Chapter 0

Notation and Preview

WebYale = http://www.stat.yale.edu/~pollard
WebParis = http://www.ihp.jussieu.fr/~pollard
UGMTP = *User's Guide to Measure-Theoretic Probability*

Let $\mathcal{X}$ be a set equipped with a sigma-field $\mathcal{A}$, and $\mathcal{Y}$ be a set equipped with a sigma-field $\mathcal{B}$. Write $\mathcal{M}^+(\mathcal{X},\mathcal{A})$ for the set of all $\mathcal{A}$-measurable functions on $\mathcal{X}$ taking values in $[0,\infty]$, and $\mathbb{L}^+(\mathcal{X},\mathcal{A})$ for the set of all nonnegative, finite measures on $\mathcal{A}$.

For a measure μ on $\mathcal{A}$ and a measurable function f (from $\mathcal{M}^+(\mathcal{X},\mathcal{A})$, or μ-integrable) write μf or $\mu^x f(x)$ for $\int f(x)\,\mu(dx)$. Identify sets with their indicator functions [UGMTP §1.4]. Identify integrals with increasing "linear functionals" on $\mathcal{M}^+(\mathcal{X},\mathcal{A})$ with the Monotone Convergence property [UGMTP §2.3].

If T is an $\mathcal{A}\backslash\mathcal{B}$-measurable map from $\mathcal{X}$ to $\mathcal{Y}$, and μ is a measure on $\mathcal{A}$, the ***image measure*** $T\mu$ is defined on $\mathcal{B}$ by $(T\mu)(B) := \mu\{x : T(x) \in B\}$ for each $B \in \mathcal{B}$. Equivalently,

$$(T\mu)^y g(y) := \mu^x g(T(x)) \qquad \text{for } g \in \mathcal{M}^+(\mathcal{Y},\mathcal{B}).$$

The $\mathcal{L}^1$ distance between two finite measures, μ and ν, on $\mathcal{A}$ is defined as

$$\|\mu-\nu\|_1 := \sup\nolimits_{|f|\le 1} |\mu f - \nu f|,$$

the supremum running over all measurable functions f that are bounded in absolute value by 1. If both μ and ν are probability measures, then

$$\tfrac{1}{2}\|\mu-\nu\|_1 = \sup\nolimits_{A\in\mathcal{A}} |\mu A - \nu A| = \sup\nolimits_{0\le f\le 1} |\mu f - \nu f|,$$

a quantity that is often called the total variation distance between the measures [UGMTP §3.3].

Markov kernels

A Markov kernel, or randomization, from $(\mathcal{X},\mathcal{A})$ to $(\mathcal{Y},\mathcal{B})$ is a family of probability measures $K := \{K_x : x \in \mathcal{X}\}$ such that $x \mapsto K_x B$ is $\mathcal{A}$-measurable, for each $B \in \mathcal{B}$. For each f in $\mathcal{M}^+(\mathcal{X}\times\mathcal{Y}, \mathcal{A}\otimes\mathcal{B})$, the function $x \mapsto K_x^y f(x,y) := \int f(x,y)\,K_x(dy)$ is $\mathcal{A}$-measurable. If μ is a measure on $\mathcal{A}$ then a measure $\mu\otimes K$ can be defined on $\mathcal{A}\otimes\mathcal{B}$ by

$$(\mu\otimes K)\, f := \mu^x \left(K_x^y f(x,y)\right).$$

It has marginals μ and λ, with λ the measure on $\mathcal{B}$ defined by

$$\lambda^y g(y) := \mu^x \left(K_x^y g(y) \right) \qquad \text{for } g \in \mathcal{M}^+(\mathcal{Y}, \mathcal{B}).$$

I will also write $K\mu$ or $\mu^x K_x$ for λ. The map $\mu \mapsto K\mu$ from $\mathbb{L}^+(\mathcal{X}, \mathcal{A})$ to $\mathbb{L}^+(\mathcal{Y}, \mathcal{B})$ is "linear", and it takes probability measures to probability measures.

If μ is a probability measure, the pair (x, y) generated by

$$x \sim \mu \qquad \text{and} \qquad y|x \sim K_x$$

has joint distribution $\mu \otimes K$. The y has marginal distribution $\mu^x K_x$.

Decision theory

Call a family of probability measures $\mathcal{P} := \{\mathbb{P}_\theta : \theta \in \Theta\}$, all defined on the same sigma-field $\mathcal{A}$ on a sample space $\mathcal{X}$, a ***statistical model*** (or statistical experiment). Let $\mathcal{T}$ be some set, equipped at least with a sigma-field $\mathcal{C}$. A ***decision procedure*** is a measurable map T from $\mathcal{X}$ to $\mathcal{T}$. (If $\mathcal{T} = \Theta$, then T is usually called an estimator for the parameter θ.) A randomized procedure is defined as a Markov kernel τ from $(\mathcal{X}, \mathcal{A})$ to $(\mathcal{T}, \mathcal{C})$.

A map ℓ from $\mathcal{T} \times \Theta$ into $[-\infty, \infty]$ is called a ***loss function***. Typically I will assume ℓ is either nonnegative or bounded, so that there are no problems with the next definition. The risk function for a procedure T is defined as

$$R(T, \theta) := \mathbb{P}_\theta^x \ell\left(T(x), \theta\right) = (T\mathbb{P}_\theta)^t \ell(t, \theta) \qquad \text{for } \theta \in \Theta.$$

The risk function for a randomized procedure τ is defined as

$$R(\tau, \theta) := \mathbb{P}_\theta^x \tau_x^t \ell\left(t, \theta\right) = (\tau\mathbb{P}_\theta)^t \ell(t, \theta) \qquad \text{for } \theta \in \Theta.$$

1. Preview of Le Cam distance

Let $\mathcal{P} := \{\mathbb{P}_\theta : \theta \in \Theta\}$ and $\mathcal{Q} := \{\mathbb{Q}_\theta : \theta \in \Theta\}$ be two statistical models, indexed by the same parameter set Θ. Suppose each $\mathbb{P}_\theta$ is defined on $(\mathcal{X}, \mathcal{A})$, and each $\mathbb{Q}_\theta$ is defined on $(\mathcal{Y}, \mathcal{B})$. Le Cam defined the quantity $\delta(\mathcal{P}, \mathcal{Q})$ to be the smallest ϵ for which there is a randomization K (which must not depend on θ) from $(\mathcal{X}, \mathcal{A})$ to $(\mathcal{Y}, \mathcal{B})$ for which

$$\tfrac{1}{2} \sup_\theta \|\mathbb{Q}_\theta - K\mathbb{P}_\theta\|_1 \leq \epsilon$$

REMARK. The factor of 1/2 makes the definition fit well with other plausible ways to define δ, in a sense that I will explain later. Actually Le Cam did not restrict his randomizations to be Markov kernels, but allowed what I will be calling ***generalized randomizations***, that is, linear maps from $\mathbb{L}^+(\mathcal{X}, \mathcal{A})$ to $\mathbb{L}^+(\mathcal{Y}, \mathcal{B})$ that take probability measures onto probability measures.

If $\epsilon := \delta(\mathbb{P}, \mathcal{Q})$ is small, then we can almost reproduce the $\mathcal{Q}$ model from the $\mathcal{P}$ model by randomization:

$$\text{if } x \sim \mathbb{P}_\theta \qquad \text{and} \qquad y|x \sim K_x$$

then the distribution of y is close to $\mathbb{Q}_\theta$ (in the $\mathcal{L}^1$, or total variation, sense). For measurable functions g on $\mathcal{Y}$ with $0 \le g \le 1$, we have

$$|\mathbb{Q}_\theta^y g(y) - \mathbb{P}_\theta^x K_x^y g(y)| \le \epsilon \qquad \text{for every } \theta.$$

Now suppose τ is a randomized procedure defined for the $\mathcal{Q}$ model. Then we can define a randomized procedure ρ for $\mathcal{P}$ by a two-step construction:

$$\text{for } x \sim \mathbb{P}_\theta, \quad \text{generate } y|x \sim K_x, \quad \text{then generate } t \sim \tau_y.$$

That is, ρ_x is the probability measure τK_x on $\mathcal{C}$:

$$\rho_x^t h(t) = K_x^y \tau_y^t h(t) \qquad \text{for } h \in \mathbb{M}^+(\mathcal{T}, \mathcal{C}).$$

and

$$\mathbb{P}_\theta^x \rho_x^t h(t) = \mathbb{P}_\theta^x K_x^y \tau_y^t h(t) \qquad \text{for every } \theta.$$

If $0 \le h \le 1$ then the function $g(y) := \tau_y^t h(t)$ also takes values in $[0, 1]$, and so the right-hand side lies within ϵ of $\mathbb{Q}_\theta^y g(y) = \mathbb{Q}_\theta^y \tau_y^t h(t)$. In particular, if ℓ is a loss function taking values in the range $[0, 1]$, then

$$|\mathbb{P}_\theta^x \rho_x^t \ell(t, \theta) - \mathbb{Q}_\theta^y \tau_y^t \ell(t, \theta)| \le \epsilon \qquad \text{for every } \theta.$$

That is, $|R(\rho, \theta) - R(\tau, \theta)| \le \epsilon$ for every θ.

In effect, the randomization K has carried the problem of evaluating randomized procedures for $\mathcal{Q}$ back to an analogous problem for $\mathcal{P}$, with less than an ϵ of error if the loss function takes values in $[0, 1]$.

If we also had $\delta(\mathcal{Q}, \mathcal{P})$ small, then there would be a similar transfer of problems for $\mathcal{P}$ back to problems for $\mathcal{Q}$.

If the quantity $\Delta(\mathcal{P}, \mathcal{Q}) := \max(\delta(\mathcal{P}, \mathcal{Q}), \delta(\mathcal{Q}, \mathcal{P}))$ is close to zero, then there is an approximate correspondence (via randomizations) between solutions to decision theoretic problems for $\mathcal{P}$ and decision theoretic problems for $\mathcal{Q}$. Such a correspondence is very helpful if one of the experiments is much easier to work with than the other.