Advanced Probability Theory **Fall 2023**

Lecture 1: Relationships Between Measures

Lecturer: ***Ioannis Karatzas*** *Scribes: Heyuan Yao*

1.1 Absolutely Continuity and Singularity

Suppose μ, ν are two measures defined on the same measurable space $(\Omega, \mathcal{F})$.

- We say that ν **is absolutely continuous with respect to** μ, and write $\nu < \mu$ (or $\nu \ll \mu$) if

$$A \in \mathcal{F}, \mu(A) = 0 \text{ imply } \nu(A) = 0 \tag{1.1}$$

  Exercise: This is the case, for instance, when there exists some $h : \Omega \to [0, \infty)$ in $\mathbb{L}^1(\mu)$, such that $\nu(A) = \int_A h d\mu$, $\forall A \in \mathcal{F}$.

  It is a major result of measure theory that, under appropriate conditions, this is always the case.

- We say that ν **and** μ **are equivalent,** and write $\mu \sim \nu$, if both $\nu < \mu$ and $\mu < \nu$ hold.

  This is the case if $h > 0$ in the above display: for then we have also $\mu(A) = \int_A \frac{1}{h} d\nu$.

- We sat that μ **and** ν **are singular,** and write $\mu \perp \nu$, if there exists a set $A \in \mathcal{F}$ with $\mu(A) = \nu(A^C) = 0$.

  For instance, with $(\Omega, \mathcal{F}) = ([0, 1], \mathcal{B}([0, 1]))$, consider $\mu = \lambda =$ Lebesgue measure, and $\nu =$ measure induced on $\mathcal{B}([0, 1])$ by the Cantor function F, $\nu((a, b]) = F(b) - F(a)$. Then $\mu(A) = 0$ if A is the Cantor set, but $\nu(A) = 1,\ \nu(A^C) = 0$

Theorem 1.1 (Lebesgue Decomposition Theorem) *Suppose $(\Omega.\mathcal{F})$ is a measurable space, and μ, ν $\sigma-$finite measure on it. then there exist measures ν_{ac}, ν_s with*

$$\nu = \nu_{ac} + \nu_s \quad \nu_{ac} < \mu,\ \nu_s \perp \mu,$$

and this decomposition is unique.

For instance, let $\lambda|_{[a,b]}$ denote Lebesgue measure on an interval $[a, b]$. Take $\mu = \lambda|_{[0,2]}$, $\nu = \lambda|_{[1,3]}$. Then $\nu_{ac} = \lambda|_{[1,2]}$, $\nu_s = \lambda|_{(2,3]}$.

Theorem 1.2 (Radon-Nykodým Theorem) *Suppose μ (resp. ν) is a $\sigma-$finite (resp, finite) measure on $(\Omega, \mathcal{F})$, and $\nu < \mu$. Then there exists a unique, up to $\mu - a.e.$ equivalence, function $h : \Omega \to [0, \infty)$ in $\mathbb{L}^1(\mu)$, such that*

$$\nu(A) = \int_A h d\mu, \quad A \in \mathcal{F}. \tag{1.2}$$

This function h is called the "Radon-Nikodým derivative" of ν with respective to μ, and is denoted

$$h = \frac{d\nu}{d\mu}.$$

We often write $d\nu = hd\mu$. This notation suggests correct intuitive conclusions. For instance:

$$\int_\Omega f h d\mu = \int_\Omega f \frac{d\nu}{d\mu} d\mu = \int_\Omega f d\nu$$

for every measurable $f : \Omega \to [0, \infty)$, so that $\underline{fh \in \mathbb{L}^1(\mu) \Leftrightarrow f \in \mathbb{L}^1(\nu)}$.

1.2 Convex Analysis and Jensen Inequality

A function $F : (a, b) \to \mathbb{R}$ is said to be **convex** if

$$F(\lambda x + (1 - \lambda) y) \leq \lambda F(x) + (1 - \lambda) F(y) \tag{1.3}$$

for every $(x, y) \in (a, b)^2,\ 0 \leq \lambda \leq 1$.

The following figure shows an example of a convex function I drew.

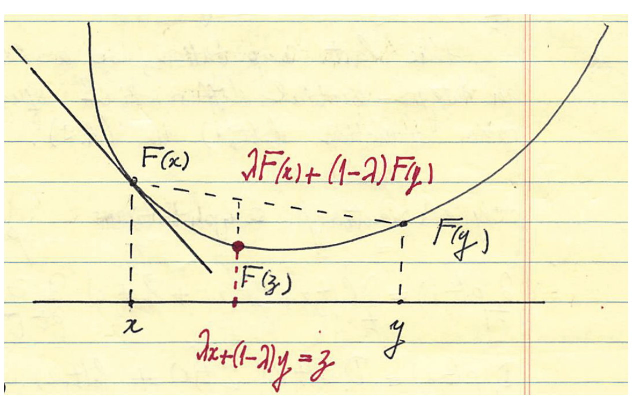

Figure 1.1: Convex Function

And we can easily derive that

$$F(\sum_{k=1}^{K} \lambda_k y_k) \leq \sum_{k=1}^{K} F(\lambda_k y_k)$$

for every $(y_1, ...y_K) \in (a,b)^K$, $K \in N$ $\lambda_1, ..., \lambda_K \geq 0$, $\sum_{k=1}^{K} \lambda_k = 1$. Equivalently: Suppose X is a random variable with $\mathbb{P}(X = y_k) = \lambda_l$, $k = 1, ..., K$. Then, this reads: $\underline{F(\mathbb{E}(X)) \leq \mathbb{E}(F(X))}$.

It turns out that this inequality holds more generally.

Theorem 1.3 (Jensen Inequality) *Suppose $X : \Omega \to (a,b)$ is an integrable random variable, and that $F : (a,b) \to \mathbb{R}$ is convex, for some $-\infty \leq a < b \leq \infty$. Then*

$$F(\mathbb{E}(X)) \leq \mathbb{E}(F(X))$$

Proof: For every $\xi \in (a,b)$, there is an affine function $L(x) = \alpha x + \beta$, $x \in (a,b)$ with $L(\cdot) \leq F(\cdot)$ and $L(\xi) = F(\xi)$.

Take $\xi = \mathbb{E}(X)$, notice

$$\mathbb{E}[F^-(X)] \leq \mathbb{E}[L^-(X)] \leq |\alpha|\mathbb{E}(|X|) + |\beta| < \infty.$$

This means that $\mathbb{E}(F(X))$ is well-defined.

Now clearly

$$\mathbb{E}[F(X)] \leq \mathbb{E}[L(X)] = L[\mathbb{E}(X)] = F[\mathbb{E}(X)].$$

■

1.3 Discrepancy of Two Measures

1.3.0.1 Total Variation Distance

How do we define measure "distance" between two measures (i.e., two distributions of mass, piles of sand, et cetera)? Here is the simplest such distance, total variation.

Definition 1.4 *Suppose μ, ν are arbitrary measures on $(\Omega, \mathcal{F})$; their Total Variation Distance is*

$$\|\mu - \nu\|_{TV} := \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)|.$$

Exercise: Suppose μ, ν are probability measures, and absolutely continuous w.r.t. some third probability measure λ:

$$\mu(A) = \int_A f d\lambda, \quad \nu(A) = \int_A g d\lambda$$

for some $f, g : \Omega \to [0, \infty)$ in $\mathbb{L}^1(\lambda)$. With $h = f - g$, we have

$$\|\mu - \nu\|_{TV} = \frac{1}{2}\int_\Omega |h| d\lambda = \int_\Omega h^+ d\lambda.$$

1.3.1 Relative Entropy

Suppose μ, ν are probability measures on $(\Omega, \mathcal{F})$. The relative entropy $\mathcal{D}(\nu|\mu)$ of ν w.r.t. μ is defined as $\mathcal{D}(\nu|\mu) = \infty$, if $\nu \perp \mu$.

On the other hand, if $\nu < \mu$, i.e. $\nu(A) = \int_A h d\mu$ for some $h : \Omega \to [0, \infty)$ in $\mathbb{L}^1(\mu)$, the relative entropy in defined as

$$\begin{aligned}\mathcal{D}(\nu|\mu) &:= \int_\Omega log h d\nu = \int_\Omega h\, log h d\mu = \int_\Omega F(h) d\mu \\ &= \int_\Omega log \frac{d\nu}{d\mu} d\nu = \int_\Omega \frac{d\nu}{d\mu}\, log \frac{d\nu}{d\mu} d\mu = \int_\Omega F(\frac{d\nu}{d\mu}) d\mu\end{aligned}$$

Unlike the total variation distance, this definition is not symmetric in μ, ν. We claim $\mathcal{D}(\nu|\mu) > 0$.

Proof: There is nothing to prove, if $\nu \perp \mu$.

Whereas, if $\nu < \mu$, Jensen gives

$$\mathcal{D}(\nu|\mu) = \mathbb{E}^\mu[F(h)] \geq F[\mathbb{E}^\mu(h)] = F[\int_\Omega h d\mu] = f[1] = 0. \tag{1.4}$$

We have used here the convexity of $F(x) = x\, log x$: $F'(x) = 1 + log x$, $F''(x) = \frac{1}{x} > 0$ ■

The following theorem reveals that, small entropy implies closeness in the total variation distance.

Theorem 1.5 (Pinsker-Csiszár Inequality) *For μ, ν two probability measure,*

$$2\|\mu - \nu\|_{TV}^2 \leq \mathcal{D}(\nu|\mu).$$

The entropy $H(\mu)$ of a probability measure μ on $(\mathbb{R}, \mathcal{B}(\mathbb{R})$ is defined as

$$H(\mu) := \begin{cases} \infty & \text{, if } \mu \perp \lambda = \text{Lebesgue measure,} \\ \int_{\mathbb{R}} f log(\frac{1}{f})d\lambda = \int_{\mathbb{R}} f(x)log(\frac{1}{f(x)})dx, & \text{if } \mu < \lambda \text{ with density } \frac{d\mu}{d\lambda},\ \mu(A) = \int_A f(x)dx. \end{cases}$$

Suppose now: $\nu(A) = \int_A f(x)dx$ has zero mean and unit variance: $\int_{\mathbb{R}} xf(x) = 0$, $\int_{\mathbb{R}} x^2 f(x)dx = 1$. Suppose also: $\mu(A) = \int_A \phi(x)dx$, $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{w}}$ standard normal.

Then,

$$\begin{aligned} \mathcal{D}(\nu|\mu) &= \int_{\mathbb{R}} log\left(\frac{f(x)}{g(x)}\right) f(x)dx \\ &= \int_{\mathbb{R}} log[f(x)]f(x)dx + \int_{\mathbb{R}} log(\frac{1}{\phi(x)})dx \\ &= \int_{\mathbb{R}} (\frac{x^2}{2} + log\sqrt{2\pi})f(x)dx - H(\nu) \\ &= \int_{\mathbb{R}} (\frac{x^2}{2} + log\sqrt{2\pi})\phi(x)dx - H(\nu) \quad \text{(Recall both } \mu \text{ and } \nu \text{ has the same second moment.)} \\ &= \int_{\mathbb{R}} log(\frac{1}{\phi})\phi(x)dx - -H(\nu) \\ &= H(\mu) - H(\nu) \geq 0. \end{aligned}$$

And we conclude that, <u>among all distributions with mean zero and variance 1, the Gaussian has the biggest entropy.</u>

1.3.2 The Information Theoretic Proof of CLT

Consider now a sequence $X_1, X_2, ...$ of I.I.D. random variables with $\mathbb{E}(X^2) < \infty$ and $m = \mathbb{E}X_1$, $\sigma = \sqrt{\mathbb{V}ar(X_1)}$. We denote by μ_n the distribution of $Z_n := \frac{1}{\sigma\sqrt{n}} \sum_{j=1}^{n} (X_j - m)$. This distribution has mean zero and variance 1.

We denote by μ the distribution of a standard Gaussian r.v. Z.

It was conjectured by Shannon (1949), and proved by Artstein et al. (2005), that

$$\lim_n \uparrow H(\mu_n) = H(\mu),$$

i.e. <u>the entropy of $(Z_n)_{n\in\mathbb{N}}$ INCREASES to the entropy of the standard Gaussian.</u>

But then, this means that $\mathcal{D}(\nu|\mu) = H(\mu) - H(\mu_n) \geq 0$ decreases to zero, as $n \to \infty$; and be the Pinsker-

Csiszár Inequality

$$2\|\mu - \nu\|_{TV}^2 \leq \mathcal{D}(\nu|\mu),$$

so does $\|\mu - \nu\|_{TV}$.