

STATS310A - Lecture 18

Persi Diaconis
Scribed by Michael Howes

11/30/21

Contents

1	Announcements	1
2	Helly's selection theorem	1
3	Tightness	2
4	The continuity theorem	3
5	Uniqueness theorem	5
6	Generalizations	5

1 Announcements

- Thursday's lecture will also be on Zoom.
- Wednesday's office hours will be on Zoom.

Here goes.

2 Helly's selection theorem

Theorem 1 (Helly's selection theorem). *If $\{F_n\}_{n=1}^\infty$ are any cumulative distribution functions on \mathbb{R} , then there exists a subsequence n_k and a monotone, right continuous function F such that $F_{n_k}(x) \rightarrow F(x)$ for all x such that F is continuous at x .*

Before we prove the above theorem it is important to note that F might not be a cumulative distribution function.

Proof. Let $\{r_i\}_{i=1}^\infty$ be an enumeration of \mathbb{Q} . We can form the array

$$\begin{array}{cccc} F_1(r_1) & F_2(r_1) & F_3(r_1) & \dots \\ F_1(r_2) & F_2(r_2) & F_3(r_2) & \dots \\ F_1(r_3) & F_2(r_3) & F_3(r_3) & \dots \\ \vdots & \vdots & \vdots & \ddots \end{array}$$

Each row is bounded since $F_n(x) \in [0, 1]$ for all n and $x \in \mathbb{R}$. Thus Cantor's diagonal argument implies that there exists a subsequence n_k and a function $G : \mathbb{Q} \rightarrow \mathbb{R}$ such that $F_{n_k}(r) \rightarrow G(r)$ for all $r \in \mathbb{Q}$.

Note that if $r < s$, then $F_{n_k}(r) \leq F_{n_k}(s)$ for all k and so $G(r) \leq G(s)$. Now define

$$F(x) = \inf\{G(r) : r > x, r \in \mathbb{Q}\}.$$

Since G is non-decreasing, F is also non-decreasing. We will now show that F is right continuous. Given x and $\varepsilon > 0$, find $r > x$ such that $G(r) < F(x) + \varepsilon$. If $x < y < r$, then

$$F(x) \leq F(y) \leq G(r) < F(x) + \varepsilon.$$

Thus, F is right continuous. Now we just need to prove that if x is a continuity point of F , then $F_{n_k}(x) \rightarrow F(x)$.

This is elementary but (slightly) tedious. Given x a continuity point and $\varepsilon > 0$, choose $y < x$ such that $F(x) - \varepsilon < F(y)$. Next choose rational numbers r and s so that $y < r < x < s$ and

$$G(s) < F(x) + \varepsilon.$$

It follows that

$$F(x) - \varepsilon < F(y) \leq G(r) \leq G(s) < F(x) + \varepsilon.$$

We also have $F_{n_k}(r) \leq F_{n_k}(x) \leq F_{n_k}(s)$ for all k and so

$$\begin{aligned} F(x) - \varepsilon &\leq G(r) \\ &= \lim_k F_{n_k}(r) \\ &\leq \lim_k F_{n_k}(x) \\ &\leq \overline{\lim}_k F_{n_k}(x) \\ &\leq \lim_k F_{n_k}(s) \\ &= G(s) \\ &\leq F(x) + \varepsilon. \end{aligned}$$

Thus $\lim_k F_{n_k}(x)$ and $\overline{\lim}_k F_{n_k}(x)$ are both within ε of $F(x)$. Since ε was arbitrary we can conclude that $\lim_k F_{n_k}(x) = F(x)$. \square

Example 1. As mentioned before, the limiting function F need not be a cumulative distribution function. For example,

- If F_n is the cumulative distribution function of a point mass at n , then $F_n(x) \rightarrow 0$ for all x .
- If F_n is the cumulative distribution function of a point mass at $-n$, then $F_n(x) \rightarrow 1$ for all x .

Neither of these limits are cumulative distribution functions.

The kind of convergence in the statement of the Helly's selection theorem is called *vague convergence*.

3 Tightness

How can we be sure that the limit function F in Helly's selection theorem is a distribution? It turns out that the key property is tightness.

Definition 1. A family of probability distributions $\{\mu_n\}$ on \mathbb{R} is *tight* if for all $\varepsilon > 0$, there exists $a < b$ such that $\mu_n([a, b]) > 1 - \varepsilon$ for all n .

We will sometimes say $\{\mu_n\}$ are “almost compactly supported” to mean $\{\mu_n\}$ is tight.

Theorem 2. Let $\{\mu_n\}$ be a family of probability distributions on \mathbb{R} . Then $\{\mu_n\}$ is tight if and only if for every subsequence n_k , there exists a further subsequence n_{k_i} and a probability distribution μ such that $\mu_{n_{k_i}} \Rightarrow \mu$ as $i \rightarrow \infty$.

In the remaining lectures, we will only use that if $\{\mu_n\}$ is tight, then for every subsequence n_k , there exists a further subsequence n_{k_i} and a probability distribution μ such that $\mu_{n_{k_i}} \Rightarrow \mu$ as $i \rightarrow \infty$. Thus we will only prove this direction of the above theorem.

Proof. Let $\{\mu_n\}$ be a tight family of probability distributions with corresponding cumulative distribution functions F_n . Let n_k be a subsequence. By Helly's selection theorem, there exists a further subsequence n_{k_i} and a monotone right-continuous function F such that $F_{n_{k_i}}(x) \rightarrow F(x)$ for all x such that F is continuous at x .

We wish to show that F is a cumulative distribution function for some probability measure μ as this will imply that $\mu_n \Rightarrow \mu$. To show that F is a cumulative distribution function, it suffices to show that $\lim_{x \rightarrow \infty} F(x) = 1$ and $\lim_{x \rightarrow -\infty} F(x) = 0$. We know that $F(x) \in [0, 1]$ for all x since each $F_{n_{k_i}}$ is a cumulative distribution function. Furthermore since $\{\mu_n\}$ is tight, for every $\varepsilon > 0$ there exist $a < b$ such that F is continuous at a and b and for all i

$$F_{n_{k_i}}(b) - F_{n_{k_i}}(a) > 1 - \varepsilon.$$

By taking a limit we have $F(b) - F(a) \geq 1 - \varepsilon$ which is sufficient to conclude that F has the correct limits. \square

Remark 1. If $\int_{\mathbb{R}} |x| \mu_n(dx)$ is uniformly bounded in n , then the family $\{\mu_n\}$ is tight.

Likewise, if $\int_{\mathbb{R}} f(|x|) \mu_n(dx)$ is uniformly bounded in n for some unbounded monotone function $f: \mathbb{R}_+ \rightarrow \mathbb{R}_+$, then $\{\mu_n\}$ is tight. Both of these claims follow by Markov's inequality for monotonically increasing functions.

Remark 2. All of what we have done works for a complete separable metric space \mathcal{X} . We have to work with $\mathcal{B}(\mathcal{X})$ the Borel σ -algebra on \mathcal{X} which is the σ -algebra generated by the open subsets of \mathcal{X} . A sequence of probabilities μ_n on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ converges weak* to μ if for all bounded and continuous functions f on \mathcal{X} , we have

$$\int_{\mathcal{X}} f(x) \mu_n(dx) \rightarrow \int_{\mathcal{X}} f(x) \mu(dx).$$

In this setting, we say that $\{\mu_n\}$ is tight if for all $\varepsilon > 0$, there exists a compact set $K \subseteq \mathcal{X}$ so that

$$\mu_n(K) > 1 - \varepsilon,$$

for all n . Some references for this topic are:

- Billingsley "Convergence of probability measures."
- Kallenberg "Probability Theory" (3rd edition).
- Dudley "Real Analysis and Probability."

All three are great books.

4 The continuity theorem

The below theorem states that pointwise convergence of characteristic functions is exactly convergence in distribution. This was a missing link in Laplace's argument for the central limit theorem.

Theorem 3. Let $\{F_n\}, F$ be cumulative distribution functions with characteristic functions ϕ_n, ϕ , then $F_n \Rightarrow F$ if and only if for all $t \in \mathbb{R}$, $\phi_n(t) \rightarrow \phi(t)$.

Proof. Let μ_n and μ be the probability distributions corresponding to F_n and F . The functions $x \mapsto \cos(tx)$ and $x \mapsto \sin(tx)$ are both bounded and continuous. Thus if $F_n \Rightarrow F$, then

$$\phi_n(t) = \int_{\mathbb{R}} (\cos(tx) + i \sin(tx)) \mu_n(dx) \rightarrow \int_{\mathbb{R}} (\cos(tx) + i \sin(tx)) \mu(dx) = \phi(t).$$

Now suppose that $\phi_n(t) \rightarrow \phi(t)$ for all t . We will show later that this implies that $\{\mu_n\}$ is tight. Now suppose that $F_n \not\Rightarrow F$. Then there exists some $x \in \mathbb{R}$ such that F is continuous at x but $F_n(x) \not\rightarrow F(x)$. Thus there exists a subsequence n_k and $\varepsilon > 0$ such that $|F_{n_k}(x) - F(x)| > \varepsilon$ for all k . Since we will show that ϕ_n is tight, this implies that there exists a cumulative distribution function G and a further subsequence n_{k_i} such that $F_{n_{k_i}} \Rightarrow G$. Note that we cannot have $G = F$ as this will imply that G is continuous at x and hence $F_{n_{k_i}}(x) \rightarrow G(x) = F(x)$.

Let ϕ_G be the characteristic function of G . Since $F_{n_{k_i}} \Rightarrow G$, we have $\phi_{n_{k_i}}(t) \rightarrow \phi_G(t)$ and thus $\phi_G(t) = \phi(t)$. By the uniqueness theorem (which we state below and will prove next lecture) this implies that $G = F$, a contradiction.

It thus remains to show that $\{\mu_n\}$ is tight. For $u > 0$, consider the quantity

$$\frac{1}{u} \int_{-u}^u (1 - \phi(t)) dt.$$

By Fubini's theorem we have

$$\begin{aligned} \frac{1}{u} \int_{-u}^u 1 - \phi(t) dt &= \int_{-\infty}^{\infty} \frac{1}{u} \int_{-u}^u 1 - e^{itx} dt \mu(dx) \\ &= 2 \int_{-\infty}^{\infty} \left[1 - \frac{\sin(ux)}{ux} \right] \mu(dx) \\ &\geq 2 \int_{\{x: |x| > 2/u\}} 1 - \frac{\sin(ux)}{ux} \mu(dx) \\ &\geq 2 \int_{\{x: |x| > 2/u\}} 1 - \frac{1}{ux} \mu(dx) \\ &\geq 2 \int_{\{x: |x| > 2/u\}} \frac{1}{2} \mu(dx) \\ &= \mu \left(\left\{ x : |x| > \frac{2}{u} \right\} \right) \end{aligned}$$

Now $\phi(t)$ is continuous and $\phi(0) = 1$. Thus for all $t > 0$, there exists $u > 0$ sufficiently small such that $|1 - \phi(t)| < \frac{\varepsilon}{2}$, given $|t| < u$. This implies that

$$\left| \frac{1}{u} \int_{-u}^u 1 - \phi(t) dt \right| < \varepsilon.$$

We know that $\phi_n(t) \rightarrow \phi(t)$ for all $t \in \mathbb{R}$. Thus by the bounded convergence theorem, there exists n_0 such that if $n \geq n_0$, then

$$\mu_n \left(\left\{ x : |x| > \frac{2}{u} \right\} \right) \leq \frac{1}{u} \int_{-u}^u 1 - \phi_n(t) dt \leq 2\varepsilon.$$

By taking u smaller we can ensure that $u > 0$ and

$$\mu_n \left(\left\{ x : |x| > \frac{2}{u} \right\} \right) \leq 2\varepsilon,$$

for $n = 1, 2, \dots, n_0 - 1$. Thus we have shown that $\{\mu_n\}$ is tight. \square

If you'd like to learn more about Laplace and his proof of the central limit theorem, search for "Steve Stigler, Laplace."

Remark 3. Two comments.

- The continuity theorem is a substantial theorem that uses topology and Helly's selection theorem.
- Our proof relies on the uniqueness theorem which we have not yet proved.

5 Uniqueness theorem

Theorem 4 (Inversion theorem). *Let μ be a probability on \mathbb{R} with characteristic function ϕ . If $a < b$, then,*

$$\mu((a, b)) + \frac{1}{2}\mu(\{a, b\}) = \lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \phi(t) dt.$$

The uniqueness theorem is a corollary.

Corollary 1. *If μ and ν are probability measures and $\phi_\mu = \phi_\nu$, then $\mu = \nu$.*

Proof. The family of sets

$$\mathcal{P} = \{(a, b) : \mu(\{a\}) = \mu(\{b\}) = \nu(\{a\}) = \nu(\{b\}) = 0\},$$

is a π -system that generates the Borel σ -algebra. Since $\phi_\mu = \phi_\nu$, the inversion theorem implies that μ and ν agree on \mathcal{P} . By the trusty old $\pi - \lambda$ theorem, this implies that $\mu = \nu$. \square

The inversion theorem does not give a formula for μ in terms of ϕ since it involves a limit. If we put additional assumptions on ϕ , then we do get a formula for μ .

Theorem 5. *If $\int_{-\infty}^{\infty} |\phi(t)| dt < \infty$, then μ has a bounded density f and*

$$f(y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ity} \phi(t) dt.$$

We will prove these theorems on Thursday.

6 Generalizations

There are versions of characteristic functions for measures on lots of spaces other than \mathbb{R} . For example one can work with measures on groups and do Fourier analysis on non-commutative groups. See "Group Representations in Probability and Statistics" by Persi. We won't talk about characteristic functions on non-commutative groups here but we will talk about characteristic functions on \mathbb{R}^d .

Definition 2. Let μ be a probability distribution on \mathbb{R}^d . The *characteristic function* of μ is the function $\phi_\mu : \mathbb{R}^d \rightarrow \mathbb{C}$ given by

$$\phi_\mu(t) = \mathbb{E}_\mu[e^{it \cdot x}],$$

where $t \cdot x$ denotes the dot product between t and x .

All of the theorems we studied on \mathbb{R} , hold for characteristic functions on \mathbb{R}^d . We also have the following result.

Proposition 1 (Cramer-Wold device). *If $X \in \mathbb{R}^d$ is a random vector and we know the distribution of $\sum_{j=1}^d a_j X_j$ for all $a \in \mathbb{R}^d$, then we know the distribution of X .*

Proof. If we know the distribution of $t \cdot X$ for all $t \in \mathbb{R}^d$, then we know $\phi(t) = \mathbb{E}[e^{it \cdot X}]$ for all t . Thus we know the characteristic function of X and hence the distribution of X . \square

Example 2. The Cramer-Wold device can be used to prove multivariate central limit theorems for $(X_1^{(n)}, \dots, X_d^{(n)})$ from univariate central limit theorems for $t \cdot X^{(n)}$. For example, if X_1, X_2, \dots are i.i.d. in \mathbb{R}^d with mean μ and covariance matrix Σ . If $S_n = \frac{1}{n} \sum_{j=1}^n (X_j - \mu)$, then

$$\sqrt{n}S_n \Rightarrow \mathcal{N}_d(0, \Sigma),$$

by the univariate central limit theorem and the Cramer-Wold device. There are also multivariate central limit theorems for triangular arrays.

Remark 4. Some final comments:

- Please look in the book to see how similar the proof of the central limit theorem is to the one that we proved in class. They both use a swapping argument.
- Persi would have liked to have talked about infinitely divisible laws but there wasn't time this quarter.
- There are all kinds of tricks, maneuvers and proofs that are achievable with characteristic functions and the full power of complex analysis. The best source for this is W. Feller - "An introduction to probability and its applications" Vol II (2nd edition) chapter 15. For example, these tools can prove the following:

Theorem 6. Suppose X, Y are independent and there exist non-zero $a, b, c, d \in \mathbb{R}$, such that $(aX + bY, cX + dY)$ has the same distribution as (X, Y) , then X and Y are normal.