7. EXAMPLES OF PROBABILITY MEASURES ON THE LINE

There are many important probability measures that occur frequently in probability and in the real world. We give some examples below and expect you to familiarize yourself with each of them.

Example 28. The examples below have CDFs of the form $F(x) = \int_{-\infty}^{x} f(t)dt$ where f is a non-negative integrable function with $\int f = 1$. In such cases f is called the *density* or pdf (probability density function). Clearly F is continuous and non-decreasing and tends to 0 and 1 at ∞ and $-\infty$ respectively. Hence, there do exist probability measures on $\mathbb{R}$ with the corresponding density.

(1) *Normal distribution*. For fixed $a \in \mathbb{R}$ and $\sigma^2 > 0$, $N(a, \sigma^2)$ is the p.m. on $\mathbb{R}$ with density $\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-a)^2/2\sigma^2} du$. F is clearly increasing and continuous and $F(-\infty) = 0$. That $F(+\infty) = 1$ is not so obvious but true!

(2) *Gamma distribution* with shape parameter $\alpha > -1$ and scale parameter $\lambda > 0$ is the p.m. with density $f(x) = \frac{\lambda^{\alpha-1}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$ for $x > 0$.

(3) *Exponential distribution*. Exponential(λ) is the p.m. with density $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ and $f(x) = 0$ if $x < 0$. This is a special case of Gamma distribution, but important enough to have its own name.

(4) *Beta distribution*. For parameters $a > -1$, $b > -1$, the Beta(a,b) distribution is the p.m. with density $B(a,b)^{-1} x^{a-1}(1-x)^{b-1}$ for $x \in [0,1]$. Here $B(a,b)$ is the beta function, equal to $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}$. (Why does it integrate to 1?).

(5) *Uniform distribution* on $[a,b]$ is the p.m. with density $f(x) = \frac{1}{b-a}$ for $x \in [a,b]$. For example, with $a = 0, b = 1$, this is a special case of the Beta distribution.

(6) *Cauchy distribution*. This is the p.m. with density $\frac{1}{\pi(1+x^2)}$ on the whole line. Unlike all the previous examples, this distribution has "heavy tails"

You may have seen the following discrete probability measures. They are very important too and will recur often.

Example 29. The examples below have CDFs of the form $F(x) = \sum_{u_i \leq x} p(x_i)dt$, where $\{x_i\}$ is a fixed countable set, and $p(x_i)$ are non-negative numbers that add to one. In such cases p is called the pmf (probability density function). and from what we have shown, there do exist probability measures on $\mathbb{R}$ with the corresponding density or CDF.

(1) *Binomial distribution*. Binomial(n,p), with $n \in \mathbb{N}$ and $p \in [0,1]$, has the pmf $p(k) = \binom{n}{k} p^k q^{n-k}$ for $k = 0, 1, \ldots, n$.

(2) *Bernoulli distribution*. $p(1) = p$ and $p(0) = 1 - p$ for some $p \in [0,1]$. Same as Binomail$(1,p)$.

(3) *Poisson(λ)* distribution with parameter $\lambda \geq 0$ has p.m.f $p(k) = e^{-\lambda} \frac{\lambda^k}{k!}$ for $k = 0, 1, 2, \ldots$.

(4) *Geometric(p)* distribution with parameter $p \in [0,1]$ has p.m.f $p(k) = q^k p$ for $k = 0, 1, 2, \ldots$.

8. A METRIC ON THE SPACE OF PROBABILITY MEASURES ON $\mathbb{R}^d$

What kind of space is $\mathcal{P}(\mathbb{R}^d)$ (the space of p.m.s on $\mathbb{R}^d$)? It is clearly a convex set (this is true for p.m.s on any sample space and σ-algebra).

We saw that for every Borel p.m. on $\mathbb{R}^d$ there is associated a unique CDF. This suggests a way of defining a distance function on $\mathcal{P}(\mathbb{R}^d)$ using their CDFs. Let $D(\mu,\nu) = \sup_{x\in\mathbb{R}^d} |F_\mu(x) - F_\nu(x)|$. Since CDFs are bounded between 0 and 1, this is well-defined and one can easily check that it gives a metric on $\mathcal{P}(\mathbb{R}^d)$.

Is this the metric we want to live with? For $a \in \mathbb{R}^d$, we denote by δ_a the p.m. for which $\delta_a(A) = 1$ if $A \ni a$ and 0 otherwise (although this p.m. can be defined on all subsets, we just look at it as a Borel measure). If $a \neq b$, it is easy to see that $D(\delta_a, \delta_b) = 1$. Thus, even when $a_n \to a$ in $\mathbb{R}^d$, we do not get convergence of δ_{a_n} to δ_a. This is an undesirable feature and hence we would like a weaker metric.

Definition 30. For $\mu,\nu \in \mathcal{P}(\mathbb{R}^d)$, define the Lévy distance between them as (here $\mathbf{1} = (1,1,\ldots,1)$)

$$d(\mu,\nu) := \inf\{u > 0 \,:\, F_\mu(x+u\mathbf{1}) + u \geq F_\nu(x),\ F_\nu(x+u\mathbf{1}) + u \geq F_\mu(x)\ \forall x \in \mathbb{R}^d\}.$$

If $d(\mu_n,\mu) \to 0$, we say that μ_n converges weakly to μ and write $\mu_n \xrightarrow{d} \mu$. [...breathe slowly and meditate on this definition for a few moments...]

First of all, $d(\mu,\nu) \leq 1$. That d is indeed a metric is an easy exercise. If $a_n \to a$ in $\mathbb{R}^d$, does δ_{a_n} converge to δ_a? Indeed $d(\delta_a,\delta_b) = (\max_i |b_i - a_i|) \wedge 1$ and hence $d(\delta_{a_n}, a) \to 0$.

Exercise 31. Let $\mu_n = \frac{1}{n}\sum_{k=1}^n \delta_{k/n}$. Show directly by definition that $d(\mu_n, m) \to 0$. What about $D(\mu_n,\mu)$?

How does convergence in the metric d show in terms of CDFs?

Proposition 32. *$\mu_n \xrightarrow{d} \mu$ if and only if $F_{\mu_n}(x) \to F_\mu(x)$ for all continuity points x of F_μ.*

Proof. Suppose $\mu_n \xrightarrow{d} \mu$. Let $x \in \mathbb{R}^d$ and fix $u > 0$. Then for large enough n, we have $F_\mu(x+u\mathbf{1}) + u \geq F_{\mu_n}(x)$, hence $\limsup F_{\mu_n}(x) \leq F_\mu(x+u\mathbf{1}) + u$ for all $u > 0$. By right continuity of F_μ, we get $\limsup F_{\mu_n}(x) \leq F_\mu(x)$. Further, $F_{\mu_n}(x) + u \geq F_\mu(x - u\mathbf{1})$ for large n, hence $\liminf F_{\mu_n}(x) \geq F_\mu(x-u)$ for all u. If x is a continuity point of F_μ, we can let $u \to 0$ and get $\liminf F_{\mu_n}(x) \geq F_\mu(x)$. Thus $F_{\mu_n}(x) \to F_\mu(x)$.

For simplicity let $d = 1$. Suppose $F_n \to F$ at all continuity points of F. Fix any $u > 0$. Find continuity points (of F) $x_1 < x_2 < \ldots < x_m$ such that $x_{i+1} \leq x_i + u$. This can be done because continuity points are dense. Fix N so that $d(\mu_n,\mu) < u$ for $n \geq N$. Henceforth, let $n \geq N$.

If $x \in \mathbb{R}$, then either $x \in [x_{j-1}, x_j]$ for some j or else $x < x_1$ or $x > x_1$. First suppose $x \in [x_{j-1}, x_j]$. Then

$$F(x+u) \geq F(x_j) \geq F_n(x_j) - u \geq F_n(x) - u, \qquad F_n(x+u) \geq F_n(x_j) \geq F(x_j) - u \geq F(x) - u.$$

If $x < x_1$, then $F(x+u) + u \geq u \geq F(x_1) \geq F_n(x_1) - u$. Similarly the other requisite inequalities, and we finally have

$$F_n(x+2u) + 2u \geq F(x) \text{ and } F(x+2u) + 2u \geq F_n(x).$$

Thus $d(\mu_n,\mu) \leq u$. Hence $d(\mu_n,\mu) \to 0$. ∎

9. Compact subsets of $\mathcal{P}(\mathbb{R}^d)$

Often we face problems like the following. A functional $L : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ is given, and we would like to find the p.m. μ that minimizes $L(\mu)$. By definition, we can find nearly optimal p.m.s μ_n satisfying $L(\mu_n) - \frac{1}{n} \leq \inf_\nu L(\nu)$. Then we might expect that if some subsequence μ_{n_k} converged to a p.m. μ, then that μ might be the optimal solution we are searching for. Thus we are faced with the question of characterizing compact subsets of $\mathcal{P}(\mathbb{R}^d)$, so that existence of convergent subsequences can be asserted.

Looking for a convergent subsequence: Let μ_n be a sequence in $\mathcal{P}(\mathbb{R}^d)$. We would like to see if a convergent subsequence can be extracted. Write F_n for F_{μ_n}. For any fixed $x \in \mathbb{R}^d$, $F_n(x)$ is a bounded sequence of reals and hence we can find a subsequence $\{n_k\}$ such that $F_{n_k}(x)$ converges.

Fix a dense subset $S = \{x_1, x_2, \ldots\}$ of $\mathbb{R}^d$. Then, by the observation above, we can find a subsequence $\{n_{1,k}\}_k$ such that $F_{n_{1,k}}(x_1)$ converges to some number in $[0,1]$ that we shall denote $G(x_1)$. Then extract a further subsequence $\{n_{2,k}\}_k \subset \{n_{1,k}\}_k$ such that $F_{n_{2,k}}(x_2) \to G(x_2)$, another number in $[0,1]$. Of course, we also have $F_{n_{2,k}}(x_1) \to G(x_1)$. Continuing this way, we get subsequences $\{n_{1,k}\} \supset \{n_{2,k}\} \supset \ldots \{n_{\ell,k}\} \ldots$ such that for each ℓ, as $k \to \infty$, we have $F_{n_{\ell,j}}(x_j) \to G(x_j)$ for each $j \leq \ell$.

The *diagonal sbsequence* $\{n_{\ell,\ell}\}$ is ultimately the subsequence of each of the above obtained subsequences and therefore, $F_{n_{\ell,\ell}}(x_j) \to G(x_j)$ for all j.

To define the limiting function on the whole line, set $F(x) := \inf\{G(x_j) \,:\, j \text{ for which } x_j > x\}$. F is well defined, takes values in $[0,1]$ and is non-decreasing. It is also right-continuous, because if $y_n \downarrow y$, then for any j for which $x_j > y$, it is also true that $x_j > y_n$ for sufficiently large n. Thus $\liminf_{n\to\infty} G(y_n) \leq \inf_{x_j > y} G(x_j) = F(y)$. Lastly, if y is any continuity point of F, then for any $\delta > 0$, we can find i, j such that $y - \delta < x_i < y < x_j < y + \delta$. Therefore

$$F(y-\delta) \leq G(x_i) = \lim F_{n_{\ell,\ell}}(x_i) \leq \liminf F_{n_{\ell,\ell}}(y) \leq \limsup F_{n_{\ell,\ell}}(y) \leq \lim F_{n_{\ell,\ell}}(x_j) = G(x_j) \leq F(y+\delta).$$

The equalities are by prperty of the subsequence $\{n_{\ell,\ell}\}$, the inner two inequalities are obvious, and the outer two inequalities follow from the definition of F in terms of G (and the fact that G is nondecreasing). Since F is continuous at y, we get $\lim F_{n_{\ell,\ell}}(y) = F(y)$.

If only we could show that $F(+\infty) = 1$ and $F(-\infty) = 0$, then F would be the CDF of some p.m. μ and we would immediately get $\mu_n \xrightarrow{d} \mu$. But this is false in general!

Example 33. Consider δ_n. Clearly $F_{\delta_n}(x) \to 0$ for all x if $n \to +\infty$ and $F_{\delta_n}(x) \to 1$ for all x if $n \to -\infty$. Even if we pass to subsequences, the limiting function is identically zero or identically one, and neither of these is a CDF of a p.m. The problem is that mass escapes to infinity. To get weak convergence to a probability measure, we need to impose a condition to avoid this sort of situation.

Definition 34. A family $\{\mu_\alpha\}_{\alpha \in I} \subset \mathcal{P}(\mathbb{R}^d)$ is said to be *tight* if for any $\varepsilon > 0$, there is a compact set $K_\varepsilon \subset \mathbb{R}^d$ such that $\mu_\alpha(K_\varepsilon) \geq 1 - \varepsilon$ for all $\alpha \in I$.

Example 35. Suppose the family has only one p.m. μ. Since $[-n,n]^d$ increase to $\mathbb{R}^d$, given $\varepsilon > 0$, for a large enough n, we have $\mu([-n,n]^d) \geq 1 - \varepsilon$. Hence $\{\mu\}$ is tight. If the family is finite, tightness is again clear.

Take $d = 1$ and let μ_n be p.m.s with $F_n(x) = F(x-n)$ (where F is a fixed CDF), then $\{\mu_n\}$ is not tight. This is because given any $[-M,M]$, if n is large enough, $\mu_n([-M,M])$ can be made arbitrarily small. Similarly $\{\delta_n\}$ is not tight.

Theorem 36 (Helly's selection principle)**.** *(a) A sequence of probability measures on $\mathbb{R}^d$ is tight if and only if every subsequence has a further subsequence that converges weakly. (b) Equivalently a subset of $\mathcal{P}(\mathbb{R}^d)$ is precompact if and only if it is tight.*

Proof. (a) If μ_n is a tight sequence in $\mathcal{P}(\mathbb{R}^d)$, then any subsequence is also tight. By the earlier discussion, given any subsequence $\{n_k\}$, we may extract a further subsequence $n_{\ell,k}$ and find a non-decreasing right continuous function F (taking values in $[0,1]$) such that $F_{n_{\ell,k}}(x) \to F(x)$ for all continuity points x of F. Fix $A > 0$ such that $\mu_n[-A,A] \geq 1 - \varepsilon$ and such that A is a continuity point of F. Then $F(A) = \lim_{k\to\infty} F_{n_{\ell,k}}(A) \geq 1 - \varepsilon$. Thus $F(+\infty) = 1$. Similarly one can show that $F(-\infty) = 0$. This shows that $F = F_\mu$ for some $\mu \in \mathcal{P}(\mathbb{R}^d)$ and thus $\mu_{n_{\ell,k}} \xrightarrow{d} \mu$ as $k \to \infty$.

Conversely, if the sequence $\{\mu_n\}$ is not tight, then for any $A > 0$, we can find an infinite sequence n_k such that $\mu_{n_k}(-A,A) < 1 - \varepsilon$ (why?). Then, either $F_{n_k}(A) < 1 - \frac{\varepsilon}{2}$ for infinitely many k or $F_{n_k}(-A) < \frac{\varepsilon}{2}$. Thus, for any $A > 0$, we have $F(A) < 1 - \frac{\varepsilon}{2}$ or $F(-A) < \frac{\varepsilon}{2}$. Thus F is not a CDF of a p.m., and we see that the subsequence $\{n_k\}$ has no further subsequence than can converge to a probability measure.

(b) Standard facts about convergence in metric spaces and part (a).

■