Notes on "Convergence of Probability Measures" by Billingsly

1 Weak Convergence in Metric Spaces

1.1 Measures on Metric Spaces

Definition 1. Our general framework here will be a metric space S equipped with a distance ρ, which defines the usual topology of open and closed sets. Along with this we will get $\mathcal{S}$, the Borel sigma algebra of subsets. We are interested in studying probability measures here, that is measures $\mathbf{P}$ on $\mathcal{S}$ which are non-negative, countable additive set functions satisfying $\mathbf{P}(\mathcal{S}) = 1$. We will also use the shorthand that for functions $f : S \to \mathbb{R}$, we define $\mathbf{P}f = \int f d\mathbf{P} = \mathbb{E}(f) \in \mathbb{R}$

Definition 2. We write that a sequence of probability measures $\mathbf{P}_n \Rightarrow \mathbf{P}$ if $\mathbf{P}_n f \to \mathbf{P}f$ for every bounded, continuous real function f on S and we say $\mathbf{P}$ *converges weakly to* $\mathbf{P}$

Theorem 3. *Every probability measure on S is what is called "regular", meaning that for every $\mathcal{S}-$set A and every $\epsilon > 0$, there is are open and closed sets G and F that sandwhich A so that $F \subset A \subset G$ and $\mathbf{P}(G - F) < \epsilon$*

Proof. This is a "Borel sigma algebra proof". One first verifies that the collection of sets A with the above property form a sigma algebra (easy), and then checks that the closed sets have this property too. Indeed, for A closed, take $F = A$ and then to get G consider the collection of open sets $A^\delta := \{x : \rho(x, A) < \delta\}$ (recall that distance from a point to a set is taken with an inf). Now, since $\cap A^{\frac{1}{n}} = A$ and is decreasing, hence $\lim_{n\to\infty} \mathbf{P}A^{\frac{1}{n}} = \mathbf{P}A$, so that $\mathbf{P}\left(A^\delta - A\right) < \epsilon$ for δ sufficiently small. □

Theorem 4. *If $\mathbf{P}$and $\mathbf{Q}$ are probability measures on S so that $\mathbf{P}F = \mathbf{Q}F$ for every closed set F, then $\mathbf{P}A = \mathbf{Q}A$ for every $A \in \mathcal{S}$.*

Proof. The collection of sets where they agree is seen to be a sigma algebra, so containing all the closed sets is enough to get equality everywhere. □

Remark 5. This theorem shows us that the probability measure is determined entirely by its action on open and closed sets (We will see a lot more results of this flavor). The next result tells us that knowing the action of $\mathbf{P}f$ for bounded continuous f is also sufficient.

Theorem 6. *Suppose* $\mathbf{P}$,$\mathbf{Q}$ *are probability measures on* S *with* $\mathbf{P}f = \mathbf{Q}f$ *for every bounded continuous* f*. Then* $\mathbf{P} = \mathbf{Q}$*.*

Proof. We will only need to consider the functions of the form: $f(x) = (1 - \rho(x, F)/\epsilon)^+$ where F is a closed set, $\epsilon > 0$ and we use the shorthand $z^+ = \max(z, 0) \geq 0$. One easily verifies the inequality:

$$1_F \leq f = (1 - \rho(x, F)/\epsilon)^+ \leq 1_{F^\epsilon}$$

(1_F is the indicator function of the set F). Here, $F^\epsilon = \{x : \rho(x, F) < \epsilon\}$ is the enlargment of F by a bit. Now, integrating the above inequality gives, $\mathbf{P}F \leq \mathbf{P}f$. By hypothesis, $\mathbf{P}f = \mathbf{Q}f$ and itegrating the above inequality again gives $\mathbf{Q}f \leq \mathbf{Q}F^\epsilon$. Stringing these together gives $\mathbf{P}F \leq \mathbf{Q}F^\epsilon$. Taking ϵ to zero gives $\mathbf{P}F \leq \mathbf{Q}F$ (like in the last theorem, $\mathbf{Q}F^\epsilon \to \mathbf{Q}F$). Of course, the argument works equally well to show $\mathbf{Q}F \leq \mathbf{P}F$, so we get the desired result. □

Definition 7. We say that a probability measure $\mathbf{P}$ on S is *tight* for every $\epsilon > 0$, there exists a compact set$K$ so that $\mathbf{P}K > 1 - \epsilon$. This notion of tight is a bridge between the idea of comapact and the probability measure on the space.

Theorem 8. $\mathbf{P}$ *is tight if and only if* $\mathbf{P}A = \sup_{K \subset A} \mathbf{P}K$ *for every* $A \in \mathcal{S}$*. (*K*'s are compact here)*

Proof. Suppose $\mathbf{P}$ is tight. For any $\epsilon > 0$, find K so that $\mathbf{P}K > 1 - \epsilon/2$ and F closed, G open so that $F \subset A \subset G$ and $\mathbf{P}(G - F) < \epsilon/2$. Then $\mathbf{P}(A - F) < \epsilon/2$ too. Now the set $K \cap F$ is compact (as K compact and F closed) and has $A - K \cap F = (A \cap K^c) \dot\cup (A - F)$, so $\mathbf{P}(A - K \cap F) \leq \mathbf{P}K^c + \mathbf{P}(A - F) < \epsilon/2 + \epsilon/2 = \epsilon$. Hence $\mathbf{P}A \leq \sup_{K \subset A} \mathbf{P}K + \epsilon$. Since this holds for any ϵ, we get $\mathbf{P}A \leq \sup_{K \subset A} \mathbf{P}K$. The reverse inequality is trivial from $K \subset A$, so we have our result.

Conversly, if $\mathbf{P}A = \sup_{K \subset A} \mathbf{P}K$ holds, feeding in $A = S$ give tightness. □

Theorem 9. *If* S *is seperable and complete, then each probability measure on* S *is tight.*

Proof. Fix any $\epsilon > 0$. Since S is seperable, for each $k \in \mathbb{N}$, by taking open $\frac{1}{k}$-balls centered at the countable dense subset, we get a sequence of open sets $A_{k,1}, A_{k,2}, \ldots$ that cover S. Now, since $\cup_i A_{k,i} \uparrow S$, we can find for each k, an n_k so that $\mathbf{P}(\cup_{i \leq n_k} A_{i,k}) > 1 - \frac{\epsilon}{2^k}$. Now, consider the set $\cap_{k \in \mathbb{N}} \cup_{i \leq n_k} A_{k,i}$. Since $\mathbf{P}(\cup_{i \leq n_k} A_{i,k})^c < \frac{\epsilon}{2^k}$,hence $\mathbf{P}(\cup_k (\cup_{i \leq n_k} A_{i,k})^c) < \sum_k \frac{\epsilon}{2^k} = \epsilon$, and so $\mathbf{P}(\cap_{k \in \mathbb{N}} \cup_{i \leq n_k} A_{k,i}) > 1 - \epsilon$. Finally, we remark that the closure of $\cap_{k \in \mathbb{N}} \cup_{i \leq n_k} A_{k,i}$ is compact, as it is closed (its a closure of a set in a complete metric space) and totally bounded, since $\cap_{k \in \mathbb{N}} \cup_{i \leq n_k} A_{k,i} \subset \cup_{i \leq n_{k_0}} A_{k_0,i}$ which can be covered by a finite number (n_{k_0} of them) of $\frac{1}{k}$-balls. Hence it is compact, so this set gives us tightness. □

Definition 10. A collection of sets $\mathcal{A} \subset \mathcal{S}$ is called a *seperating class* if two probability measures which agree on $\mathcal{A}$ must agree on all of $\mathcal{S}$. This is called

a seperating class, since the values of $\mathbf{P}$ on $\mathcal{A}$ are enough to "seperate" $\mathbf{P}$ from any other probability measure. For example, as we have seen, the open sets are a seperating class.

Definition 11. A $\pi-$system is a collection of sets which is closed under finite intersections. e.g. the half open intervals $(-\infty, a]$ are a $\pi-$system on $\mathbb{R}$.

Remark 12. If $\mathcal{A}$ is $\pi-$system which generates $\mathcal{S}$ (the Borel sigma algebra), then $\mathcal{A}$ is a seperating class. e.g. the half open intervals $(-\infty, a]$ on $\mathbb{R}$.

Example 13. On $\mathbb{R}^k$, the half open "intervals" are the sets $(-\infty, a_1]\times(-\infty, a_2]\times \ldots \times (-\infty, a_k]$. These are a $\pi-$system that generates the Borel sigma algebra and so is a seperating class. This means the cumulative distribution functions: $F(a_1, a_2, \ldots a_k) = \mathbf{P}(-\infty, a_1] \times (-\infty, a_2] \times \ldots \times (-\infty, a_k]$ complete determine $\mathbf{P}$. Since $\mathbb{R}^k$ is seperable and complete, one can see any probabilty measure is tight by the previous theorem. Another way to see tightness is to see that $\mathbb{R}^k = \cup_i \bar{B}(0, i)$ so that $\mathbf{P}\left(B(0, i)\right) \uparrow 1$. This arguement is saying that $\sigma-$compact spaces are tight.

Example 14. On the space of real valued seuqences, $\mathbb{R}^\infty = \{x_1, x_2, \ldots\}$, take the metric:

$$\rho(x, y) = \sum_i b(x_i, y_i)/2^i$$

Where $b(x_i, y_i) = 1 \wedge |x_i - y_i| = \min(1, |x_i - y_i|)$. This metrices pointwise convergence. If a sequence $x^n \in \mathbb{R}^\infty$ has $\rho(x^n, x) \to 0$, then notice that $b(x_i^n, x_i) \leq 2^i \rho(x_i^n, x_i) \to 0$, so $x_i^n \to x_i$. Conversly, if $x_i^n \to x_i$ for each i given any $\epsilon > 0$, take n_0 so large so that $\sum_{i>n_0} 1/2^i < \epsilon/2$ and then take n_1 so large so that $|x_i^n - x_i| < \epsilon/2$ for every $1 \leq i \leq n_0$ and $n > n_1$ (since there are only finitely many i here, $1 \leq i \leq n_0$ this is ok!) But then, for $n > n_1$ we have:

$$\begin{aligned}
\rho(x^n, x) &= \sum_i b(x_i, y_i)/2^i \\
&\leq \sum_{i \leq n_0} b(x_i, y_i)/2^i + \sum_{i > n_0} 1/2^i \\
&\leq \sum_{i \leq n_0} (\epsilon/2)/2^i + (\epsilon/2) \\
&< \epsilon
\end{aligned}$$

So $\rho(x^n, x) \to 0$ if and only if $x_i^n \to x_i$ for each i. One can see that the natural projections onto the first k coordinates $\pi_k : \mathbb{R}^\infty \to \mathbb{R}^k$ are continuous then since convergence on $\mathbb{R}^k$ is also characterized by coordinate-wise convergence. Hence the sets:

$$N_{k,\epsilon}(x) = \{y : |y_i - x_i| < \epsilon, 1 \leq i \leq k\} = \pi_k^{-1}(B(x, \epsilon))$$

Are open. Moreover, $y \in N_{k,\epsilon}(x)$ implies $\rho(y, x) \leq \epsilon + \frac{1}{2^k}$, so by choosing k large enough and ϵ small enough, we can, for any r, have $N_{k,\epsilon}(x) \subset B(x, r)$.

By choosing only rational x, we can then see that the space is seperable. Completness of the space follows since any Caucy sequeunce will be coordinate-wise Cauchy. Since $\mathbb{R}$ is complete, we will then have coordinatewise convergence, which we've already shown is equivalent to our metric. Hence $\mathbb{R}^\infty$is tight! Notice that $\mathbb{R}^\infty$is *not* $\sigma-$compact (this can be proved by the Baire category theorem) so the fact that probability measures are tight is not obvious here. Conisder the finite dimensional sets of the form $\pi_k^{-1}H$ for $H \in \mathcal{S}$. Notice that these form a $\pi-$system, since the intersection of two such sets can be written as $\pi_{k_1}^{-1}H_1 \cap \pi_{k_2}^{-1}H_2 = \pi_k^{-1}(H_1 \cap H_2)$ (some manipulation with adding extra coordinates to make k_1 and k_2 compatible may be nessasary). Morevoer, the fact that the subsets $N_{k,\epsilon}$, which are exactly such finitie dimensional sets, form a basis for the space (they can be found within every open ball) shows that these finite dimensional sets generate the whole sigma algebra. This is saying precisly that the finite dimensional sets are a *seperating class.*

Example 15. The space of continuous functions on $[0,1]$,with the sup-norm distance, can also be shown to be seperable and complete. Seperability can be seen by considering piecewise linear functions with rational values that live on finer and finer grids. (The fact that these are dense uses uniform continuity). Completeness is a standard exercise, one sees that every Cauchy sequence must be Cauchy at each coordinate, and that the convergence is uniform. So again, we have tightness.

Example 16. (Again on $C[0,1]$ with the sup-norm distance.) We will show that the finite dimensional sets form a seperating class. Let $\pi_{t_1,t_2,\ldots t_k} : C[0,1] \to \mathbb{R}^k$ by $\pi_{t_1\ldots t_k} = (x(t_1), x(t_2), \ldots x(t_k))$ These are continuous, so the sets $\pi^{-1}_{t_1,\ldots t_k}H, H \in \mathbb{R}^k$ are Borel sets. As in the $\mathbb{R}^\infty$example, consider the set of finite dimensional sets of the form $\pi^{-1}_{t_1,\ldots t_k}H$. As before, by refining the indices $t's$ a little bit, we see that these sets form a $\pi-$system. By continuity of the functions, we can write in $C[0,1]$ that $B(x,\epsilon) = \cap_r\{y : |x(r) - y(r)| < \epsilon\}$ where r ranges over the rationals in $\mathbb{Q} \cap [0,1]$. Since the rationals are countable, these balls are in the sigma algebra generated by the finite dimensional sets! That is to say, the finite dimensional sets are again a *seperating class.*

1.2 Properties of Weak Convergence

Example 17. Write δ_x to be the probability measure on S that has unit mass at $x \in S$. That is $\delta_x(A) = \mathtt{1}_A(x)$ and $\delta_x(f) = \int f d\delta_x = f(x)$ If we have $x_n \to x_0$ in S, then for continuous f we have that:

$$\delta_{x_n} f = f(x_n) \to f(x_0) = \delta_{x_0}$$

and therefore, since this holds for all continuous f, $\delta_{x_n} \Rightarrow \delta_{x_0}$ by definition of $\Rightarrow$. Conversly, if $x_n \nrightarrow x_0$, the there exists $\epsilon > 0$ so that $\rho(x_n, x_0) > \epsilon$ infinetly often, and then for the function $f(x) = (1 - \rho(x, x_0)/\epsilon)^+$, we will have that $\delta_{x_n} f = f(x_n) = 0$ infinetly often while $\delta_{x_0} f = f(x_0) = 1$, so of course $\delta_{x_n} f \not\Rightarrow \delta_{x_0} f$. Hence $\delta_{x_n} \Rightarrow \delta_{x_0}$ if and only if $x_n \to x_0$.

Example 18. Take $S = [0,1]$ and $\mathbf{P} = \mathcal{L}[0,1]$ to be the usual Lebesgue measure. Suppose that we have a sequence of probability measures which are constructed as a sum of many point masses. That is each $\mathbf{P}_n = \frac{1}{m_n}\sum_{1\leq k\leq m_n}\delta_{x_{n,k}}$, and that the points are asymptotically evenly distributed over $[0,1]$ in the sense that in any interval $J \subset [0,1]$

$$\mathbf{P}_n J = \frac{\#\{k : x_{n,k} \in J\}}{m_n} \to \mathcal{L}(J) = \mathbf{P}J$$

This condition is enough to see that $\mathbf{P}_n \Rightarrow \mathbf{P}$. One neat way to see this more rigourosly is to use the Theory of Riemann integrals, for if f is any continuous function on $[0,1]$, then it is Riemann integrable. Take a partition $J_1, J_2, \ldots J_r$ of $[0,1]$ fine enough so that the upper (sups!) and lower (infs!) Riemann sums disagree by at most ϵ, and we will have that:

$$\mathbf{P}_n f = \frac{1}{m_n}\sum f(x_{n,k}) \leq \frac{1}{m_n}\sum \sup\{f(x) : x \in J_i\}\cdot\#\{k : x_{n,k} \in J\} \to \sum \sup\{f(x) : x \in J_i\}\mathcal{L}(J_i) \leq \mathbf{P}f$$

The other inequality holds too for the lower Riemann sums, and so we have a sandwhich from which we conclude that $\mathbf{P}_n f \to \mathbf{P}f$. Since this holds for every continuous f, we conclude that $\mathbf{P}_n \Rightarrow \mathbf{P}$.

Definition 19. For a probability measure $\mathbf{P}$ on S, a set $A \in \mathcal{S}$ whose boundary ∂A satifies $\mathbf{P}(\partial A) = 0$ is called a $\mathbf{P}-$continuity set.

Theorem 20. *(Portmanteau Theorem) The following are equivalent ways to say $\mathbf{P}_n \Rightarrow \mathbf{P}$:*

(i)/(ii) $\mathbf{P}_n \Rightarrow \mathbf{P}$ that is:$\mathbf{P}_n f \to \mathbf{P}f$ for every bounded continuous f

(iii)/(iv) $\limsup \mathbf{P}_n F \leq \mathbf{P}F$ for every closed set F / $\liminf \mathbf{P}_n G \geq \mathbf{P}G$ for every open set G

(v) $\mathbf{P}_n A \to \mathbf{P}A$ for all $\mathbf{P}-$continuity sets A.

Proof. (i)/(ii) $\Rightarrow$ (iii)/(iv): For F closed, let $f = (1 - \rho(\cdot, F)/\epsilon)^+$ so that $\mathtt{1}_F \leq f \leq \mathtt{1}_{F^\epsilon}$ as in a previous argument. Since this is bounded and continuous, we have by (ii) that $\lim \mathbf{P}_n f = \mathbf{P}f$. But then, $\limsup \mathbf{P}_n F \leq \limsup \mathbf{P}_n f = \mathbf{P}f \leq \mathbf{P}F^\epsilon$. Since F is closed, we know $\cap_\epsilon F^\epsilon = F$, so taking $\epsilon \to 0$ with this inequality gives the desired result in (iii). Taking complements shows (iii) and (iv) are really the same thing. □

Proof. $(iii)/(iv) \Rightarrow (v)$: Given a set A, recall that the boundary of A can be written as $\partial A = \bar{A} - \overset{\circ}{A}$, so $\mathbf{P}(\partial A) = 0$ implies that $\mathbf{P}(\bar{A}) = \mathbf{P}(\overset{\circ}{A}) = \mathbf{P}(A)$. Since $\bar{A}$ is closed and $\overset{\circ}{A}$ is open, we have by $(iii), (iv)$ that:

$$\begin{aligned}
\mathbf{P}\bar{A} &\geq \limsup \mathbf{P}_n \bar{A} \\
&\geq \limsup \mathbf{P}_n A \\
&\geq \liminf \mathbf{P}_n A \\
&\geq \liminf \mathbf{P}_n \overset{\circ}{A} \\
&\geq \mathbf{P}\overset{\circ}{A}
\end{aligned}$$

Since $\mathbf{P}(\bar{A}) = \mathbf{P}(\mathring{A}) = \mathbf{P}(A)$, these are all equalities, and moreover $\lim \mathbf{P}_n A = \mathbf{P}A$, which is (v). $\square$

Proof. $(v) \Rightarrow (i), (ii)$

Since f is bounded, by linearity,we may assume that $0 \leq f \leq 1$. Now, $\mathbf{P}f = \int f d\mathbf{P} = \int_0^1 \mathbf{P}\{f > t\}dt$ (this a Fubini-Tonelli type statement), the same equality holds with $\mathbf{P}_n$. Now, for continuous f, $\partial\{f > t\} \subset \{f = t\}$ (find sequences in $\{f > t\}$ and $\{f \leq t\}$ converging to any point in $\partial\{f > t\}$, and by continuity, we will have that such a point has both $f \geq t$ and $f \leq t$), so $\{f > t\}$ is a $\mathbf{P}-$continuity whenever $\mathbf{P}\{f = t\} = 0$. Of course, $\mathbf{P}\{f = t\} \neq 0$ for at most countably many t, say at $t_1, t_2, \ldots$, and everywhere else $\{f > t\}$ is a $\mathbf{P}-$continuity set. By condition 5, everywhere except for $t = t_1, t_2, \ldots$ we will have $\mathbf{P}_n\{f > t\} \to \mathbf{P}\{f > t\}$. That is to say that $\mathbf{P}_n\{f > t\} \to \mathbf{P}\{f > t\}$ for $\mathcal{L}-$almost every t in $[0, 1]$. By the bounded convergence theorem then:

$$\mathbf{P}_n f = \int_0^1 \mathbf{P}_n\{f > t\}dt \to \int_0^1 \mathbf{P}\{f > t\} = \mathbf{P}f$$

$\square$

Definition 21. A collection of Borel sets, $\mathcal{A}$ is called a *convergence determening class* is $\mathbf{P}_n A \to \mathbf{P}A$ for all $\mathbf{P}-$continuity sets $A \in \mathcal{A}$ implies $\mathbf{P}_n \Rightarrow \mathbf{P}$. One can prove some wacky theorems that tell you that certain classes are convergence determening. See pages 17-19 of the book.

Example 22. The finite dimensional sets of $\mathbb{R}^\infty$ are convergence determening. (Details ommited)

Example 23. The finite dimensional sets of $C[0, 1]$ are *not* convergence determening. To see this, take a sequence f_n of functions which gets to zero pointwise after finite n, but for which $f_n \nrightarrow 0$ uniformly (e.g. f_n has a spike at $\frac{1}{n}$, and is zero everywhere after $\frac{2}{n}$). Let $\mathbf{P}_n = \delta_{f_n}$ since $f_n \nrightarrow 0$ in this metric space, $\delta_{f_n} \not\Rightarrow \delta_0$. However, for all the finite dimensional subsets $\pi^{-1}_{t_1,t_2,\ldots}H$, we have that $\mathbf{P}_n\left(\pi^{-1}_{t_1,t_2,\ldots}H\right) = \mathtt{1}_{\pi^{-1}_{t_1,t_2,\ldots}H}(f_n) = \mathtt{1}_{H_1}\left(f(t_1)\right) \cdot \ldots \cdot \mathtt{1}_{H_n}\left(f(t_n)\right) \overset{(\text{for } n\text{large})}{=} \mathtt{1}_{H_1}(0) \cdot \ldots \cdot \mathtt{1}_{H_n}(0) = \mathtt{1}_{\pi^{-1}_{t_1,t_2,\ldots}H}(0) = \mathbf{P}\left(\pi^{-1}_{t_1,t_2,\ldots}H\right)$ since f_n gets to zero at the points $t_1, t_2, \ldots$ for large enough n by the pointwise stipulation we made. So $\mathbf{P}_n A \to \mathbf{P}A$ for every finite dimensional A and yet $\mathbf{P}_n \not\Rightarrow \mathbf{P}$. This means these are NOT a convergence determening class in $C[0, 1]$; this highlites a fundemental difference between $\mathbb{R}^\infty$and $C[0, 1]$

Theorem 24. *If every subsequence of* $\mathbf{P}_n$*, call it*$\mathbf{P}_{n_i}$*,has a further subsquence, call it* $\mathbf{P}_{n_{i_j}}$*, so that*$\mathbf{P}_{n_{i_j}} \Rightarrow \mathbf{P}$*, then* $\mathbf{P}_n \Rightarrow \mathbf{P}$.

Proof. (By contrapositive). If $\mathbf{P}_n \not\Rightarrow \mathbf{P}$ then there is a function f and $\epsilon > 0$ so that $|\mathbf{P}_n f - \mathbf{P}f| > \epsilon$ infitently often. Taking this infinetly often as our subsequence, we see that it's impossible to have a sub-sub-sequence with $\mathbf{P}_{n_{i_j}} \not\Rightarrow \mathbf{P}$, as the function f provides an obstruction. $\square$

1.2.1 The Mapping Theorem

Theorem 25. *Suppose $h : S \to S'$ is a continous function between two metric spaces. For $\mathbf{P}$ a probability measure on S, we get an induced probability measure on S', namely $\mathbf{P}h^{-1}$ by $\mathbf{P}h^{-1}(A') = \mathbf{P}\left(h^{-1}(A')\right)$. If $\mathbf{P}_n \Rightarrow \mathbf{P}$ then $\mathbf{P}_n h^{-1} \Rightarrow \mathbf{P}h^{-1}$.*

Proof. For any bounded continuous f, the function $f \circ h$ is again bounded and conditunous. Hence $\mathbf{P}_n \Rightarrow \mathbf{P}$ gives $\mathbf{P}_n(f \circ h) \to \mathbf{P}(f \circ h)$. By a change of variables for probability spaces however we see that, $\mathbf{P}(f \circ h) = \int f \circ h(x) d\mathbf{P}(dx) = \int f(h(x)) d\mathbf{P}(dx) = \int f(y) d\mathbf{P}(h^{-1}(dy)) = \mathbf{P}h^{-1}(f)$. So indeed $\mathbf{P}_n(f \circ h) \to \mathbf{P}(f \circ h)$ is the same as $\mathbf{P}_n h^{-1}(f) \to \mathbf{P}h^{-1}(f)$. Since this holds for every bounded continous f, we see $\mathbf{P}_n h^{-1} \Rightarrow \mathbf{P}h^{-1}$. □

Example 26. The projections $\pi_k : \mathbb{R}^\infty \to \mathbb{R}^k$ are continuous, so if $\mathbf{P}_n \Rightarrow \mathbf{P}$ on $\mathbb{R}^\infty$, then the measures $\mathbf{P}_n \pi_k^{-1} \Rightarrow \mathbf{P}\pi_k^{-1}$ for every k. The converse is also true. First, we can show directly that $\partial\left(\pi_k^{-1} H\right) = \pi_k^{-1}(\partial H)$. (one direction is trivial, the other not so hard using sequences), so that the $\mathbf{P}$-continuity sets of $\mathbb{R}_f^\infty$ (finite dimensional sets) are precisly those which are $\mathbf{P}\pi_k^{-1}-$continuity sets in $\mathbb{R}^k$ for every k. Hence $\mathbf{P}_n \pi_k^{-1} \Rightarrow \mathbf{P}\pi_k^{-1}$ for every k means that $\mathbf{P}_n A \to \mathbf{P}A$ whenever A is a finite dimensional $\mathbf{P}-$continuity set. Since the finite dimensional sets are a convergence-determening class, we have $\mathbf{P}_n \Rightarrow \mathbf{P}$.

Example 27. The projections $\pi_{t_1,t_2,\ldots t_k} : C[0,1] \to \mathbb{R}^k$ are continuous, so if $\mathbf{P}_n \Rightarrow \mathbf{P}$ on $\mathbb{R}^\infty$, then the measures $\mathbf{P}_n \pi_{t_1,\ldots}^{-1} \Rightarrow \mathbf{P}\pi_{t_1,\ldots}^{-1}$ for every k. The converse is NOT true though, the same example with $\delta_{f_n} \not\Rightarrow \delta_0$ with pointwise convergence works in the same way.

Theorem 28. *Let h be any measurable function $h : S \to S'$, and let $D_h \in \mathcal{S}$ be the set of discontinuities of h. If $\mathbf{P}_n \Rightarrow \mathbf{P}$ and $\mathbf{P}(D_h) = 0$, then $\mathbf{P}_n h^{-1} \Rightarrow \mathbf{P}h^{-1}$*
.

Proof. We use the characterization from the Portmeanteu theorem $\mathbf{P}_n \Rightarrow \mathbf{P}$ iff $\limsup \mathbf{P}_n F \leq \mathbf{P}F$ for every closed F. To start, we first remark that $D_h^c \cap \overline{h^{-1}F} \subset h^{-1}\left(\overline{F}\right)$ for any F, since for $x \in D_h^c \cap \overline{h^{-1}F}$, there is a sequence $x_n \to x$ so that $hx_n \in F$, but since x is a continuity point of h, then $hx_n \to hx$ means that $hx \in \overline{F}$. For any closed set F we have (overbar is closure):

$$\begin{aligned}
\limsup \mathbf{P}_n h^{-1}(F) &\leq \limsup \mathbf{P}_n \overline{(h^{-1}F)} \\
&\leq \mathbf{P}\left(\overline{(h^{-1}F)}\right) \\
&= \mathbf{P}\left(\overline{(h^{-1}F)} \cap D_h^c\right) \text{ (since } D_h^c \text{is a null event)} \\
&\leq \mathbf{P}\left(h^{-1}\overline{F}\right) \text{ (above remark)} \\
&= \mathbf{P}h^{-1}F \text{ (since } F \text{is closed)}
\end{aligned}$$

□

1.3 Convergence in Distribution

So far we have been talking about probability measures on metric spaces. A different way to think about the same thing is to consider metric space valued random variables which come from an arbitarty probability space $X : (\Omega, \mathtt{P}, \mathcal{F}) \to (S, \mathcal{S})$. Of course, such random variables induces a measure on $(S, \mathcal{S})$ in the usual way:

$$\mathbf{P}A = \mathtt{P}\left(X^{-1}A\right) = \mathtt{P}\left\{X \in A\right\}$$

This is also called the law of X and is sometimes denoted $\mathbf{P} = \mathcal{L}(X)$. This captures all the information we would want about X, so when we think of random variables we can think of them in two ways: either as a measurable function on a probability space, or as a measure on the metric space S on which X takes values. In this language:

$$\mathbf{E}\left[f(x)\right] = \int_{\Omega} f(X(w))\mathtt{P}(dw) = \int_{S} f(x)\mathbf{P}(dx) = \mathbf{P}f$$

Definition 29. We say that a sequence of random variables X_n converges to X in *distribution* if $\mathcal{L}(X_n) \Rightarrow \mathcal{L}(X)$. For convenience we write $X_n \Rightarrow X$.

Theorem 30. *(Portmenteau Theorem) In this setting the different equivalent ways to think about weak convergence look like:*

(i),(ii) $X_n \Rightarrow X$*, that is* $\mathbf{E}\left(f(X_n)\right) \to \mathbf{E}\left(f(X)\right)$ *for all bounded continuous* f*.*

(iii),(iv) $\limsup \mathtt{P}\left\{X_n \in F\right\} \leq \mathtt{P}\left\{X \in F\right\}$ *for every closed set* F */* $\liminf \mathtt{P}\left\{X_n \in G\right\} \geq \mathtt{P}\left\{X \in G\right\}$ *for every open set* G

(v) $\mathtt{P}\left\{X_n \in A\right\} \to \mathtt{P}\left\{X \in A\right\}$ *for all* $X-$*continuity sets* A*.*

Definition 31. Sometimes we will conflate our two notations, so we will write things like:

$$\begin{aligned} \mathbf{P}_n &= \mathbf{P} \\ X_n &\Rightarrow X \\ X_n &\Rightarrow \mathbf{P} \\ \mathbf{P}_n &\Rightarrow X \end{aligned}$$

Where X can be read as $\mathcal{L}(X)$ if you ever get confused.

1.3.1 Convergence of Probability

Definition 32. Say $a \in S$. We say that X_n *converges to* a *in probability* if for every $\epsilon > 0$ we have:

$$\mathtt{P}\left\{\rho(X_n, a) < \epsilon\right\} \to 1$$

Proposition 33. X_n converges to a in probability if and only if $X_n \Rightarrow a$

Proof. Suppose X_n converges to a in probability. To see that $X_n \Rightarrow a$, we use the "open-set-lim-inf" criteria of the Portmenteau theorem. Let G be any open set. If $a \in G$, then find $\epsilon > 0$ so that $B(a,\epsilon) \subset G$. For this ϵ we have that $\mathtt{P}\{X_n \in B(a,\epsilon)\} = \mathtt{P}\{\rho(X_n,a) < \epsilon\} \to 1$, but $B(a,\epsilon) \subset G$, so $\mathtt{P}\{X_n \in G\} \to 1$ too. Hence $\liminf \mathtt{P}\{X_n \in G\} = 1 = \mathtt{P}\{a \in G\}$. If $a \notin G$, then we have the trivial inequality $\liminf \mathtt{P}\{X_n \in G\} \geq 0 = \mathtt{P}\{a \in G\}$.

Conversly, suppose $X_n \Rightarrow a$. For every $\epsilon > 0$, $B(a,\epsilon)$ is an open set so choosing $G = B(a,\epsilon)$, we have by the portmenteau theorem that $1 \geq \liminf \mathtt{P}\{X_n \in B(a,\epsilon)\} \geq 1 = \mathtt{P}\{a \in B(a,\epsilon)\}$, hence $\lim \mathtt{P}\{X_n \in B(a,\epsilon)\} = 1$ so we have convergence in probability. □

Theorem 34. *Suppose that (X_n, Y_n) are random elements of $S \times S$. If $X_n \Rightarrow X$ and $\rho(X_n, Y_n) \Rightarrow 0$ then $Y_n \Rightarrow X$*

Proof. We use the "closed-set-lim-sup" criteria. Let F be any closed set, and $F_\epsilon = \{x : \rho(x,F) \leq \epsilon\}$. Then:

$$\mathtt{P}\{Y_n \in F\} \leq \mathtt{P}\{\rho(X_n,Y_n) \geq \epsilon\} + \mathtt{P}\{X_n \in F_\epsilon\}$$

Since $F_\epsilon$is closed, we take lim sup to get:

$$\begin{aligned} \limsup \mathtt{P}\{Y_n \in F\} &\leq \limsup \mathtt{P}\{\rho(X_n,Y_n) \geq \epsilon\} + \limsup \mathtt{P}\{X_n \in F_\epsilon\} \\ &\leq 1 - \liminf \mathtt{P}\{\rho(X_n,Y_n) < \epsilon\} + \limsup \mathtt{P}\{X_n \in F_\epsilon\} \\ &\leq 1 - 1 + \mathtt{P}\{X \in F_\epsilon\} = \mathtt{P}\{X \in F_\epsilon\} \end{aligned}$$

Since F is closed, taking $\epsilon \to 0$ gives $\mathtt{P}F_\epsilon \downarrow \mathtt{P}F$ and gives the result. □

1.3.2 Local vs Integral Laws

Proposition 35. *Suppose $\mathbf{P}_n$ and $\mathbf{P}$ are absolutly continuous with respect to some other measure μ, and have densities f_n and f respectivly. If $f_n(x) \to f(x)$ for $\mu-$almost every x. Then $\mathbf{P}_n \Rightarrow \mathbf{P}$. The converse statment is not true.*

Proof. (sketch) Have (this is related to the total variation stuff we looked at in Markov Mixing)

$$\sup_{A \in \mathcal{S}} |\mathbf{P}A - \mathbf{P}_n A| \leq \int_S |f(x) - f_n(x)| \, \mu(dx) \to 0$$

So of course, $\mathbf{P}_n A \to \mathbf{P}A$ for every $\mathbf{P}$continuity set. □

1.3.3 Integration to the Limit

If $X_n \Rightarrow X$, when does $\mathbf{E}(X_n) \to \mathbf{E}(X)$?

Theorem 36. *If $X_n \Rightarrow X$, then $\mathbf{E}|X| \leq \liminf \mathbf{E}|X_n|$*

Proof. Since $|\cdot|$ is a continuous function, by the mapping theorem, $|X_n| \Rightarrow |X|$. By the same type of argument in the proof of $(v) \Rightarrow (i)$ in the Portmanteau theorem, $\mathtt{P}\{|X_n| > t\} \to \mathtt{P}\{|X| > t\}$ for all but countably many t. The result now follows by Fatous lemma:

$$\mathbf{E}\,|X| = \int_0^\infty \mathtt{P}\{|X| > t\}\,dt \leq \liminf \int_0^\infty \mathtt{P}\{|X_n| > t\}\,dt = \liminf \mathbf{E}\,|X_n|$$

□

Definition 37. We say that the sequence of random variables X_n is *uniformly integrable* if:

$$\lim_{\alpha\to\infty} \sup_n \int_{|X_n|>\alpha} |X_n| d\mathtt{P} = 0$$

This holds if the X_n are uniformly bounded, as for αlarger than the bound, all of these are 0.

Proposition 38. *If X_n are uniformly integrable, then* $\sup_n \mathbf{E}(X_n) < \infty$.

Proof. Take α_0 so large so that the $\sup_n \int_{|X_n|>\alpha_0} |X_n| d\mathtt{P} \leq 1$. Now have:

$$\begin{aligned}
\sup_n \mathbf{E}(X_n) &\leq \sup_n \int_{|X_n|>\alpha_0} |X_n| d\mathtt{P} + \sup_n \int_{|X_n|\leq\alpha_0} |X_n| d\mathtt{P} \\
&\leq 1 + \int \alpha_0 d\mathtt{P} \\
&= 1 + \alpha_0 < \infty
\end{aligned}$$

□

Theorem 39. *If X_n are uniformly integrable, and $X_n \Rightarrow X$, then X is integrable and* $\mathbf{E}(X_n) \to \mathbf{E}(X)$

Proof. Since the $\mathbf{E}\,|X_n|$ are bounded, we know by our Fatou-type lemma that $\mathbf{E}\,|X| \leq \liminf \mathbf{E}\,|X_n| < \infty$ is integrable. By the mapping theorem with the continuous maps $(\cdot)^+, (\cdot)^-$, we have that $X_n^+ \Rightarrow X^+$ and $X_n^- \Rightarrow X^-$. Now write:

$$\begin{aligned}
\mathbf{E}\left(X_n^+\right) &= \int_0^\alpha \mathtt{P}\left\{t < X_n^+ < \alpha\right\} dt + \int_{X_n\geq\alpha} X_n^+ d\mathtt{P} \\
\mathbf{E}\left(X^+\right) &= \int_0^\alpha \mathtt{P}\left\{t < X^+ < \alpha\right\} dt + \int_{X_n\geq\alpha} X^+ d\mathtt{P}
\end{aligned}$$

The last term in these equations $\int_{X_n\geq\alpha} X_n^+ d\mathtt{P}$, $\int_{X_n\geq\alpha} X^+ d\mathtt{P}$ tend to zero as $\alpha \to \infty$ by the uniform integrability condition. Hence, to see $\mathbf{E}(X_n^+) \to$

$\mathbf{E}(X^+)$, it suffices to check that as α large that $\int_0^\alpha \mathtt{P}\{t < X_n^+ < \alpha\}\, dt \to \int_0^\alpha \mathtt{P}\{t < X^+ < \alpha\}\, dt$. By choosing an α with $\mathtt{P}\{X = \alpha\} = 0$, this follows by the bounded convergence theorem (as it did in the Portmenteau theorem). The same can be said of X^-, so we get $\mathbf{E}(X_n) \to \mathbf{E}(X)$ as desired. □

Proposition 40. *If there exists $\epsilon > 0$ so that $\sup_n \mathbf{E}\left(|X_n|^{1+\epsilon}\right) < \infty$,then X_n is uniformly integrable.*

Proof. Have:

$$\begin{aligned} \int\limits_{X_n \geq \alpha} X_n d\mathtt{P} &\leq \int\limits_{X_n \geq \alpha} \frac{X_n^{1+\epsilon}}{\alpha^\epsilon} d\mathtt{P} \\ &\leq \frac{1}{\alpha^\epsilon} \mathbf{E}\left(|X_n|^{1+\epsilon}\right) \\ &\to 0 \end{aligned}$$

□

1.4 Skipped Section on Permutations

1.5 Prohorov's Theorem

Definition 41. Let Π be a family of probability measures on S. We call Π*reletivly compact* if every sequence of elements of Πcontails a weakly convergent subsequence. I.e. $\forall \{P_n\}_n \in \Pi, \exists P_{n_i} s.t. P_{n_i} \Rightarrow P$. We will be mostly concerned with the case wehre $\Pi$is itself a sequence, in this setting relativly compact if every subsequence has a further subsequence which weakly converges to something.

If Recal the folowing theorem

Theorem 42. *If $\mathbf{P}_n$is relativly compact, and if the limiting probability measure is the same for every subsequence, then $\mathbf{P}_n \Rightarrow \mathbf{P}$. In other words, if every subsequence of $\mathbf{P}_n$, call it$\mathbf{P}_{n_i}$,has a further subsquence, call it $\mathbf{P}_{n_{i_j}}$, so that$\mathbf{P}_{n_{i_j}} \Rightarrow \mathbf{P}$, then $\mathbf{P}_n \Rightarrow \mathbf{P}$.*

Proof. (By contrapositive). If $\mathbf{P}_n \not\Rightarrow \mathbf{P}$ then there is a function f and $\epsilon > 0$ so that $|\mathbf{P}_n f - \mathbf{P} f| > \epsilon$ infitently often. Taking this infinetly often as our subsequence, we see that it's impossible to have a sub-sub-sequence with $\mathbf{P}_{n_{i_j}} \not\Rightarrow \mathbf{P}$, as the function f provides an obstruction.

Why is relative compactness useful? Here are some examples. □

Example 43. Suppose we are on$C[0,1]$ and we know that the for some sequence $\mathbf{P}_n$, the finite dimensional distributions converge to some other distribution $\mathbf{P}$. i.e. we have that $\mathbf{P}_n \pi_{t_1,\ldots t_k}^{-1} \Rightarrow \mathbf{P}\pi_{t_1,\ldots t_k}^{-1}$for every $t_1, \ldots t_k$. We have seen already that this does not nessasarily mean that $\mathbf{P}_n \Rightarrow \mathbf{P}$ (this is the statement that the finite dimensional sets are not *convergence determining*. (for example, the if we take the pointwise convergent sequence of continuous spikes going to zero, the

point masses at these functions does not weakly converge, as the functions do not uniformly converge). However, if we know in addition to this that the family $\mathbf{P}_n$ is relativly compact, then we have a candidate $\mathbf{Q}$ so that $\forall \mathbf{P}_{n_i}, \exists \mathbf{P}_{n_{i_j}} \Rightarrow \mathbf{Q}$. Now, by the mapping theorem, it must be that $\mathbf{P}_{n_{i_j}}\pi^{-1}_{t_1,\ldots,t_k} \Rightarrow \mathbf{Q}\pi^{-1}_{t_1,\ldots,t_k}$. By uniqieness of weak convergence, we have then that $\mathbf{P}\pi^{-1}_{t_1,\ldots t_k} = \mathbf{Q}\pi^{-1}_{t_1,\ldots,t_k}$. But since the finite dimensional sets are a *seperating class,* hence $\mathbf{P} = \mathbf{Q}$, wh. Finally, by the last theorem, since every subsequence $\mathbf{P}_{n_i}$ has a further subsequence $\mathbf{P}_{n_{i_j}}$ converging to $\mathbf{P}$, by contrapositive we prove that $\mathbf{P}_n \Rightarrow \mathbf{P}$. In other words, finite dimenisoanl convergence + weak compactness gives weak convergence.

Example 44. Similar to the above, if $\mathbf{P}_n$ is relativly compact, and we know that $\mathbf{P}_n\pi^{-1}_{t_1,\ldots t_k} \Rightarrow \mu_{t_1,\ldots t_k}$ for a family of measures μ, there is a $\mathbf{P}$so that $\mathbf{P}\pi^{-1}_{t_1,\ldots t_k} = \mu_{t_1,\ldots t_k}$

1.5.1 Tightness

How do we prove relative compactness? On the real line, let F_n be the distribution functions for $\mathbf{P}_n$. By the Helly selection theorem, every subsequence F_{n_i} has a further subsequence for which there is a nondecreasing, right continuos F so that $F_{n_{i(m)}} \to F$ pointwise for all continuity points of F. However, F might fail to be a proper distribution function for the reason that it doesnt have total mass 1, i.e. $\lim_{x\to\infty} F(x) \neq 1$ or $\lim_{x\to-\infty} F(x) \neq 0$. e.g. δ_n has $F(x) = Heaviside(x-n) \to 0$ as $n \to \infty$. Another example is the uniform distriution on $[-n, n]$. A condition that prevents this from happening is uniform tightness:

Definition 45. A family Π is *tight* or *uniformly tight* if for every$\epsilon > 0$, there exists a compact K such that $\mathbf{P}K > 1 - \epsilon$ for every $\mathbf{P} \in \Pi$.

Theorem 46. *(Prohorov's Theorem) If* Π*is tight, then it is relativly compact.*

Corollary 47. *If* $\{\mathbf{P}_n\}$ *is tight, and if each convergent subsequence of* $\mathbf{P}_n$ *converges to* $\mathbf{P}$*, then* $\mathbf{P}_n \Rightarrow \mathbf{P}$

Proof. By Prohorov's theorem, $\{\mathbf{P}_n\}$ is relativly compact. Hence every susequence has a subsubsequence which converges. By hypothesis, it must converge to $\mathbf{P}$. By the earlier theorem with the proof by contrapositive, $\mathbf{P}_n \Rightarrow \mathbf{P}$. □

1.5.2 The proof of Prohorov's Theorem

Its pretty technical, so I will skip it for now.