^Δ is conservative from $T_t^\Delta 1 = 1$ and the positivity of T_t^Δ,

$$\sup_{f \leq 1} T_t^\Delta f \leq T_t^\Delta 1 = 1$$

$\square$

Now we can show that a unique transition kernel can be associated with any Feller semigroup. By the application Daniell-Kolmogorov Theorem an associated Markov process exists as well.

Proposition 18.45. *Let S be locally compact separable metric space and let T_t be a Feller semigroup on S, then there exists a unique homogeneous transition kernel μ_t on S^Δ such that*

$$T_t f(x) = \int f(s) \mu_t(x, ds) \text{ for all } f \in C_0(S) \text{ and } x \in S$$

and such that $\mu_t(\Delta, \{\Delta\}) = 1$ for all $t \geq 0$.

Moreover there exists a homogeneous Markov process X with transition kernel μ_t and semigroup T_t.

Proof. Let T_t^Δ be the conservative Feller semigroup on $C(S^\Delta)$ constructed in Proposition 18.44. For fixed $x \in S^\Delta$ and $t \geq 0$, $T_t f(x)$ defines a positive linear functional on $C(S^\Delta)$. By the Riesz-Markov Theorem 15.63 we know that there exists a Radon measure $\mu_t(x, \cdot)$ such that $T_t f(x) = \int f(s) \, \mu_t(x, ds)$. Since T_t^Δ is conservative we have $1 = T_t 1 = \mu_t(x, S)$ and therefore $\mu_t(x, \cdot)$ is a probability measure for all $x \in S^\Delta$ and $t \geq 0$.

To see that μ_t is a probability kernel first note that by Proposition 18.10 we know that $T_t f(x) = \int f(s) \, \mu_t(x, ds)$ is a continuous function of x for every $f \in C(S^\Delta)$ and $t \geq 0$. Given a open set $U \subset S^\Delta$ we pick a metric d on S and approximate $\mathbf{1}_U(s) = \lim_{n \to \infty} nd(s, U^c) \wedge 1$ use Dominated Convergence to see that $\mu_t(x, U) = \lim_{n \to \infty} \int (nd(s, U^c) \wedge 1) \, \mu(x, ds)$ and therefore $\mu_t(x, U)$ is measurable in x (Lemma 2.14). Finally let $t \geq 0$ be fixed and define $\mathcal{C} = \{A \in \mathcal{B}(S) \mid \mu_t(x, A) \text{ is measurable}\}$. If $A, B \in \mathcal{C}$ and $A \subset B$ then $\mu_t(x, B \setminus A) = \mu_t(x, B) - \mu_t(x, A)$ is measurable hence $B \setminus A \in \mathcal{C}$ and if $A_1 \subset A_2 \subset \cdots$ with $A_n \in \mathcal{C}$ then $\mu_t(x, \cup_n A_n) = \lim_{n \to \infty} \mu_t(x, A_n)$ by Lemma 2.30 hence $\cup_n A \in \mathcal{C}$. It follows that $\mathcal{C}$ is a λ-system and it is clear that the set of open sets is a π-system therefore $\mathcal{C} = \mathcal{B}(S)$ by the π-λ Thereom 2.27.

The Chapman-Kolmogorov relations follow from Proposition 18.1. To see $\mu_t(\Delta, \{\Delta\}) = 1$ note that for every $f \in C_0(S)$ we have

$$\int f(s) \, \mu_t(\Delta, ds) = T_t^\Delta f(\Delta) = f(\Delta) + T_t(f - f(\Delta))(\Delta) = 0$$

so just approximate $\mathbf{1}_S$ by an sequence of nonnegative $f \in C_0(S)$ and use Fatou's Lemma (TODO: Explicity create such a sequence).

Since a compact metric space S^Δ is complete (Theorem 1.29) it is Polish and therefore Borel (Theorem 15.15). Applying Theorem 13.11 we conclude that there is a Markov process with homogeneous transition kernels μ_t. $\square$

As mentioned we have to be a bit more clever to show that we may find a cadlag version of Markov process X constructed in the last result. The basic idea is that we show that many functions of X are in fact supermartingales and therefore have cadlag versions; once enough functions X are known to have a cadlag version then we may approximate the identity and conclude that X itself has a cadlag version. Before stating the Theorem we need two preliminary lemmas. The first constructs the supermartingales in question and the latter gives a probabilistic interpretation of the strong continuity of the Feller semigroup; the probabilistic continuity will be used in the limiting process of showing X is cadlag.

LEMMA 18.46. *Let S be locally compact separable metric space, let T_t be a Feller semigroup on S, X be the Markov process on S^Δ defined by T_t and initial distribution ν and $\mathcal{F}$ be the filtration induced by X. For every $f \in C_0(S)$ such that $f \geq 0$ define*

$$Y_t = e^{-t} R_1 f(X_t) = e^{-t} \int_0^\infty e^{-s} T_s f(X_t)\, ds$$

then Y_t is a $\mathcal{F}$-supermartinagle.

PROOF. Using the Markov property, the definitions of T_t and R_1, a change of integration variables and the non-negativity of f,

$$\begin{aligned}
\mathbf{E}\left[Y_t \mid \mathcal{F}_s\right] &= \mathbf{E}\left[e^{-t} R_1 f(X_t) \mid \mathcal{F}_s\right] \\
&= \mathbf{E}_{X_s}\left[e^{-t} R_1 f(X_{t-s})\right] \\
&= T_{t-s} e^{-t} R_1 f(X_s) \\
&= T_{t-s} e^{-t} \int_0^\infty e^{-u} T_u f(X_s)\, du \\
&= e^{-t} \int_0^\infty e^{-u} T_{t+u-s} f(X_s)\, du \\
&= e^{-s} \int_0^\infty e^{-(t+u-s)} T_{t+u-s} f(X_s)\, du \\
&= e^{-s} \int_{t-s}^\infty e^{-u} T_u f(X_s)\, du \\
&\leq e^{-s} \int_0^\infty e^{-u} T_u f(X_s)\, du \\
&= e^{-s} R_1 f(X_s) = Y_s
\end{aligned}$$

TODO: Make sure everything here is sensible with respect to working on S versus on S^Δ. The main point is that one can compute the resolvent of f using either T_t or the extension T_t^Δ; however since $f(\Delta) = 0$ we know that $T_t f = T_t^\Delta f$.

TODO: What about the use of $\mathcal{F}_+^X$ in EK? ∎

LEMMA 18.47. *Let (S, ρ) be compact separable metric space, let X^x be a Markov process on S starting at $x \in S$ and with a Feller transition semigroup T_t, then $\lim_{h \to 0^+} \sup_x \boldsymbol{E}_x\left[\rho(X^x_{t+h}, X^x_t) \wedge 1\right] = 0$ for all $t \geq 0$. In particular for every initial distribution μ, $X^\nu_{t+h} \xrightarrow{P} X^\nu_t$ as $h \downarrow 0$ for all $t \geq 0$.*

PROOF. Since S is compact and separable it follows that $C(S)$ is separable (Lemma 15.17). Pick a countable dense set $f_1, f_2, \ldots \in C(S)$.

CLAIM 18.47.1. Let x_n be a sequence in S then $\lim_{n\to\infty} x_n = x$ if and only if $\lim_{n\to\infty} f_m(x_n) = f_m(x)$ for all $m \in \mathbb{N}$.

Clearly if $x_n \to x$ then $f_m(x_n) \to f_m(x)$ for all f_m since f_m is continuous. On the other hand, suppose that x_n does not converge. Then by compactness there exists $x \in S$ and a subsequence $N \subset \mathbb{N}$ such that x_n converges to x along N. Since x_n does not have a limit there exists an open neighborhood of x such that $x_n \notin U$ infinitely often; again by compactness we may pass to a convergent subsequence $N' \subset \mathbb{N}$ and by construction x_n converges to y along N' and $x \neq y$. The function $\rho(x, \cdot)$ is continuous and $\rho(x,x) \neq \rho(x,y)$. Pick f_m such that $\sup_{s \in S} |\rho(x,s) - f_m(s)| < \rho(x,y)/2$ in particular by the triangle inequality $|f_m(x) - f_m(y)| \geq \rho(x,y) - |\rho(x,y) - f_m(y)| - |f_m(x)| > 0$ and $f_m(x) \neq f_m(y)$. By continuity of f_m, $f_m(x_n) \to f_m(x)$ along N and $f_m(x_n) \to f_m(y)$ along N' and therefore $f_m(x_n)$ does not converge.

By the claim, it follows that ρ is topologically equivalent to the metric

$$\rho'(x,y) = \sum_{m=1}^{\infty} 2^{-m}(|f_m(x) - f_m(y)| \wedge 1)$$

Now suppose that $f \in C(S)$, $x \in S$ and $t, h \geq 0$ we compute

$$\begin{aligned}
&\mathbf{E}_x\left[(f(X_t) - f(X_{t+h}))^2\right] = \mathbf{E}_x\left[f^2(X_t) - 2f(X_t)f(X_{t+h}) + f^2(X_{t+h})\right] \\
&= \mathbf{E}_x\left[f^2(X_t) - 2f(X_t)\mathbf{E}\left[f(X_{t+h}) \mid \mathcal{F}_t\right] + \mathbf{E}\left[f^2(X_{t+h}) \mid \mathcal{F}_t\right]\right] \\
&= \mathbf{E}_x\left[f^2(X_t) - 2f(X_t)T_h f(X_t) + T_h f^2(X_t)\right] \\
&\leq \sup_{x \in S}\left|f^2(x) - 2f(x)T_h f(x) + T_h f^2(x)\right| \\
&\leq \sup_{x \in S}\left|2f^2(x) - 2f(x)T_h f(x)\right| + \sup_{x \in S}\left|T_h f^2(x) - f^2(x)\right| \\
&\leq 2\sup_{x \in S}|f(x)| \sup_{x \in S}|f(x) - T_h f(x)| + \sup_{x \in S}\left|T_h f^2(x) - f^2(x)\right|
\end{aligned}$$

From the strong continuity of T_t we conclude $\lim_{h \to 0} \sup_{x \in S} \mathbf{E}_x\left[(f(X_t) - f(X_{t+h}))^2\right] = 0$ for each fixed $f \in C(S)$. By Cauchy-Schwartz or Jensen's Inequality we know that $\mathbf{E}_x\left[|f(X_t) - f(X_{t+h})|\right]^2 \leq \mathbf{E}_x\left[(f(X_t) - f(X_{t+h}))^2\right]$ and therefore we also have $\lim_{h \to 0} \sup_{x \in S} \mathbf{E}_x\left[|f(X_t) - f(X_{t+h})|\right] = 0$ for each fixed $f \in C(S)$. In particular this is true for each of the f_m and therefore by Tonelli's Theorem (Corollary 2.44) and Dominated Convergence we get

$$\begin{aligned}
\lim_{h \to 0^+} \sup_x \mathbf{E}_x\left[\rho'(X^x_{t+h}, X^x_t)\right] &= \lim_{h \to 0^+} \sup_x \sum_{m=1}^{\infty} 2^{-m}\mathbf{E}_x\left[\left|f_m(X^x_{t+h}) - f_m(X^x_t)\right| \wedge 1\right] \\
&\leq \lim_{h \to 0^+} \sum_{m=1}^{\infty} 2^{-m} \sup_x \mathbf{E}_x\left[\left|f_m(X^x_{t+h}) - f_m(X^x_t)\right| \wedge 1\right] \\
&= \sum_{m=1}^{\infty} 2^{-m} \lim_{h \to 0^+} \sup_x \mathbf{E}_x\left[\left|f_m(X^x_{t+h}) - f_m(X^x_t)\right| \wedge 1\right] = 0
\end{aligned}$$

TODO: Now argue that this implies the result for ρ not just ρ'.

To see the last statement by Lemma 13.13 and Dominated Convergence we get

$$\begin{aligned}\lim_{h \to 0^+} \mathbf{E}_\nu \left[\rho(X_{t+h}, X_t) \wedge 1\right] &= \lim_{h \to 0^+} \int \mathbf{E}_x \left[\rho(X_{t+h}, X_t) \wedge 1\right] \nu(dx) \\ &= \int \lim_{h \to 0^+} \mathbf{E}_x \left[\rho(X_{t+h}, X_t) \wedge 1\right] \nu(dx) = 0\end{aligned}$$

so we use Lemma 5.9. □

THEOREM 18.48. *Let S be locally compact separable metric space, let T_t be a Feller semigroup on S and let X be the Markov process on S^Δ defined by T_t and initial distribution ν. Then X has a cadlag version $\tilde{X}$ such that Δ is an absorbing state for $\tilde{X}$. If T_t is conservative then X has a cadlag version $\tilde{X}$ with values in S.*

PROOF. By Lemma 18.46 we know that $e^{-t} R_1 f(X_t)$ is a supermartingale for any $f \in C_0(S)$ with $f \geq 0$. Thus applying Theorem 9.76 we know that there is a P_ν-null set in $N_f \in \mathcal{F}^X_\infty$ such that the restriction of $e^{-t} R_1 f(X_t)$ to $\mathbb{Q}_+$ has left and right limits for all $t \geq 0$ outside of N_f; multiplying by the continuous function e_t, the same statement holds for $R_1 f(X_t)$. Writing an arbitrary $f \in C_0(S)$ as $f = f_+ - f_-$ with $f_\pm \in C_0(S)$ and $f_\pm \geq 0$ and using the linearity of R_1 we see that $R_1 f(X_t)$ almost surely has all left and right limits along $\mathbb{Q}_+$ for arbitrary $f \in C_0(S)$. By definition of the resolvent and the Hille-Yosida theorem we know that $\Re(R_1 f)$ is dense in $C_0(S)$ (or is it $C(S^\Delta)$?). Thus given an arbitrary $f \in C_0(S)$ we may find $f_n \in C_0(S)$ such that $\lim_{n\to\infty} \sup_{x \in S} \| R_1 f_n(x) - f(x) \|$. For every $t \geq 0$, $q \in \mathbb{Q}_+$ and $n \in \mathbb{N}$ we write

$$\begin{aligned}|f(X_q) - f(X_t)| &\leq |f(X_q) - R_1 f_n(X_q)| + |R_1 f_n(X_q) - R_1 f_n(X_t)| + |R_1 f_n(X_t) - f(X_t)| \\ &\leq 2 \sup_{x \in S} \| R_1 f_n(x) - f(x) \| + |R_1 f_n(X_q) - R_1 f_n(X_t)|\end{aligned}$$

Each $R_1 f_n$ has left and right limits along $\mathbb{Q}_+$ outside a null set N_n. Letting either $q \uparrow t$ or $q \downarrow t$ and then taking $n \to \infty$ we see that $f(X_t)$ has left and right limits along $\mathbb{Q}_+$ outside the null set $\cup_n N_n$. Note also that $C_0(S)$ is separable (Corollary 15.46) so we may find a countable dense subset $f_1, f_2, \ldots \in C_0(S)$ and letting $N = \cup_n N_{f_n}$ it follows by repeating the same limiting argument as above that for all $f \in C_0(S)$, $R_1 f$ has left and right limits along $\mathbb{Q}$ outside of N (where N is now independent of f).

CLAIM 18.48.1. X has left and right limits along $\mathbb{Q}_+$ outside of the null set N.

By a slight modification of the argument in Lemma 18.47 one sees that the compactness of S^Δ implies that if we are given a sequence $x_1, x_2, \ldots \in S^\Delta$ such that $f(x_n)$ converges for every $f \in C_0(S)$ then it follows that x_n converges in S^Δ (observe that if there are two convergent subsequences which converge to different elements of S^Δ then we can separate the two points with an element $f \in C_0(S)$). From this fact, for every $\omega \notin N$ and we know that all increasing or decreasing sequences $q_n \in \mathbb{Q}_+$ we have $\lim_{n\to\infty} f(X_{q_n}(\omega))$ exists and thus $\lim_{n\to\infty} X_{q_n}(\omega)$ exists.

Now we can define $\tilde{X}_t \equiv \Delta$ on N and $\tilde{X}_t = \lim_{\substack{q \to t^+ \\ q \in \mathbb{Q}}} X_q$ off of N and by the proof of Theorem 9.76 we know that $\tilde{X}_t$ is cadlag.

CLAIM 18.48.2. $\tilde{X}$ is a version of X.

The claim means simply that for every $t \geq 0$ we have $\lim_{\substack{q \to t^+ \\ q \in \mathbb{Q}}} X_q = X_t$ a.s. We know that the limit in question exists almost surely (i.e. off of N) so it suffices to show that $\lim_{\substack{q \to t^+ \\ q \in \mathbb{Q}}} X_q = X_t$ a.s. along a subsequence. Lemma 18.47 we know that $X_q \xrightarrow{P} X_t$ and the convergence of the subsequence follows by Lemma 5.10.

Now let $f \in C_0(S)$ with $f > 0$ on S (e.g. $f(x) = \rho(\Delta, x)$). Note that $R_1 f > 0$; by strong continuity for each $x \in S$ we pick $\delta > 0$ such that $\|T_t f - f\| < f(x)/2$ for $0 \leq t \leq \delta$ and if follows that $|T_t f(x)| \geq f(x) - |T_t f(x) - f(x)| \geq f(x)/2 > 0$ for all $0 \leq t \leq \delta$ and therefore by positivity of T_t we have $R_1 f(x) \geq \int_0^\delta e^{-u} T_u f(x)\, du \geq f(x)(1 - e^{-\delta})/2 > 0$. It is also clear that $f(\Delta) = 0$ implies $R_1 f(\Delta) = 0$. Therefore $Y_t = e^{-t} R_1 f(\tilde{X}_t)$ is a non-negative cadlag supermartingale such that $Y_t = 0$ is equivalent to $\tilde{X}_t = \Delta$. If we apply Lemma 14.61 to Y_t and translate the conclusion in terms of $\tilde{X}_t$ we see that if we define $\tau = \inf\{t \mid \tilde{X}_{t-} \wedge \tilde{X}_t = \Delta\}$ then $\tilde{X}_t \equiv \Delta$ a.s. on $[\tau, \infty)$. Setting $\tilde{X}$ to be identically Δ on the null set where the conclusion fails we that by a further modification of $\tilde{X}$ we can assume Δ is absorbing everywhere.

If T_t is conservative on $C_0(S)$ and we assume that $\nu(\{\Delta\}) = 0$ then it follows that $\tilde{X}_t \in S$ a.s. for every $t \geq 0$ (TODO: Why?) Therefore we must have $\tau > t$ a.s. for all $t \geq 0$ which implies that $\tau = \infty$ a.s. Pick an arbitrary $x \in S$ and make an additional modification of $\tilde{X}$ to set $\tilde{X}_t \equiv x$ on $\tau < \infty$ and therefore $\tau = \infty$ everywhere. This implies that $\tilde{X}_t$ and $\tilde{X}_{t-}$ take values in S. $\square$

TODO: Show that we get a homogeneous cadlag Markov family out of this construction. At a minimum we can see that we get probability measures $\mathbf{P}_x$ for each $x \in S^\Delta$ and this is a kernel from S to $S^{[0,\infty)}$ (by Lemma 13.13).

Theorem 18.49. *Let X be a Feller family, $A \in (\mathcal{S}^\Delta)^{[0,\infty)}$ and let τ be an $\mathcal{F}_+^X$-optional time, then*

$$\boldsymbol{P}\{\theta_\tau X \in A \mid \mathcal{F}_{\tau+}^X\} = \boldsymbol{P}_{X_\tau}\{A\} \text{ a.s. on } \tau < \infty$$

Let X be the canonical Feller process, ν be an initial distribution, τ an $\mathcal{F}_+^X$-optional time, ξ a non-negative random variable then

$$\boldsymbol{E}_\nu\left[\xi \circ \theta_\tau \mid \mathcal{F}_{\tau+}^X\right] = \boldsymbol{E}_{X_\tau}[\xi] \;\; \boldsymbol{P}_\nu \text{ a.s. on } \tau < \infty$$

Proof. By Proposition 13.31 it suffices to assume that $\tau < \infty$ almost surely. Let $A \in (\mathcal{S}^\Delta)^{[0,\infty)}$. Define $\tau_n = 2^{-n}\lfloor 2^n \tau + 1 \rfloor$ so that by Lemma 9.71 we know that τ_n are $\mathcal{F}^X$-optional times and $\tau_n \downarrow \tau$. Since $\tau < \tau_n$ we also know that $\mathcal{F}_{\tau+}^X \subset \mathcal{F}_{\tau_n}^X$ for all $n \in \mathbb{N}$. Since τ_n is countably valued X is strong Markov at τ_n (Theorem 13.29) and

$$\mathbf{P}\{\theta_{\tau_n} X \in A \mid \mathcal{F}_{\tau_n}^X\} = \mathbf{P}_{X_{\tau_n}}\{A\} \text{ a.s. on } \tau_n < \infty$$

It is useful to avoid appeal to Theorem 13.29 so that the role of the Feller properties can be appreciated even in the case of countably valued τ (not sure I agree because the same continuity argument seems to appear when approximating by countably valued). So let τ be $\mathcal{F}_+^X$-optional and let T be the countable range of τ, $f \in C(S^\Delta)$, $s \geq 0$ and $A \in \mathcal{F}_{t+}^X$. Thus, for every $t \geq 0$, $\epsilon > 0$ we have $A \cap \{\tau = t\} \in \mathcal{F}_{t+}^X \subset \mathcal{F}_{t+\epsilon}^X$. Let $0 \leq \epsilon \leq s$ and use the tower property of conditional expectation and Proposition 18.2 to get

$$\begin{aligned}
\mathbf{E}\left[f(X_{\tau+s});A\right] &= \sum_{t\in T}\mathbf{E}\left[f(X_{t+s});A\cap\{\tau=t\}\right] \\
&= \sum_{t\in T}\mathbf{E}\left[\mathbf{E}\left[f(X_{t+s})\mid\mathcal{F}^X_{t+\epsilon}\right];A\cap\{\tau=t\}\right] \\
&= \sum_{t\in T}\mathbf{E}\left[T_{s-\epsilon}f(X_{t+\epsilon});A\cap\{\tau=t\}\right] \\
&= \mathbf{E}\left[T_{s-\epsilon}f(X_{\tau+\epsilon});A\right]
\end{aligned}$$

By strong continuity of $T_t f$ and right continuity of X_t we know that $\lim_{\epsilon\to 0}T_{s-\epsilon}f(X_{\tau+\epsilon}) = T_s f(X_\tau)$ (TODO: Show this in more detail; actually we don't need this since we only need countably valued $\mathcal{F}^X$-optional times) and therefore by Dominated Convergence we get $\mathbf{E}\left[f(X_{\tau+s});A\right] = \mathbf{E}\left[T_s f(X_\tau);A\right]$. Since $T_s f(X_\tau)$ is $\mathcal{F}^X_{\tau+}$-measurable we conclude

$$\mathbf{E}\left[f(X_{\tau+s})\mid\mathcal{F}^X_{\tau+}\right] = T_s f(X_\tau)$$

To extend this fact to arbitrary τ with $\tau<\infty$,

In what follows we are using a claim that $\sigma<\tau$ implies $\mathcal{F}_{\sigma+}\subset\mathcal{F}_\tau$. TODO: Prove it or disprove it (if it turns out not to be true then we use the ϵ argument above taken from EK).

Let $\tau_n = 2^{-n}\lfloor 2^n\tau+1\rfloor$ so that τ_n are countably valued $\mathcal{F}^X$-optional times such that $\tau<\tau_n$ and $\tau_n\downarrow\tau$(Lemma 9.71). Now we use continuity of $T_s f$, Dominated Convergence for conditional expectations and the $\mathcal{F}^X_{\tau+}$-measurability of $T_s f(X_\tau)$ (since X is cadlag, it is progressive by Lemma 9.89 thus X_τ is $\mathcal{F}^X_{\tau+}$-measurable by Lemma 9.90) to see

$$\begin{aligned}
&\mathbf{E}\left[f(X_{\tau+s})\mid\mathcal{F}^X_{\tau+}\right] \\
&= \mathbf{E}\left[\lim_{n\to\infty}f(X_{\tau_n+s})\mid\mathcal{F}^X_{\tau+}\right] && \text{right continuity of } X\text{, continuity of } f \\
&= \lim_{n\to\infty}\mathbf{E}\left[f(X_{\tau_n+s})\mid\mathcal{F}^X_{\tau+}\right] && \text{Dominated Convergence} \\
&= \lim_{n\to\infty}\mathbf{E}\left[\mathbf{E}\left[f(X_{\tau_n+s})\mid\mathcal{F}^X_{\tau_n}\right]\mid\mathcal{F}^X_{\tau+}\right] && \text{chain rule of conditional expectations} \\
&= \lim_{n\to\infty}\mathbf{E}\left[T_s f(X_{\tau_n})\mid\mathcal{F}^X_{\tau+}\right] && \text{Theorem 13.29 and Proposition 18.2} \\
&= \mathbf{E}\left[\lim_{n\to\infty}T_s f(X_{\tau_n})\mid\mathcal{F}^X_{\tau+}\right] && \text{Dominated Convergence} \\
&= \mathbf{E}\left[T_s f(X_\tau)\mid\mathcal{F}^X_{\tau+}\right] && \text{right continuity of } X\text{, continuity of } T_s f \\
&= T_s f(X_\tau) && \mathcal{F}^X_{\tau+}\text{-measurability of } T_s f(X_\tau) \\
&= \int f(u)\,\mu_s(X_\tau,du)
\end{aligned}$$

Now we may apply Proposition 13.30 to complete the proof.

TODO: Show the canonical case... □

We first need to introduce the formalize notation and concepts surrounding bounded pointwise limits and closure.

DEFINITION 18.50. Let S be a metric space and let $B_b(S)$ be the space of bounded measurable functions then given $f, f_1, f_2, \ldots \in B_b(S)$ we say that $f = \text{bp-}\lim_{n\to\infty} f_n$ if and only if $sup_n \sup_{x\in S} |f(x)| < \infty$ and $\lim_{n\to\infty} f_n(x) = f(x)$ for all $x \in S$. We say that a set $M \subset S$ is *bp-closed* if $f_1, f_2, \ldots \in M$ and $f = \text{bp-}\lim_{n\to\infty} f_n$ implies $f \in M$.

TODO: On $C_0(S)$ we know (or should prove) that the dual $C_0^*(S)$ is the space of finite signed Radon measures. It turns out that a sequence $f = \text{bp-}\lim f_n$ if and only if f_n converges weakly to f (i.e. $\int f_n\, d\mu \to \int f\, d\mu$ for all finite signed Radon measures μ). This is not true for arbitrary nets however.

DEFINITION 18.51. A multivalued operator $A \subset B_b(S) \times B_b(S)$ is *conservative* if and only if $(1, 0)$ is contained in the bp-closure of A.

TODO: Understand the relationship between this definition of conservative and the more elementary statement that $T_t 1 = 1$ for all $t \geq 0$. Understand the relationship between this definition and the Kallenberg definition of conservativeness of a the semigroup (NOT the generator) that says $\sup_{f\leq 1} T_t f(x) = 1$ for all $x \in S$. Understand the relationship between convervativeness and the statement that a sub-Markov transition semigroup is Markov (i.e. the property that every measure has total mass 1). Yet another defintion of conservative that is specific to Feller semigroups: if $f_n \uparrow 1$, $f_n \in C_0(S)$ then $T_t f_n \uparrow 1$ (this is essentially Kallenberg's definition). I believe that the following is true: $(1, 0) \in bp - closure(A)$ implies $T_t 1 = 1$ for all $t \geq 0$. Also if S is compact then $T_t 1 = 1$ for all $t \geq 0$ if and only if $1 \in \mathcal{D}(A)$ and $A1 = 0$. Note also the fact that $1 \notin C_0(S)$ if S is not compact so the definition in terms of $T_t 1 = 1$ not correct; it can be rescued by proving another theorem (which is a consequence of Riesz representation) that says any semigroup of continuous operators on $C_0(S)$ can be extended to a semigroup of continuous operators on $B_b(S)$.

LEMMA 18.52. *Let A be the generator of a strongly continuous contraction semigroup on a subspace of $X \subset B_b(S)$ then if A is conservative then $T_t 1 = 1$ for all $t \geq 0$ (TODO: This statement requires the extension of T_t to $C_b(S)$ or at least a space that contains 1; this should follow from the construction of the transition semigroup constructed from the Feller semigroup).*

PROOF. By the Kolmogorov backward equation Proposition 18.11 we know that

$$\{(f, Af) \mid f \in \mathcal{D}(A)\} \subset \{(f, g) \in B_b(S) \times B_b(S) \mid T_t f - f = \int_0^t T_s g\, ds \text{ for all } t \geq 0\}$$

CLAIM 18.52.1. The right hand set is bp-closed

Suppose (f_n, g_n) is a sequence in the right hand set such that $\sup_n \sup_{x\in S} |f_n(x)| < \infty$, $\sup_n \sup_{x\in S} |g_n(x)| < \infty$, $\lim_{n\to\infty}(f_n(x), g_n(x)) = (f(x), g(x))$ for all $x \in S$. In particular it is the case that $\lim_{n\to\infty} f_n = f$ and $\lim_{n\to\infty} g_n = g$ in X (TODO: WRONG we only have pointwise limits not uniform limits; how do we know that $T_t f_n(x) \to T_t f(x)$ and $T_t g_n(x) \to T_t g(x)$? I believe the answer is that for contraction semigroups norm continuity and weak continuity coincide) and therefore $\lim_{n\to\infty} T_t f_n = T_t f$ and $\lim_{n\to\infty} T_t g_n = T_t g$ for all $t \geq 0$. Since T_t is a contraction operator we also have $\sup_n \sup_{x\in S} |T_t g_n(x)| < \infty$ and therefore by Dominated Convergence we also have $\lim_{n\to\infty} \int_0^t T_s g_n\, ds = \int_0^t T_s g\, ds$.

Now since A is conservative we conclude that $(1, 0)$ is in the right hand set and lemma follows. $\square$

4. Approximation of Feller Processes

In this section we begin considering the weak convergence theory of Feller Processes. Semigroup theory plays a central role and we begin by proving a convergence theorem for strongly continuous contraction semigroups that will be applied in the Feller case. We begin with two lemmas. The first extends Lemma 18.25.

LEMMA 18.53. *Let X and Y be Banach spaces, let T_t be a strongly continuous contraction semigroup on X with generator A, let S_t be a strongly continuous contraction semigroup on Y with generator B and let $\pi : X \to Y$ be a bounded linear operator. Let $v \in \mathcal{D}(A)$ assume that $\pi T_t v \in \mathcal{D}(B)$ for all $t \geq 0$ and that $B\pi T_t v : [0, \infty) \to Y$ is continuous then*

$$S_t \pi v - \pi T_t v = \int_0^t S_{t-s}(B\pi - \pi A) T_s v \, ds$$

for all $t \geq 0$ and in particular,

$$\|S_t \pi v - \pi T_t v\| = \int_0^t \|B\pi - \pi A) T_s v\| \, ds$$

PROOF. Let $v \in \mathcal{D}(A)$ and consider the term $S_{t-s} \pi T_s v$ for $0 \leq s \leq t$. Since $v \in \mathcal{D}(A)$ we know that $T_s v$ is a differentiable function of s and $\frac{d}{ds} T_s v = A T_s v$ (Proposition 18.11). Since π is a bounded linear map we know from the chain rule (Proposition 15.137) that $\pi T_s v$ is also differentiable and $\frac{d}{ds} \pi T_s v = \pi A T_s v$. Also for fixed $w \in \mathcal{D}(B)$ and $0 \leq s \leq t$ we know from Proposition 18.11 that $S_{t-s} w$ is differentiable with respect to s and $\frac{d}{ds} S_{t-s} w = -S_{t-s} B w$. We claim that we have a product rule for differentiation that shows

$$\frac{d}{ds} S_{t-s} \pi T_s v = S_{t-s} \pi A T_s v - S_{t-s} B \pi T_s v = -S_{t-s}(B\pi - \pi A) T_s v$$

TODO: Show this

Now we can just apply the Fundamental Theorem of Calculus 15.112 to see that

$$\int_0^t S_{t-s}(B\pi - \pi A) T_s v \, ds = -S_{t-s} \pi T_s v \mid_0^t = S_t \pi v - \pi T_t v$$

Also we can just apply Proposition 15.113 and the fact that S_t is a contraction to see

$$\|S_t \pi v - \pi T_t v\| \leq \int_0^t \|S_{t-s}(B\pi - \pi A) T_s v\| \, ds \leq \int_0^t \|(B\pi - \pi A) T_s v\| \, ds$$

$\square$

The second required lemma is a continuity property of the Yosida approximation.

LEMMA 18.54. *Let X and $X_1, X_2, \dots$ be Banach spaces, $T_{n,t}$ be strongly continuous contraction semigroups on X_n with generator A_n and let T_t be a strongly continuous contraction semigroup on X with generator A. Let $\pi_n : X \to X_n$ be bounded linear operators and assume that $\sup_n \|\pi_n\| < \infty$. Let D be a core for A and suppose that for every $v \in D$ there exists $v_n \in \mathcal{D}(A_n)$ such that $\lim_{n\to\infty} \|v_n - \pi_n v\| = 0$*

and $\lim_{n\to\infty} \|A_n v_n - \pi_n A v\| = 0$*. Suppose that* A_n^λ *and* A^λ *denote the Yosida approximations of* A_n *and* A *then for all* $v \in X$ *and* $\lambda > 0$ *we have*

$$\lim_{n\to\infty} \left\|A_n^\lambda \pi_n v - \pi_n A^\lambda v\right\| = 0$$

Proof. Let $\lambda > 0$ and $v \in D$. Set $w = (\lambda - A)v$ and pick $v_n \in \mathcal{D}(A_n)$ such that $\lim_{n\to\infty} v_n - \pi_n v = \lim_{n\to\infty} A_n v_n - \pi_n A v = 0$. Note that by the triangle inequality $\lim_{n\to\infty} \|(\lambda - A_n)v_n - \pi_n w\| = 0$. Recall that since

$$A^\lambda = \lambda A R_\lambda = \lambda(\lambda - (\lambda - A))R_\lambda = \lambda^2 R_\lambda - \lambda \,\mathrm{Id}$$

and from $\|R_\lambda\| \le \lambda^{-1}$ shows that $\left\|A^\lambda\right\| \le 2\lambda$. The same holds for A_n^λ where we use the notation $R_{n,\lambda}$ for the resolvent of $T_{n,t}$.

Using the above identity we compute

$$\begin{aligned}
\left\|A_n^\lambda \pi_n w - \pi_n A^\lambda w\right\| &= \left\|(\lambda^2 R_{n,\lambda} - \lambda \,\mathrm{Id})\pi_n w - \pi_n(\lambda^2 R_\lambda - \lambda \,\mathrm{Id})w\right\| \\
&= \lambda^2 \left\|(R_{n,\lambda}\pi_n - \pi_n R_\lambda)w\right\| \\
&\le \lambda^2 \left\|R_{n,\lambda}\pi_n w - v_n\right\| + \lambda^2 \left\|v_n - \pi_n R_\lambda w\right\| \\
&= \lambda^2 \left\|R_{n,\lambda}[\pi_n w - (\lambda - A_n)v_n]\right\| + \lambda^2 \left\|v_n - \pi_n v\right\| \\
&= \lambda \left\|\pi_n w - (\lambda - A_n)v_n\right\| + \lambda^2 \left\|v_n - \pi_n v\right\|
\end{aligned}$$

and therefore we see that $\lim_{n\to\infty} \left\|A_n^\lambda \pi_n w - \pi_n A^\lambda w\right\| = 0$ for $w \in (\lambda - A)(D)$.

On the other hand $(\lambda - A)(D)$ is dense in X (Proposition 18.41) and

$$\left\|A_n^\lambda \pi_n - \pi_n A^\lambda\right\| \le (\left\|A_n^\lambda\right\| + \left\|A^\lambda\right\|) \left\|\pi_n\right\| \le 4\lambda \sup_n \left\|\pi_n\right\| < \infty$$

and therefore $\lim_{n\to\infty} \left\|A_n^\lambda \pi_n w - \pi_n A^\lambda w\right\| = 0$ for $w \in X$ (let $w_m \to w$ with $w_m \in (\lambda - A)(D)$, write $\left\|A_n^\lambda \pi_n w - \pi_n A^\lambda w\right\| \le \left\|A_n^\lambda \pi_n w_m - \pi_n A^\lambda w_m\right\| + \sup_n \left\|A_n^\lambda \pi_n - \pi_n A^\lambda\right\| \left\|w_m - w\right\|$ and then let $n \to \infty$ followed by $m \to \infty$). ☐

Now we can present the semigroup convergence theorem itself.

Theorem 18.55. *Let* X *and* $X_1, X_2, \dots$ *be Banach spaces,* $T_{n,t}$ *be strongly continuous contraction semigroups on* X_n *with generator* A_n *and let* T_t *be a strongly continuous contraction semigroup on* X *with generator* A*. Let* $\pi_n : X \to X_n$ *be bounded linear operators and assume that* $\sup_n \|\pi_n\| < \infty$*. Let* D *be a core for* A *then the following are equivalent*

(i) *For every* $v \in X$ *and every* $t \ge 0$*,* $\lim_{n\to\infty} \sup_{0\le s\le t} \|T_{n,s}\pi_n v - \pi_n T_s v\| = 0$*.*
(ii) *For every* $v \in X$ *and every* $t \ge 0$*,* $\lim_{n\to\infty} \|T_{n,t}\pi_n v - \pi_n T_t v\| = 0$*.*
(iii) *For every* $v \in D$ *there exists* $v_n \in \mathcal{D}(A_n)$ *such that* $\lim_{n\to\infty} \|v_n - \pi_n v\| = 0$ *and* $\lim_{n\to\infty} \|A_n v_n - \pi_n A v\| = 0$*.*

Proof. The implication (i) implies (ii) is immediate.

To see that (ii) implies (iii), let $\lambda > 0$, $v \in \mathcal{D}(A)$ and set $w = (\lambda - A)v$. By Proposition 18.14 we have

$$v = R_\lambda w = \int_0^\infty e^{-\lambda t} T_t w \, dt$$

Now define

$$v_n = \int_0^\infty e^{-\lambda t} T_{n,t} \pi_n w \, dt = R_{n,\lambda} \pi_n w$$

and note that by (ii) we have $\lim_{n\to\infty} \|T_{n,t}\pi_n w - \pi_n T_t w\| = 0$ and since $T_{n,t}$ and T_t are contractions we have $\sup_n \|T_{n,t}\pi_n w - \pi_n T_t w\| \leq 2 \sup_n \|\pi_n\| \, \|w\| < \infty$. Now apply Proposition 15.116, Proposition 15.113 and Dominated Convergence

$$\begin{aligned}
\lim_{n\to\infty} \|v_n - \pi_n v\| &= \lim_{n\to\infty} \left\| \int_0^\infty e^{-\lambda t} T_{n,t}\pi_n w \, dt - \pi_n \int_0^\infty e^{-\lambda t} T_t w \, dt \right\| \\
&= \lim_{n\to\infty} \left\| \int_0^\infty e^{-\lambda t} T_{n,t}\pi_n w \, dt - \int_0^\infty e^{-\lambda t} \pi_n T_t w \, dt \right\| \\
&\leq \lim_{n\to\infty} \int_0^\infty e^{-\lambda t} \|T_{n,t}\pi_n w \, dt - \pi_n T_t w\| \, dt \\
&= \int_0^\infty e^{-\lambda t} \lim_{n\to\infty} \|T_{n,t}\pi_n w \, dt - \pi_n T_t w\| \, dt = 0
\end{aligned}$$

Also note that

$$\begin{aligned}
\|A_n v_n - \pi_n A v\| &= \|\lambda v_n - (\lambda - A_n) v_n - \lambda \pi_n v + \pi_n w\| \\
&\leq \|\lambda v_n - \lambda \pi_n v + \pi_n(\lambda - A) v\| + \|\pi_n w - (\lambda - A_n) v_n\| \\
&= \|\lambda v_n - \lambda \pi_n v + \pi_n(\lambda - A) v\| + \|\pi_n w - (\lambda - A_n) R_{n,\lambda} \pi_n w\| \\
&= \|\lambda v_n - \lambda \pi_n v + \pi_n(\lambda - A) v\|
\end{aligned}$$

and therefore (iii) is proven.

To see that (iii) implies (i) let A_n^λ and A^λ be the Yosida approximations of A_n and A respectively; let $T_{n,t}^\lambda = e^{tA_n^\lambda}$ and $T_t^\lambda = e^{tA^\lambda}$ be the corresponding strongly continuous contraction semigroups. Let $v \in D$ and choose $v_n \in \mathcal{D}(A_n)$ such that $\lim_{n\to\infty} \|v_n - \pi_n v\| = 0$ and $\lim_{n\to\infty} \|A_n v_n - \pi_n A v\| = 0$. Use the triangle inequality to write

$$\begin{aligned}
\|T_{n,t}\pi_n v - \pi_n T_t v\| \leq{}& \|T_{n,t}(\pi_n v - v_n)\| + \left\|T_{n,t} v_n - T_{n,t}^\lambda v_n\right\| + \left\|T_{n,t}^\lambda (v_n - \pi_n v)\right\| \\
&+ \left\|T_{n,t}^\lambda \pi_n v - \pi_n T_t^\lambda v\right\| + \left\|\pi_n (T_t^\lambda v - T_t v)\right\|
\end{aligned}$$

Now we consider estimates of each of the five terms on the right hand side. Let $t \geq 0$ be fixed. For the first term note that since $T_{n,t}$ is a contraction

$$\limsup_{n\to\infty} \sup_{0\leq s\leq t} \|T_{n,s}(\pi_n v - v_n)\| \leq \lim_{n\to\infty} \|\pi_n v - v_n\| = 0$$

and the same argument works for the third term as well. For the second term, apply Corollary 18.29, the triangle inequality, Lemma 18.54 and the fact that $\left\|A_n^\lambda\right\| \leq 2\lambda$ to see that

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{0\leq s\leq t} \left\|T_{n,s} v_n - T_{n,s}^\lambda v_n\right\| \leq t \limsup_{n\to\infty} \left\|A_n v_n - A_n^\lambda v_n\right\| \\
&\leq t \limsup_{n\to\infty} \left[\|A_n v_n - \pi_n A v\| + \left\|\pi_n A v - \pi_n A^\lambda v\right\| + \left\|\pi_n A^\lambda v - A_n^\lambda \pi_n v\right\| + \left\|A_n^\lambda \pi_n v - A_n^\lambda v_n\right\|\right] \\
&\leq t \sup_n \|\pi_n\| \left\|A v - A^\lambda v\right\|
\end{aligned}$$

For the fourth term we apply Lemma 18.53 to see that

$$\sup_{0\leq s\leq t} \left\|T^\lambda_{n,s}\pi_n v - \pi_n T^\lambda_s v\right\| \leq \int_0^t \left\|(A^\lambda_n \pi_n - \pi_n A^\lambda) T^\lambda_s v\right\| \, ds$$

By Lemma 18.54 we know that $\lim_{n\to\infty} \left\|(A^\lambda_n \pi_n - \pi_n A^\lambda) T^\lambda_s v\right\| = 0$ and moreover $\left\|(A^\lambda_n \pi_n - \pi_n A^\lambda) T^\lambda_s v\right\| \leq 4\lambda \sup_n \|\pi_n\| \|v\|$ so that by Dominated Convergence we get

$$\limsup_{n\to\infty} \sup_{0\leq s\leq t} \left\|T^\lambda_{n,s}\pi_n v - \pi_n T^\lambda_s v\right\| = 0$$

For the fifth and final term we apply Corollary 18.29 to get

$$\begin{aligned}
\limsup_{n\to\infty} \sup_{0\leq s\leq t} \left\|\pi_n (T^\lambda_s v - T_s v)\right\| &\leq \sup_n \|\pi_n\| \sup_{0\leq s\leq t} \left\|T^\lambda_s v - T_s v\right\| \\
&\leq t \sup_n \|\pi_n\| \left\|A^\lambda v - Av\right\|
\end{aligned}$$

Putting all of the estimates together we see that

$$\limsup_{n\to\infty} \sup_{0\leq s\leq t} \|T_{n,t}\pi_n v - \pi_n T_t v\| \leq 2t \sup_n \|\pi_n\| \left\|A^\lambda v - Av\right\|$$

Letting $\lambda \to \infty$ and using Lemma 18.24 we see that $\lim_{n\to\infty} \sup_{0\leq s\leq t} \|T_{n,t}\pi_n v - \pi_n T_t v\| = 0$ for $v \in D$. Since D is dense by the Hille-Yosida Theorem 18.35 for $v \in X$ we may take $v_m \in D$ such that $\lim_{m\to\infty} v_m = v$ and then

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{0\leq s\leq t} \|T_{n,t}\pi_n v - \pi_n T_t v\| \\
&\leq \limsup_{n\to\infty} \sup_{0\leq s\leq t} \left[\|T_{n,t}\pi_n v - T_{n,t}\pi_n v_m\| + \|T_{n,t}\pi_n v_m - \pi_n T_t v_m\| + \|\pi_n T_t v - \pi_n T_t v_m\|\right] \\
&\leq 2\sup_n \|\pi_n\| \|v - v_m\| + \lim_{n\to\infty} \sup_{0\leq s\leq t} \|T_{n,t}\pi_n v_m - \pi_n T_t v_m\| = 2\sup_n \|\pi_n\| \|v - v_m\|
\end{aligned}$$

Now let $m \to \infty$ and (i) follows. □

TODO: Compare Kallenberg and EK accounts of the following theorem and understand the differences and similarities. The main difference is that the Theorem in Kallenberg is really a combination of two theorems in EK. The first theorem is Theorem 18.55 and simply addresses equivalent conditions for convergence of sccs. The second theorem is what we have as Theorem 18.56 below and shows that Feller semigroup convergence implies weak convergence of Markov processes (using the standard FDD and tightness approach). EK and Kallenberg use the same argument to show FDD convergence whereas Kallenberg uses the Aldous criterion directly but EK use a more sophisticated argument for tightness using the martingale property (which may ultimately boil down to the Aldous criterion). Also EK are careful to prove the Theorem in such a way that the existence of cadlag X is not assumed but rather it is derived whereas Kallenberg has already proved that a cadlag X may be created from T_t and thus may appeal to the strong Markov property in showing the Aldous criterion is satisfied.

Theorem 18.56. *Let (S, r) be a locally compact and separable metric space and let X^n for $n \in \mathbb{N}$ be cadlag Feller processes in S with semigroup $T_{n,t}$ on $C_0(S)$ respectively. Suppose that T_t is a Feller semigroup on $C_0(S)$ and*

$$\lim_{n\to\infty} T_{n,t} f = T_t f \text{ for all } f \in C_0(S) \text{ and } t \geq 0$$

If the distributions of $X^n(0)$ converges to a limit $\nu \in \mathcal{P}(S)$ then there exists a cadlag Feller process X with semigroup T_t and initial distribution ν and $X^n \xrightarrow{d} X$.

Proof. For $n \in \mathbb{N}$ let A_n be the generator of $T_{n,t}$.

First we assume that S is compact and that the $T_{n,t}$ and T_t are conservative. TODO: Remove this assumption.

Claim 18.56.1. For all $m \in \mathbb{Z}_+$, $0 = t_0 < t_1 < \cdots < t_m$ and $f_i \in C(S)$ for $i = 0, \ldots, m$ we have

$$\lim_{n\to\infty} \mathbf{E}\left[f_0(X^n_{t_0}) \cdots f_m(X^n_{t_m})\right] = \mathbf{E}\left[f_0(X_{t_0}) \cdots f_m(X_{t_m})\right]$$

To see the claim, note that the case $m = 0$ is simply the statement that $X^n_0 \xrightarrow{d} X_0$ which is an assumption. For $m > 0$, we apply Proposition 18.2

$$\begin{aligned}
\lim_{n\to\infty} \mathbf{E}\left[f_0(X^n_{t_0}) \cdots f_m(X^n_{t_m})\right] &= \lim_{n\to\infty} \mathbf{E}\left[f_0(X^n_{t_0}) \cdots f_{m-1}(X^n_{t_{m-1}})\mathbf{E}\left[f_m(X^n_{t_m}) \mid X^n_{t_{m-1}}\right]\right] \\
&= \lim_{n\to\infty} \mathbf{E}\left[f_0(X^n_{t_0}) \cdots f_{m-1}(X^n_{t_{m-1}}) T_{n,t_m - t_{m-1}} f_m(X^n_{t_{m-1}})\right] \\
&= \lim_{n\to\infty} \mathbf{E}\left[f_0(X^n_{t_0}) \cdots f_{m-1}(X^n_{t_{m-1}}) T_{t_m - t_{m-1}} f_m(X^n_{t_{m-1}})\right] \\
&= \mathbf{E}\left[f_0(X_{t_0}) \cdots f_{m-1}(X_{t_{m-1}}) T_{t_m - t_{m-1}} f_m(X_{t_{m-1}})\right] \\
&= \mathbf{E}\left[f_0(X_{t_0}) \cdots f_{m-1}(X_{t_{m-1}})\mathbf{E}\left[f_m(X_{t_m}) \mid X_{t_{m-1}}\right]\right] \\
&= \mathbf{E}\left[f_0(X_{t_0}) \cdots f_{m-1}(X_{t_{m-1}}) f_m(X_{t_m})\right]
\end{aligned}$$

where in the third line we have the fact that $T_{n,t}f$ converges to $T_t f$ for every $f \in C(S)$ to see

$$\begin{aligned}
&\lim_{n\to\infty} \left|\mathbf{E}\left[f_0(X^n_{t_0}) \cdots f_{m-1}(X^n_{t_{m-1}})(T_{n,t_m - t_{m-1}} - T_{t_m - t_{m-1}}) f_m(X^n_{t_{m-1}})\right]\right| \\
&\leq \|f_0\| \cdots \|f_{m-1}\| \lim_{n\to\infty} \left\|(T_{n,t_m - t_{m-1}} - T_{t_m - t_{m-1}}) f_m\right\| = 0
\end{aligned}$$

and in the fourth line applied the induction hypothesis to the continuous functions $f_0, \ldots, f_{m-2}, f_{m-1} T_{t_m - t_{m-1}} f_m$.

Because S is separable we know that the finite dimensional distributions of X^n converge to those of X. (TODO: prove this; basically we use induction and the standard approximation of indicators by continuous functions to show that $\lim_{n\to\infty} \mathbf{P}\{(X^n_{t_0}, \ldots, X^n_{t_m}) \in A_1 \times \cdots \times A_m\}$ and then use separability of S to conclude that we have a generating π-system and monotone classes apply).

Now we need to show tightness of the family X^n. We appeal to the Aldous criterion so we must show that for every bounded sequence of optional times τ_n and deterministic sequence δ_n such that $\lim_{n\to\infty} \delta_n = 0$ we have $r(X^n_{\tau_n}, X^n_{\tau_n + \delta_n}) \xrightarrow{P} 0$. It suffices to show that every subsequence $N' \subset \mathbb{N}$ has a futher subsequence $N'' \subset N'$ such that $r(X^n_{\tau_n}, X^n_{\tau_n + \delta_n}) \xrightarrow{P} 0$ along N'' (for then we can use Lemma 5.10 to find yet another subsequence $N''' \subset N''$ such that $r(X^n_{\tau_n}, X^n_{\tau_n + \delta_n}) \xrightarrow{a.s.} 0$ along N''' and using Lemma 5.10 in the opposite direction we conclude that $r(X^n_{\tau_n}, X^n_{\tau_n + \delta_n}) \xrightarrow{P} 0$; alternatively just note that convergence in probability is convergence with respect to the Ky Fan metric).

Let $\nu_n = \mathcal{L}(X^n_{\tau_n})$ for $n \in \mathbb{N}$. Then by compactness of S the family ν_n is automatically tight and therefore relatively compact by Prohorov's Theorem 15.20.

Given any subsequence N' then there is a further subsequence N'' and a probability measure ν such that $\nu_n \stackrel{d}{\to} \nu$. In what follows we implicit working along the subsequence N'' and mention it no further.

Let f and g be continuous functions on S then since δ_n converges it is bounded and by Theorem 18.55 and strong continuity of T we have

$$\begin{aligned}\lim_{n\to\infty} \|T_{n,\delta_n} g - g\| &\leq \lim_{n\to\infty} \left[\|T_{n,\delta_n} g - T_{\delta_n} g\| + \|T_{\delta_n} g - g\|\right] \\ &\leq \lim_{n\to\infty} \sup_{0\leq h\leq \sup_n \delta_n} \|T_{n,h} g - T_h g\| + \lim_{n\to\infty} \|T_{\delta_n} g - g\| = 0\end{aligned}$$

and by the Strong Markov Property Theorem 18.49 and Proposition 18.2

$$\begin{aligned}\mathbf{E}\left[f(X^n_{\tau_n}) g(X^n_{\tau_n+\delta_n})\right] &= \mathbf{E}\left[\mathbf{E}\left[f(X^n_{\tau_n}) g(X^n_{\tau_n+\delta_n}) \mid \mathcal{F}_{\tau_n}\right]\right] \\ &= \mathbf{E}\left[\mathbf{E}_{X^n_{\tau_n}}\left[f(X^n_0) g(X^n_{\delta_n})\right]\right] \\ &= \mathbf{E}\left[\mathbf{E}_{X^n_{\tau_n}}\left[f(X^n_0)\mathbf{E}\left[g(X^n_{\delta_n}) \mid X^n_0\right]\right]\right] \\ &= \mathbf{E}\left[\mathbf{E}_{X^n_{\tau_n}}\left[f(X^n_0) T_{n,\delta_n} g(X^n_0)\right]\right] \\ &= \mathbf{E}\left[\mathbf{E}\left[f(X^n_{\tau_n}) T_{n,\delta_n} g(X^n_{\tau_n}) \mid \mathcal{F}_{\tau_n}\right]\right] \\ &= \mathbf{E}\left[f(X^n_{\tau_n}) T_{n,\delta_n} g(X^n_{\tau_n})\right]\end{aligned}$$

Putting these two facts together with the expectation rule Lemma 3.7 and the prior claim we see that

$$\begin{aligned}\lim_{n\to\infty} \mathbf{E}\left[f(X^n_{\tau_n}) g(X^n_{\tau_n+\delta_n})\right] &= \lim_{n\to\infty} \mathbf{E}\left[f(X^n_{\tau_n}) T_{n,\delta_n} g(X^n_{\tau_n})\right] \\ &= \lim_{n\to\infty} \mathbf{E}\left[f(X^n_{\tau_n}) T_{n,\delta_n} g(X^n_{\tau_n}) - f(X^n_{\tau_n}) g(X^n_{\tau_n})\right] + \lim_{n\to\infty} \mathbf{E}\left[f(X^n_{\tau_n}) g(X^n_{\tau_n})\right] \\ &= \lim_{n\to\infty} \int f(s) g(s)\, \nu_n(ds) \\ &= \int f(s) g(s)\, \nu(ds)\end{aligned}$$

Thus by the same argument we used in concluding that fdd's converged we have $\mathcal{L}(X^n_{\tau_n}, X^n_{\tau_n+\delta_n}) \stackrel{w}{\to} \nu$ where we are regarding ν as the measure on $S\times S$ concentrated on the diagonal. By the Continuous Mapping Theorem 5.45 and the fact that ν is concentrated on the diagonal,

$$r(X^n_{\tau_n}, X^n_{\tau_n+\delta_n}) \stackrel{d}{\to} 0$$

and we conclude that $r(X^n_{\tau_n}, X^n_{\tau_n+\delta_n}) \stackrel{P}{\to} 0$ by Lemma 5.33 . Now apply Theorem 17.32. ☐

4.1. Approximating Feller Processes by Markov Chains. Now we turn to the consideration of approximating Feller processes by Markov chains. Underlying the result is a theorem about the approximation of strongly continuous contraction semigroups by powers of contraction operators. Before stating the theorem we call out an estimate that will be required in the proof.

LEMMA 18.57. *Let B be a contraction operator on a Banach space X then for all $v \in X$ and $n \in \mathbb{Z}_+$ we have*

$$\left\| B^n v - e^{n(B-\mathrm{Id})} v \right\| \leq \sqrt{n} \left\| Bv - v \right\|$$

PROOF. First let $n \geq m$ and note that

$$\|B^n v - B^m v\| = \left\| \sum_{k=0}^{n-m-1} B^{m+k}(Bv - v) \right\| \leq \sum_{k=0}^{n-m-1} \left\| B^{m+k} \right\| \|Bv - v\| \leq (n-m) \|Bv - v\|$$

Expanding in power series and using this estimate and Cauchy Schwartz we get

$$\begin{aligned}
\left\| B^n v - e^{n(B-\mathrm{Id})} v \right\| &= e^{-n} \left\| e^n B^n v - e^{nB} v \right\| = e^{-n} \left\| \sum_{k=0}^{\infty} \frac{n^k}{k!} (B^n v - B^k v) \right\| \\
&= e^{-n} \sum_{k=0}^{\infty} \frac{n^k}{k!} |n-k| \|Bv - v\| \\
&\leq e^{-n} \|Bv - v\| \left(\frac{n^k}{k!} (n-k)^2 \right)^{1/2} \left(\frac{n^k}{k!} \right)^{1/2} \\
&= \|Bv - v\| \left(e^{-n} \frac{n^k}{k!} (n-k)^2 \right)^{1/2} = \sqrt{n} \|Bv - v\|
\end{aligned}$$

where in the last line we have used the formula for the variance of a Poisson random variable of rate n (Proposition 13.63) □

THEOREM 18.58. *Let X and $X_1, X_2, \dots$ be Banach spaces, let T_n be a contraction operator on X_n, $\epsilon_n \geq 0$ such that $\lim_{n\to\infty} \epsilon_n = 0$ and let T_t be a strongly continuous contraction semigroup on X with generator A. Let $\pi_n : X \to X_n$ be bounded linear operators and assume that $\sup_n \|\pi_n\| < \infty$. Let D be a core for A and define $A_n = \epsilon_n^{-1}(T_n - \mathrm{Id})$ then the following are equivalent*

(i) *For every $v \in X$ and every $t \geq 0$, $\lim_{n\to\infty} \sup_{0 \leq s \leq t} \left\| T_n^{\lfloor t/\epsilon_n \rfloor} \pi_n v - \pi_n T_s v \right\| = 0$.*
(ii) *For every $v \in X$ and every $t \geq 0$, $\lim_{n\to\infty} \left\| T_n^{\lfloor t/\epsilon_n \rfloor} \pi_n v - \pi_n T_t v \right\| = 0$.*
(iii) *For every $v \in D$ there exists $v_n \in \mathcal{D}(A_n)$ such that $\lim_{n\to\infty} \|v_n - \pi_n v\| = 0$ and $\lim_{n\to\infty} \|A_n v_n - \pi_n A v\| = 0$.*

PROOF. Before we proceed first note that e^{tA_n} is in fact a contraction operator for all $t \geq 0$,

$$\left\| e^{tA_n} \right\| = \left\| e^{t\epsilon_n^{-1}(T_n - \mathrm{Id})} \right\| = e^{-t\epsilon_n^{-1}} \left\| e^{t\epsilon_n^{-1} T_n} \right\| \leq e^{-t\epsilon_n^{-1}} e^{t\epsilon_n^{-1} \|T_n\|} \leq 1$$

Clearly (i) implies (ii). To see that (ii) implies (iii) let $\lambda > 0$ be given, $v \in \mathcal{D}(A)$ and set $w = (\lambda - A)v$ (so that $v = \int_0^\infty e^{-\lambda t} T_t w \, dt$). For $n \in \mathbb{N}$ define

$$v_n = \epsilon_n \sum_{k=0}^{\infty} e^{-\lambda k \epsilon_n} T_n^k \pi_n w \in X_n$$

TODO: We need to show that $\|v_n\|$ is bounded (or maybe we use Tonneli to push the limit inside the sum and don't worry about it); there may be something to this.

From (ii) and Dominated Convergence we have

$$\begin{aligned}
\lim_{n\to\infty} \|v_n - \pi_n v\| &= \lim_{n\to\infty} \left\| v_n - \int_0^\infty e^{-\lambda t} \pi_n T_t w \, dt \right\| \\
&\leq \lim_{n\to\infty} \left\| v_n - \int_0^\infty e^{-\lambda t} T_n^{\lfloor t/\epsilon_n \rfloor} \pi_n v \, dt \right\| \\
&= \lim_{n\to\infty} \left\| v_n - \sum_{k=0}^\infty \int_{k\epsilon_n}^{(k+1)\epsilon_n} e^{-\lambda t} T_n^k \pi_n v \, dt \right\| \\
&= \lim_{n\to\infty} \left\| v_n - () \sum_{k=0}^\infty e^{-\lambda k \epsilon_n} T_n^k \pi_n v \right\| \\
&= \lim_{n\to\infty} (1 - \frac{1 - e^{-\lambda \epsilon_n}}{\lambda}) \|v_n\| = 0
\end{aligned}$$

We also get

$$\begin{aligned}
(\lambda - A_n) v_n &= (\lambda - \epsilon_n^{-1}(T_n - \mathrm{Id})) \epsilon_n \sum_{k=0}^\infty e^{-\lambda k \epsilon_n} T_n^k \pi_n w \\
&= \lambda \epsilon_n \sum_{k=0}^\infty e^{-\lambda k \epsilon_n} T_n^k \pi_n w - \sum_{k=0}^\infty e^{-\lambda k \epsilon_n} T_n^{k+1} \pi_n w + \sum_{k=0}^\infty e^{-\lambda k \epsilon_n} T_n^k \pi_n w \\
&= \lambda \epsilon_n \pi_n w + e^{-\lambda \epsilon_n} \lambda \epsilon_n \sum_{k=0}^\infty e^{-\lambda k \epsilon_n} T_n^{k+1} \pi_n w - \sum_{k=0}^\infty e^{-\lambda k \epsilon_n} T_n^{k+1} \pi_n w + \\
&\pi_n w + e^{-\lambda \epsilon_n} \sum_{k=0}^\infty e^{-\lambda k \epsilon_n} T_n^{k+1} \pi_n w \\
&= \lambda \epsilon_n \pi_n w + \pi_n w + (e^{-\lambda \epsilon_n} \lambda \epsilon_n - 1 + e^{-\lambda \epsilon_n}) \sum_{k=0}^\infty e^{-\lambda k \epsilon_n} T_n^{k+1} \pi_n w
\end{aligned}$$

Now take a difference, a limit and apply Tonelli's Theorem

$$\begin{aligned}
\lim_{n\to\infty} \|(\lambda - A_n) v_n - \pi_n w\| &\leq \lim_{n\to\infty} \|\epsilon_n \lambda \pi_n w\| + \lim_{n\to\infty} \sum_{k=0}^\infty \left| e^{-\lambda \epsilon_n} \lambda \epsilon_n - 1 + e^{-\lambda \epsilon_n} \right| e^{-\lambda k \epsilon_n} \left\| T_n^{k+1} \pi_n w \right\| \\
&\leq \lambda \sup_n \|\pi_n\| \|w\| \lim_{n\to\infty} \epsilon_n + \sup_n \|\pi_n\| \|w\| \sum_{k=0}^\infty \lim_{n\to\infty} \left| e^{-\lambda \epsilon_n} \lambda \epsilon_n - 1 + e^{-\lambda \epsilon_n} \right| e^{-\lambda k \epsilon_n} \\
&= 0
\end{aligned}$$

Therefore

$$\lim_{n\to\infty} \|A_n v_n - \pi_n A v\| \leq \lim_{n\to\infty} (\|(\lambda - A_n) v_n - \pi_n (\lambda - A) v\| + \lambda \|v_n - \pi_n v\|) = 0$$

To see that (iii) implies (i), let $v \in X$ and pick $v_n \in X_n$ such that $\lim_{n\to\infty} \|v_n - \pi_n v\| = 0$ and $\lim_{n\to\infty} \|A_n v_n - \pi_n A v\| = 0$. For $n \in \mathbb{N}$ adding and subtracting terms and using the triangle equality we get

$$\begin{aligned}\left\| T_n^{\lfloor t/\epsilon_n \rfloor} \pi_n v - \pi_n T_t v \right\| \leq & \left\| T_n^{\lfloor t/\epsilon_n \rfloor} \pi_n v - T_n^{\lfloor t/\epsilon_n \rfloor} v_n \right\| + \left\| T_n^{\lfloor t/\epsilon_n \rfloor} v_n - e^{\epsilon_n \lfloor t/\epsilon_n \rfloor A_n} v_n \right\| + \\ & \left\| e^{\epsilon_n \lfloor t/\epsilon_n \rfloor A_n} v_n - e^{\epsilon_n \lfloor t/\epsilon_n \rfloor A_n} \pi_n v \right\| + \left\| e^{\epsilon_n \lfloor t/\epsilon_n \rfloor A_n} \pi_n v - e^{tA_n} \pi_n v \right\| + \\ & \left\| e^{tA_n} \pi_n v - \pi_n T_t v \right\|\end{aligned}$$

We consider each of the terms on the right hand side in sequence. For the first term simply note that since T_n is a contraction operator

$$\limsup_{n \to \infty} \sup_{0 \leq s \leq t} \left\| T_n^{\lfloor s/\epsilon_n \rfloor} \pi_n v - T_n^{\lfloor t/\epsilon_n \rfloor} v_n \right\| \leq \lim_{n \to \infty} \| \pi_n v - v \| = 0$$

For the second term we use Lemma 18.57 to see

$$\begin{aligned}\limsup_{n \to \infty} \sup_{0 \leq s \leq t} \left\| T_n^{\lfloor s/\epsilon_n \rfloor} v_n - e^{\epsilon_n \lfloor s/\epsilon_n \rfloor A_n} v_n \right\| &= \limsup_{n \to \infty} \sup_{0 \leq s \leq t} \left\| T_n^{\lfloor s/\epsilon_n \rfloor} v_n - e^{\lfloor s/\epsilon_n \rfloor (T_n - \mathrm{Id})} v_n \right\| \\ &\leq \limsup_{n \to \infty} \sup_{0 \leq s \leq t} \sqrt{\lfloor s/\epsilon_n \rfloor} \, \| (T_n - \mathrm{Id}) v_n \| \\ &= \limsup_{n \to \infty} \sqrt{\lfloor t/\epsilon_n \rfloor} \epsilon_n \, \| A_n v_n \| \\ &\leq \limsup_{n \to \infty} \sqrt{\lfloor t/\epsilon_n \rfloor} \epsilon_n (\sup_n \| \pi_n \| \, \| A v \| + \| A_n v_n - \pi_n A v \|) \\ &= 0\end{aligned}$$

For the third term using the fact that e^{tA_n} is a contraction operator

$$\limsup_{n \to \infty} \sup_{0 \leq s \leq t} \left\| e^{\epsilon_n \lfloor s/\epsilon_n \rfloor A_n} v_n - e^{\epsilon_n \lfloor s/\epsilon_n \rfloor A_n} \pi_n v \right\| \leq \limsup_{n \to \infty} \| v_n - \pi_n v \| = 0$$

For the fourth term we use the Fundamental Theorem of Calculus, the fact that e^{tA_n} is a contraction operator and the fact that $\| \pi_n A v - A_n \pi_n v \| \to 0$ to see

$$\begin{aligned}\limsup_{n \to \infty} \sup_{0 \leq s \leq t} \left\| e^{\epsilon_n \lfloor s/\epsilon_n \rfloor A_n} \pi_n v - e^{sA_n} \pi_n v \right\| &= \limsup_{n \to \infty} \sup_{0 \leq s \leq t} \left\| \int_{\epsilon_n \lfloor s/\epsilon_n \rfloor}^{s} \frac{d}{du} e^{uA_n} \pi_n v \, du \right\| \\ &= \limsup_{n \to \infty} \sup_{0 \leq s \leq t} \left\| \int_{\epsilon_n \lfloor s/\epsilon_n \rfloor}^{s} e^{uA_n} A_n \pi_n v \, du \right\| \\ &\leq \limsup_{n \to \infty} \epsilon_n \left\| e^{uA_n} \right\| \| A_n \pi_n v \| \\ &\leq \limsup_{n \to \infty} \epsilon_n (\sup_n \| \pi_n \| \, \| A v \| + \| \pi_n A v - A_n \pi_n v \|) = 0\end{aligned}$$

Lastly for the fifth term we see that

$$\limsup_{n \to \infty} \sup_{0 \leq s \leq t} \left\| e^{sA_n} \pi_n v - \pi_n T_s v \right\| = 0$$

from Theorem 18.55 applied to the semigroups e^{tA_n} and T_t. □

TODO: Trotter product formula and Chernoff product formula. The Trotter product formula is inspired by the Lie product formula for matrix groups.

Possible exercise: According to Bratelli (who attributes to Hille) the Trotter product formula implies the Weierstrass approximation theorem (I think Goldstein has this as well). Let X be Banach space of bounded uniformly continuous functions and let $T_t f(x) = f(x+t)$ be the translation semigroup with generator $A = \frac{d}{dx}$. Then let $A_n = (T_{1/n} - 1)/(1/n)$ and the Trotter product formula then shows

$$f(t) = \lim_{n \to \infty} \sum_{m=1}^{\infty} t^m (A_n^m f)(0)/m!$$

where the convergence is uniform on compacts. Apparently this implies the result. Also chase down and understand how Goldstein uses the Chernoff product formula to show the central limit theorem.

CHAPTER 19

Stochastic Approximation

This chapter covers some of the basic results in the theory of stochastic approximation and in doing so provides some applications of discrete time martingale theory and weak convergence theory to optimization problems. The statement of the stochastic approximation problem that one often encounters is so abstract and general that it can be difficult to understand how it could be relevant to any particular problem. Indeed it is common to see stochastic approximation defined as the study of discrete time stochastic processes of the form

$$\theta_{n+1} = \theta_n + \epsilon_n Y_n$$

where Y_n is a random vector.

To motivate the form of the problem statement, let us tie this into the problem of optimization specifically gradient descent. Given a function f we have a globally convergent algorithm for minimization given by $x_{n+1} = x_n - \alpha_n \nabla f(x_n)$ where α_n is a sequence of real numbers that satisfies Armijo conditions. Now suppose that we don't have the ability to measure $-\nabla f(x_n)$ exactly but that we have some noise corrupted version thereof. If we call the observed approximate gradient Y_n, then the gradient descent algorithm has the form of a stochastic approximation problem and we can ask whether we still have convergence in a appropriate stochastic sense (e.g. almost sure). In line with this specific case, we often think of the process Y_n as being a sequence of observations and though it doesn't have any real mathematical meaning, we shall use the terminology in what follows.

As we've mentioned in our discussion of optimization, in practice constrained optimization is at least as important as unconstrained optimization and therefore we should look for how to incorporate constraints into stochastic approximation. The way we shall do this at this point is to assume that the sequence θ_n is constrained to lie in some closed set F and to maintain the constraint at each iteration by a brute force projection (say in L^2 norm) onto the set F. Thus in the constrained case we are considering a stochastic process

$$\theta_{n+1} = \Pi_F \left[\theta_n + \epsilon_n Y_n\right]$$

where Y_n is a random vector and Π_F represents projection onto F. It is common to define the projection correction term $Z_n = \epsilon_n^{-1}\{\Pi_F \left[\theta_n + \epsilon_n Y_n\right] - \theta_n - \epsilon_n Y_n\}$ so that we may write

$$\theta_{n+1} = \theta_n + \epsilon_n Y_n + \epsilon_n Z_n$$

In order to discuss the hypotheses that one might need to make on the stochastic process Y_n, it is convenient to assume a structural form for Y_n. Let $\mathcal{F}_n = \sigma(\theta_0, Y_j; j < n)$ be a filtration $\mathcal{F}$. For our first results we shall assume that there exists functions g_n, an $\mathcal{F}$-martingale difference sequence δM_n and a stochastic process β_n such that $Y_n = g_n(\theta_n) + \delta M_n + \beta_n$. The reader should think of these terms in the following way. The term $g_n(\theta_n)$ represents the mean/true value of the process (e.g. the value of the gradient in the steepest descent case), the term δM_n represents a noise term and β_n represents a bias term in the observation. The reason why the bias term β_n is called out as being different from $g_n(\theta_n)$ is that we shall be assuming that it becomes asymptotically small.

One of the key techniques in proving theorems in stochastic approximation is the ODE method. The idea is that one can view the process θ_n as a discretization of an ordinary differential equation that is described by the conditional means of Y_n.

In many of the proofs we will considering the continuous time limits of discrete time processes. To do this we be making interpolations of the discrete time processes and want to have recourse to compactness results that will give conditions under which limits exist. A natural tool for this would be to use the Arzela-Ascoli Theorem for continuous functions or the Skorohod topology versions of that for cadlag functions. As it turns out neither of these is the exact fit for what we do since we'll be considering cadlag processes that converge to continuous functions and we want uniform convergence on compact sets. So what we want is a slight extension of Arzela-Ascoli Theorem.

The following is a version of the Arzela Ascoli Theorem 19.1 that gives a sufficent criteria for a sequence of possibly discontinuous functions to have a continuous limit. The referenced version of the Arzela Ascoli Theorem doesn't apply in a non-trivial way since the criterion for equicontinuity $\lim_{\delta\to 0}\sup_{f\in A} m(T,f,\delta) = 0$ implies that every f is continuous. However, note that if f_n is a sequence of functions in $C([0,\infty);\mathbb{R}^d)$ then $\{f_n\}$ is equicontinuous if and only if $\lim_{\delta\to 0}\limsup_{n\to\infty} m(T,f_n,\delta) = 0$. This criterion does not imply that each f_n is continuous and turns out to be a useful extension of equicontinuity for sequences of non-continuous functions.

THEOREM 19.1 (Extended Arzela-Ascoli Theorem). *Let $f_n : [0,\infty) \to \mathbb{R}^d$ be measurable functions such that*

(i) $\sup_n |f_n(0)| < \infty$

(ii) $\lim_{\delta\to 0}\limsup_{n\to\infty} m(T,f_n,\delta) = 0$ *for all* $T > 0$.

then there exists $f \in C([0,\infty),\mathbb{R}^d)$ such that f_n converges to f uniformly on compact sets.

PROOF. First note that if (i) and (ii) hold for the sequence f_n then the conditions also hold for any subsequence of f_n. Suppose that f_n satisfy (i) and (ii) and let $T > 0$. By (ii) there exists a $\delta > 0$ such that

$$\limsup_{n\to\infty} m(T,f_n,\delta) < 1$$

hence there exists an $N \in \mathbb{N}$ such that $m(T,f_n,\delta) < 1$ for all $n \geq N$. Now pick $m \in \mathbb{N}$ such that $m\delta < T \leq (m+1)\delta$ and just as in Theorem 15.25 by considering the grid $0, \delta, 2\delta, \dots, m\delta, T$ we can write the telescoping sum

$$f_n(T) - f_n(0) = f_n(T) - f_n(m\delta) + \sum_{k=1}^{m} f_n(k\delta) - f_n((k-1)\delta)$$

and use the triangle inequality to conclude that $|f_n(T)| \leq |f_n(0)| + m + 1$ for every $n \geq N$. Coupled with (i) this shows that $\sup_{n\geq N}|f_n(T)| < \infty$. By local compactness of $\mathbb{R}$ we see that $f_n(T)$ converges along a subsequence of $\{n, n+1, \dots\}$. Using the observation that f_n along this subsequence still satisfies (i) and (ii) we see that we can enumerate $T \in \mathbb{Q}_+$ and use induction and a diagonal subsequence argument to get a single subsequence of f_n that converges for all $T \in \mathbb{Q}_+$. Define $f(T)$ for $T \in \mathbb{Q}_+$ as the limit of this subsequence of f_n.

The proof that $f \in C([0,\infty);\mathbb{R}^d)$ and that f_n converges to f uniformly on compact sets is almost exactly the same as the proof in Theorem 15.25. The only difference is that the condition (ii) applied the sequence f_n only constrains terms $|f_n(s) - f_n(t)|$ for n sufficiently large. Examining the proof of Theorem 15.25 one will see that this is all that is required. $\square$

We also need a partial converse, namely that a convergence sequence of functions is equicontinuous in the extended sense.

PROPOSITION 19.2. *Let $f_n : \mathbb{R} \to \mathbb{R}^d$ be a sequence of measurable functions that converges to a continuous function with convergence uniform on compact sets, then f_n is equicontinuous in the extended sense.*

PROOF. Let f be the limit of f_n. Pick an N such that $\|f_n(0) - f(0)\| < 1$ for all $n \geq N$ then it follows that $\|f_n(0)\| \leq \|f_0(0)\| \vee \cdots \vee \|f_{N-1}(0)\| \vee \|f_N(0)\| + 1$ for all $n \in \mathbb{N}$. Now let $T > 0$ and $\epsilon > 0$ be given and use the fact that f is uniformly continuous on $[-T,T]$ to pick a $\delta > 0$ such that $m(T,f,\delta) < \epsilon/3$. Now by uniform convergence of f_n to f on $[-T,T]$ we pick $N > 0$ such that $\sup_{-T \leq t \leq T} \|f_n(t) - f(t)\| < \epsilon/3$ for $n \geq N$. Thus for $n \geq N$,

$$\begin{aligned} m(T,f_n,\delta) &= \sup_{\substack{|s-t|<\delta \\ -T \leq s,t \leq T}} \|f_n(s) - f_n(t)\| \\ &\leq \sup_{\substack{|s-t|<\delta \\ -T \leq s,t \leq T}} \|f_n(s) - f(s)\| + \|f(s) - f(t)\| + \|f_n(t) - f(t)\| < \epsilon \end{aligned}$$

and it follows that $\limsup_{n\to\infty} m(T,f_n,\delta) < \epsilon$ and thus $\lim_{\delta\to 0}\limsup_{n\to\infty} m(T,f_n,\delta) = 0$. $\square$

We shall now assume that we are in the situation of having a constraint set F defined by continuously differentiable function $c_i(x)$ which satisfy the LICQ.

TODO: Use the KKT conditions applied to $\min_{x\in F} \|x - (\theta_n + \epsilon_n Y_n)\|^2$ to show that Z_n is in the normal cone

THEOREM 19.3. *Suppose we are given a process Y_n, a constraint set F, a random variable θ_0 and a deterministic sequence ϵ_n. Define the process*

$$\theta_{n+1} = \Pi_F\left[\theta_n + \epsilon_n Y_n\right]$$

and suppose that there are measurable functions $g_n(\theta)$ such that if we write $\boldsymbol{E}[Y_n \mid \theta_0, Y_i; 0 \leq i \leq n-1] = g_n(\theta_n) + \beta_n$ such that

(i) $\sup_n \boldsymbol{E}\left[Y_n^2\right] < \infty$
(ii) *ϵ_n for $n \in \mathbb{Z}$ is a sequence with $\epsilon_n = 0$ for $n < 0$, $\epsilon_n \geq 0$ for $n \geq 0$, $\lim_{n\to\infty}\epsilon_n = 0$, $\sum_{n=0}^\infty \epsilon_n = \infty$ and $\sum_{n=0}^\infty \epsilon_n^2 < \infty$.*
(iii) *Suppose the $g_n(\theta)$ are uniformly continuous in n and there is a continuous function $\overline{g}(\theta)$ such that for each $\theta \in F$ we have*

$$\lim_{n\to\infty}\left|\sum_{i=n}^{m(t_n+t)} \epsilon_i\{g_i(\theta) - \overline{g}(\theta)\}\right| = 0$$

(iv) $\beta_n \overset{a.s.}{\to} 0$

Then there is a set A of probability zero such that for $\omega \notin A$ the set of functions $\{\theta^n(\omega,\cdot), Z^n(\omega,\cdot); n < \infty\}$ is equicontinuous. If $(\theta(\omega,\cdot), Z(\omega,\cdot))$ is the limit of some convergent subsequence then the pair satisfies the projected ODE

$$\dot{\theta} = \overline{g}(\theta) + z,\ z \in \mathcal{N}(\theta)$$

and $\theta_n(\omega)$ converges to a limit set of the projected ODE in F.

Proof. Let $\mathcal{F}_n = \sigma(\theta_0, Y_i; 0 \leq i \leq n)$ be the filtration defined by Y_n and the initial condition θ_0. We write

$$\begin{aligned}
\theta_{n+1} &= \Pi_F\left[\theta_n + \epsilon_n Y_n\right] \\
&= \theta_n + \epsilon_n Y_n + \epsilon_n Z_n \\
&= \theta_n + \epsilon_n \mathbf{E}\left[Y_n \mid \mathcal{F}_{n-1}\right] + \epsilon_n\left(Y_n - \mathbf{E}\left[Y_n \mid \mathcal{F}_{n-1}\right]\right) + \epsilon_n Z_n \\
&= \theta_n + \epsilon_n g_n(\theta_n) + \epsilon_n \beta_n + \epsilon_n \delta M_n + \epsilon_n Z_n
\end{aligned}$$

where we have defined $\delta M_n = Y_n - \mathbf{E}\left[Y_n \mid \mathcal{F}_{n-1}\right]$. Note that $\epsilon_n \delta M_n$ is an $\mathcal{F}$-martingale difference sequence and therefore by Proposition 9.58 the process $M_n = \sum_{j=0}^n \epsilon_j \delta M_j$ is an $\mathcal{F}$-martingale. Furthermore by Jensen's Inequality for conditional expectations (Theorem 8.36) and the fact that Y_n is L^2-bounded we also know that δM_n is an L^2 martingale difference sequence hence by Proposition 9.59 we know that $\mathbf{E}\left[\delta M_n \delta M_m\right] = 0$. For every fixed $m \in \mathbb{Z}_+$ we know that the process $(M_{n+m} - M_m)^2$ is a submartingale with respect to the shifted filtration $\tilde{\mathcal{F}}_n = \mathcal{F}_{n+m}$ and if we apply Doob's Maximal Inequality (Lemma 9.44) we get for every $\lambda > 0$ and $m < n$,

$$\begin{aligned}
\mathbf{P}\{\sup_{m \leq j \leq n} |M_j - M_m| \geq \lambda\} &= \mathbf{P}\{\sup_{m \leq j \leq n} (M_j - M_m)^2 \geq \lambda^2\} \\
&\leq \lambda^{-2} \mathbf{E}\left[(M_n - M_m)^2\right] \\
&= \lambda^{-2} \sum_{i=m+1}^n \sum_{j=m+1}^n \epsilon_i \epsilon_j \mathbf{E}\left[\delta M_i \delta M_j\right] \\
&= \lambda^{-2} \sum_{j=m+1}^n \epsilon_j^2 \mathbf{E}\left[\delta M_j^2\right] \\
&\leq 2\lambda^{-2} \sup_n \mathbf{E}\left[Y_n^2\right] \sum_{j=m+1}^\infty \epsilon_j^2
\end{aligned}$$

By continuity of measure, we can let $n \to \infty$ and then $m \to \infty$ and use the hypothesis that $\sum_{n=0}^\infty \epsilon_n^2 < \infty$ to conclude that for every $\lambda > 0$ we have

$$\lim_{m \to \infty} \mathbf{P}\{\sup_{m \leq j} |M_j - M_m| \geq \lambda\} = 0 \tag{36}$$

TODO: Could we have just appealed to an off the shelf Martingale Convergence theorem here; not sure this is a interesting kind of convergence because are looking at a subsequence that starts at a point that goes to infinity???? We have just proven that $\sup_{m \leq j} |M_j - M_m| \xrightarrow{P} 0$.

Now we move to the interpolated process. Recall that we define $t_0 = 0$ and $t_n = \sum_{i=0}^{n-1} \epsilon_i$ for $n \in \mathbb{N}$. We define $m(t) = n$ for $t_n \leq t < t_{n-1}$. Using $m(t)$ we define the interpolated processes for $t \geq 0$,

$$M^n(t) = \sum_{i=n}^{m(t_n+t)-1} \epsilon_i \delta M_i \quad B^n(t) = \sum_{i=n}^{m(t_n+t)-1} \epsilon_i \beta_i \quad Z^n(t) = \sum_{i=n}^{m(t_n+t)-1} \epsilon_i Z_i$$

and for $t < 0$,

$$M^n(t) = -\sum_{i=m(t_n+t)}^{n-1} \epsilon_i \delta M_i \quad B^n(t) = -\sum_{i=m(t_n+t)}^{n-1} \epsilon_i \beta_i \quad Z^n(t) = -\sum_{i=m(t_n+t)}^{n-1} \epsilon_i Z_i$$

and note that $M^n(t) = M^0(t_n + t) - M^0(t_n)$ and similarly with B^n and Z^n. Moreover $M^0(t) = M_{m(t)-1}$. Furthermore we define

$$\overline{G}^n(t) = \sum_{i=n}^{m(t_n+t)-1} \epsilon_i \overline{g}(\theta_n) \qquad \tilde{G}^n(t) = \sum_{i=n}^{m(t_n+t)-1} \epsilon_i \left(g_n(\theta_n) - \overline{g}(\theta) \right)$$

so that we have

$$\theta^n(t) = \theta_n + \overline{G}^n(t) + \tilde{G}^n(t) + M^n(t) + Z^n(t) + B^n(t)$$

Claim 19.3.1. Almost surely for all $T > 0$, $\lim_{n\to\infty} \sup_{-T \leq t \leq T} M^n(t) = 0$.

Let $T > 0$ be given. By the definition of M^n and the triangle inequality we get for every $n \in \mathbb{N}$ and $m < m(t_n - T)$ we have

$$\begin{aligned}
\sup_{-T \leq t \leq T} |M^n(t)| &= \sup_{-T \leq t \leq T} \left| M^0(t_n + t) - M^0(t_n) \right| \\
&= 2 \sup_{-T \leq t \leq T} \left| M^0(t_n + t) - M^0(t_n - T) \right| \\
&\leq 2 \sup_{m(t_n - T) - 1 \leq j} \left| M_j - M_{m(t_n - T)-1} \right| \\
&\leq 4 \sup_{m \leq j} |M_j - M_m|
\end{aligned}$$

If $\lim_{n\to\infty} \sup_{-T \leq t \leq T} M^n(t)! = 0$ then there is a $\lambda > 0$ and a subsequence $n_j$such that $\sup_{-T \leq t \leq T} \left|M^{n_j}(t)\right| \geq \lambda$ for all $j \in \mathbb{N}$. Since we know that $\lim_{j\to\infty} m(t_{n_j} - T) = \infty$, we know $\sup_{m \leq j} |M_j - M_m| \geq \lambda/4$ for all m and therefore the claim follows from Equation (36).

Claim 19.3.2. Almost surely for all $T > 0$, $\lim_{n\to\infty} \sup_{-T \leq t \leq T} B^n(t) = 0$.

We actually want this under a couple of different hypotheses: $\beta_n \overset{a.s.}{\to} 0$ and $\sum_{n=0}^{\infty} \epsilon_n |\beta_n| < \infty$ a.s. In the latter case

$$\sup_{-T \leq t \leq T} |B^n(t)| \leq \sum_{i=m(t_n-T)}^{m(t_n+T)-1} \epsilon_i |\beta_i| \leq \sum_{i=m(t_n-T)}^{\infty} \epsilon_i |\beta_i|$$

so the result follows from the fact that $\lim_{n\to\infty} m(t_n - T) = \infty$. TODO: What about the former case?

Claim 19.3.3. $\theta_{n+1} - \theta_n \overset{a.s.}{\to} 0$.

Using the Markov Inequality, $\sup_n \mathbf{E}\left[Y_n^2\right] < \infty$ and $\sum_{n=0}^\infty \epsilon_n^2 < \infty$ we get for any $\lambda > 0$,

$$\begin{aligned}\sum_{n=0}^\infty \mathbf{P}\{\epsilon_n \left|Y_n\right| \geq \lambda\} &\leq \sum_{n=0}^\infty \frac{\epsilon_n^2 \mathbf{E}\left[Y_n^2\right]}{\lambda^2} \\ &\leq \frac{\sup_n \mathbf{E}\left[Y_n^2\right]}{\lambda^2} \sum_{n=0}^\infty \epsilon_n^2 < \infty\end{aligned}$$

and therefore the Borel Cantelli Theorem 4.23 implies that $\mathbf{P}\{\epsilon_n \left|Y_n\right| \geq \lambda \text{ i.o.}\} = 0$ and therefore by Lemma 5.4 we conclude that $\epsilon_n \left|Y_n\right| \overset{a.s.}{\to} 0$. Thus the definition of Π_F and the fact that $\theta_n \in F$, we see that (we are using the argument that $\Pi_F(\theta_n + \epsilon_n Y_n) = \arg\min_{x \in F} \left|\theta_n + \epsilon_n Y_n - x\right|$ hence since $\theta_n \in F$,

$$\left|\theta_n + \epsilon_n Y_n - \Pi_F(\theta_n + \epsilon_n Y_n)\right| \leq \left|\theta_n + \epsilon_n Y_n - \theta_n\right| = \epsilon_n \left|Y_n\right|$$

is this always true or does it require some assumption like prox-regularity????)

$$\begin{aligned}\lim_{n\to\infty} \left|\theta_{n+1} - \theta_n\right| &= \lim_{n\to\infty} \left|\Pi_F\left[\theta_n + \epsilon_n Y_n\right] - \theta_n\right| \\ &\leq \lim_{n\to\infty} \left\{\left|\Pi_F\left[\theta_n + \epsilon_n Y_n\right] - \theta_n - \epsilon_n Y_n\right| + \epsilon_n \left|Y_n\right|\right\} \\ &\leq 2 \lim_{n\to\infty} \epsilon_n \left|Y_n\right| = 0\end{aligned}$$

almost surely.

By prior claims and Proposition 19.2 we know that almost surely, $M^n(t)$ and $B^n(t)$ are each equicontinuous in the extended sense.

Claim 19.3.4. Z^n is almost surely equicontinuous in the extended sense.

Here is the hyperrectangle case. We work pathwise so lets assume that $\omega \in \Omega$ is fixed.

Claim 19.3.5. If $Z^n(\omega)$ is not equicontinuous in the extended sense then there is an $\epsilon > 0$, a sequence $n_k \in \mathbb{N}$ with $\lim_{k\to\infty} n_k = \infty$ and a sequence $\delta_k > 0$ with $\lim_{k\to\infty} \delta_k = 0$ such that $\left|Z^{n_k}(\omega, \delta_k)\right| \geq \epsilon$.

If Z^n is not equicontinuous in the extended sense then there is a $T > 0$ such that $\lim_{\delta\to 0} \limsup_{n\to\infty} m(T, Z^n, \delta) > 0$ i.e. an $\epsilon > 0$, a sequence δ_k with $\delta_k \to 0$, a sequence n_k with $n_k \to \infty$ and s_k, u_k with $-T \leq s_k < u_k \leq T$ and $u_k - s_k < \delta_k$ such that $\left|Z^{n_k}(u_k) - Z^{n_k}(s_k)\right| \geq \epsilon$. Recalling $Z^n(t) = Z^0(t_n + t) - Z^0(t_n)$ we see that $\left|Z^{n_k}(u_k) - Z^{n_k}(s_k)\right| \geq \epsilon$ is equivalent to $\left|Z^0(t_{n_k} + u_k) - Z^0(t_{n_k} + s_k)\right| \geq \epsilon$ and therefore if we define m_k such that . By redefining the sequence n_k to be $\tilde{n}_k = m(t_{n_k} + s_k)$, $\tilde{s}_k = s_k + t_{n_k} - t_{\tilde{n}_k}$ and $\tilde{u}_k = u_k + t_{n_k} - t_{\tilde{n}_k}$ we get $\left|Z^{\tilde{n}_k}(\tilde{u}_k) - Z^{\tilde{n}_k}(\tilde{s}_k)\right| \geq \epsilon$ where by definition $t_{\tilde{n}_k} \leq s_k < t_{\tilde{n}_k+1} = t_{\tilde{n}_k} + \epsilon_{\tilde{n}_k+1}$ and thus $0 \leq \tilde{s}_k < \epsilon_{\tilde{n}_k+1}$ but still

$$\lim_{k\to\infty} \tilde{n}_k = \lim_{k\to\infty} m(t_{\tilde{n}_k}) \geq \lim_{k\to\infty} m(t_{n_k} - T) = \infty$$

from which it also follows that $\lim_{k\to\infty} \tilde{s}_k = 0$ and therefore $\lim_{k\to\infty} \tilde{u}_k \leq \lim_{k\to\infty} \tilde{s}_k + \delta_k = 0$. Since $0 \leq \tilde{s}_k < \epsilon_{\tilde{n}_k+1}$ we also have $Z^{\tilde{n}_k}(\tilde{s}_k) = Z^{\tilde{n}_k}(0) = 0$ so the sequences $\tilde{n}_k$ and $\tilde{u}_k$ satisfy the claim.

TODO: Note that δ_k doesn't necessarily go to 0 faster than $\epsilon_{n_k} + \epsilon_{n_k+1}$ so $Z^{n_k}(\delta_k)$ may be a sum of multiple jumps $\epsilon_{n_k} Z_{n_k} + \cdots + \epsilon_{n_k+m_k} Z_{n_k+m_k}$. This makes the geometry a tad confusing for me. Can we reduce to a case in which $0 \leq \delta_k + \epsilon_{n_k} < \epsilon_{n_k+1}$?

TODO: I still don't see the geometry here. Relevant facts:

- $\theta_{n+1} \in \text{int}(F)$ implies $Z_n = 0$ (this follows from the next item and the fact that $N_F(\theta_{n+1}) = \{0\}$ if $\theta_{n+1} \in \text{int}(F)$
- $Z_n \in N_F(\theta_{n+1})$
- $\epsilon_n Z_n \overset{a.s.}{\to} 0$
- $M^n(t) \overset{a.s.}{\to} 0$ uniformly on compacts
- $B^n(t) \overset{a.s.}{\to} 0$ uniformly on compacts

In prose, a asymptotic jump in Z^{n_k} cannot be into the interior because that would imply that $Z_{n_k} = 0$. The claim is that in the limit, the jump therefore must be from the boundary of F (I don't see this yet but I suspect it follows from $\theta_{n+1} - \theta_n \overset{a.s.}{\to} 0$) to another point on the boundary of F (I get this follows from $Z_n \in N_F(\theta_{n+1})$) and that this contradicts the fact that $Z_n \in N_F(\theta_{n+1})$ (I don't see this yet). The basic intuition is this:

- asymptotically since $\epsilon_n Z_n \to 0$ a.s. we know that any jump $Z^n(0)$ in the limit must be from the boundary of F
- jumps in Z_n are always to the boundary and not to the interior thus in the limit this is true and therefore the asymptotic jump is from a boundary point to another boundary point
- since $Z_n \in N_F(\theta_{n+1})$ and the normal cone points *into* F it is impossible for the jump to be between boundary points

So the relevant geometry here just precludes some kind of limit of Z_n pointing out of the normal cone (this seems like it will be true with regularity assumptions for then we know that any limit of proximal normals is a limiting normal but with appropriate regularity assumptions we know that proximal normals are the same thing as limiting normals).

TODO: Finish ☐

1. Dynamical Systems Approach

This section follows Benaim's notes (which in turn summarize a bunch of the Benaim and Hirsch work).

DEFINITION 19.4. Let (S, r) be a metric space a continuous map $\Phi : [0, \infty) \times S \to S$. We write $\Phi_t(x) = \Phi(t, x)$ for $0 \leq t < \infty$ and $\Phi^x(t) = \Phi(t, x)$ for $x \in S$. Φ is said to be a *semiflow* if

(i) $\Phi_0 = \text{Id}$
(ii) $\Phi_s \circ \Phi_t = \Phi_{t+s}$ for all $0 \leq t, s < \infty$.

A continuous map $\Phi : (-\infty, \infty) \times S \to S$ with properties (i) and (ii) is called a *flow*. Given $x \in S$ we often refer to Φ^x as an *orbit*.

DEFINITION 19.5. Let (S, r) be a metric space and let Φ be a semiflow on S, then $X \in C([0, \infty); S)$ is said to be an *asymptotic pseudotrajectory* of Φ if for all $T > 0$ we have

$$\lim_{t \to \infty} \sup_{0 \leq h \leq T} r(X(t + h), \Phi_h(X(t))) = 0$$

If the image $X([0, \infty))$ has compact closure we often say that X is *precompact*.

Our first goal is formulate some conditions that are equivalent to an X being an asymptotic pseudotrajectory. The idea is to investigate the orbits of X under a canonical flow on the space $C((-\infty,\infty);S)$. The first step is to define the *translation flow* on $C((-\infty,\infty);S)$. Recall that $C((-\infty,\infty);S)$ is a metric space with metric

$$d(f,g) = \sum_{n=1}^{\infty} 2^{-n} \left(\sup_{-n \leq t \leq n} r(f(t), g(t)) \wedge 1 \right)$$

PROPOSITION 19.6. *Let (S,r) be a metric space define $\Theta : (-\infty,\infty) \times C((-\infty,\infty);S) \to C(-\infty,\infty);S)$ by $\Theta(t,f)(s) = f(t+s)$, then Θ is a flow.*

PROOF. The only non-trivial part is the proof that Θ is continuous. Suppose that $(t,f),(t_n,f_n) \in (-\infty,\infty) \times C((-\infty,\infty);S)$ and $\lim_{n\to\infty}(t_n,f_n) = (t,f)$. Let $\epsilon > 0$ and $T > 0$ be given. f is uniformly continuous on $[-T-|t|-1, T+|t|+1]$ so we may pick a $\delta > 0$ such that

$$\sup_{\substack{|u|,|v| \leq T+|t|+1 \\ |u-v|<\delta}} r(f(u), f(v)) < \epsilon/2$$

Since $\lim_{n\to\infty} t_n = t$ and $\lim_{n\to\infty} f_n = f$ we may pick $N > 0$ such that $|t_n - t| < \delta \wedge 1$ and $\sup_{|s| \leq T+|t|+1} r(f_n(s), f(s)) < \epsilon/2$ for all $n \geq N$ Therefore for all $n \geq N$

$$\begin{aligned}
\sup_{-T\leq s\leq T} r(\Theta(t_n,f_n)(s), \Theta(t,f)(s)) &= \sup_{-T\leq s\leq T} r(f_n(t_n+s), f(t+s)) \\
&\leq \sup_{-T\leq s\leq T} r(f_n(t_n+s), f(t_n+s)) + \sup_{-T\leq s\leq T} r(f(t_n+s), f(t+s)) \\
&\leq \sup_{|s|\leq T+|t|+1} r(f_n(s), f(s)) + \sup_{|s|\leq T+|t|+1} r(f(s), f(s)) < \epsilon
\end{aligned}$$

∎

TODO: Is there a translation flow on $D((-\infty,\infty);S)$ in the uniform topology? in the Skorohod topology?

If we are given a semiflow $\Phi : [0,\infty) \times S \to S$ it is often convenient to consider it as a flow $\Phi : (-\infty,\infty) \times S \to S$ by defining $\Phi(-t,x) = \Phi(0,x) = x$. TODO: Exercise to show this is a flow (specifically continuity)

In order to compare the flow Φ with the flow Θ we use the following

PROPOSITION 19.7. *Let (S,r) be a metric space and let Φ be a semiflow or flow on S. If we define $H(x) = \Phi^x \in C((-\infty,\infty);S)$ and $S_\Phi = \Re(H)$ then H is a homeomorphism of S and S_Φ and moreover*

$$\Theta_t(H(x)) = H(\Phi_t(x))$$

where we assume that $t \geq 0$ if Φ is a semiflow and $-\infty < t < \infty$ is Φ is a flow. Therefore S_Φ is a closed subset of $C((-\infty,\infty);S)$ and is invariant under Θ. The map $\hat{\Phi} : C((-\infty,\infty);S) \to S_\Phi$ defined by $\hat{\Phi}(X) = H(X(0)) = \Phi^{X(0)}$ is continuous retraction.

PROOF. TODO ∎

We can now reformulate the definition of asymptotic pseudotrajectory in terms of behavior under the translation flow.

LEMMA 19.8. *Let (S, r) be a metric space, $X \in C([0, \infty); S)$ and Φ a semiflow on S then X is an asymptotic pseudotrajectory of Φ if and only if*

$$\lim_{t \to \infty} d(\Theta_t(X), \hat{\Phi}(\Theta_t(X))) = 0$$

PROOF. By Lemma 15.22 we know that $\lim_{t \to \infty} d(\Theta_t(X), \hat{\Phi}(\Theta_t(X))) = 0$ if and only if $\lim_{t \to \infty} \sup_{0 \leq s \leq T} r(\Theta_t(X)(s), \hat{\Phi}(\Theta_t(X))(s))$ for every $T > 0$. Substituting definitions of Θ and $\hat{\Phi}$ we see that this is equivalent to $\lim_{t \to \infty} \sup_{0 \leq s \leq T} r(X(t + s), \Phi_s(X_t))$ for every $T > 0$. □

THEOREM 19.9. *Let (S, r) be a metric space, $X \in C((-\infty, \infty); S)$ such that $\Re(X)$ has compact closure in S and $\Phi : [0, \infty) \times S \to S$ be a semiflow. The X is an asymptotic pseudotrajectory of Φ is and only if X is uniformly continuous and every limit point of $\Theta_t(X)$ is in S_Φ. In either case $\{\Theta_t(X)\}$ is relatively compact in $C((-\infty, \infty); S)$.*

PROOF. To see (i) implies (ii) first assume that X is an asymptotic pseudotrajectory; for every $T > 0$ we have $\lim_{t \to \infty} \sup_{0 \leq s \leq T} r(X_{t+s}, \Phi_s(X_t)) = 0$. Let K be the closure of $\Re(X)$ so that K is compact in S.

CLAIM 19.9.1. Let $\epsilon > 0$ be given and then there exists $\delta > 0$ such that $r(\Phi_t(y), y) < \epsilon$ for all $|t| < \delta$

Let $x \in S$ be arbitrary. Consider

$$\Phi^{-1}(B(x, \epsilon/2)) = \{(t, y) \mid r(\Phi_t(y), x) < \epsilon/2\}$$

Since Φ_0 is identity, we know that $(0, x) \in \Phi^{-1}(B(x, \epsilon/2))$ and therefore we can find a $0 < \delta_x < \epsilon/2$ such that $B((0, x), \delta_x) \subset \Phi^{-1}(B(x, \epsilon/2))$. Moreover we assume are using the metric $|\cdot| \vee r(\cdot, \cdot)$ on $[0, \infty) \times S$ so that $B((0, x), \delta_x) = (-\delta_x, \delta_x) \times B(x, \delta_x)$ From these two facts it follows that for all $y \in B(x, \delta_x)$ then for all $|t| < \delta_x$

$$r(\Phi_t(y), y) \leq r(\Phi_t(y), x) + r(x, y) < \epsilon$$

Now the set of $B(x, \delta_x)$ with $x \in K$ is an open cover of K hence there is a finite subcover $B(x_1, \delta_{x_1}), \ldots, B(x_n, \delta_{x_n})$. Let $\delta = \delta_{x_1} \wedge \cdots \wedge \delta_{x_n}$. Every $y \in K$ belongs to some $B(x_j, \delta_{x_j})$ and therefore $r(\Phi_t(y), y) < \epsilon$ for $|t| < \delta_{x_j}$ an a fortiori for $|t| < \delta$.

CLAIM 19.9.2. X is uniformly continuous

Let $\epsilon > 0$ be given and pick $\delta > 0$ as in the previous claim. Because X is an asymptotic pseudotrajectory we may pick a $t_0 > 0$ such that for all $t \geq t_0$ we have $\sup_{0 \leq h \leq \delta} r(X_{t+h}, \Phi_h(X_t)) < \epsilon$. Thus

$$\sup_{0 \leq h \leq \delta} r(X_{t+h}, X_t) \leq \sup_{0 \leq h \leq \delta} r(X_{t+h}, \Phi_h(X_t)) + \sup_{0 \leq h \leq \delta} r(\Phi_h(X_t), X_t) < 2\epsilon$$

and uniformly continuity follows.

Suppose that Y is a limit point of $\Theta_t(X)$ and $Y \notin S_\Phi$. Since S_Φ is closed it follows that $d(Y, S_\Phi) > 0$. Let n_k be a subsequence such that $\lim_{k \to \infty} \Theta_{n_k}(X) = Y$ and observe

$$\lim_{k \to \infty} d(\Theta_{t_k}(X), \hat{\Phi}(\Theta_{t_k}(X))) \geq \lim_{k \to \infty} \{d(Y, \hat{\Phi}(\Theta_{t_k}(X))) - d(\Theta_{t_k}(X), Y)\} \geq d(Y, S_\Phi) > 0$$

which is a contradiction.

Now to see that (ii) implies (i) suppose that X is uniformly continuous and all the limit points of $\Theta_t(X)$ are in S_Φ. First we show that (ii) implies (iii).

CLAIM 19.9.3. The family $\{\Theta_t(X)\}$ is relatively compact.

First we establish equicontinuity of $\{\Theta_t(X)\}$. Let $T > 0$ and $\epsilon > 0$ be given. By uniform continuity there exists $\delta > 0$ such that $r(X(t), X(s)) < \epsilon$ for all $|t - s| < \delta$. Therefore

$$\sup_{\substack{-T\leq u<v\leq T \\ v-u<\delta}} r(\Theta_t(X)(u), \Theta_t(X)(v)) = \sup_{\substack{-T\leq u<v\leq T \\ v-u<\delta}} r(X(t+u), X(t+v)) \leq \sup_{\substack{-\infty<u<v<\infty \\ v-u<\delta}} r(X(u), X(v)) < \epsilon$$

Since $\Re(X)$ is relatively compact it follows that $\{\Theta_t(X)(s) \mid t \geq 0\}$ is relatively compact for every s and therefore we may apply the Arzela-Ascoli Theorem 15.25 to conclude that $\{\Theta_t(X)\}$ is relatively compact in $C((-\infty, \infty); S)$.

Now suppose that $\lim_{t\to\infty} d(\Theta_t(X), \hat{\Phi}(\Theta_t(X))) \neq 0$ then there exists an $\epsilon > 0$ and a sequence t_k such that $d(\Theta_{t_k}(X), \hat{\Phi}(\Theta_{t_k}(X))) \geq \epsilon$. By relative compactness by passing to a subsequence and by using the assumption that every limit point of $\Theta_t(X)$ is in S_Φ we may assume that there exists $Y \in S_\Phi$ such that $\lim_{k\to\infty} \Theta_{t_k}(X) = Y$. Since $\hat{\Phi}$ is a continuous retraction onto S_Φ it follows that

$$\lim_{k\to\infty} \hat{\Phi}(\Theta_{t_k}(X)) = \hat{\Phi}(Y) = Y = \lim_{k\to\infty} \Theta_{t_k}(X)$$

which is a contradiction. □

DEFINITION 19.10. Suppose we have a continuous function $F : \mathbb{R}^d \to \mathbb{R}^d$, a sequence of vectors $x_n \in \mathbb{R}^d$ for $n \in \mathbb{Z}_+$, $U_n \in \mathbb{R}^d$ for $n \in \mathbb{N}$ and $\gamma_n \in \mathbb{R}_+$ for $n \in \mathbb{N}$ such that $\sum_{n=1}^\infty \gamma_n = \infty$ and $\lim_{n\to\infty} \gamma_n = 0$ satisfying

$$x_{n+1} - x_n = \gamma_{n+1}(F(x_n) + U_{n+1})$$

we define a sequence

$$\tau_0 = 0 \text{ and } \tau_n = \sum_{j=1}^n \gamma_j \text{ for } n \geq 1$$

interpolations of x_n $X, \overline{X} : \mathbb{R}_+ \to \mathbb{R}^d$

$$X(\tau_n + s) = x_n + s\frac{x_{n+1} - x_n}{\tau_{n+1} - \tau_n} = x_n + s\gamma_{n+1}^{-1}(x_{n+1} - x_n), \text{ and } \overline{X}(\tau_n + s) = x_n \text{ for } n \in \mathbb{N} \text{ and } 0 \leq s < \gamma_{n+1}$$

the "inverse" of the sequence τ_n, $m : \mathbb{R}_+ \to \mathbb{Z}_+$

$$m(t) = \sup\{k \geq 0 \mid t \geq \tau_k\}$$

and continuous time interpolations $\overline{U} : \mathbb{R}_+ \to \mathbb{R}^d$ and $\overline{\gamma} : \mathbb{R}_+ \to \mathbb{R}_+$ defined by

$$\overline{U}(\tau_n + s) = U_{n+1}, \text{ and } \overline{\gamma}(\tau_n + s) = \gamma_{n+1} \text{ for } n \in \mathbb{N} \text{ and } 0 \leq s < \gamma_{n+1}$$

We note the following simple facts

PROPOSITION 19.11. *For all* $t \in \mathbb{R}_+$, $\overline{X}(t) = X(\tau_{m(t)}) = x_{m(t)}$, $\overline{U}(t) = U_{m(t)+1}$, $\overline{\gamma}(t) = \gamma_{m(t)+1}$, $\tau_{m(t)} \leq t < \tau_{m(t)+1}$ *and*

$$X(t) = X(0) + \int_0^t (F(\overline{X}(s)) + \overline{U}(s))\, ds$$

PROOF. By the assumption $\sum_{n=1}^\infty \gamma_n$ we know that $\lim_{n\to\infty} \tau_n = \infty$. By non-negativity of γ_n we know that the sequence τ_n is non-decreasing. Therefore by definition we know that $m(t)$ is the unique non-negative integer such that $\tau_{m(t)} \leq t < \tau_{m(t)+1}$. Note that it may be that for some $n < m(t)$ we also have $\tau_n = \tau_{m(t)}$ but then $\tau_n = \tau_{n+1}$ and it is not the case that $\tau_n \leq t < \tau_{n+1}$. For such n is also true that $\gamma_{n+1} = 0$ and thus the values at n are not used in any of the interpolations. In particular writing $t = \tau_{m(t)} + s$ with $0 \leq s < \gamma_{m(t)+1}$ by the definitions of X and $\overline{X}$ we see that

$$\overline{X}(t) = \overline{X}(\tau_{m(t)} + s) = x_{m(t)} = X(\tau_{m(t)})$$

$\overline{U}(t)\overline{U}(\tau_{m(t)} + s) = U_{m(t)+1}$ and $\overline{\gamma}(t)\overline{\gamma}(\tau_{m(t)} + s) = \gamma_{m(t)+1}$.

To see the last equality we express the integral using the fact that the integrand is piecewise constant and that from the discussion above we may write $t = \tau_{m(t)} + s$ with $0 \leq s < \gamma_{m(t)+1}$ and use the recursion defining x_n

$$\begin{aligned}
\int_0^t (F(\overline{X}(s)) + \overline{U}(s))\, ds &= \sum_{n=0}^{m(t)-1} \gamma_{n+1}(F(x_n) + U_{n+1}) + s(F(x_{m(t)}) + U_{m(t)+1}) \\
&= \sum_{n=0}^{m(t)-1} (x_{n+1} - x_n) + s(F(x_{m(t)}) + U_{m(t)+1}) \\
&= x_{m(t)} - x_0 + s\frac{x_{m(t)+1} - x_{m(t)}}{\gamma_{m(t)+1}} \\
&= X(t) - X(0)
\end{aligned}$$

□

Thinking of the recursion $x_{n+1} = x_n + \gamma_{n+1}(F(x_n) + U_{n+1})$ as a perturbed version of the Euler method for solving an ordinary differential equation we pose the question of how well the interpolations $X(t)$ approximate solutions to the differential equation $\frac{dX}{dt} = F(X(t))$. In particular we seek asymptotic decay conditions of the sequence of perturbations $\gamma_{n+1}U_{n+1}$ that allow us to prove that $X(t)$ is an asymptotic pseudotrajectory. The following condition on the noise sequence is due to Kushner and Clark (TODO: Reference).

Definition 19.12. We say that γ_{n+1} and U_{n+1} satisfy the *Kushner-Clark criterion* if for every $T > 0$ we have

$$\lim_{n\to\infty} \sup\{\left\|\sum_{i=n}^{k-1} \gamma_{i+1}U_{i+1}\right\| \mid k = n+1, \dots, m(\tau_n + T)\} = 0$$

Proposition 19.13. *Let γ_{n+1} and U_{n+1} be given and define $\tau_0 = 0$, $\tau_n = \sum_{i=1}^n \gamma_i$ for $n \in \mathbb{N}$ and*

$$\Delta(t, T) = \sup_{0\leq h\leq T} \left\| \int_t^{t+h} \overline{U}(s)\, ds \right\|$$

then

(i) *for every $\delta > 0$ we have $\Delta(t, T) \leq 2\Delta(t - \delta, T + \delta)$*
(ii) *for every $n \in \mathbb{Z}_+$ we have*

$$\sup\left\{\left\|\sum_{i=n}^{k-1} \gamma_{i+1}U_{i+1}\right\| \mid k = n+1, \dots, m(\tau_n + T)\right\} = \Delta(\tau_n, m(\tau_n + T) - \tau_n) \leq \Delta(\tau_n, T)$$

If $\lim_{n\to\infty} \gamma_n = 0$ and $\sum_{n=1}^\infty \gamma_n = \infty$ then the Kushner-Clark condition is equivalent to $\lim_{t\to\infty} \Delta(t, T) = 0$ for all $T > 0$.

Proof. To see (i) let $t \geq 0$, $T, \delta > 0$ and $0 \leq h \leq T$ be given then

$$\begin{aligned}\left\| \int_t^{t+h} \overline{U}(s)\, ds \right\| &= \left\| \int_{t-\delta}^{t+h} \overline{U}(s)\, ds - \int_{t-\delta}^{t} \overline{U}(s)\, ds \right\| \\ &\leq \left\| \int_{t-\delta}^{(t-\delta)+h+\delta} \overline{U}(s)\, ds \right\| + \left\| \int_{t-\delta}^{t} \overline{U}(s)\, ds \right\| \\ &\leq 2 \sup_{0 \leq h \leq T+\delta} \left\| \int_{t-\delta}^{(t-\delta)+h} \overline{U}(s)\, ds \right\| = 2\Delta(t-\delta, T+\delta)\end{aligned}$$

To see (ii) we write

$$\sum_{i=n}^{k-1} \gamma_{i+1} U_{i+1} = \sum_{i=n}^{k-1} (\tau_{i+1} - \tau_i) \overline{U}(\tau_i) = \int_{\tau_n}^{\tau_k} \overline{U}(s)\, ds$$

We note that $\int_{\tau_n}^t \overline{U}(s)\, ds$ is linear for $\tau_k \leq t < \tau_{k+1}$ so by the convexity of norms we see that

$$\sup_{\tau_k \leq t < \tau_{k+1}} \left\| \int_{\tau_n}^{t} \overline{U}(s)\, ds \right\| = \left\| \int_{\tau_n}^{\tau_k} \overline{U}(s)\, ds \right\| \vee \left\| \int_{\tau_n}^{\tau_{k+1}} \overline{U}(s)\, ds \right\|$$

Combining this fact with the fact that $m(\tau_n + T) \leq \tau_n + T$ we get

$$\begin{aligned}&\sup \left\{ \left\| \sum_{i=n}^{k-1} \gamma_{i+1} U_{i+1} \right\| \mid k = n+1, \dots, m(\tau_n + T) \right\} \\ &= \sup \left\{ \left\| \int_{\tau_n}^{\tau_k} \overline{U}(s)\, ds \right\| \mid k = n+1, \dots, m(\tau_n + T) \right\} \\ &= \sup \left\{ \left\| \int_{\tau_n}^{\tau_k} \overline{U}(s)\, ds \right\| \mid k = n, \dots, m(\tau_n + T) \right\} \\ &= \sup \left\{ \left\| \int_{\tau_n}^{t} \overline{U}(s)\, ds \right\| \mid \tau_n \leq t \leq \tau_{m(\tau_n+T)} \right\} \\ &= \Delta(\tau_n, m(\tau_n + T) - \tau_n) \\ &\leq \Delta(\tau_n, T)\end{aligned}$$

If $\tau_n \to \infty$ then the condition $\lim_{t\to\infty} \Delta(t, T) = 0$ for all $T > 0$ clearly implies $\lim_{n\to\infty} \Delta(\tau_n, T) = 0$ for all $T > 0$; by (ii) this implies the Kushner-Clark criterion. On the other hand assume the Kushner-Clark criterion, let $T > 0$ and $\epsilon > 0$ be given and pick $N_T > 0$ such that $\Delta(\tau_n, m(\tau_n + T + 2) - \tau_n) < \epsilon/2$ for all $n \geq N_T$. Because $\gamma_n \to 0$ we may pick N_1 large enough that $\gamma_{n+1} < 1$ for all $n \geq N_1$ from which it follows that for such n

$$T + 1 < \tau_n + T + 2 - \gamma_{m(\tau_n+T+2)+1} - \tau_n \leq \tau_{m(\tau_n+T+2)+1} - \gamma_{m(\tau_n+T+2)+1} - \tau_n = \tau_{m(\tau_n+T+2)} - \tau_n$$

For $n \geq N_1 \vee N_T$ we have $\Delta(\tau_n, T+1) \leq \Delta(\tau_n, m(\tau_n + T + 2) - \tau_n) < \epsilon/2$. Then for all $t \geq N_1 \vee \tau N_T$ there is a unique $n \geq N_1 \vee N_T$ such that $\tau_n \leq t < \tau_{n+1}$ which implies $t - \tau_n < 1$ so

$$\Delta(t, T) = \Delta(t - \tau_n + \tau_n, T) \leq 2\Delta(\tau_n, T + (t - \tau_n)) \leq 2\Delta(\tau_n, T + 1) < \epsilon$$

which shows $\lim_{t\to\infty} \Delta(t, T) = 0$. $\square$

PROPOSITION 19.14. *Let $F : \mathbb{R}^d \to \mathbb{R}^d$ be a continuous vector field such that the differential equation $\dot{x} = F(x)$ has unique integral curves. Assume that γ_{n+1} and U_{n+1} satisfy the Kushner-Clark criterion and either*

(i) $\sup_n \|x_n\| < \infty$
(ii) *F is Lipschitz and bounded on a (convex?) neighborhod of $\{x_n\}$.*

then X is an asymptotic pseudotrajectory of the flow induced by F. Furthermore if (ii) is true then

$$\sup_{0\leq h\leq T} \|X(t+h) - \Phi_h(X(t))\| \leq C\left(\Delta(t-1, T+1) + \sup_{t\leq s\leq t+T} \|\overline{\gamma}(s)\|\right)$$

where the constant C only depends on T and F.

PROOF.

CLAIM 19.14.1. X is uniformly continuous.

If (i) holds then by continuity of F we know that $\sup_n \|F(x_n)\| < \infty$. If (ii) holds then $\sup_n \|F(x_n)\| < \infty$ because F is assumed bounded on a neighborhood of $\{x_n\}$. In either case there exists a constant K such that $\sup_t \|F(X(t))\| \leq K$. From Proposition 19.11 and Proposition 19.13 we get

$$\limsup_{t\to\infty} \sup_{0\leq h\leq T} \|X(t+h) - X(t)\| = \limsup_{t\to\infty} \sup_{0\leq h\leq T} \left\| \int_t^{t+h} (F(\overline{X})(u) + \overline{U}(u))\, du \right\|$$
$$\leq \limsup_{t\to\infty} \sup_{0\leq h\leq T} \left\| \int_t^{t+h} F(\overline{X})(u)\, du \right\| + \limsup_{t\to\infty} \sup_{0\leq h\leq T} \left\| \int_t^{t+h} \overline{U}(u))\, du \right\| \leq KT$$

From this it follows that X is uniformly continuous (given $\epsilon > 0$ let $\delta_1 = \epsilon/2K \wedge 1$ and pick $s > 0$ such that $\sup_{t\geq s} \sup_{0\leq h\leq \delta_1} \|X(t+h) - X(t)\| < K\delta_1 + \epsilon/2 = \epsilon$ for all $t \geq s$. Since $[0, s+1]$ is compact we know that X is uniformly continuous on $[0, s+1]$ (Theorem 1.35) and therefore there exists $\delta_2 > 0$ such that $\sup 0 \leq t \leq s \sup_{0\leq h\leq \delta_2} \|X(t+h) - X(t)\| < \epsilon$. Let $\delta = \delta_1 \wedge \delta_2$ and it follows that $\sup_{0\leq t<\infty} \sup_{0\leq h\leq \delta} \|X(t+h) - X(t)\| < \epsilon$).

Now lets break $\Theta_t(X)$ up into some terms that we'll examine individually.

(37)
$$\Theta_t(X)(s) = X(t+s) = X(0) + \int_0^{t+s} F(\overline{X}(u))\, du + \int_0^{t+s} \overline{U}(u)\, du$$

(38)
$$= X(0) + \int_0^t (F(\overline{X}(u)) + \overline{U}(u))\, du + \int_t^{t+s} F(\overline{X}(u))\, du + \int_t^{t+s} \overline{U}(u)\, du \text{ ????}$$

(39)
$$= X(t) + \int_0^s F(X(t+u))\, du + \int_t^{t+s} (F(\overline{X})(u) - F(X(u)))\, du + \int_t^{t+s} \overline{U}(u)\, du$$

(40)
$$= \Theta_t(X)(0) + \int_0^s F(\Theta_t(X)(u))\, du + \int_t^{t+s} (F(\overline{X})(u) - F(X(u)))\, du + \int_t^{t+s} \overline{U}(u)\, du$$

(41)
$$= L_F(\Theta_t(X))(s) + A_t(s) + B_t(s)$$

where have defined

$$
\begin{aligned}
L_F(X)(s) &= X(0) + \int_0^s F(X(u))\,du \text{ for } X \in C([0,\infty);\mathbb{R}^d) \\
A_t(s) &= \int_t^{t+s} (F(\overline{X}(u)) - F(X(u)))\,du \\
B_t(s) &= \int_t^{t+s} \overline{U}(u)\,du
\end{aligned}
$$

Note that $L_F(X) = X$ if and only if X is an integral curve of the differential equation $\dot{x} = F(x)$ (TODO: Presumably this is the fixed point operator used in Picard iteration).

The last term is the easiest to handle; by Lemma 15.22 the Kushner-Clark assumption is equivalent to the statement that $\lim_{t\to\infty} B_t = 0$ in $C([0,\infty);\mathbb{R}^d)$ (i.e. uniformly on compact sets).

We now turn to estimates on the second term which addresses errors that arise as a result of mixing the linear and constant interpolations of the x_n.

Claim 19.14.2. $\lim_{t\to\infty} A_t = 0$ in $C([0,\infty);\mathbb{R}^d)$

Fix $t, T > 0$ and consider the interval $t \leq u \leq t+T$. From Proposition 19.11 we get

$$
\begin{aligned}
\left\|X(u) - \overline{X}(u)\right\| = \left\|X(u) - X(\tau_{m(u)})\right\| &= \left\|\int_{\tau_{m(u)}}^u (F(\overline{X}(s)) + \overline{U}(s))\,ds\right\| \\
&\leq K(u - \tau_{m(u)}) + \left\|\int_{\tau_{m(u)}}^u \overline{U}(s)\,ds\right\| \leq K\overline{\gamma}(u) + \left\|\int_{\tau_{m(u)}}^u \overline{U}(s)\,ds\right\|
\end{aligned}
$$

Since $\lim_{n\to\infty} \gamma_n = 0$ and $t \leq u$ we know that for t sufficiently large we have

$$
\tau_{m(u)+1} - \tau_{m(u)} = \gamma_{m(u)+1} = \overline{\gamma}(u) < 1
$$

and therefore $t \leq u < \tau_{m(u)+1} < 1 + \tau_{m(u)}$; in particular $t - 1 < \tau_{m(u)} \leq u$. Therefore we may write

$$
\begin{aligned}
\left\|\int_{\tau_{m(u)}}^u \overline{U}(s)\,ds\right\| &= \left\|\int_{t-1}^u \overline{U}(s)\,ds - \int_{t-1}^{\tau_{m(u)}} \overline{U}(s)\,ds\right\| \\
&\leq \left\|\int_{t-1}^u \overline{U}(s)\,ds\right\| + \left\|\int_{t-1}^{\tau_{m(u)}} \overline{U}(s)\,ds\right\| \leq 2\Delta(t-1, T+1)
\end{aligned}
$$

and so we get

$$
\sup_{t\leq u\leq t+T} \left\|X(u) - \overline{X}(u)\right\| \leq 2\Delta(t-1, T+1) + K \sup_{t\leq u\leq t+T} \overline{\gamma}(u) \tag{42}
$$

Under either assumption (i) or (ii), F is uniformly continuous on a neighborhood of the $\{x_n\}$. Fix $T > 0$ then for any $\epsilon > 0$ there exists a δ such that $\left\|X(u) - \overline{X}(u)\right\| < \delta$ implies $\left\|F(X(u)) - F(\overline{X}(u))\right\| < \epsilon/T$. By (42), the Kushner-Clark condition and the fact that $\gamma_n \to 0$ we know that $\sup_{t\leq u\leq t+T} \left\|X(u) - \overline{X}(u)\right\| < \delta$ for sufficiently large t. Hence

$$
\sup_{0\leq s\leq T} \|A_t(s)\| = \sup_{0\leq s\leq T} \left\|\int_t^{t+s} (F(\overline{X}(u)) - F(X(u)))\,du\right\| \leq T(\epsilon/T) = \epsilon
$$

hence $\lim_{t \to \infty} \sup_{0 \leq s \leq T} \|A_t(s)\| = 0$. Since $T > 0$ we arbitrary we see that $A_t \to 0$ in $C([0, \infty), \mathbb{R}^d)$.

We note that the operator $L_F : C([0, \infty), \mathbb{R}^d) \to C([0, \infty), \mathbb{R}^d)$ is continuous (TODO: This should be put somewhere that we discuss Picard iteration). This follows from the fact that F is continuous and therefore $X \to F \circ X$ is continuous, evaluation $X \to X(0)$ is continuous and $X \to \int_0 X(s)\, ds$ is continuous.

Now suppose that X^* is a limit point of $\{\Theta_t(X)\}$. Thus there exists a sequence t_n with $t_n \to \infty$ and $\lim_{n \to \infty} \Theta_{t_n}(X) = X^*$ in $C([0, \infty), \mathbb{R}^d$. It follows from the prior claims and the continuity of L_F that

$$\begin{aligned} X^* &= \lim_{n \to \infty} \Theta_{t_n}(X) = \lim_{n \to \infty} L_F(\Theta_{t_n}(X)) + A_{t_n} + B_{t_n} = \lim_{n \to \infty} L_F(\Theta_{t_n}(X)) \\ &= L_F(\lim_{n \to \infty} \Theta_{t_n}(X)) = L_F(X^*) \end{aligned}$$

and therefore X^* is a solution of $\dot{x} = F(x)$. Since we have assumed that F has unique integral curves it follows that in fact $X^* = \Phi^{X^*(0)} \in S_\Phi$. From Claim 19.14.1 and Theorem 19.9 it follows that X is an asymptotic pseudotrajectory. (TODO: Don't we need relative compactness of $X([0, \infty))$ which only holds under (i)?)

Now suppose that (ii) holds and that $\|F(x) - F(y)\| \leq L \|x - y\|$ on a neighborhood of the $\{x_n\}$. In this case from (42) we get for t sufficiently large and $T > 0$

$$\|A_t(s)\| = \left\| \int_t^{t+s} (F(\overline{X}(u)) - F(X(u)))\, du \right\| \leq TL(2\Delta(t-1, T+1) + K \sup_{t \leq u \leq t+T} \overline{\gamma}(u)) \text{ for } 0 \leq s \leq T$$

and we also have

$$\|B_t(s)\| = \left\| \int_t^{t+s} \overline{U}(u)\, du \right\| \leq \Delta(t, T) \leq 2\Delta(t-1, T+1)$$

By (37) and the fact that as an integral curve we have $\Phi_s(X(t)) = \Phi^{X(t)}(s) = L_F(\Phi^{X(t)})$ and Gronwall's Inequality (Proposition 14.73) for sufficiently large t, all $T > 0$ and $0 \leq s \leq T$

$$\begin{aligned} \|X(t+s) - \Phi_s(X(t))\| &= \left\| \Theta_t(X)(s) - L_F(\Theta_t(X))(x) + L_F(\Theta_t(X))(s) + L_F(\Phi^{X(t)})(s) \right\| \\ &= \left\| A_t(s) + B_t(s) + \int_0^s (F(\Theta_t(X)(u)) - F(\Phi_u(X(t))))\, du \right\| \\ &\leq \|A_t(s)\| + \|B_t(s)\| + L \int_0^s \|X(t+u) - \Phi_u(X(t))\|\, du \\ &\leq 2(TL+1)\Delta(t-1, T+1) + TLK \sup_{t \leq u \leq t+T} \overline{\gamma}(u) + L \int_0^s \|X(t+u) - \Phi_u(X(t))\|\, du \\ &\leq (2(TL+1)\Delta(t-1, T+1) + TLK \sup_{t \leq u \leq t+T} \overline{\gamma}(u))(1 + \int_0^s e^{L(s-u)}\, du) \\ &\leq (2(TL+1)\Delta(t-1, T+1) + TLK \sup_{t \leq u \leq t+T} \overline{\gamma}(u))(1 + L^{-1} e^{TL}) \end{aligned}$$

Now take the supremeum over all $0 \leq s \leq T$. $\square$

We now return to the realm of probability and we assume that the sequences x_n and U_n are now random processes. The first question is whether we can find probabilistic hypotheses that guarantee the Kushner-Clark conditions hold almost surely.

We need a small consequence of Hölder's inequality.

LEMMA 19.15. *Let $a_i \geq 0$, $b_i \in \mathbb{R}$, $p > 1$ and $0 < \delta < 1$ then*

$$\left(\sum_{i=n}^{m} |a_i b_i|\right)^p \leq \left(\sum_{i=n}^{m} a_i^{\delta p/(p-1)}\right)^{p-1} \sum_{i=n}^{m} a_i^{1-\delta)p} |b_i|^p$$

PROOF. Noting that p and $\frac{p}{p-1}$ are conjugate exponents we simply apply Hölder's inequality

$$\begin{aligned}\sum_{i=n}^{m} |a_i b_i| &= \sum_{i=n}^{m} (a_i^{\delta})(a_i^{1-\delta} |b_i|) \\ &\leq \left(\sum_{i=n}^{m} a_i^{\delta p/p-1}\right)^{p-1/p} \left(\sum_{i=n}^{m} a_i^{1-\delta)p} |b_i|^p\right) 1/p\end{aligned}$$

and take the p^{th} power. TODO: Why do we need $0 < \delta < 1$? ☐

TODO: Do we need a vector valued version of Burkholder or is the current proof provided valid in $\mathbb{R}^d$???

PROPOSITION 19.16. *Let $(\Omega, \mathcal{A}, P)$ be a probability space with a filtration $\mathcal{F}_n$. Suppose that x_n and U_n are adapted processes, U_n is a martingale difference sequence (i.e. $\boldsymbol{E}[U_{n+1} \mid \mathcal{F}_n] = 0$ for all $n \in \mathbb{Z}_+$), γ_n is a deterministic sequence such that $\lim_{n\to\infty} \gamma_n = 0$ and $\sum_{n=1}^{\infty} \gamma_n = \infty$ and*

$$x_{n+1} = x_n + \gamma_{n+1}(F(x_n) + U_{n+1})$$

If for some $q \geq 2$ we have $\sup_n \boldsymbol{E}[\|U_{n+1}\|^q] < \infty$ and $\sum_{n=1}^{\infty} \gamma_n^{1+q/2} < \infty$ then the Kushner-Clark condition holds almost surely.

PROOF.

CLAIM 19.16.1. For all $T > 0$ and $t \geq 0$ there exists a constant C_q such that

$$\mathbf{E}\left[\Delta(t,T)^q\right] \leq C_q T^{q/2-1} \sup_m \|U_m\|^q \int_t^{t+T} \overline{\gamma}^{q/2}(s)\, ds$$

The proof of the claim requires an inequality. Let $\psi_1, \psi_2, \dots$ be a sequence of non-negative numbers and define $\sigma_0 = 0$ and $\sigma_n = \sum_{i=1}^{n} \psi_i$ for $n \geq 1$. Since $\psi_{n+1} U_{n+1}$ is an $\mathcal{F}$ martingale difference sequence we know that $\sum_{i=1}^{n} \psi_{n+1} U_{n+1}$ is an $\mathcal{F}$-martingale and therefore by the right hand side of Burkholder's Inequality (Theorem 9.62) we conclude that for every $t \geq 0$ and $n \in \mathbb{Z}_+$

$$\mathbf{E}\left[\sup_{n<k\leq m(\sigma_n+T)} \left\|\sum_{i=n}^{k-1} \psi_{i+1} U_{i+1}\right\|^q\right] \leq C_q \mathbf{E}\left[\left(\sum_{i=n}^{m(\sigma_n+T)-1} \psi_{i+1}^2 \|U_{i+1}\|^2\right)^{q/2}\right]$$

If we suppose $q > 2$ then we can apply Lemma 19.15 with $p = q/2$, $\delta = (q-2)/2q$, $a_i = \psi_{i+1}^2$ and $b_i = \|U_{i+1}\|^2$ hence $(p/(p-1) = (q/2)/((q/2)-1) = q/(q-2)$ and

$1-\delta = (q+2)/2q)$ to conclude

$$
\begin{aligned}
&\mathbf{E}\left[\sup_{n<k\le m(\sigma_n+T)}\left\|\sum_{i=n}^{k-1}\psi_{i+1}U_{i+1}\right\|^q\right] \\
&\le C_q\mathbf{E}\left[\left(\sum_{i=n}^{m(\sigma_n+T)-1}\psi_{i+1}^{2\left(\frac{q-2}{2q}\right)\left(\frac{q}{q-2}\right)}\right)^{q/2-1}\sum_{i=n}^{m(\sigma_n+T)-1}\psi_{i+1}^{2\left(\frac{q+2}{2q}\right)\left(\frac{q}{2}\right)}\|U_{i+1}\|^{2\left(\frac{q}{2}\right)}\right] \\
&= C_q\mathbf{E}\left[\left(\sum_{i=n}^{m(\sigma_n+T)-1}\psi_{i+1}\right)^{q/2-1}\sum_{i=n}^{m(\sigma_n+T)-1}\psi_{i+1}^{1+q/2}\|U_{i+1}\|^q\right] \\
&= C_q\mathbf{E}\left[\left(\sigma_{m(\sigma_n+T)}-\sigma_n\right)^{q/2-1}\sum_{i=n}^{m(\sigma_n+T)-1}\psi_{i+1}^{1+q/2}\|U_{i+1}\|^q\right] \\
&\le C_q(\sigma_n+T-\sigma_n)^{q/2-1}\mathbf{E}\left[\sum_{i=n}^{m(\sigma_n+T)-1}\psi_{i+1}^{1+q/2}\|U_{i+1}\|^q\right] \\
&\le C_q T^{q/2-1}\sup_m\|U_m\|^q\sum_{i=n}^{m(\sigma_n+T)-1}\psi_{i+1}^{1+q/2}
\end{aligned}
$$

Now if we fix $t\ge 0$ and we consider $\Delta(t,T)=\sup_{t\le u\le t+T}\left\|\int_t^u \overline{U}(s)\,ds\right\|$. As in the proof of Proposition 19.13 the piecewise linearity of the integral as a function of u and the convexity of the norm implies that the supremum is attained at some $u\in\{t, m(t)+1,\dots,m(t+T),t+T\}$. Define the sequence $\psi_{m(t)}=t$, $\psi_{m(t)+1}=\tau_{m(t)+1}-t=\gamma_{m(t)+1}-(t-\tau_{m(t)})$, $\psi_i=\gamma_i$ for $i=m(t)+1,\dots,m(t+T)$ and $\psi_{m(t+T)+1}=t+T-m(t+T)$. Applying the above inequality (noting $\sigma_n=t$) we get (TODO: there is some ambiguity to clear up about m defined by the γ_i and m defined by the ψ; the salient point is these two functions are equal on the interval $[t,t+T]$).

$$
\begin{aligned}
\mathbf{E}\left[\Delta(t,T)\right] &= \mathbf{E}\left[\left\|\sup_{m(t)<k\le m(t+T)}\psi_{i+1}U_{i+1}\right\|^q\right] \\
&\le C_q T^{q/2-1}\sup_m\|U_m\|^q\sum_{i=m(t)}^{m(t+T)-1}\psi_{i+1}^{1+q/2} \\
&\le C_q T^{q/2-1}\sup_m\|U_m\|^q\sum_{i=m(t)}^{m(t+T)-1}\psi_{i+1}\gamma_{i+1}^{q/2} \\
&= C_q T^{q/2-1}\sup_m\|U_m\|^q\int_t^{t+T}\overline{\gamma}^{q/2}(s)\,ds
\end{aligned}
$$

TODO: Handle the case $q=2$ which is more direct.

For $q \geq 2$ and every $\epsilon > 0$

$$\begin{aligned}\sum_{k=0}^{\infty} \mathbf{P}\{\Delta(kT,T) > \epsilon\} &\leq \epsilon^{-q} \sum_{k=0}^{\infty} \mathbf{E}\left[\Delta(kT,T)^q\right] \\ &\leq \epsilon^{-q} C_q T^{q/2-1} \sup_m \|U_m\|^q \sum_{k=0}^{\infty} \int_{kT}^{(k+1)T} \overline{\gamma}^{q/2}(s)\, ds \\ &= \epsilon^{-q} C_q T^{q/2-1} \sup_m \|U_m\|^q \int_0^{\infty} \overline{\gamma}^{q/2}(s)\, ds \\ &= \epsilon^{-q} \sum_{n=1}^{\infty} \gamma_{n+1}^{1+q/2} < \infty\end{aligned}$$

and therefore by the Borel Cantelli Theorem 4.23 we get $\mathbf{P}\{\Delta(kT,T) > \epsilon i.o.\} = 0$ and by Lemma 5.4 $\lim_{k\to\infty} \Delta(kT,T) = 0$ almost surely.

For an arbitrary $0 \leq t < \infty$ there exists a unique $k \in \mathbb{Z}_+$ such that $kT \leq t < (k+1)T$ and for such a k we have for $0 \leq h \leq T$ $\int_t^{(t+h)\wedge(k+1)T} = \int_{kT}^{(t+h)\wedge(k+1)T} - \int_{kT}^{t}$ and therefore $\left\|\int_t^{(t+h)\wedge(k+1)T}\right\| \leq \left\|\int_{kT}^{(t+h)\wedge(k+1)T}\right\| + \left\|\int_{kT}^{t}\right\| \leq 2\sup_{0\leq h\leq T} \int_{kT}^{kT+h}$ also $\left\|\int_{(t+h)\wedge(k+1)T}^{t+h}\right\| \leq sup_{0\leq h\leq T} \left\|\int_{(k+1)T}^{(k+1)T+h}\right\|$ and so $\Delta(t,T) \leq 2\Delta(kT,T) + \Delta((k+1)T,T)$. This shows that $\lim_{t\to\infty} \Delta(t,T) = 0$ almost surely. Now we apply Proposition 19.13. $\square$

CHAPTER 20

The General Theory of Processes

1. The Debut Theorem

In studying cadlag processes in this book we have made repeated use that hitting times of open and closed sets are usually optional times. These facts were very easy to prove and appear in Lemma 9.70. In the general theory of processes we don't want to require continuity properties but at the same time we want to have optional times handy. In this section we prove that given a filtration satisfying the usual conditions, hitting times of progressively measurable sets are optional times.

THEOREM 20.1 (The Debut Theorem). *Let $(\Omega, \mathcal{A}, P)$ be a probability space with a filtration $\mathcal{F}_t$ for $0 \leq t < \infty$ that satisfies the usual conditions. Let $A \subset [0, \infty) \times \Omega$ be progressively measurable and define $\tau_A = \inf\{t \geq 0 \mid (t, \omega) \in A\}$ then τ_A is an $\mathcal{F}$-optional time.*

The proof of the theorem is a bit involved and requires some definitions and lemmas. Before turning to those let's understand the main issue that the theorem has to deal with. We have made the assumption that $\mathcal{F}$ is right continuous so we only need to show that τ_A is weakly $\mathcal{F}_t$-optional and this amounts to

$$\begin{aligned}\{\tau_A < t\} &= \{\omega \in \Omega \mid \text{ there exists } 0 \leq s < t \text{ such that } (s, \omega) \in A\} \\ &= \pi([0, t) \times \Omega \cap A) \in \mathcal{F}_t\end{aligned}$$

where $\pi : [0, t] \times \Omega \to \Omega$ is projection on the second coordinate. Now by progressive measurability of A we know that $[0, t] \times \Omega \cap A \in \mathcal{B}([0, t]) \otimes \mathcal{F}_t$ however there is nothing that tells us that the projection on the second coordinate lives in $\mathcal{F}_t$. The proof of the Debut Theorem thus boils down to two key tasks. First we must find a class of sets in $\mathcal{B}([0, t]) \otimes \mathcal{F}_t$ whose projection on Ω are in $\mathcal{F}_t$. Secondly we must show that for progressively measurable A, the intersections $A \cap [0, t] \times \Omega$ are in this class.

We now begin on the first task. We define the class of t-approximable subsets $[0, t] \times \Omega$ and show that they project to sets in $\mathcal{F}_t$. We first define a simple constructive class of sets and show that they have the properties we seek.

DEFINITION 20.2. For $0 \leq t < \infty$ let $\mathcal{K}^0(t)$ be the set of subsets $K \times A \subset [0, t] \times \Omega$ where K is compact and $A \in \mathcal{F}_t$. Let $\mathcal{K}(t)$ the set of finite unions of elements of $\mathcal{K}^0(t)$ and let $\mathcal{K}_\delta(t)$ be the set of countable intersections of sets in $\mathcal{K}(t)$.

LEMMA 20.3. *For every $A \in \mathcal{K}_\delta(t)$ there exist a nested sequence $A_1 \supset A_2 \supset \cdots$ with $A_n \in \mathcal{K}(t)$ and $A = \cap_n A_n$.*

PROOF. By definition we know that there exists $A_1, A_2, \ldots \in \mathcal{K}(t)$ such that $\cap_n A_n = A$. The result will follow if we can show that $\cap_{j=1}^n A_j \in \mathcal{K}(t)$ for all $n \in \mathbb{N}$. By induction this reduces to showing that $\mathcal{K}(t)$ is closed under pairwise intersection.

Let $A = \cup_{i=1}^n K_i \times C_i$ and $B = \cup_{j=1}^n L_j \times D_j$ with K_i and L_j all compact subsets of $[0,t]$ and $C_i, D_j \in \mathcal{F}_t$. The intersection is an elementary computation

$$\begin{aligned} A \cap B &= (\cup_{i=1}^n K_i \times C_i) \cap (\cup_{j=1}^m L_j \times D_j) = \cup_{i=1}^n \cup_{j=1}^m (K_i \times C_i) \cap (L_j \times D_j) \\ &= \cup_{i=1}^n \cup_{j=1}^m (K_i \cap L_j) \times (C_i \times D_j) \end{aligned}$$

and each $K_i \cap L_j$ is compact and $C_i \times D_j \in \mathcal{F}_t$. □

The next result shows members of $\mathcal{K}_\delta(t)$ project into $\mathcal{F}_t$.

LEMMA 20.4. *If $A \in \mathcal{K}_\delta(t)$ then $\pi(A) \in \mathcal{F}_t$. If $A, A_1, A_2, \dots \in \mathcal{K}_\delta(t)$, $A_1 \supset A_2 \supset \cdots$ and $\cap_n A_n = A$ then $\pi(A) = \cap_n \pi(A_n)$.*

PROOF. First consider $\mathcal{K}^0(t)$ and $\mathcal{K}(t)$. If $K \times A \in \mathcal{K}^0(t)$ then by assumption, $\pi(K \times A) = A \in \mathcal{F}_t$. If $A \in \mathcal{K}(t)$ then write $A = \cup_{i=1}^n K_i \times A_i$ and note that $\pi(A) = \cup_{i=1}^n A_i \in \mathcal{F}_t$.

To handle $A \in \mathcal{K}_\delta(t)$ we introduce the notation

$$S(A)(\omega) = \{s \in [0,t] \mid (s,\omega) \in A\} = \pi_1([0,t] \times \{\omega\} \cap A)$$

where π_1 is the projection onto $[0,t]$. Note that if $A \subset B$ it follows that $S(A)(\omega) \subset S(B)(\omega)$ for all $\omega \in \Omega$ and moreover if $A = \cap_{n=1}^\infty A_n$ then $S(A)(\omega) = \cap_{n=1}^\infty S(A_n)(\omega)$ (if $s \in S(A_n)(\omega)$ for all $n \in \mathbb{N}$ then $(s,\omega) \in \cap_{n=1}^\infty A_n$ which implies $s \in S(A)(\omega)$).

Clearly if $A = K \times C \in \mathcal{K}^0(t)$ then

$$S(A)(\omega)$$

□

DEFINITION 20.5. A set $A \in \mathcal{B}([0,t]) \otimes \mathcal{F}_t$ is said to be *t-approximable* if for every $\epsilon > 0$ there exists $B \in \mathcal{K}_\delta(t)$ such that $B \subset A$ and

$$\mathbf{P}^*\{\pi(A)\} \le \mathbf{P}^*\{\pi(B)\} + \epsilon$$

Note that at this point we must use outer probabilities since we don't know about the measurability of $\pi(A)$ and $\pi(B)$.

2. The Section Theorem

3. The Doob-Meyer Decomposition

The Doob-Meyer decomposition for continuous time stochastic processes turns out to be significantly more subtle that one might guess given how easily most of the results for discrete time martingales translated to the continuous time setting once the regularization Theorem 9.76 was proven. We finally turn to the extension here.

One of the obvious approaches that one could take in proving a decomposition for continuous time is discretize the continuous time process, take the Doob Decomposition of the discrete time process and then try to take a limit. As it turns out, finding the correct setting in which to prove that such a limit exists takes some effort. One approach is to show that the decompositions converge the weak L^1 norm. With this approach it is not too hard to show that the decompositions converge, but it is hard to prove properties of the limiting process.

Here we use an approach that gives normal L^1 convergence at the expense of using convex combinations. We develop some general machinery for proving such limits exist. We start with a simple motivational fact : very general convex combinations of convergent sequences also converge.

DEFINITION 20.6. Let A be an arbitrary index sets, $\{v_\alpha\}_{\alpha in A}$ be a set of elements in a vector space X then the *convex hull* $\text{conv}\{v_\alpha\}$ is the set of all vectors of the form $\sum_{n=1}^N c_n v_{\alpha_n}$ for $N \in \mathbb{Z}$, $\alpha_n \in A$, $c_n \geq 0$ for all $n = 1, \dots, N$ and $\sum_{n=1}^N c_n = 1$.

PROPOSITION 20.7. *Let $v, v_1, v_2, \dots$ be elements of a normed vector space such that $\|v_n - v\| < \epsilon$ for some $\epsilon > 0$, then for every $w \in \text{conv}\{v_1, v_2, \dots\}$ we have $\|w - v\| < \epsilon$. In particular if $\lim_{n\to\infty} v_n = v$ then for every sequence $w_n \in \text{conv}\{v_n, v_{n+1}, \dots\}$ we have $\lim_{n\to\infty} w_n = v$.*

PROOF. Let $w = c_1 v_1 + \dots + c_N v_N$ with $c_n \geq 0$ for $n = 1, \dots, N$ and $\sum_{n=1}^N c_n = 1$. Then by the triangle inequality

$$\|w - v\| = \left\| \sum_{n=1}^N c_n (v_n - v) \right\| \leq \sum_{n=1}^N c_n \|v_n - v\| < \epsilon \sum_{n=1}^N c_n = \epsilon$$

Now if $\lim_{n\to\infty} v_n = v$ and let $w_n \in \text{conv}\{v_n, v_{n+1}, \dots\}$ for every $n \in \mathbb{N}$. For every $\epsilon > 0$ there exists $N > 0$ such that $\|v_n - v\| < \epsilon$ for all $n \geq N$ so by the first part of the proposition $\|w_n - v\| < \epsilon$ for all $n \geq N$ and it follows that $\lim_{n\to\infty} w_n = v$. □

The key observation is a type of converse of the trivial observation above: with simply hypotheses we can show there exist convergent convex combinations of sequences. We start with a simple case.

LEMMA 20.8. *Let H be a Hilbert space and suppose that we are given a bounded sequence $v_1, v_2, \dots \in H$, then there exists a convergent sequence w_n where w_n is finite convex combination of $v_n, v_{n+1}, \dots$.*

PROOF. For each $n \in \mathbb{N}$ let K_n be the convex hull of $\{v_n, v_{n+1}, \dots\}$ (the set of finite convex combinations of $v_n, v_{n+1}, \dots$) and define

$$A_n = \inf\{\|g\| \mid g \in K_n\}$$
$$A = \sup_n A_n$$

Note that the A_n are an increasing sequence and moreover since $v_n \in K_n$ we know that $A_n \leq \|v_n\|$ and therefore

$$A = \sup_n A_n = \lim_{n\to\infty} A_n \leq \sup_n \|v_n\| < \infty$$

CLAIM 20.8.1. For each $n \in \mathbb{N}$ select $w_n \in K_n$ such that $w_n \leq A_n + \frac{1}{n} \leq A + \frac{1}{n}$, then w_n is a Cauchy sequence.

Let $\epsilon > 0$ be given and pick $N > 0$ such that $\frac{1}{N} < \epsilon$ and $A_n > A - \epsilon$ for all $n \geq N$. Note that for all $n, m \geq N$ we have $(w_n + w_m)/2 \in K_N$ and therefore $\|(w_m + w_n)/2\| \geq A_N > A - \epsilon$. Therefore for all $m, n \geq N$

$$\begin{aligned}\|w_m - w_n\|^2 = 2\|w_m\|^2 + 2\|w_n\| - \|w_m + w_n\| &\leq 2(A + \frac{1}{m})^2 + 2(A + \frac{1}{n})^2 - 4(A-\epsilon)^2 \\ &\leq 4(A + \frac{1}{N})^2 - 4(A-\epsilon)^2 = 8A(\frac{1}{N} + \epsilon) + (\frac{1}{N^2} - \epsilon^2) < 16A\epsilon\end{aligned}$$

and the claim is proven.

Now since H is complete it follows that w_n converges in H. □

We now use a truncation procedure to extend the result from Hilbert spaces to uniformly integrable sequences in L^1 spaces.

PROPOSITION 20.9. *Let $\xi_1, \xi_2, \ldots$ be a uniformly integrable sequence of random variables then there exists $\eta, \eta_1, \eta_2, \ldots$ with η_n a finite convex combination of $\xi_n, \xi_{n+1}, \ldots$ and $\eta_n \stackrel{L^1}{\to} \eta$.*

PROOF. For each $N \in \mathbb{Z}$ we define the truncated sequence $\xi_{n,\leq N} = \xi_n \cdot \mathbf{1}_{\xi_n \leq N}$. Since each sequence $\xi_{n,\leq N}$ is pointwise bounded, it is also L^2 bounded and we can apply Lemma 20.8 to get a convergent convex combination. However with a bit of care we can do more.

CLAIM 20.9.1. For each $n \in \mathbb{N}$ there exist an $M_n \in \mathbb{Z}$ and $c_n^n, \ldots, c_{M_n}^n$ with $c_j^n \geq 0$ for $j = n, \ldots, M_n$ and $\sum_{j=n}^{M_n} c_j^n = 1$ such that $\sum_{j=n}^{M_n} c_j^n \xi_{n,\leq N}$ converges in L^2 for every $N \in \mathbb{Z}$.

Applying Lemma 20.8 to the sequence $\xi_{n,\leq 1}$ we get $N_n^1 \in \mathbb{Z}$ and convex coefficients $c_n^{1,n}, \ldots, c_{N_n^1}^{1,n}$ such that $\sum_{j=n}^{N_n^1} c_j^{1,n} \xi_{j,\leq 1}$ converges in L^2. Now consider the sequence $\sum_{j=n}^{N_n^1} c_j^{1,n} \xi_{j,\leq 2}$ and observe that it is also pointwise bounded hence L^2 bounded. Thus we can find positive integers M_n and convex coefficients $d_n^n, \ldots, d_{N_n^2}^n$ such that $\sum_{m=n}^{M_n} d_m^n \sum_{j=m}^{N_m^1} c_j^{1,m} \xi_{j,\leq 2}$ converges. Moreover by Proposition 20.7 we know that the sequence $\sum_{m=n}^{M_n} d_m^n \sum_{j=m}^{N_m^1} c_j^{1,m} \xi_{j,\leq 1}$ also converges (in fact to the same limit as $\sum_{j=n}^{N_n^1} c_j^{1,n} \xi_{j,\leq 1}$). Defining $N_n^2 = N_{M_n}^1$ and $c_j^{2,n} =$ we see that

$$\sum_{j=n}^{N_n^2} c_j^{2,n} \xi_{j,\leq 2} =$$

and similarly with $\xi_{j,\leq 1}$.

TODO: Finish □

4. Exercises

EXERCISE 1. Let $f : \mathbb{R} \to \mathbb{R}$ be a right continuous function, show that f is Borel measurable.

PROOF. It suffices to show that $f^{-1}(t, \infty)$ is Borel measurable for all $t \in \mathbb{R}$. Let $x \in f^{-1}(t, \infty)$ then by right continuity there exists some y_x with $x < y_x$ such that $[x, y_x) \subset f^{-1}(t, \infty)$. Clearly we may write $f^{-1}(t, \infty) = \cup_{x \in f^{-1}(t,\infty)} [x, y_x)$. We now show that we can make this a countable union of intervals. For a fixed $q \in \mathbb{Q}$ consider the set $A_q = \cup_{\substack{x \in f^{-1}(t,\infty) \\ q \in [x, y_x)}} [x, y_x)$. It is easy to see that A_q is either empty or an interval (either open or half open) by taking least upper bounds and greatest lower bounds of the intervals in the union. Thus each A_q is measurable. More over each $[x, y_x)$ contains a rational number so it follows that $[x, y_x) \subset A_q$ for some $q \in rationals$. From this it follows that $f^{-1}(t, \infty) = \cup_{q \in \mathbb{Q}} A_q$ which is a countable union of measurable sets and therefore measurable. $\square$

EXERCISE 2. 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

$$\lim_{x \to \infty} a(x) = \lim_{x \to \infty} \frac{1}{2\sqrt{1 - g(x)}} = \infty$$

$\square$

EXERCISE 3. 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

$$\begin{aligned}\mathbf{E}\left[e^{s(\xi - \lambda)}\right] &= e^{-s\lambda} \sum_{k=0}^{\infty} \frac{s^k \mathbf{E}\left[\xi^k\right]}{k!} \\ &\geq e^{-s\lambda} \mathbf{E}\left[\xi^q\right] \lambda^{-q} \sum_{k=0}^{\infty} \frac{s^k \lambda^k}{k!} = \mathbf{E}\left[\xi^q\right] \lambda^{-q}\end{aligned}$$

$\square$

EXERCISE 4. Let ξ be a nonnegative integer valued random variable. Show $\mathbf{P}\{\xi \neq 0\} \leq \mathbf{E}[\xi]$ and

$$\mathbf{P}\{\xi = 0\} \leq \frac{\mathbf{Var}(\xi)}{\mathbf{Var}(\xi) + (\mathbf{E}[\xi])^2}$$

PROOF. For the first inequality,

$$\mathbf{P}\{\xi \neq 0\} = \sum_{k=1}^{\infty} \mathbf{P}\{\xi = k\} \leq \sum_{k=1}^{\infty} k\mathbf{P}\{\xi = k\} = \mathbf{E}[\xi]$$

For the second inequality, use Cauchy-Schwartz

$$\begin{aligned}(\mathbf{E}[\xi])^2 &\leq (\mathbf{E}[\mathbf{1}_{\xi>0}\xi])^2 \\ &\leq \mathbf{E}\left[\xi^2\right]\mathbf{P}\{\xi > 0\}\end{aligned}$$

Now use $\mathbf{P}\{\xi > 0\} = 1 - \mathbf{P}\{\xi = 0\}$ and $\mathbf{Var}(\xi) = \mathbf{E}\left[\xi^2\right] - (\mathbf{E}[\xi])^2$ and rearrangement of terms to get the result. $\square$

EXERCISE 5. 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. $\square$

EXERCISE 6. 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

$$f^{-1}(\sigma(\mathcal{C})) \subset f^{-1}(f_*(\sigma(f^{-1}(\mathcal{C})))) \subset \sigma(f^{-1}(\mathcal{C}))$$

$\square$

EXERCISE 7. Let $f(x) = |x|$. Show that $f_*(\mathcal{B}(\mathbb{R}))$ is a strict σ-subalgebra of $\mathcal{B}(\mathbb{R})$.

Exercise 8. 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 9. 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 10. 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 4.23) 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 11. 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 4.29 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 4.16. □

EXERCISE 12. 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 13. Let ξ, ξ_n be random elements in a metric space S such that $\xi_n \xrightarrow{P} \xi$, let A_n be events such that $\mathbf{P}\{A_n\} = 1$ and let η_n be random elements in S such that $\eta_n = \xi_n$ on A_n, show that $\eta_n \xrightarrow{P} \xi$.

PROOF. Fix $\epsilon > 0$ and note that

$$\lim_{n\to\infty} \mathbf{P}\{d(\eta_n, \xi) > \epsilon\} = \lim_{n\to\infty} \mathbf{P}\{d(\eta_n, \xi) > \epsilon; A_n\} + \lim_{n\to\infty} \mathbf{P}\{d(\eta_n, \xi) > \epsilon; A_n^c\} \leq \lim_{n\to\infty} \mathbf{P}\{d(\xi_n, \xi) > \epsilon\} + \lim_{n\to\infty} \mathbf{P}\{A$$

☐

EXERCISE 14. Suppose $\xi, \xi_1, \xi_2, \ldots$ are random variables with $\xi_n \xrightarrow{a.s.} \xi$ and $\xi < \infty$ a.s. Let $\eta = \sup_n |\xi_n|$ and show that $\eta < \infty$ a.s.

PROOF. TODO ☐

EXERCISE 15 (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 4.13 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 16 (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 $\xi^t, \xi^t_n \in S_t$ for $n \in \mathbb{N}$ such that $\xi^t_n \xrightarrow{a.s.} \xi^t$. If for each fixed $n \in \mathbb{N}$ the ξ^t_n 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 4.18 and the independence of the $\xi^{t_j}_n$ we have $\mathbf{E}\left[f_1(\xi^{t_1}_n) \cdots f(\xi^{t_m}_n)\right] = \mathbf{E}\left[f_1(\xi^{t_1}_n)\right] \cdots \mathbf{E}\left[f(\xi^{t_m}_n)\right]$ for each $n \in \mathbb{N}$. But now we can use the boundedness and continuity of the f_j

$$\begin{aligned}
&\mathbf{E}\left[f_1(\xi^{t_1}) \cdots f_m(\xi^{t_m})\right] \\
&= \mathbf{E}\left[\lim_{n\to\infty} f_1(\xi^{t_1}_n) \cdots f_m(\xi^{t_m}_n)\right] && \text{by continuity} \\
&= \lim_{n\to\infty} \mathbf{E}\left[f_1(\xi^{t_1}_n) \cdots f_m(\xi^{t_m}_n)\right] && \text{boundedness of } f_j \text{ and Dominated Convergence} \\
&= \lim_{n\to\infty} \mathbf{E}\left[f_1(\xi^{t_1}_n)\right] \cdots \mathbf{E}\left[f_m(\xi^{t_m}_n)\right] && \text{independence} \\
&= \mathbf{E}\left[f_1(\xi^{t_1})\right] \cdots \mathbf{E}\left[f_m(\xi^{t_m})\right] && \text{continuity and Dominated Convergence}
\end{aligned}$$

We now prove a slight extension of Lemma 4.18 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 5.41). Let $f_n(x) = 1 \wedge nd(x,U^c)$ and observe that $f_n \uparrow \mathbf{1}_U$. Now suppose $U_j \subset S_{t_j}$ are open sets for $j = 1, \dots, m$ 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

$$\begin{aligned}
\mathbf{P}\{\xi^{t_1} \in U_1 \cap \dots \cap \xi^{t_m} \in U_m\} &= \lim_{n\to\infty} \mathbf{E}\left[f_n^1(\xi^{t_1}) \cdots f_n^m(\xi^{t_m})\right] \\
&= \lim_{n\to\infty} \mathbf{E}\left[f_n^1(\xi^{t_1})\right] \cdots \mathbf{E}\left[f_n^m(\xi^{t_m})\right] \\
&= \mathbf{P}\{\xi^{t_1} \in U_1\} \cdots \mathbf{P}\{\xi^{t_m} \in U_m\}
\end{aligned}$$

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 4.13 it suffices to show independence on open sets. $\square$

A simpler subcase of the above

Exercise 17. 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 4.13. $\square$

Exercise 18. Let $\xi_1, \xi_2, \dots$ 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 4.18, we have

$$\mathbf{E}\left[\prod_{k=1}^\infty \xi_k\right] = \lim_{n\to\infty} \mathbf{E}\left[\prod_{k=1}^n \xi_k\right] = \lim_{n\to\infty} \prod_{k=1}^n \mathbf{E}\left[\xi_k\right] = \prod_{k=1}^\infty \mathbf{E}\left[\xi_k\right]$$

$\square$

Exercise 19. Provide an example of uncorrelated but non-independent random variables.

Proof. See Example 4.21. $\square$

Exercise 20. 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 $\mathbb{R}$). $\square$

EXERCISE 21. Let $\xi_1, \xi_2, \ldots$ 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 2.44) 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 $0 \leq \xi_n \leq 1$.

$$
\begin{aligned}
0 < \mathbf{E}\left[e^{-\sum_{n=1}^\infty \xi_n}\right] \qquad &= \mathbf{E}\left[\prod_{n=1}^\infty e^{-\xi_n}\right] = \prod_{n=1}^\infty \mathbf{E}\left[e^{-\xi_n}\right] \\
\leq \prod_{n=1}^\infty (1 - a\mathbf{E}[\xi_n]) \qquad & \qquad \text{where } a = 1 - e^{-1} \text{ by Lemma C.1} \\
\leq \prod_{n=1}^\infty e^{-a\mathbf{E}[\xi_n]} \qquad & \qquad \text{since } 1 + x \leq e^x \text{ by Lemma C.1} \\
= e^{-a\sum_{n=1}^\infty \mathbf{E}[\xi_n]}
\end{aligned}
$$

which shows that $\sum_{n=1}^\infty \mathbf{E}[\xi_n] < \infty$. ☐

Exercise 22. Let $\mu : S \times \mathcal{T} to [0,1]$ be a probability kernel and let $f : U \times T \to \mathbb{R}$ be measurable then $\int f(u,t)\, \mu(s,dt)$ is $\mathcal{U} \times \mathcal{S}$ measurable.

Proof. Assume first that f is the characteristic function of a set $A \times B \in \mathcal{U} \otimes \mathcal{T}$. Then

$$\int \mathbf{1}_{A\times B}(u,t)\, \mu(s,dt) = \mathbf{1}_A(u)\mu(s,B)$$

which is clearly measurable since μ is a kernel. We know that sets $A \times B$ are a π-system generating $\mathcal{U} \otimes \mathcal{T}$ so can argue with monotone classes to extend to general characteristic functions. To be specific, let

$$\mathcal{C} = \{C \in \mathcal{U} \otimes \mathcal{T} \mid \int \mathbf{1}_C(u,t)\, \mu(s,dt) \text{ is measurable}\}$$

If $A \in \mathcal{C}$ and $B \in \mathcal{C}$ with $A \subset B$ then $\int \mathbf{1}_{B\setminus A}(u,t)\, \mu(s,dt) = \int \mathbf{1}_B(u,t)\, \mu(s,dt) - \int \mathbf{1}_A(u,t)\, \mu(s,dt)$ is measurable so that $B \setminus A \in \mathcal{C}$. Similarly if $A_1 \subset A_2 \subset \cdots$ with $A_n \in \mathcal{C}$ then defining $A = \cup_{n=1}^\infty A_n$ we have by Monotone Convergence $\int \mathbf{1}_A(u,t)\, \mu(s,dt) = \lim_{n\to\infty} \int \mathbf{1}_{A_n}(u,t)\, \mu(s,dt)$ is a limit of measurable function hence is measurable. This shows that $\mathcal{C}$ is a λ-system and therefore by the π-λ Theorem 2.27 we know that $\mathcal{C} \subset \mathcal{U} \otimes \mathcal{T}$. By linearity it follows that $\int f(u,t)\, \mu(s,dt)$ is measurable for all simple functions.

Now given an arbitrary non-negative measurable $f : U \times T \to [0,\infty)$ we find an increasing sequence of simple functions $f_n \uparrow f$, note that for each fixed $u \in U$ it remains true that the sections $f_n(u,\cdot) \uparrow f(u,\cdot)$ and thus we can use Monotone Convergence to see that $\int f(u,t)\, \mu(s,dt) = \lim_{n\to\infty} \int f_n(u,t)\, \mu(s,dt)$ for every $(u,s) \in U \times S$ so that $\int f(u,t)\, \mu(s,dt)$ is measurable. For an arbitrary measurable function f just write $f = f_+ - f_-$ and use the result for non-negative measurable functions. ☐

Exercise 23. Suppose ξ is a random variable, let $\mathcal{F}$ be a σ-algebra and let A be a measurable set. Show that $\mathbf{E}[\xi \mid \mathcal{F}, A] = \frac{\mathbf{E}[\xi; A|\mathcal{F}]}{\mathbf{P}\{A|\mathcal{F}\}}$ 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 $\mathbf{E}[\xi \mid \mathcal{F}, A] = \mathbf{1}_A \frac{\mathbf{E}[\xi|\mathcal{F}]}{\mathbf{P}\{A|\mathcal{F}\}}$ almost surely.

Pick $F \in \mathcal{F}$ and calculate

$$\begin{aligned}
\mathbf{E}\left[\mathbf{1}_A \frac{\mathbf{E}\left[\xi \mid \mathcal{F}\right]}{\mathbf{P}\{A \mid \mathcal{F}\}}; A \cap F\right] &= \mathbf{E}\left[\mathbf{E}\left[\frac{\xi; F}{\mathbf{P}\{A \mid \mathcal{F}\}} \mid \mathcal{F}\right]; A\right] && \text{by pushout} \\
&= \mathbf{E}\left[\mathbf{E}\left[\frac{\xi; F}{\mathbf{P}\{A \mid \mathcal{F}\}} \mid \mathcal{F}\right] \mathbf{P}\{A \mid \mathcal{F}\}\right] \\
&= \mathbf{E}\left[\mathbf{E}\left[\xi; F \mid \mathcal{F}\right]\right] && \text{by pushout} \\
&= \mathbf{E}\left[\xi; F\right] = \mathbf{E}\left[\xi; A \cap F\right] && \text{by tower property}
\end{aligned}$$

and trivially

$$\mathbf{E}\left[\mathbf{1}_A \frac{\mathbf{E}\left[\xi \mid \mathcal{F}\right]}{\mathbf{P}\{A \mid \mathcal{F}\}}; A^c \cap F\right] = 0 = \mathbf{E}\left[\xi; A^c \cap F\right]$$

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. $\square$

EXERCISE 24. 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 $\mathbf{E}\left[\xi \mid \mathcal{F}\right] = \sum_{\mathbf{P}\{A_n\} \neq 0} \frac{\mathbf{E}[\xi; A_n]}{\mathbf{P}\{A_n\}} \mathbf{1}_{A_n}$ 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

$$0 = \mathbf{E}\left[\sum_{\mathbf{P}\{A_n\} \neq 0} \frac{\mathbf{E}\left[\xi; A_n\right]}{\mathbf{P}\{A_n\}} \mathbf{1}_{A_n}; A_m\right] = \mathbf{E}\left[\xi; A_m\right]$$

$\square$

EXERCISE 25. 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

$$\begin{aligned}
\mathbf{P}\{f(\xi) \in A \mid \mathcal{F}\}(\omega) &= \mathbf{E}\left[\mathbf{1}_{f^{-1}(A)}(\xi) \mid \mathcal{F}\right] \\
&= \int \mathbf{1}_{f^{-1}(A)}(s)\, d\mu(\omega, s) && \text{by Theorem 8.35} \\
&= \mu(\omega, f^{-1}(A))
\end{aligned}$$

and we are done. $\square$

EXERCISE 26. 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. $\square$

EXERCISE 27. Let ξ be an integrable random variable for which $\mathbf{E}[\xi \mid \mathcal{F}] \overset{d}{=} \xi$. Show that in fact $\mathbf{E}[\xi \mid \mathcal{F}] = \xi$ a.s.

PROOF. Here is a simple and conceptual proof in the case that $\mathbf{E}[\xi \mid \mathcal{F}]$ (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 $A_1 = \{\mathbf{E}[\xi \mid \mathcal{F}] = y_1\}$. By definition of conditional expectation $\mathbf{E}[\xi; A_1] = \mathbf{E}[\mathbf{E}[\xi \mid \mathcal{F}]; A_1] = y_1 \mathbf{P}\{A_1\}$. Because y_1 is the minimum value of ξ it follows that we must have $\xi = y_1$ identically on A_1. Since $\xi \overset{d}{=} \mathbf{E}[\xi \mid \mathcal{F}]$, 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 $f_n(\xi) \overset{d}{=} f_n(\mathbf{E}[\xi \mid \mathcal{F}])$ but not necessarily that $f_n(\xi) \overset{d}{=} \mathbf{E}[f_n(\xi) \mid \mathcal{F}]$ 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 8.34 be have a regular version ν for $\mathbf{P}\{\xi \in A \mid \mathcal{F}\}$. By Theorem 8.35 we know that $\mathbf{E}[f(\xi) \mid \mathcal{F}] = \int s \, d\mu(s)$.

Because $\xi \overset{d}{=} \mathbf{E}[\xi \mid \mathcal{F}]$ we also know that $f(\xi) \overset{d}{=} f(\mathbf{E}[\xi \mid \mathcal{F}])$ 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

$$\begin{aligned}f(\mathbf{E}[\xi \mid \mathcal{F}]) &= \mathbf{E}[f(\mathbf{E}[\xi \mid \mathcal{F}]) \mid \mathcal{F}] \\ &= \int s d\mu \circ f^{-1}(s) && \text{by Theorem 8.35} \\ &= \int f(s) \, d\mu(s) && \text{by Expectation Rule} \\ &= \mathbf{E}[f(\xi) \mid \mathcal{F}] && \text{by Theorem 8.35}\end{aligned}$$

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. $\square$

EXERCISE 28. Prove or disprove the following statement. Suppose $\xi \overset{d}{=} \eta$, 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 $\xi \overset{d}{=} \eta$ (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 $A = \{1\}$. $\square$

EXERCISE 29. Find $\xi, \eta, \mathcal{F}$ such that $\xi \overset{d}{=} \eta$ 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. $\square$

EXERCISE 30. Suppose $\xi, \tilde{\xi}$ are integrable random variables and $\eta, \tilde{\eta}$ are random elements in $(T, \mathcal{T})$ such that $(\xi, \eta) \overset{d}{=} (\tilde{\xi}, \tilde{\eta})$. Show that $\mathbf{E}[\xi \mid \eta] \overset{d}{=} \mathbf{E}\left[\tilde{\xi} \mid \tilde{\eta}\right]$.

PROOF. First, note the intuition behind the statement. As a result of $(\xi, \eta) \overset{d}{=} (\tilde{\xi}, \tilde{\eta})$ we can also conclude that $\xi \overset{d}{=} \tilde{\xi}$ and $\eta \overset{d}{=} \tilde{\eta}$. 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 8.34 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 $B = \{t \in T \mid \mu(t, A) > \tilde{\mu}(t, A)\}$.

$$\begin{aligned}
0 &= \mathbf{P}\{\xi \in A; \eta \in B\} - \mathbf{P}\{\tilde{\xi} \in A; \tilde{\eta} \in B\} && \text{by hypothesis} \\
&= \mathbf{E}\left[\int \mathbf{1}_{A \times B}(s, \eta)\, d\mu(\eta, s) - \int \mathbf{1}_{A \times B}(s, \tilde{\eta})\, d\tilde{\mu}(\eta, s)\right] && \text{by Theorem 8.35} \\
&= \mathbf{E}\left[\mathbf{1}_B(\eta)\mu(\eta, A) - \mathbf{1}_B(\tilde{\eta})\tilde{\mu}(\tilde{\eta}, A)\right] \\
&= \int \mathbf{1}_B(t)\mu(t, A) - \mathbf{1}_B(t)\tilde{\mu}(t, A)\, d\mathcal{L}(\eta)(t) && \text{by Lemma 2.55 and } \mathcal{L}(\eta) = \mathcal{L}(\tilde{\eta}).
\end{aligned}$$

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 8.35 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. $\square$

EXERCISE 31. 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 8.34 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. Conceptually, the conditional distribution captures all of the dependence between a random element and a σ-algebra (think of the case $\mathcal{F} = \sigma(\zeta)$ for a random element ζ to make this even more concrete).

By Lemma 8.20 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 32. 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 8.21, $\xi \perp\!\!\!\perp_\eta (\zeta, \gamma)$ 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 8.21 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 8.21 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 33. Suppose we have σ-algebras $\mathcal{F}$, $\mathcal{G}_1$, $\mathcal{G}_2$, $\mathcal{H}$ with $\mathcal{G}_1 \subset \mathcal{G}_2$. If $\mathcal{F} \perp\!\!\!\perp_{\mathcal{G}_1} \mathcal{H}$ is it true that $\mathcal{F} \perp\!\!\!\perp_{\mathcal{G}_2} \mathcal{H}$? 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 $\mathbf{P}\{B \mid C\} = 1/2$.

Note that primary conceptual point here is that given two independent events (here "first toss is heads" and "second toss is tails") one can condition that there is a relationship between them (here "first toss equals the second toss") and destroy independence. □

EXERCISE 34. 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 4.14. To see the lack of joint independence consider $\mathbf{P}\{A \cap B \cap C\} = 0$. □

EXERCISE 35. 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 36. 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 37. 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 38. 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 39. 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 $M^\xi_n = \mathbf{E}\left[\xi \mid \mathcal{F}_n\right]$ be the martingale generated by ξ. We claim

$$\mathbf{E}\left[\xi \mid \mathcal{F}_\sigma\right] = M^\xi_\sigma$$

TODO: Dude, this is just optional stopping (at least for the uniformly integrable case); is that supposed to be available? 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

$$\begin{aligned}\mathbf{E}\left[\mathbf{1}_A \xi\right] &= \sum_{n=0}^\infty \mathbf{E}\left[\mathbf{1}_A \mathbf{1}_{\{\sigma=n\}} \xi\right] = \sum_{n=0}^\infty \mathbf{E}\left[\mathbf{1}_A \mathbf{E}\left[\mathbf{1}_{\{\sigma=n\}} \xi \mid \mathcal{F}_n\right]\right] = \mathbf{E}\left[\mathbf{1}_A \sum_{n=0}^\infty \mathbf{E}\left[\mathbf{1}_{\{\sigma=n\}} \xi \mid \mathcal{F}_n\right]\right] \\ &= \mathbf{E}\left[\mathbf{1}_A M^\xi_\sigma\right]\end{aligned}$$

Using this formula twice we have

$$\begin{aligned}\mathbf{E}\left[\mathbf{E}\left[\xi \mid \mathcal{F}_\tau\right] \mid \mathcal{F}_\sigma\right] &= \mathbf{E}\left[M^\xi_\tau \mid \mathcal{F}_\sigma\right] \\ &= \sum_{n=0}^\infty \mathbf{1}_{\{\sigma=n\}} \mathbf{E}\left[M^\xi_\tau \mid \mathcal{F}_n\right] \\ &= \sum_{n=0}^\infty \sum_{m=0}^\infty \mathbf{1}_{\{\sigma=n\}} \mathbf{E}\left[\mathbf{E}\left[\mathbf{1}_{\{\tau=m\}} \xi \mid \mathcal{F}_m\right] \mid \mathcal{F}_n\right]\end{aligned}$$

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,

$$\begin{aligned}
\mathbf{E}\left[\mathbf{E}\left[\xi \mid \mathcal{F}_\tau\right] \mid \mathcal{F}_\sigma\right] &= \sum_{n=0}^\infty \sum_{m=n+1}^\infty \mathbf{1}_{\{\sigma=n\}}\mathbf{E}\left[\mathbf{E}\left[\mathbf{1}_{\{\tau=m\}}\xi \mid \mathcal{F}_m\right] \mid \mathcal{F}_n\right] + \\
&\sum_{m=0}^\infty \sum_{n=m}^\infty \mathbf{1}_{\{\sigma=n\}}\mathbf{E}\left[\mathbf{E}\left[\mathbf{1}_{\{\tau=m\}}\xi \mid \mathcal{F}_m\right] \mid \mathcal{F}_n\right] \\
&= \sum_{n=0}^\infty \sum_{m=n+1}^\infty \mathbf{1}_{\{\sigma=n\}}\mathbf{1}_{\{\tau=m\}}\mathbf{E}\left[\xi \mid \mathcal{F}_n\right] + \\
&\sum_{m=0}^\infty \sum_{n=m}^\infty \mathbf{1}_{\{\sigma=n\}}\mathbf{1}_{\{\tau=m\}}\mathbf{E}\left[\xi \mid \mathcal{F}_m\right] \\
&= \sum_{m=0}^\infty \sum_{n=0}^\infty \mathbf{1}_{\{\sigma=n\}}\mathbf{1}_{\{\tau=m\}}\mathbf{E}\left[\xi \mid \mathcal{F}_{m\wedge n}\right] \\
&= M^\xi_{\sigma\wedge\tau} = \mathbf{E}\left[\xi \mid \mathcal{F}_{\sigma\wedge\tau}\right]
\end{aligned}$$

□

Exercise 40. Let σ and τ be $\mathcal{F}$-optional times on either $\mathbb{Z}_+$ or $\mathbb{R}_+$. Show that $\sigma+\tau$ is $\mathcal{F}$-optional.

Proof. For the case of $\mathbb{Z}_+$ valued optional times we pick $n \geq 0$ and note that

$$\{\sigma+\tau=n\} = \cup_{m=0}^n \{\sigma=m\} \cap \{\tau = n-m\}$$

which is in $\mathcal{F}_n$ since for $0 \leq m \leq n$ we have $\{\sigma = m\} \in \mathcal{F}_m \subset \mathcal{F}_n$ and $\{\tau = n-m\} \in \mathcal{F}_{n-m} \subset \mathcal{F}_n$.

Pick $t \in$ and note that it suffices to show

$$\{\sigma+\tau > t\} = \{\sigma > t\} \cup \cup_{\substack{q<t \\ q\in\mathbb{Q}}} \{\sigma > q\} \cap \{\tau > t-q\}$$

by reasoning similar to the discrete case. To see this equality for one inclusion note that for all $q \in \mathbb{Q}$ we have $\{\sigma > q\} \cap \{\tau > t - q\} \subset \{\sigma + \tau > t\}$. By positivity of τ we know that $\{\sigma > t\} \subset \{\sigma + \tau > t\}$.

For the other inclusion suppose $\sigma(\omega) + \tau(\omega) > t$. If $\sigma(\omega) \leq t$ then by density of rationals we can pick $q \in \mathbb{Q}$ such that $t - \tau(\omega) < q < \sigma(\omega) \leq t$ and we have $\omega \in \{\sigma > q\} \cap \{\tau > t - q\}$. If $\sigma(\omega) > t$ then it follows that $\omega \in \{\sigma > t\}$ so we are done. $\square$

EXERCISE 41. Reprove the Kolmogorov Maximal Inequality Proposition 5.17 using the Doob Maximal Inequality Lemma 9.44.

PROOF. The random walk $\sum_{k=1}^n \xi_k$ is a martingale, by assumption $\mathbf{E}\left[\left(\sum_{k=1}^n \xi_k\right)^2\right] = \sum_{k=1}^n \mathbf{E}\left[\xi_k^2\right] < \infty$ thus by Proposition 9.24 we know that $\left(\sum_{k=1}^n \xi_k\right)^2$ is a submartingale therefore

$$\begin{aligned}\mathbf{P}\{\sup_{1 \leq m \leq n} \left|\sum_{k=1}^m \xi_k\right| \geq \lambda\} &= \mathbf{P}\{\sup_{1 \leq m \leq n} \left(\sum_{k=1}^m \xi_k\right)^2 \geq \lambda^2\} \\ &\leq \lambda^{-2}\mathbf{E}\left[\left(\sum_{k=1}^n \xi_k\right)^2\right] = \lambda^{-2} \sum_{k=1}^n \mathbf{E}\left[\xi_k^2\right] \leq \lambda^{-2} \sum_{k=1}^\infty \mathbf{E}\left[\xi_k^2\right]\end{aligned}$$

Now take let $n \to \infty$ and use continuity of measure. $\square$

EXERCISE 42. 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 10.7. $\square$

EXERCISE 43. Let X be a right continuous submartingale then almost surely X is cadlag.

PROOF. By Theorem 9.76 we know that there is a null set A such that the process $Z_t = \mathbf{1}_{A^c} \lim_{q \to t^+} X_q$ is a cadlag process (in fact a cadlag $\overline{\mathcal{F}}^+$-submartingale). As X is almost surely right continuous, it follows that almost surely $Z = X$ and we conclude that almost surely X has cadlag sample paths. $\square$

EXERCISE 44. Suppose we are given σ-algebras $\mathcal{G}, \mathcal{H}, \mathcal{F}_1, \mathcal{F}_2, \dots$ and define $\mathcal{F}_\infty = \bigvee_n \mathcal{F}_n$. If $\mathcal{G} \perp\!\!\!\perp_{\mathcal{F}_n} \mathcal{H}$ for all $n \in \mathbb{N}$ then $\mathcal{G} \perp\!\!\!\perp_{\mathcal{F}_\infty} \mathcal{H}$.

PROOF. By definition of conditional independence we see that for every $G \in \mathcal{G}$ and $H \in \mathcal{H}$ we have $\mathbf{P}\{G \cap H \mid \mathcal{F}_n\} = \mathbf{P}\{G \mid \mathcal{F}_n\}\mathbf{P}\{H \mid \mathcal{F}_n\}$. By Levy-Jessen Theorem 9.56 we conclude

$$\mathbf{P}\{G \cap H \mid \mathcal{F}_\infty\} = \lim_{n \to \infty} \mathbf{P}\{G \cap H \mid \mathcal{F}_n\} = \lim_{n \to \infty} \mathbf{P}\{G \mid \mathcal{F}_n\}\mathbf{P}\{H \mid \mathcal{F}_n\} = \mathbf{P}\{G \mid \mathcal{F}_\infty\}\mathbf{P}\{H \mid \mathcal{F}_\infty\}$$

which shows the result. $\square$

EXERCISE 45. Let B_t be a standard Brownian motion and define $\tau = \inf\{t > 0 \mid B_t = 1\}$. Show that $B_\tau = 1$ almost surely and $\mathbf{E}\left[\tau^c\right] < \infty$ for all $0 \leq c < 1/2$.

PROOF. Note that $\tau < \infty$ almost surely since $\limsup_{t\to\infty} B_t = \infty$ almost surely and B_t is continuous. For any $\lambda \geq 0$ note that $\{\tau \geq t\} = \{\sup_{0\leq s\leq t} B_s \leq 1$. Since the law of $\sup_{0\leq s\leq t} B_s$ is the same as the law of $|B_t|$ by Lemma 3.8 we get

$$\begin{aligned}\mathbf{E}\left[\tau^c\right] &= c^{-1}\int_0^\infty t^{c-1}\mathbf{P}\{\tau \geq t\}\,dt = \frac{2}{c\sqrt{2\pi}}\int_0^\infty \int_0^{1/\sqrt{t}} t^{c-1}e^{-x^2/2}\,dx\,dt \\ &= \frac{2}{c\sqrt{2\pi}}\int_0^\infty \int_0^{1/x^2} t^{c-1}e^{-x^2/2}\,dt\,dx = \frac{2}{\sqrt{2\pi}}\int_0^\infty x^{-2c}e^{-x^2/2}\,dx\end{aligned}$$

For $0 \leq c < 1/2$ an integration by parts shows that this integral is finite.

Note also that integration by parts also shows that $\mathbf{E}\left[\tau^c\right] = \infty$ for $c \geq 1/2$ (as we know must be true because of the BDG/Optional Stopping argument above). $\square$

EXERCISE 46. Let B_t be a standard Brownian motion show that for every $c \in \mathbb{R}$ the process $M_t = e^{cB_t - \frac{c^2 t}{2}}$ is a martingale.

PROOF. Adaptedness follows from the fact that B_t is $\mathcal{F}_t$-measurable and e^x is continuous hence Borel measurable. First to see that M_t is integrable we compute by Lemma 3.7 and completing the square

$$\mathbf{E}\left[e^{cB_t}\right] = \frac{1}{\sqrt{2\pi t}}\int_{-\infty}^\infty e^{cx}e^{-x^2/2t}\,dx = \frac{e^{c^2t/2}}{\sqrt{2\pi t}}\int_{-\infty}^\infty e^{-(x-ct)^2/2t}\,dx = e^{c^2t/2} < \infty$$

If we take $0 \leq s < t < \infty$ then using the pullout rule of conditional expectation, the fact that $B_t - B_s$ is independent of $\mathcal{F}_s$ and the above computation of the expectation to see that

$$\begin{aligned}\mathbf{E}\left[e^{cB_t - \frac{c^2t}{2}} \mid \mathcal{F}_s\right] &= e^{-\frac{c^2t}{2}}\mathbf{E}\left[e^{c(B_t - B_s)}e^{cB_s} \mid \mathcal{F}_s\right] = e^{-\frac{c^2t}{2}}\mathbf{E}\left[e^{c(B_t-B_s)}\right]e^{cB_s} \\ &= e^{-\frac{c^2t}{2}}e^{\frac{c^2(t-s)}{2}}e^{cB_s} = e^{cB_s - \frac{c^2s}{2}}\end{aligned}$$

$\square$

EXERCISE 47. Let B_t be a standard Brownian motion, show that $\inf\{t > 0 \mid B_t > 0\} = 0$ a.s. (Hint: Use Blumenthal's 0-1 Law).

PROOF. Let $\tau = \inf\{t > 0 \mid B_t > 0\}$. Clearly the event

$$\{\tau = 0\} = \cap_{n=1}^\infty \cup_{\substack{0<q<1/n \\ q\in\mathbb{Q}}} \{B_q > 0\}$$

is $\mathcal{F}_0^+$-measurable so by Lemma 12.18 we know that it has probability 0 or 1. It therefore suffices to show that $\mathbf{P}\{\tau = 0\} \neq 0$. To see this note that for each $s > 0$ we have $\{\tau \leq s\} \supset \{B_s > 0\}$ hence $\mathbf{P}\{\tau \leq s\} \geq 1/2$ and by continuity of measure we know that $\mathbf{P}\{\tau = 0\} = \lim_{s\downarrow 0}\mathbf{P}\{\tau \leq s\} \geq 1/2$. $\square$

EXERCISE 48 (Law of Large Numbers for Brownian Motion). Let B_t be a standard Brownian motion show that $M_t = t^{-1}B_t$ is a backward martingale. From this conclude that $t^{-1}B_t \overset{a.s.}{\to} 0$ and $t^{-1}B_t \overset{L^p}{\to} 0$ for all $p > 0$.

PROOF. Adaptedness and integrability are immediate. For the backward martingale property, let $s < t$ and we first find the density function for the conditional distribution $\mathbf{P}\{B_s \in \cdot \mid B_t\}$. To find the joint density (B_s, B_t) we note that

$(B_s, B_t) = (x, y)$ if and only if $(B_s, B_t - B_s) = (x, y - x)$ so by the independence of B_s and $B_t - B_s$ and completing the square we get

$$\begin{aligned}\mathbf{P}\{B_s = x; B_t = y\} &= \frac{1}{\sqrt{2\pi s}} e^{-x^2/2s} \frac{1}{\sqrt{2\pi(t-s)}} e^{-(y-x)^2/2(t-s)} \\ &= \frac{1}{\sqrt{2\pi s}} \frac{1}{\sqrt{2\pi(t-s)}} e^{-\frac{t}{2s(t-s)}(x - \frac{s}{t}y)^2} e^{-y^2/2t}\end{aligned}$$

So we see that the conditional density B_s given B_t is Gaussian with mean $\frac{s}{t} B_t$. Thus $\mathbf{E}\left[s^{-1} B_s \mid B_t\right] = t^{-1} B_t$. By the extended Markov property (Lemma 13.2) we know that $B_s \perp\!\!\!\perp_{B_t} \bigvee_{u \geq t} \sigma(B_u)$ and therefore $\mathbf{E}\left[s^{-1} B_s \mid B_t\right] = \mathbf{E}\left[s^{-1} B_s \mid \bigvee_{u \geq t} \sigma(B_u)\right]$ (Lemma 8.20) which shows the backward martingale property.

Now we need to show that $\cap_{t>0} \bigvee_{u \geq t} \sigma(B_u)$ is a trivial σ-field (Lemma 12.18); from that it follows that for all $t > 1$,

$$t^{-1} B_t = \mathbf{E}\left[B_1 \mid \bigvee_{u \geq t} \sigma(B_u)\right]$$

and by Jessen-Levy and triviality we have

$$\mathbf{E}\left[B_1 \mid \bigvee_{u \geq t} \sigma(B_u)\right] \xrightarrow{a.s.} \mathbf{E}\left[B_1 \mid \cap_{t>0} \bigvee_{u \geq t} \sigma(B_u)\right] = \mathbf{E}\left[B_1\right] = 0$$

TODO: Finish the L^p argument; presumably we need L^p boundedness. ∎

Exercise 49 (Kallenberg Exercise 13.19). Let X be a Brownian bridge, show that $(1-t)^{-1} X_t$ is a martingale with respect to the induced filtration and is not L^1 bounded on $[0, 1)$.

Proof. Since X_t is Gaussian with variance $t - t^2$ we have

$$\frac{1}{1-t} \mathbf{E}\left[|X_t|\right] = \frac{1}{1-t} \frac{2}{\sqrt{2\pi(t-t^2)}} \int_0^\infty x e^{-x^2/2(t-t^2)}\, dx = \sqrt{\frac{2t}{\pi(1-t)}}$$

which shows that $(1-t)^{-1} X_t$ is integrable on $[0, 1)$ but not L^1 bounded.

Let $\mathcal{F}_t = \sigma((1-s)^{-1} X_s; 0 \leq s \leq t)$. If we let B be a Brownian motion then for $0 \leq s < t < 1$ we have

$$\frac{1}{(1-t)(1-s)} \mathbf{E}\left[X_t X_s\right] = \frac{s(1-t)}{(1-t)(1-s)} = \frac{s}{1-s}$$

which is does not depend on t. Since $(1-t)^{-1} X_t$ is Gaussian this implies that for every $0 \leq r_1 < \cdots < r_m \leq s < t < 1$ we have $((1-r_1)^{-1} X_{r_1}, \ldots, (1-r_m)^{-1} X_{r_m}, (1-s)^{-1} X_s, (1-t)^{-1} X_t)$ is a Gaussian random vector and the same is true of $((1-r_1)^{-1} X_{r_1}, \ldots, (1-r_m)^{-1} X_{r_m}, (1-t)^{-1} X_t - (1-s)^{-1} X_s)$ (Example 7.20). By Proposition 7.21 we know that $(1-t)^{-1} X_t - (1-s)^{-1} X_s \perp\!\!\!\perp ((1-r_1)^{-1} X_{r_1}, \ldots, (1-r_m)^{-1} X_{r_m})$ and thus for all $0 \leq s < t < 1$ by Lemma 4.13 we have $(1-t)^{-1} X_t - (1-s)^{-1} X_s \perp\!\!\!\perp \mathcal{F}_s$. The martingale property follows in the standard way,

$$\begin{aligned}\mathbf{E}\left[(1-t)^{-1} X_t \mid \mathcal{F}_s\right] &= \mathbf{E}\left[(1-t)^{-1} X_t - (1-s)^{-1} X_s \mid \mathcal{F}_s\right] + \mathbf{E}\left[(1-s)^{-1} X_s \mid \mathcal{F}_s\right] \\ &= \mathbf{E}\left[(1-t)^{-1} X_t - (1-s)^{-1} X_s\right] + (1-s)^{-1} X_s = (1-s)^{-1} X_s\end{aligned}$$

□

EXERCISE 50. Let X be a Markov process in $(S,\mathcal{S})$ on time scale T with transition kernel $\mu_{s,t}$ and let Y be a Markov process in $(U,\mathcal{U})$ on time scale T with transition kernel $\nu_{s,t}$. Show that if X and Y are independent that (X,Y) is a Markov process in $(S\times U, \mathcal{S}\otimes\mathcal{U})$ on time scale T with transition kernel $\mu_{s,t}\otimes\nu_{s,t}$ (note that kernel $\mu_{s,t}\otimes\nu_{s,t}$ is just the pointwise product measure).

PROOF. TODO: Finish

Pick $t < u \in T$. Let $C \in \mathcal{S}$ and $D \in \mathcal{U}$ and compute using the independence of X and Y, let

$$\begin{aligned}
&\mathbf{P}\{(X_u,Y_u) \in A\times B; (X_t,Y_t)\in C\times D\}\\
&= \mathbf{P}\{X_u \in A; X_t \in C\}\mathbf{P}\{Y_u \in B; Y_t \in D\}\\
&= \mathbf{E}\left[\mathbf{P}\{X_t \mid X_u \in A\}; X_t \in C\right]\mathbf{E}\left[\mathbf{P}\{Y_t \mid Y_u \in B\}; Y_t \in D\right]\\
&= \mathbf{E}\left[\mathbf{P}\{X_t \mid X_u \in A\}; X_t \in C; \mathbf{P}\{Y_t \mid Y_u \in B\}; Y_t \in D\right]
\end{aligned}$$

which since sets of the form $(X_t,Y_t)\in C\times D$ are a generating π-system of $\sigma(X_t,Y_t)$ we the claim is shown by Lemma 8.8.

Finally we conclude that $\mathbf{P}\{(X_t,Y_t) \mid (X_u,Y_u)\in\cdot\} = \mu_{t,u}\otimes\nu_{t,u}$ by the uniqueness of product measure. □

EXERCISE 51 (Kallenberg Exercise 17.1). Show that if M is a local martingale and ξ is a $\mathcal{F}_0$-measurable random variable then $N_t = \xi M_t$ is also a local martingale.

PROOF. Let τ_n be a localizing sequence for $(M - M_0)$ so that $(M-M_0)^{\tau_n}$ is a martingale for all $n \in \mathbb{N}$ and $\tau_n \uparrow \infty$ almost surely. Let $\sigma_n = \tau_n \mathbf{1}_{|\xi|\leq n}$. Since ξ is almost surely finite then it follows that $\sigma_n \uparrow \infty$ almost surely (specifically on the intersection of the event that $\tau_n \uparrow \infty$ and $|\xi| < \infty$). Moreover since $\{\sigma_n \leq t\} = \{\tau_n \leq t\} \cup \{|\xi| \leq n\}$ and ξ is $\mathcal{F}_0$-measurable it follows that σ_n are $\mathcal{F}$-optional times. Now note that ξM is $\mathcal{F}$-adapted since ξ is $\mathcal{F}_0$-measurable and for each $n\in\mathbb{N}$ and $0 \leq t < \infty$ we have

$$\mathbf{E}\left[|\xi M_{t\wedge\sigma_n} - \xi M_0|\right] = \mathbf{E}\left[|\xi|\, |M_{t\wedge\tau_n} - M_0|\,; |\xi| \leq n\right] \leq n\mathbf{E}\left[(M-M_0)^{\tau_n}\right] < \infty$$

and moreover by the pullout rule of conditional expectation and the fact that τ_n localizes M,

$$\begin{aligned}
\mathbf{E}\left[(\xi M_{\sigma_n\wedge t} - \xi M_0) \mid \mathcal{F}_s\right] &= \mathbf{E}\left[\xi(M_{\sigma_n\wedge t} - M_0); |\xi| \leq n \mid \mathcal{F}_s\right]\\
&= \xi\mathbf{1}_{|\xi|\leq n}\mathbf{E}\left[M_{\tau_n\wedge t} - M_0 \mid \mathcal{F}_s\right]\\
&= \xi\mathbf{1}_{|\xi|\leq n}(M_{\tau_n\wedge s} - M_0)\\
&= \xi M_{\sigma_n\wedge s} - \xi M_0
\end{aligned}$$

showing that $(M-M_0)^{\sigma_n}$ is a martingale. □

EXERCISE 52 (Kallenberg Exercise 17.2). Show that a local martingale M with $M \geq 0$ for all $0 \leq t < \infty$ almost surely and $\mathbf{E}[M_0] < \infty$ is a supermartingale. Give an example to show that M is not necessarily a martingale.

PROOF. Let τ_n be a localizing sequence for M. We know that $(M-M_0)^{\tau_n}$ is a martingale but by the integrability and $\mathcal{F}_0$-measurability of M_0 we see that in fact M^{τ_n} is a martingale. Now since $M \geq 0$, $M_{\tau_n \wedge t} \overset{a.s.}{\to} M_t$ for all $0 \leq t < \infty$ and Fatou's Lemma for conditional expectations we have for all $0 \leq s < t < \infty$

$$\mathbf{E}\left[M_t \mid \mathcal{F}_s\right] \leq \liminf_{n\to\infty} \mathbf{E}\left[M_{\tau_n \wedge t} \mid \mathcal{F}_s\right] = \liminf_{n\to\infty} M_{\tau_n \wedge t} = M_s$$

almost surely. Choosing $s = 0$ and taking expectations shows that $\mathbf{E}\left[M_t\right] \leq \mathbf{E}\left[M_0\right] < \infty$ and therefore M is a supermartingale.

TODO: Do the last part. $\square$

Exercise 53 (Kallenberg Exercise 17.4). Let M_n be a sequence of continuous local martingales starting at zero and let τ_n be a sequence of optional times, then $(M_n^*)_{\tau_n} \overset{P}{\to} 0$ if and only if $[M_n]_{\tau_n} \overset{P}{\to} 0$. State and prove a corresponding result for stochastic integrals.

Proof. Define $\tilde{M}_n = M_n^{\tau_n}$ and apply Lemma 14.28 to conclude that $\tilde{M}_n^* \overset{P}{\to} 0$ if and only if $[\tilde{M}_n]_\infty \overset{P}{\to} 0$. Now note that

$$\tilde{M}_n^* = \sup_{0 \leq t < \infty} \left|(\tilde{M}_n)_t\right| = \sup_{0 \leq t < \infty} |(M_n)_{\tau_n \wedge t}| = (M_n)^*_{\tau_n}$$

and by Theorem 14.26 we get

$$[\tilde{M}_n]_\infty = [M_n^{\tau_n}]_\infty = [M_n]^{\tau_n}_\infty = [M_n]_{\tau_n}$$

The corresponding result for stochastic integrals says that given continuous local martingales M_n, processes $V_n \in L(M_n)$ and optional times τ_n we have $\left(\int V_n \, dM_n\right)^*_{\tau_n} \overset{P}{\to} 0$ if and only if $\int_0^{\tau_n} V_n^2(s)\, d[M_n](s) \overset{P}{\to} 0$. This follows from what we have just proven and Lemma 14.38 to see

$$[\int V_n \, dM_n]_{\tau_n} = \int_0^{\tau_n} V_n(s)\, d[M_n](s)$$

$\square$

Exercise 54 (Kallenberg Exercise 17.6). Let B_t be a Brownian motion starting at zero and τ be an optional time. Show that $\mathbf{E}\left[\tau^{1/2}\right] < \infty$ implies $\mathbf{E}\left[B_\tau\right] = 0$ and $\mathbf{E}\left[\tau\right] < \infty$ implies $\mathbf{E}\left[B_\tau^2\right] = \mathbf{E}\left[\tau\right]$.

Proof. For τ bounded by a constant T these are both consequences of Optional Stopping. For then since B_t is a martingale we have

$$\mathbf{E}\left[B_\tau\right] = \mathbf{E}\left[\mathbf{E}\left[B_T \mid \mathcal{F}_\tau\right]\right] = \mathbf{E}\left[B_T\right] = 0$$

and since $B_t^2 - t$ is a martingale we have

$$\mathbf{E}\left[B_\tau^2\right] - \mathbf{E}\left[\tau\right] = \mathbf{E}\left[\mathbf{E}\left[B_T^2 - T \mid \mathcal{F}_\tau\right]\right] = \mathbf{E}\left[B_T^2 - T\right] = 0$$

Now consider the sequence of bounded optional times $\tau \wedge n$. If we have $\mathbf{E}\left[\tau^{1/2}\right] < \infty$ the we can apply the BDG inequality (Lemma 14.32) to the stopped process B^τ (which is a priori only a continuous local martingale) to see that there is a constant $c_1 > 0$ such that $\mathbf{E}\left[\sup_{0 \leq t \leq \tau} |B_t|\right] \leq c\mathbf{E}\left[\tau^{1/2}\right] < \infty$ and therefore since $|B_{\tau \wedge n}| \leq \sup_{0 \leq t \leq \tau} |B_t|$ we can use Dominated Convergence to see that $\mathbf{E}\left[B_\tau\right] = \lim_{n\to\infty} \mathbf{E}\left[B_{\tau \wedge n}\right] = 0$. Similarly when $\mathbf{E}\left[\tau\right] < \infty$, we get a constant $c_2 > 0$ such that $\mathbf{E}\left[\sup_{0 \leq t \leq \tau} \left|B_t^2\right|\right] \leq c_2 \mathbf{E}\left[\tau\right] < \infty$ and therefore Dominated Convergence gives us

$$\mathbf{E}\left[B_\tau^2\right] = \lim_{n\to\infty} \mathbf{E}\left[B_{\tau\wedge n}^2\right] = \lim_{n\to\infty} \mathbf{E}\left[\tau \wedge n\right] = \mathbf{E}\left[\tau\right]$$

While we're at it, we can provide a different proof that $\mathbf{E}[B_\tau] = 0$ under the weaker assumption $\mathbf{E}[\tau] < \infty$ that doesn't rely on the BDG inequalities. As above, it suffices to show that $|B_{\tau\wedge t}|$ is dominated by an integrable random variable. The proof here is taken from Peres and Morters. For each integer $k \geq 0$ consider

$$M_k = \sup_{0 \leq t \leq 1} |B_{t+k} - B_k|$$

TODO: Finish □

EXERCISE 55 (Kallenberg Exercise 17.8). Let X and Y be continuous semimartingales show that $[X+Y]_t^{1/2} \leq [X]_t^{1/2} + [Y]_t^{1/2}$ for all $0 \leq t < \infty$ almost surely.

PROOF. From bilinearity and the Cauchy Schwartz inequality Lemma 14.34 we have

$$[X+Y]_t = [X]_t + 2[X,Y]_t + [Y]_t \leq [X]_t + 2\,|[X,Y]_t| + [Y]_t \leq [X]_t + 2[X]_t^{1/2}[Y]_t^{1/2} + [Y]_t = ([X]_t^{1/2} + [Y]_t^{1/2})^2$$

for all $0 \leq t < t$ almost surely. The result follows by noting the non-negativity of quadratic variation and taking square rooots. □

EXERCISE 56 (Kallenberg Exercise 17.11). Let X be a continuous semimartingale and let $U, V \in L(X)$ be such that $U = V$ a.s. on a set $A \in \mathcal{F}_0$. Use Lemma 14.42 to show that $\int U\,dX = \int V\,dX$ a.s on A.

PROOF. Define

$$\tau(\omega) = \begin{cases} \infty & \text{where } \omega \in A \\ 0 & \text{when } \omega \notin A \end{cases}$$

and note that τ is an optional time since $A \in \mathcal{F}_0$. Also note that

$$\mathbf{1}_{[0,\tau(\omega)]}(t) \cdot U_t(\omega) = \begin{cases} U_t(\omega) & \text{when } \omega \in A \text{ or } \omega \notin A \text{ and } t = 0 \\ 0 & \text{when } \omega \notin A \text{ and } t > 0 \end{cases}$$

and similarly with V and therefore we conclude $\mathbf{1}_{[0,\tau]} \cdot U = \mathbf{1}_{[0,\tau]} \cdot V$ almost surely and therefore $\int \mathbf{1}_{[0,\tau]} U\,dM = \int \mathbf{1}_{[0,\tau]} V\,dM$ almost surely by Lemma 14.38. From Lemma 14.42 and the fact that $\int U\,dX$ starts at zero we get almost surely

$$\mathbf{1}_A \int_0^t U\,dM = \mathbf{1}_A \int_0^t U\,dM + \mathbf{1}_{A^c} \int_0^0 U\,dM = \int_0^{t\wedge\tau} U\,dM = \int_0^t \mathbf{1}_{[0,\tau]} U\,dM$$

and similarly with V and the result follows. □

EXERCISE 57 (Kallenberg Exercise 17.13). Let X be Brownian bridge then $X_{t\wedge 1}$ is a semimartingale.

PROOF. We know from Exercise 49 that $Y_t = (1-t)^{-1}X_t$ is a martingale on $[0,1)$. Now apply the local Ito Lemma to the semimartingale $(1 \wedge t, (1-t)^{-1}X_t)$ using the function $f(t,x) = (1-t)x$ on the domain $(-\infty, 1) \times \mathbb{R}$ to see that

$$X_{t\wedge 1} = f(t, Y_t) - f(0, Y_0) = \int_0^t (1-s)\,dY_s - \int_0^t Y_s\,ds$$

where the first stochastic integral is continuous local martingale and the second random Stieltjes integral is of finite variation.

TODO: Get all the details around the local Ito stuff spelled out. □

EXERCISE 58. Let M be a continuous local martingale in $\mathbb{R}^r$ and let $V \in L(X)$ be an $d \times r$ matrix valued process in $L(M)$, then

$$\left[\left(\int V_s\, dM_s\right)^{(i)}, \left(\int V_s\, dM_s\right)^{(i)}\right]_t = \sum_{k=1}^r \sum_{l=1}^r \int_0^t V_s^{ik} V_s^{jl}\, d[M^k, M^l]_s \text{ for all } 1 \leq i, j \leq d$$

which we write stylistically as $[\int V\, dM] = \int V d[M] V^T$.

PROOF. This just follows from bilinearity of quadratic covariation and Lemma 14.38

$$\begin{aligned}\left[\left(\int V_s\, dM_s\right)^{(i)}, \left(\int V_s\, dM_s\right)^{(i)}\right]_t &= \left[\sum_{k=1}^r \int V_s^{ik}\, dM_s^k, \sum_{l=1}^r \int V_s^{jl}\, dM_s^l\right]_t \\ &= \sum_{k=1}^r \sum_{l=1}^r \left[\int V_s^{ij}\, dM_s^k, \int V_s^{jl}\, dM_s^l\right]_t \\ &= \sum_{k=1}^r \sum_{l=1}^r \int V_s^{ik} V_s^{jl}\, d[M^k, M^l]_s\end{aligned}$$

□

EXERCISE 59. Let (S, d) be a totally bounded metric space. Prove that S is separable and that every uniformly continuous function $f : S \to \mathbb{R}$ is bounded.

PROOF. For each $n \in \mathbb{Z}$ we may find a finite set $x_1^n, \dots, x_{m_n}^n$ such that $B(x_j^n; 1/n)$ covers S. The union of x_j^n for $n \in \mathbb{Z}$ and $j = 1, \dots, m_n$ is a countable dense subset of S. Assume that f is uniformly continuous, pick an $\epsilon > 0$ such that $d(x, y) < \epsilon$ implies $|f(x) - f(y)| < 1$. Now pick $x_1, \dots, x_n$ such that $B(x_1; \epsilon), \dots, B(x_n; \epsilon)$ covers S. Let $K = |f(x_1)| \vee \cdots \vee |f(x_n)|$ and note that for every x we may pick x_j such that $x \in B(x_j; \epsilon)$ and therefore $|f(x)| \leq K + 1$. □

EXERCISE 60. Let (S, r) and (T, r') be metric spaces and $g : S \to T$ be a continuous. Define $g_* : D([0, \infty); S) \to D([0, \infty); T)$ by $g_*(f)(t) = g(f(t))$ then g_* is continuous in the J_1 topology.

PROOF. Since the J_1 topology is metrizable it suffices to show that g_* takes convergence sequences convergent sequences. Let $f, f_1, f_2, \ldots \in D([0, \infty); S)$ be such that $f_n \to f$. Then by Proposition 17.26 we know that for every $T > 0$ and every $t, t_1, t_2, \ldots \in [0, T]$ with $\lim_{n\to\infty} t_n = t$ we have

(i) $\lim_{n\to\infty} r(f_n(t_n), f(t)) \wedge r(f_n(t_n), f(t-)) = 0$.
(ii) If $\lim_{n\to\infty} r(f_n(t_n), f(t)) = 0$ then for every sequence s_n such that $t_n \leq s_n \leq T$ and $\lim_{n\to\infty} s_n = t$ we have $\lim_{n\to\infty} r(f_n(s_n), f(t)) = 0$
(iii) If $\lim_{n\to\infty} r(f_n(t_n), f(t-)) = 0$ then for every sequence s_n such that $0 \leq s_n \leq t_n$ and $\lim_{n\to\infty} s_n = t$ we have $\lim_{n\to\infty} r(f_n(s_n), f(t-)) = 0$

Using continuity of g at $f(t)$ and $f(t-)$ for every $\epsilon > 0$ there exists $\delta > 0$ such that $r(x, f(t)) < \delta$ implies $r'(g(x), g(f(t))) < \epsilon$ and $r(x, f(t-)) < \delta$ implies $r'(g(x), g(f(t-))) < \epsilon$. By (i) we may find N such that for all $n \geq N$ we have $r(f_n(t_n), f(t)) \wedge r(f_n(t_n), f(t-)) < \delta$; thus $r'(g(f_n(t_n)), g(f(t))) \wedge r'(g(f_n(t_n)), g(f(t-))) < \epsilon$ which shows $\lim_{n\to\infty} r'(g(f_n(t_n)), g(f(t))) \wedge r'(g(f_n(t_n)), g(f(t-))) = 0$.

Also that $\lim_{s\to t-} g(f(s)) = g(f(t-))$ since g is continuous at $f(t-)$: given $\epsilon > 0$ we take $\delta > 0$ such that $r(x, f(t-)) < \delta$ implies $r'(g(x), g(f(t-))) < \epsilon$; now take $\rho > 0$ such that $t - \rho < s < t$ implies $r(f(s), f(t-)) < \delta$. For the moment we fix $T > 0$ and $0 \leq t \leq T$. We consider two cases separately.

Case 1: $g(f(t)) = g(f(t-))$. In this conditions (i), (ii) and (iii) reduce to the assertion that for all $t_n \to t$ we have $g(f_n(t_n)) \to g(f(t))$. From the continuity of g given $\epsilon > 0$ then $g^{-1}(B(g(f(t)), \epsilon))$ is open in S and contains both $f(t)$ and $f(t-)$ (it doesn't matter whether $f(t) = f(t-)$ or not). We may find a $\delta > 0$ such that $B(f(t), \delta) \subset g^{-1}(B(g(f(t)), \epsilon))$ and $B(f(t-), \delta) \subset g^{-1}(B(g(f(t)), \epsilon))$. By the property (i) of f and the f_n we may find $N > 0$ such that $f_n(t_n) \in B(f(t), \delta) \cup B(f(t-), \delta)$ for all $n \geq N$ and we are done. I think this even easier because for this case (i) is equivalent to (i), (ii) and (iii) for $g_*(f)$ and $g_*(f_n)$ and we have already shown that (i) implies (i).

Case 2: $g(f(t)) = g(f(t-))$ and $f(t) \neq f(t-)$. We already know that (i) holds for $g_*(f)$ and $g_*(f_n)$. Suppose $t_n \to t$ and $g(f_n(t_n)) \to g(f(t))$. Then because $f_n \to f$ we know that $r(f_n(t_n), f(t)) \wedge r(f_n(t_n), f(t-)) \to 0$. If it is not true that $r(f_n(t_n), f(t)) \to 0$ then we can find a subsequence n_k such that $f_{n_k}(t_{n_k}) \to f(t-)$ but by continuity of g we conclude $g(f_{n_k}(t_{n_k})) \to g(f(t-))$ which is a contradiction (recall that $g(f(t-)) = \lim_{s \to t^-} g \circ f(s)$ by continuity of g). Therefore we get $f_n(t_n) \to f(t)$ and for any $s_n \geq t_n$ with $s_n \to t$ we conclude $f_n(s_n) \to f(t)$ hence $g(f_n(t_n)) \to g(f(t))$ by continuity of g. By a similar argument, if we assume that $g(f_n(t_n)) \to g(f(t-))$ then we conclude that $f_n(t_n) \to f(t-)$ and therefore for all $s_n \leq t_n$ with $s_n \to t$ we have $g(f_n(s_n)) \to g(f(t-))$ and thus (iii) holds.

Now we apply Proposition 17.26 in the opposite direction to conclude that $g \circ f_n \to g \circ f$ in the J_1 topology. □

EXERCISE 61. Define $\psi : D([0, \infty); \mathbb{R}) \to D([0, \infty); \mathbb{R})$ by $\psi(f)(t) = \sup_{0 \leq s \leq t} f(s)$. Show that ψ is continuous in the J_1 topology.

PROOF. $\psi(f)$ is non-decreasing and finite therefore it cadlag. To see continuity suppose that $f_n \to f$ in the J_1 topology. Pick λ_n such that $\gamma(\lambda_n) \to 0$ and $\lim_{n \to \infty} \sup_{0 \leq t \leq T} |f_n(t) - f(\lambda_n(t))| = 0$ for all $T > 0$. Fix $T > 0$ and note that for any $\delta > 0$, $n \in \mathbb{N}$ and $0 \leq t \leq T$ we may pick $0 \leq w_n, u_n \leq t \leq T$ such that $\sup_{0 \leq s \leq t} f_n(s) \leq f_n(u_n) + \delta$ and $\sup_{0 \leq s \leq t} f(\lambda_n(s)) \leq f(\lambda_n(w_n)) + \delta$. From these two inequalities and the fact that $\psi(f)(\lambda_n(t)) = \sup_{0 \leq s \leq \lambda_n(t)} f(s) = \sup_{0 \leq s \leq t} f(\lambda_n(s))$, we get

$$\begin{aligned}
\psi(f_n)(t) &\leq f_n(u_n) + \delta \leq f(\lambda_n(u_n)) + \delta + \sup_{0 \leq t \leq T} |f_n(t) - f(\lambda_n(t))| \\
&\leq \sup_{0 \leq s \leq t} f(\lambda_n(s)) + \delta + \sup_{0 \leq t \leq T} |f_n(t) - f(\lambda_n(t))| \\
&= \psi(f)(\lambda_n(t)) + \delta + \sup_{0 \leq t \leq T} |f_n(t) - f(\lambda_n(t))|
\end{aligned}$$

and

$$\begin{aligned}
\psi(f)(\lambda_n(t)) &\leq f(u) + \delta \leq f_n(w_n) + \delta + \sup_{0 \leq t \leq T} |f_n(t) - f(\lambda_n(t))| \\
&\leq \sup_{0 \leq s \leq t} f_n(s) + \delta + \sup_{0 \leq t \leq T} |f_n(t) - f(\lambda_n(t))| \\
&= \psi(f_n)(t) + \delta + \sup_{0 \leq t \leq T} |f_n(t) - f(\lambda_n(t))|
\end{aligned}$$

thus $\sup_{0 \leq t \leq T} |\psi(f_n)(t) - \psi(f)(\lambda_n(t))| \leq \delta + \sup_{0 \leq t \leq T} |f_n(t) - f(\lambda_n(t))|$ and the result follows by taking the limit as $n \to \infty$ and then letting $\delta \to 0$. □

EXERCISE 62. Suppose that T_t is a strongly continuous contraction semigroup for which $T_t v$ is differentiable for all $v \in X$, show that T_t is of the form e^{tA} for a bounded operator $A : X \to X$.

PROOF. TODO: I think on uses the closed graph theorem to show that the generator is bounded, then use the fact that generators uniquely determine the semigroup (Corollary 18.28). $\square$

APPENDIX A

Techniques

This section is a place to collect some of the recurring proof techniques that one should be familiar with.

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 2.55 and Lemma 2.57.

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 2.71 and Lemma 4.13.

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 4.31 and Lemma 5.10.

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 5.23 and Lemma 5.10.

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.

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 3.8. A good example of how to do this is with Lemma 10.7.

TODO: Passing from L^p convergence to almost sure convergence. Note that we easily get almost sure convergence along a subsequence.

4. Proving Inequalities

4.1. Using Calculus. If one wants to show that $f(x) \geq 0$ on an interval $[a, b]$ one of the easiest ways to show the inequality is to find the minimum of $f(x)$ on $[a, b]$ and to show this value is bigger than zero. Finding the minimum is a lot easier if $f(x)$ is differentiable. A common special case one can easily show $f(a) \geq 0$ and $f(b) \geq 0$ and show that $f(x)$ is increasing or decreasing on $[a, b]$ by showing $f'(x)$ is positive or negative. The problem with this technique is that it is really a proof technique and requires that one knows the answer beforehand (e.g. one usually wants to show $g(x) \leq h(x)$ and the bound $h(x)$ is what you are trying to figure out). Sometimes the technique can be used to guess the answer by taking a simpler known inequality and antidifferentiating (see Lemma 7.23 for a non-trivial example).

4.2. Using Taylor's Theorem. Taylor's Theorem is also a good way of both guessing and proving inequalities; if one can show that the remainder term (in either integral or Lagrange form usually) is of a particular sign over an interval an inequality follows.

APPENDIX B

Integrals

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

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

APPENDIX C

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 C.1. *The following inequalities hold:*

(i) $1 + x \leq e^x$ *for all* $x \in \mathbb{R}$.
(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* $x \in \mathbb{R}$.
(v) $|\sin(x)| < |x|$ *for all* $x \neq 0$.
(vi) $1 - \frac{x^2}{2} \leq \cos(x)$ *for all* $x \in \mathbb{R}$.
(vii) $x + \log(1 - x) \leq 0$ *for all* $x \in [0, 1)$.
(viii) $e^{-x} \leq 1 - (1 - e^{-1})x$ *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. Alternative this can be seen by Taylor's Theorem. One writes using the Lagrange form of the remainder $e^x = 1 + x + \frac{x^2}{2} e^c$ for some c. Since the remainder is positive the result follows.

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 see (v), because the function $x - \sin(x)$ is odd, it suffices to show that it is strictly positive for $x > 0$. Clearly $x - \sin(x) > 0$ for $x > 1$. For $0 < x < 1$ we just use Taylor's Theorem with Lagrange remainder to see that $\sin(x) = x - \frac{x^2}{2}\cos(c)$ for some $0 < c < x < 1$. The remainder is negative so the result follows.

To show (vi), 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 $x = 0$ by (v) and moreover $f'(x)$ changes sign at $x = 0$ which shows that $f(0) = 0$ is a strict global minimum.

To show (vii), 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 (viii), 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: $f(0) = f(1) = 0$. □

When dealing with characteristic functions, it is useful to have estimates for the function e^{ix}. We collect a few useful ones here.

Theorem C.2. *The following inequalities hold:*

(i) $\left|e^{ix} - 1 - ix\right| \leq \frac{x^2}{2}$ *for all* $x \in \mathbb{R}$.
(ii) $\left|e^{ix} - 1 - ix + \frac{x^2}{2}\right| \leq x^2 R(x)$ *for all* $x \in \mathbb{R}$ *where* $|R(x)| \leq 1$ *and* $\lim_{x\to 0} R(x) = 0$.

Proof. To see (i) we use Taylor's Theorem with the Lagrange form of the remainder to write $e^{ix} - 1 - ix = -\frac{x^2}{2}e^{ic}$ for some $c \in \mathbb{R}$. Now take absolute values and use the fact that $\left|e^{ic}\right| = 1$.

To see (i) we use Taylor's Theorem with the integral form of the remainder to write $e^{ix} - 1 - ix = -\int_0^x (x-t)e^{it}\,dt$. Now we write

$$\int_0^x (x-t)e^{it}\,dt = \int_0^x (x-t)(e^{it}-1)\,dt + \int_0^x (x-t)\,dt = \int_0^x (x-t)(e^{it}-1)\,dt + \frac{x^2}{2}$$

so that $x^2 R(x) = -\int_0^x (x-t)(e^{it}-1)\,dt$. Observe that on the one hand

$$\begin{aligned} |R(x)| &= \frac{1}{x^2}\left|\int_0^x (x-t)(e^{it}-1)\,dt\right| \leq \frac{\sup_{0\leq t\leq x}\left|e^{it}-1\right|}{x^2}\int_0^x (x-t)\,dt \\ &= \sup_{0\leq t\leq x}\left|e^{it}-1\right| \end{aligned}$$

and thus continuity of e^{ix} implies $\lim_{x\to 0} R(x) = 0$. On the other hand

$$\left|\int_0^x (x-t)(e^{it}-1)\,dt\right| \leq \int_0^x (x-t)\left|e^{it}-1\right|\,dt \qquad \leq 2\int_0^x (x-t)\,dt = x^2$$

which shows $|R(x)| \leq 1$. □

Theorem C.3 (Arithmetic Mean Geometric Mean Inequality). *Let $x_1, \dots, x_n$ be non-negative real numbers and let $p_1, \dots, p_n$ be non-negative real numbers such that $\sum_{j=1}^n p_j = 1$ then*

$$x_1^{p_1}\cdots x_n^{p_n} \leq p_1 x_1 + \cdots + p_n x_n$$

Proof. TODO: □

Proposition C.4. *Let $x_1, \dots, x_n$ be non-negative real numbers and $m \in \mathbb{N}$ then*

$$\left(\frac{x_1 + \cdots + x_n}{n}\right)^m \leq \frac{x_1^m + \cdots + x_n^m}{n}$$

PROOF. We first validate the result for $m = 2$. To see this case observe that by Theorem C.3 we have for arbitrary non-negative integers α and β and non-negative reals x, y

$$x^\alpha y^\beta = (x^{\alpha+\beta})^{\frac{\alpha}{\alpha+\beta}} + (y^{\alpha+\beta})^{\frac{\beta}{\alpha+\beta}} \leq \frac{\alpha}{\alpha+\beta} x^{\alpha+\beta} + \frac{\beta}{\alpha+\beta} y^{\alpha+\beta}$$

Therefore by the Binomial Theorem and the fact that $\binom{m}{k} = \binom{m}{m-k}$

$$\begin{aligned}
(x_1 + x_2)^m &= \sum_{k=0}^m \binom{m}{k} x_1^k x_2^{m-k} \leq \sum_{k=0}^m \binom{m}{k} \left(\frac{k}{m} x_1^m + \frac{m-k}{m} x_2^m \right) \\
&= \begin{cases} (x_1^m + x_2^m) \sum_{k=0}^{\lfloor \frac{m}{2} \rfloor} \binom{m}{k} \text{ if } m \text{ is odd} \\ (x_1^m + x_2^m) \{ \sum_{k=0}^{\frac{m}{2}-1} \binom{m}{k} + \frac{1}{2}\binom{m}{m/2} \} \text{ if } m \text{ is even} \end{cases} \\
&= 2^{m-1} (x_1^m + x_2^m)
\end{aligned}$$

Now an easy induction shows that the result holds for any $n = 2^k$:

$$\begin{aligned}
\left(\frac{x_1 + \cdots + x_{2^k}}{2^k} \right) &\leq \frac{1}{2} \left\{ \left(\frac{x_1 + \cdots + x_{2^{k-1}}}{2^{k-1}} \right) + \left(\frac{x_{2^{k-1}+1} + \cdots + x_{2^k}}{2^{k-1}} \right) \right\} \\
&\leq \frac{x_1^m + \cdots + x_{2^k}^m}{2^k}
\end{aligned}$$

It remains to extend the result to arbitrary n. Suppose that $2^{k-1} \leq n < 2^k$ and define

$$A = \left(\frac{x_1^m + \cdots + x_n^m}{n} \right)^{1/m}$$

they by the result for 2^k we get

$$\left(\frac{x_1^m + \cdots + x_n^m + (2^k - n)A}{2^k} \right)^m \leq \frac{x_1^m + \cdots + x_n^m + (2^k - n)A^m}{2^k} = A^m$$

If we take m^{th} roots and collect terms involving A we get $\frac{x_1 + \cdots + x_k}{k} \leq A$. Now take m^{th} power and use the definition of A to get the result.

TODO: I actually think we need this result for non-integral m. Get the full blown power mean inequality from Steele and then fix up the BDG vector inequality. □

THEOREM C.5.